



Variational AdaBoost knowledge distillation for skin lesion classification in dermatology images

Xiangchun Yu^{1,2} · Guoliang Xiong^{1,2} · Jianqing Wu^{1,2} · Jian Zheng^{1,2} · Miaomiao Liang^{1,2} · Liujin Qiu³ · Lingjuan Yu^{1,2} · Qing Xu^{1,2}

Received: 30 November 2023 / Accepted: 16 May 2024 / Published online: 22 June 2024
© The Author(s) 2024

Abstract

Knowledge Distillation has shown promising results for classifying skin lesions in dermatology images. Traditional knowledge distillation typically involves the student model passively mimicking the teacher model's knowledge. We propose utilizing AdaBoost to enable the student to actively mine the teacher's learning representation for skin lesion classification. This paradigm allows the student to determine the “granularity” in mining the teacher's knowledge. As the student's learning process progresses, it can become challenging to pinpoint specific learning difficulties, especially with potential interference from the teacher. To address this issue, we introduce a variational difficulty mining strategy to reduce the impact of such interference. This strategy involves the distillation module capturing more nuanced classification difficulties by extracting information from the node's l th hops. By maximizing the mutual information between the teacher and student, we effectively filter out noise interference from these nuanced difficulties. Our proposed framework, Variational AdaBoost Knowledge Distillation (VAdaKD), allows the student to actively mine and leverage the teacher's knowledge for improved skin lesion classification. Our proposed method performs satisfactorily on three benchmark datasets: the Dermnet dataset, ISIC 2019 dataset, and HAM10000 dataset, respectively. Specifically, our method shows an improvement of 2–3% over the baseline on the Dermnet dataset and outperforms the best results of the other compared methods by 1%. Experimental results and visualization performance indicate that our proposed method effectively captures the learning difficulties and achieves better visualized t-distributed stochastic neighbor embedding classification results. Our code is available at <https://github.com/25brilliant/VAdaKD>.

Keywords AdaBoost · Knowledge distillation · Variational difficulty mining strategy · Skin lesion classification · Dermatology images

✉ Jianqing Wu
jianqing.wu@jxust.edu.cn

Xiangchun Yu
yuxc@jxust.edu.cn

Guoliang Xiong
brilliant2125@163.com

Jian Zheng
gzzj_yanjiu@163.com

Miaomiao Liang
liangmiaom@jxust.edu.cn

Liujin Qiu
951486061@qq.com

Lingjuan Yu
yljsmile@163.com

Qing Xu
xuqing@jxust.edu.cn

Introduction

Knowledge distillation has shown promising results for classifying skin lesions in dermatology images [1–5]. Knowledge distillation encompasses three types of knowledge: logits-based knowledge, intermediate feature-based knowledge, and relationship-based knowledge [6]. Logits-based knowledge distillation [7, 8] enables the student to mimic the teacher's soft output by adjusting the temperature parameter.

¹ School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China

² Jiangxi Province Key Laboratory of Multidimensional Intelligent Perception and Control, Jiangxi, China

³ Department of Ganzhou Cancer Hospital, Ganzhou 341000, China

Knowledge transfer through logits is accomplished by minimizing the Kullback–Leibler (KL) divergence between the student’s and teacher’s final outputs. Intermediate feature-based knowledge distillation [9–11] is utilized to transfer the intermediate features of the teacher to the student as knowledge. While the approaches above focus on transferring knowledge between individual instances, they overlook the valuable knowledge embedded in the relationships between multiple instances. Consequently, relationship-based knowledge distillation [14, 21] can yield superior performance by transferring the relationships between multiple instances to the student as a form of knowledge with more substantial representational power. *Motivations:* However, in traditional knowledge distillation, the student typically passively mines information from the teacher, which can restrict the students’ learning potential. This limitation becomes more pronounced when there is a knowledge gap between student and teacher models in skin lesion classification. The passive learning paradigm impedes the student’s ability to mine the teacher’s knowledge fully. Furthermore, the teacher model might contain unnecessary information when classifying dermatology images or retain misleading pseudo-label interference. Indiscriminately mimicking the teacher’s knowledge can hinder students’ potential for classifying dermatology images. A more effective approach would be for the student to actively mine knowledge from the teacher based on their specific challenges in classifying dermatology images.

Boosting algorithms [22, 23] sequentially learn multiple weak classifiers (i.e., base classifiers) and then combine them with weights to obtain a robust classifier. AdaBoost algorithm [24, 30] is the most typical Boosting algorithm, where sample weights are readjusted based on the error of the base classifier. Misclassified samples are given higher weights and emphasized in the following base classifier, following this iterative process. The weight of each base classifier is estimated based on its performance. These weighted base classifiers are then integrated to generate a robust classifier. Utilizing AdaBoost regulations during deep neural network training has been shown to improve the representation power of network models [25–33]. For instance, Taherkhani et al. [27] proposed AdaBoost-CNN by combining AdaBoost with a convolutional neural network (CNN), successfully addressing the multi-class imbalanced sample classification issue. Shakeel et al. [29] utilized the AdaBoost concept in ensemble neural networks and analyzed lung features to classify normal and abnormal lung categories. Huang et al. [34] developed a BoostResNet network using the Boosting principle to train ResNet via a step-by-step method and documented its efficacy in enhancing model training through theoretical validation. In addition, Sun et al. [30] introduced the AdaBoost algorithm to graph convolutional neural networks (GCNs) [30, 31] to integrate the information of neighboring nodes of different orders in an AdaBoost manner. This

approach effectively overcomes the over-smoothness problem provoked by aggregating the graph convolutional layers multiple times. *Motivations:* The student often encounters diverse challenges in knowledge distillation. It is a better paradigm for the student to identify their learning difficulties first and then actively mine relevant knowledge from the teacher to facilitate the distillation process. Therefore, exploring the potential of AdaBoost in helping the student sense the complexity of classifying dermatology images is valuable. Being able to sense difficulty levels could assist the student in deciding the appropriate level of granularity needed when formulating teacher knowledge based on their current comprehension.

Zhou et al. [32] proposed a weighted soft-label distillation framework, WSLD. This framework assigns a dynamic weight to the distillation loss to determine the extent to which the teacher’s soft-label information can be utilized based on the cross-entropy loss of both the student and the teacher. However, a student’s learning is a sequential process and will encounter varying difficulties at different stages. The WSLD framework only mines logits-based knowledge, neglecting the representational knowledge inherent in each stage and failing to utilize intermediate feature-based and relationship-based knowledge. Meanwhile, this method cannot avoid noise interference from the teacher. *Innovations:* This research introduces AdaBoost and a variational difficulty mining strategy (VDMS) to knowledge distillation and proposes a distillation framework called Variational AdaBoost Knowledge Distillation (VAdaKD). The framework aims to help the student determine the “granularity” of mining the teacher’s knowledge by considering the learning difficulties of skin lesion classification in dermatology images. Specifically, we apply AdaBoost to treat all stages of the student model as a sequential learning process. We also introduce an intermediate auxiliary classifier for each stage, where the weights of input samples from each stage correspond to the degree of learning difficulty. The paper proposes actively leveraging the teacher’s knowledge based on the student’s learning difficulty at each stage, facilitating targeted knowledge transfer. Finally, the paper performs linear weighting based on the weights of the base classifiers to obtain the final distillation loss. In this research, we incorporate three forms of knowledge into the distillation loss: logits-based, intermediate feature-based, and relationship-based, respectively. However, as the student’s learning process progresses, it becomes increasingly challenging to identify the degree of learning difficulties in subsequent learning stages due to noise interference from the teacher. Therefore, we first adopt the idea of GCN to construct the nearest-neighbor relationship matrix A . This matrix helps us calculate the information of the current node’s l th hop and then recount it to the student. The goal is to enable the student

to perceive the nuanced classification difficulties by leveraging the multi-hop information among dermatology samples while maintaining the same nearest-neighbor relationship with the teacher. Next, we eliminate noise interference from these nuanced difficulties by maximizing the mutual information between the teacher and student. *Contributions:* The main contributions of this paper can be summarized as follows:

- (1) This paper proposes a Variational AdaBoost Knowledge Distillation framework, VAdaKD, to address the limitations of conventional knowledge distillation methods, where the student passively learns the teacher's knowledge. VAdaKD offers a more active paradigm for knowledge distillation, allowing the student to determine the "granularity" in mining the teacher's knowledge within this framework.
- (2) VAdaKD employs a two-step strategy to improve the efficiency of AdaBoost-based knowledge distillation for categorizing dermatology images. Initially, the student is empowered to mine the teacher's learning representation through AdaBoost actively. Subsequently, a variational difficulty mining strategy (VDMS) is introduced to reduce the influence of noise from the teacher by maximizing the mutual information shared between the teacher and student.
- (3) Finally, we formulate the weighted distillation loss with sample-level weights to effectively incorporate three types of knowledge. Our research involves extensive experiments on three well-known dermatological datasets, namely the Dermnet ISIC2019 and HAM10000, respectively. The results of our experiments clearly show the efficacy of our proposed VAdaKD method. Additionally, the visualization results confirm that VAdaKD excels in identifying learning challenges and reducing interference from teacher noise at various stages, consequently enhancing the classification accuracy of dermatology images.

In contrast to previous work on knowledge distillation, our proposed VAdaKD introduces a novel paradigm for actively mining the teacher's knowledge, leading to a more comprehensive perception of the knowledge in the teacher while minimizing the presence of unnecessary information. The highlights of this paper are outlined below:

- (1) Propose a distillation framework, VAdaKD, for actively mining the teacher's learning representation for skin lesion classification.
- (2) Design a GCN-based difficulty mining to perceive the more nuanced classification difficulties.
- (3) Construct the weighted distillation loss with sample-level weights to effectively engage three forms of knowledge.

Section "[Introduction](#)" of this paper outlines the limitations of traditional knowledge distillation methods and proposes potential solutions. Section "[Related work](#)" reviews relevant literature, while Sect. "[Methodology](#)" introduces our proposed method. Experimental results and visualizations on three datasets are presented in Sect. "[Experiments](#)" to demonstrate the efficacy of our method. The results of the research are discussed in Sect. "[Conclusion](#)".

Related work

Knowledge distillation

Knowledge distillation, renowned for its ability to condense models, enables knowledge transfer from a cumbersome teacher model to a compact student model. This allows for efficiently deploying the developed skin lesion classification and diagnostic model on lightweight mobile devices. Hinton et al. [7] proposed to enable a student to learn the hidden knowledge of the teacher by reducing the KL divergence of the last layer's soft output with temperature. Zhao et al. [8] developed a framework called Decoupled Knowledge Distillation (DKD) that separates the information of target and non-target classes present in logits. They concluded that the effectiveness of logits-based knowledge distillation is due to the information provided by the non-target classes. Hossain et al. [40] proposed LumiNet to reconstruct finer inter-class relationships, enabling the student model to learn richer knowledge. RCO [13] introduced a route-constrained optimization strategy to narrow the knowledge gap between the teacher and the student through hint learning with route constraints. In addition to extracting logits-based information from the hierarchical backbone of the teacher, there is also the possibility of mining the intermediate feature-based representations. FitNets [9] achieved this by minimizing the L_2 loss of the intermediate features between the student and teacher. AT [10] focused the attention on the intermediate features, thereby encouraging the student to learn more discriminative feature representations. FT [11] utilized a paraphraser and translator to extract factors from the teacher and student, respectively, which are then operated as distillable intermediate-layer representations. Yang et al. [12] incorporated auxiliary classifiers from all stages of the backbone to transfer intermediate-layer knowledge. Shu et al. [42] enhanced the student network's attention towards the most salient regions in each channel by computing KL divergence between the student's and teacher's channel probability maps. However, the aforementioned knowledge distillation methods only focus on individual-instance knowledge distillation. PKT [14] matched the probability distributions of the feature space between the student's and teacher's data samples for knowledge transfer, and it effectively explored

the relationships between different individuals. RKD [15], SP [16], FSP [17], and IRG [19] engaged in various relationship matrices to facilitate knowledge transfer. Additionally, CCKD [21] leveraged both individual and inter-individual relationships to enrich knowledge distillation. In addition, the relationships between different layers in the distillation framework also provide valuable representational information. Lee et al. [18] used a correlation matrix as a feature map and extracted vital information from the feature map using singular value decomposition. Passalis et al. [20] suggested that the student can simulate the information flow in the teacher to facilitate knowledge transfer. Huang et al. [41] found that predicting by matching KL divergence may underperform when the teacher and student have a considerable gap. They proposed a correlation-based DITS loss to better capture the teacher's intrinsic inter-class relationships. However, the current distillation framework primarily focuses on the passive mimicry of feature representations from the teacher without considering how the student can actively extract beneficial knowledge in skin lesion classification.

AdaBoost in neural networks

The sequence of stages in the backbone of a student model can be considered as an ordered learning setting. AdaBoost is a sequential learning algorithm that trains base classifiers at different stages. It assigns higher weights to misclassified samples to prioritize them in the subsequent stages. Eventually, all the base classifiers are combined with respective base classifier weights to form a robust classifier. Researchers have successfully applied the AdaBoost algorithm to train models like convolutional neural networks, ensemble learning, and graph convolutional networks, and these applications have suggested satisfactory results [22–30]. CNN's performance is hindered when multi-class unbalanced samples are encountered. To address this, Taherkhani et al. [27] introduced AdaBoost-CNN, a method that effectively tackles the classification of multi-class unbalanced samples by combining AdaBoost with CNN.

Furthermore, to improve the joint optimization of ensemble neural networks, Shakeel et al. [29] utilized AdaBoost to ensemble neural networks and analyze the features of medical image data more effectively. Huang et al. [34] employed AdaBoost to train ResNet in a progressive training manner and showed the potential of boosting theory in ResNet under weak learning conditions. Besides, to address the issue of over-smoothness resulting from multiple aggregations of graph convolutional layers in the GCN, Sun et al. [30] proposed AdaGCN, which utilizes AdaBoost to integrate information from neighboring nodes of high orders. How-

ever, there is an urgent need to investigate how AdaBoost can actively enhance the students' ability to mine the teacher's learning representation in knowledge distillation. This research seeks to integrate AdaBoost into a knowledge distillation framework and develop a new paradigm for skin lesion classification in dermatology images. The goal is to enable the student model to determine the "granularity" of mining teacher's knowledge.

Active knowledge distillation

Traditional knowledge distillation methods involve the student model passively mimicking the teacher. WSLD [32] introduced a weighted soft-label approach and assigned a dynamic weight to the distillation loss based on the student's and teacher's learning on the supervised task. CTKD [33] proposed a dynamic temperature hyperparameter distillation framework. This framework increases distillation loss by adjusting the temperature adversarially, allowing the student to conduct knowledge transfer from easy to complex. While WSLD and CTKD provide a form of active knowledge distillation, they cannot perceive the nuanced difficulty inherent in the learning process, and it is difficult to eliminate noise interference in the teacher.

Additionally, WSLD and CTKD only distill at the back-end and do not utilize intermediate features or relationship knowledge. While AdaBoost-CNN [27], BoostResNet [34], and AdaGCN [30] have involved AdaBoost in deep neural networks, its application in knowledge distillation remains to be further explored. Additionally, as the learning process progresses, the learning difficulties become increasingly difficult to recognize due to noise interference from the teacher, and exploring how to discover these difficulties at the current learning stage effectively is a research topic that has yet to be investigated. This research engages AdaBoost with the sequence of the stages in the backbone of a student model as a progressive learning process. The weights of the input samples in each stage represent the degree of learning difficulty. The classifying knowledge in the teacher is actively mined according to the student's learning difficulty to carry out a targeted knowledge transfer. To better mine the nuanced difficulties in the subsequent stages of skin lesion classification, we first incorporate the GCN [30, 31] to capture nuanced higher-order information from the nearest-neighboring nodes. GCN involves extracting and recounting the information from the current node's l th hop to the student during the learning process. We also ensure that the nearest-neighbor relationship between the student and teacher remains consistent. Next, we aim to eliminate noise interference from these nuanced difficulties by maximizing the mutual information between the teacher and student.

Dermatologic diagnostic systems in clinical applications

To further advance the use of dermatologic diagnostic systems in the clinic, Haberman et al. [43] propose to modify the dermatologic diagnostic system (DIAG). They strive to provide clinical acceptability and enhance diagnostic potential. Brooks et al. [44] introduce the skin disease diagnostic prompting system, DERMIS. The DERMIS system generates lists of plausible diagnoses based on probabilities computed from Bayes' theorem and is designed to be used exclusively by non-dermatologists (e.g., general practitioners). Liu et al. [45] have developed a deep learning system, DLS, to aid in the differential diagnosis of dermatological diseases. This system offers differential diagnosis for 26 specific dermatological conditions, thereby significantly enhancing the accuracy and effectiveness of dermatological diagnosis and treatment in primary care settings. In order to promote the use of healthcare mobile apps in telemedicine, Hameed et al. [46] proposed the detection of 4 dermatological diseases using mobile apps on the Android platform. This approach eliminates the necessity for patients to visit the clinic physically. A reported 1.9 billion people worldwide are affected by skin diseases, with the shortage of dermatologists leading to many seeking dermatologic care from general practitioners, resulting in lower diagnostic accuracy [45]. Additionally, the utilization of heuristic/evolutionary search optimization AI-based algorithms, such as the Gravitational Search Algorithm [47–49] and Inclined Planes System Optimizations [50, 51], to tackle hyperparameter optimization during network training is a promising research area. This paper introduces supervised loss, AdaBoost loss, and distillation loss, suggesting that enhancing the performance of dermatological diagnostic models could be achieved by carefully selecting optimal values for loss term weights and the temperature coefficient. Developing efficient and stable systems [52–54] to deploy lightweight dermatologic diagnostic models on mobile devices will aid in early screening and improve rural diagnosis and treatment.

Methodology

Framework design

The framework of the Variational AdaBoost Knowledge Distillation, VAdaKD proposed in this paper is illustrated in Fig. 1. The figure illustrates the evolution of sample weights throughout the training process and demonstrates how the student model leverages these weights to mine and learn knowledge from the teacher model actively. Both the teacher and student models are assumed to have L stages. We consider the stages in the backbone as an ordered learning

process. We incorporate intermediate auxiliary classifiers for each stage, resulting in a total of L base classifiers. The base classifier for the l th stage is denoted as $p_l(\cdot)$. The function $p_l(\cdot)$ comprises a convolutional layer, a global average pooling layer, and a fully-connected layer. AdaBoost learns base classifiers sequentially, and the weights of the input dermatology samples for the l th base classifier are denoted as $\{w_{l-1}^i\}_{i=1}^B$, where B represents the number of samples in the input batch. The base classifier weight can be calculated as α_l . The weights of the input dermatology samples at each stage measure the learning difficulty in skin lesion classification. These sample-level weights, represented by $\{w_{l-1}^i\}_{i=1}^B$, are utilized in the distillation module to identify which dermatology samples the student finds difficult to learn at that stage. The student can actively mine knowledge from the teacher regarding the learning difficulties in skin lesion classification. However, as the student's learning process progresses, it becomes increasingly challenging to identify learning difficulties due to noise interference from the teacher. We introduce a variational difficulty mining strategy (VDMS) that aims to minimize the impact of noise by maximizing the mutual information between the teacher and student. The final distillation loss is then obtained by linearly weighting the individual base classifiers using the base classifier weights α_l . The losses in the VAdaKD encompass the cross-entropy loss for the student task, the AdaBoost training loss for the student task, and the Variational AdaBoost-based distillation loss, respectively.

The distillation loss comprises three types of knowledge: logits-based knowledge, intermediate feature-based knowledge, and relationship-based knowledge, respectively. However, it becomes increasingly challenging to identify learning difficulties due to noise interference from the teacher. VDMS first incorporates the GCN to form a nearest-neighbor relationship matrix A . This matrix is then used to calculate the information of the current node's l th hop and repeat it to the student. Next, VDMS eliminates noise interference from these nuanced difficulties by maximizing the mutual information between the teacher and student. The general framework of AdaKD, as proposed in this paper, is depicted in Fig. 1.

Pre-training Teacher Model

The teacher model for skin lesion classification consists of a backbone network, denoted as $f^t(\cdot)$, with L stages. The auxiliary (i.e., base) classifiers, represented by $\{p_l^t(\cdot)\}_{l=1}^L$, are contained within the backbone. The pre-training process of the teacher model is divided into two phases. In the first phase, the teacher model with L stages is trained, resulting in the trained backbone $f^t(\cdot)$. In the second phase, the weights of the backbone $f^t(\cdot)$ are frozen, and the parameters of the auxiliary classifiers $\{p_l^t(\cdot)\}_{l=1}^L$ are updated. According to

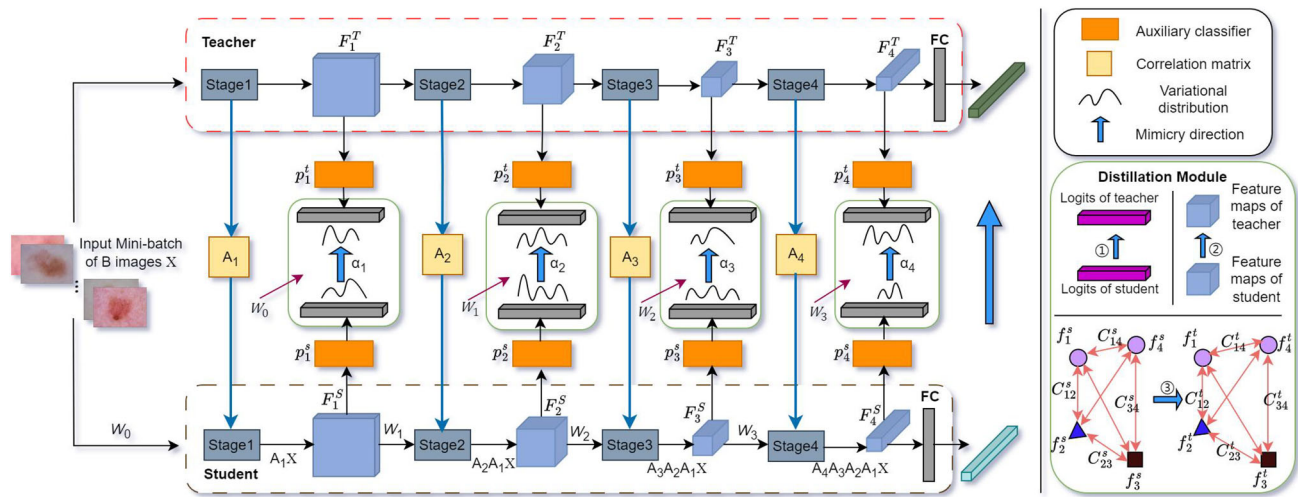


Fig. 1 The proposed framework, AdaBoost Knowledge Distillation, VAdaKD in this paper. VAdaKD empowers the student to actively mine the teacher's learning representation in skin lesion classification using AdaBoost. The weights of the dermatology input samples, denoted as $\{w_l^i\}_{i=1}^B$ for each stage, represent the learning difficulty for each sample x_i at the l -th stage. Based on these weights, the three forms of knowledge are selectively transferred to the student. Among them, F_l^S and F_l^T

represent the intermediate features of the student and teacher at the l -th stage. f_i^s and f_i^t represent the outputs of the i -th dermatology sample for the student and teacher, respectively. C represents the correlation between two specific dermatology samples. Finally, the final distillation loss is obtained by linearly weighting each form of knowledge according to the base classifier weights α_l . We introduce a VDMS that eliminates noise interference from nuanced difficulties by maximizing the mutual information between the teacher and student

the AdaBoost boosting theory, the weights of the input dermatology samples for the l th base classifier are represented as $\{w_{l-1}^i\}_{i=1}^B$, where B represents the number of samples in the input batch. Then, based on the error rate of each base classifier, we calculate the weight of the base classifier α_l . Finally, we obtain the AdaBoost loss by linearly weighting all base classifiers by weight $\{\alpha_l\}_{l=1}^L$. Note that we use the cross-entropy loss in both training phases to update the parameters for true-label supervision. It is worth noting that, at the l th stage, the error rate of the auxiliary classifier is represented by Eq. (1),

$$err_l = \frac{\sum_{i=1}^B w_{l-1}^i \mathbb{I}(y_i \neq p_l^t(x_i))}{\sum_{i=1}^B w_{l-1}^i}, \quad (1)$$

where y_i represents the true label (i.e., Ground Truth) of the dermatology input samples x_i , and the batch size is denoted as B . The weight α_l of the auxiliary classifier for the l -th stage is shown in Eq. (2),

$$\alpha_l = \log \frac{1 - err_l}{err_l} + \log(K - 1), \quad (2)$$

where K represents the number of categories. In order to ensure that α_l is positive, the condition $(1 - err_l) > 1/K$ needs to be satisfied. We then proceed to update the weights of the input samples x_i , giving a higher weight to the misclassified sample. Finally, w_l^i represents the weight of the dermatology input sample for the $l + 1$ th base classifier, as

shown in Eq. (3). This process is iterative, and the initial weight of the sample x_i is set to $w_0^i = 1/B$.

$$w_l^i \leftarrow w_{l-1}^i \cdot \exp(\alpha_l \cdot \mathbb{I}(y_i \neq p_l^t(x_i))). \quad (3)$$

As depicted in Fig. 2, the initial training phase involves using the true labels of skin lesion classification to supervise training the teacher model's backbone. In the subsequent training phase, the true labels are still used for supervision, but the weights of the backbone are frozen. AdaBoost is also employed to train the auxiliary classifiers introduced at all stages. Finally, the outputs of each classifier are integrated into a final prediction utilizing the corresponding weights.

Training student model

We refer to the backbone of the student model for skin lesion classification as $f^s(\cdot)$. Additionally, we have auxiliary classifiers within this backbone, denoted as $\{p_l^s(\cdot)\}_{l=1}^L$, where $p_l^s(\cdot)$ represents the auxiliary classifier of the l th stage in the student. The student is trained under the guidance of the pre-trained teacher. The overall loss function includes the task loss of true-label supervision, the AdaBoost loss of true-label supervision, and the weighted distillation loss with sample-level weights between the student and teacher.

Task loss of true-label supervision in skin lesion classification L_{ce} : The task loss in the student is determined by the cross-entropy loss of true-label supervision. The objective is

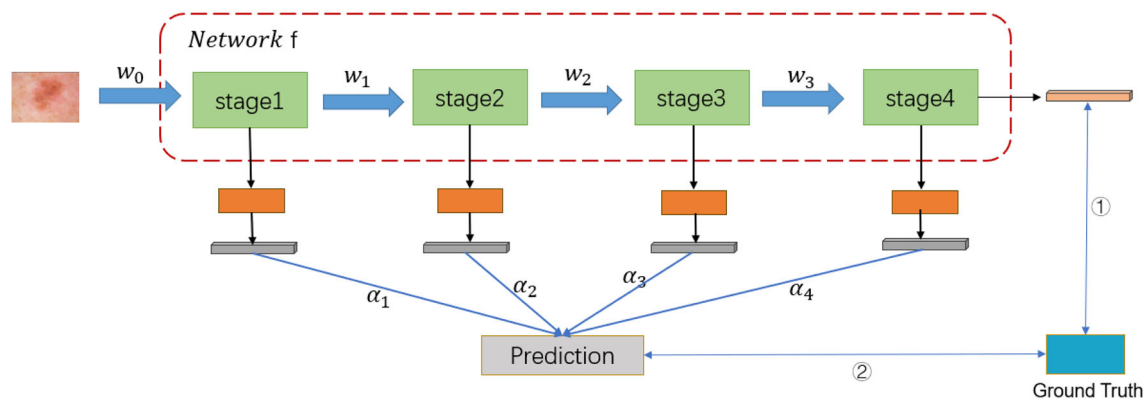


Fig. 2 Two-phase pre-training of a teacher model with L stages for skin lesion classification. In the first phase, the backbone $f^t(\cdot)$ is trained. In the second phase, the weights of the backbone $f^t(\cdot)$ are frozen, and the parameters of the auxiliary classifiers $\{p_l^t(\cdot)\}_{l=1}^L$ are updated using the AdaBoost boosting theory, where the input sample weights of the l

th base classifier are denoted as $\{w_{l-1}^i\}_{i=1}^B$. Then the weight α_l of the base classifier is calculated based on the error rate. Again, the dermatology input sample weights $\{w_l^i\}_{i=1}^B$ of the $l+1$ th auxiliary classifier are updated and adjusted by the l th auxiliary classifier

to enable $f^s(\cdot)$ to accurately classify one-hot labeled data. The task loss L_{ce} is obtained by computing the cross-entropy of the final outputs against the true labels.

AdaBoost loss of true-label supervision in skin lesion classification L_{ada_ce} : AdaBoost employs a sequential training approach and this involves training the student model using L auxiliary classifiers with prediction functions $\{p_l^s(\cdot)\}_{l=1}^L$. The input samples' weights of the l th base classifier are denoted as $\{w_{l-1}^i\}_{i=1}^B$. Similarly, the input samples' weights $\{w_l^i\}_{i=1}^B$ of the $l+1$ th base classifier are updated and adjusted by the l th base classifier. The weight of the l th base classifier α_l is then computed based on its error rate. Finally, the AdaBoost loss is obtained by linearly combining the individual base classifiers by weights $\{\alpha_l\}_{l=1}^L$. In addition, as the learning process of skin lesion classification progresses, it becomes increasingly challenging for the subsequent learning stages to mine nuanced classifying difficulties. To address the problem, the information from the current node's l th hop in the teacher is repeated to the student. Specifically, the correlation matrix $\{A_l^t(\cdot)\}_{l=1}^L \in R^{B \times B}$ of each stage in the teacher model is copied to the student model. Here, B represents the batch size, and the features of the l th stage in the student are transformed into $F_l^S = A_l^t F_l^S \in R^{B \times C \times H \times W}$. The final AdaBoost loss of true-label supervision is shown in Eq. (4),

$$L_{ada_ce} = - \sum_{i=1}^B \sum_{l=1}^L y_i \log \alpha_l p_l^s(x_i), \quad (4)$$

where y_i represents the true label of the dermatology input sample x_i , $p_l^s(x_i)$ is formulated as $p_l^s(x_i) = C_l(w_{l-1}^i \cdot F_l^S(x_i))$, the intermediate features of the input sample x_i is extracted by $F_l^S(x_i)$ and is classified by the l th auxiliary classifier module $C_l(\cdot)$ with the weight w_{l-1}^i . The value of

w_{l-1}^i represents the difficulty of classification learning for x_i in the l th base classifier, with a larger value indicating higher difficulty.

Weighted distillation loss with sample-level weights L_{mimic} : The weights of the input samples obtained by AdaBoost are represented as $\{w_{l-1}^i\}_{i=1}^B$. We can effectively engage three forms of knowledge by applying the sample-level weights to the distillation module. The distillation loss encompasses logits-based, intermediate feature-based, and relationship-based distillation loss. The logits-based distillation loss consists of two forms, the first being the KL loss of the soft outputs with temperature for the last layer of the student and teacher (as shown in Eq. (5)),

$$L_{kd} = -\tau^2 \sum_{i=1}^B p^t(x_i; \tau) \log(p^s(x_i; \tau)), \quad (5)$$

where τ represents the temperature, and we set it to 3 in this paper, and B denotes the size of the input batch. The second form of logits-based distillation loss is the distillation loss of the soft outputs with temperature for the last layer of the student–teacher correspondence auxiliary classifiers, as shown in Eq. (6),

$$L_{ada_kd} = -\tau^2 \sum_{i=1}^B \sum_{l=1}^L w_{l-1}^i \cdot p_l^t(x_i; \tau) \log(p_l^s(x_i; \tau)). \quad (6)$$

In order to effectively extract the representation information from the intermediate features $\{F_l^T\}_{l=1}^L$, this paper constructs the distillation loss using Attention Transfer [10] as shown in Eq. (7),

$$L_{att} = \sum_{i=1}^B \sum_{l=1}^L w_{l-1}^i \cdot L_{AT}(F_l^T(x_i), F_l^S(x_i)), \quad (7)$$

however, it becomes increasingly challenging to identify learning difficulties due to noise interference from the teacher as the student's learning process progresses. Therefore, we introduce the VDMS to minimize the impact of noises, and the AT-based distillation loss with VDMS is illustrated in Eq. (8),

$$L_{att_vdms} = \sum_{i=1}^B \sum_{l=1}^L w_{l-1}^i \cdot \left(\frac{(t_l^T(x_i) - \mu_l^S(x_i))^2}{2\sigma_l^2(x_i)} + \log \sigma_l^2(x_i) \right), \quad (8)$$

where $t_l^T(\cdot)$ represents the teacher's attention map after imposing attention mapping to intermediate feature $F_l^T(\cdot)$ as implemented in Attention Transfer [10]. $\mu_l^S(\cdot)$ represents the student's attention map after imposing attention mapping to intermediate feature $F_l^S(\cdot)$, and $\sigma_l^2(\cdot)$ represents the variance of the attention map.

The above two kinds of distillation losses focus solely on individual-instance knowledge distillation. However, to explore the relationship between different individuals in classifying dermatology images, we present a relationship-based distillation loss. Firstly, we extract the logits features of B dermatology samples from each auxiliary classifier of the teacher model and construct the correlation matrix $\{R_l^T(\cdot)\}_{l=1}^L$. Similarly, we construct the correlation matrix $\{R_l^S(\cdot)\}_{l=1}^L$ for the logits features of B dermatology samples from each auxiliary classifier of the student model. Then, we calculate the Mean Squared Error (MSE) loss between the correlation matrices of the student and teacher to obtain the relationship-based distillation loss, as shown in Eq. (9),

$$L_{rkd} = \sum_{l=1}^L \alpha_l \cdot L_{mse}(R_l^T, R_l^S) = \sum_{l=1}^L \alpha_l \cdot (R_l^T - R_l^S)^2, \quad (9)$$

the relationship-based distillation loss with VDMS is expressed in Eq. (10), where σ_l^2 represents the variance of the correlation matrix,

$$L_{rkd_vdms} = \sum_{l=1}^L \alpha_l \cdot \left(\frac{(R_l^T - R_l^S)^2}{2\sigma_l^2} + \log \sigma_l^2 \right). \quad (10)$$

Total Loss: The total loss of the student model includes the task loss for true-label supervision, the AdaBoost loss for true-label supervision, and the weighted distillation loss with sample-level weights between the student and the teacher, as shown in Eq. (11),

$$L_{total} = L_{ce} + \lambda_1 L_{ada_ce} + \lambda_2 L_{mimic}, \quad (11)$$

where the weighted distillation loss with sample-level weights in skin lesion classification L_{mimic} is represented by Eq. (12), where $\lambda_1, \lambda_2, \eta, \beta$, and γ are balance factors. In the distillation loss, we set the temperature hyperparameter τ to 3.

$$L_{mimic} = L_{kd} + \eta L_{ada_kd} + \gamma L_{att_vdms} + \beta L_{rkd_vdms}. \quad (12)$$

Experiments

The experimental setup of this paper consists of three parts. Firstly, in Sect. "Comparison Experiment", we compare with related knowledge distillation frameworks to validate the effectiveness of our proposed VAdaKD. It should be noted that AdaKD represents our proposed AdaBoost-based KD with GCN, and VAdaKD represents our proposed AdaBoost-based KD with VDMS. Secondly, in Sect. "Ablation Study", the ablation experiments are designed to observe the impact of the number of auxiliary classifiers and to explore the validity of each component in the distillation loss. Lastly, in subSect. "Visualization", we explore the performance of VAdaKD in mining the learning difficulties at each stage by visualizing the correlation matrix for each stage of the student model and validate the multi-classification performance achieved by VAdaKD through t-distributed stochastic neighbor embedding (t-SNE) visualization.

Our experiments assess the model's performance using four evaluation indicators: Accuracy (Acc), Recall, F1-score, and AUC. Accuracy is a widely used metric in classification tasks that measures the overall classification performance of the model. However, achieving a high accuracy in highly imbalanced data may not be meaningful. AUC helps to make up for this shortcoming and better reflects the classifier's performance. Recall is crucial in the medical field as it indicates the correct classification rate among all diseased samples, essential for timely patient treatment. F1-score is a comprehensive metric that takes into account both Recall and Precision.

Comparison experiment

Extensive experiments are conducted on three benchmark datasets, including the Dermnet dataset, ISIC 2019 dataset, and HAM10000 dataset. The values of λ_1 and λ_2 in Eq. (11) are set to 0.5 and 0.3, respectively. In Eq. (12), the value of η is set to 0.25. Furthermore, based on the experimental setup of AT [10] and SP [16], the values of γ and β are set to $5e-7$ and 0.3. The framework is implemented in PyCharm using the PyTorch library, with training performed on a single NVIDIA 3090 GPU.

Experimental results on the Dermnet dataset

The Dermnet dataset [36], available on the largest online dermatology resource site, Dermnet, consists of 23 categories of skin lesion diseases. The dataset includes a training set with 15,552 images and a test set with 3,968 images. Table 1 presents a comparison of Accuracy (Acc), Recall, F1-score, and AUC for different distillation methods on the Dermnet dataset using various teacher-student backbones. We reproduce and repeat the experiment 5 times for each distillation framework and show its standard deviation on all metrics. The teacher model is obtained by the pre-training technique provided in subSect. "Pre-training Teacher Model". For a fair comparison, all experimental results are obtained under the guidance of this teacher model. Besides, we introduce the VDMS to AT-based and relationship-based distillation loss to minimize the noise impact from the teacher. Specifically, we first impose a random mask with a mask rate of 0.6 on the student's intermediate feature, then pass through two 3×3 convolutional layers and two 1×1 convolutional layers, and finally obtain the prediction $F_l^S(\cdot)$. Then, according to Eq. (8) and Eq. (10), we get L_{att_vdms} and L_{rkd_vdms} , respectively. The results indicate that our proposed VAdaKD achieves superior performance across all three teacher-student network backbones. Our framework performs very closely to the teacher model when the teacher is ResNet34 and the student is ResNet18.

Experimental results on the ISIC2019 dataset

The ISIC2019 Challenge dataset [38, 39] consists of 8 categories of dermatology images. This experiment divides the dataset into training and test sets in an 8:2 ratio. The training set contains 20,224 images, while the test set has 5056 images. Table 2 shows a comparison of the Accuracy (Acc), Recall, F1-score, and AUC of different distillation frameworks on the ISIC2019 dataset using various teacher-student backbones. Similar to the setup on the Dermnet dataset, we introduce the VDMS to AT-based and relationship-based distillation loss to minimize the noise impact from the teacher. It can be observed from Table 2 that our proposed VAdaKD achieves almost the superior results across three different teacher-student backbones, thereby validating the effectiveness of our proposed distillation framework. Furthermore, we compare the Recall performance of our proposed VAdaKD against the comparison methods on the ISIC2019 dataset using logits-based (e.g., DKD), intermediate feature-based (e.g., CAT-KD), relationship-based (e.g., SP) knowledge distillation, and active knowledge distillation (e.g., WLS), respectively. The teacher and student are ResNet34 and ResNet18, respectively. The results show that our proposed VAdaKD performs better than other comparison methods on hard classes like DF and SCC, which highlights the

effectiveness of our proposed VAdaKD in actively identifying learning difficulties against noise interference from the teacher.

Experimental results on the HAM10000 dataset

To further validate the robustness of our proposed method, we also performed comparison experiments on the HAM10000 dataset [37], which comprises seven classes of dermatology images. The training set includes 9,187 images, while the testing set has 828 images. Table 3 presents the comparison of Accuracy (Acc), Recall, F1-score, and AUC of various distillation frameworks applied to the HAM10000 dataset with different backbone architectures. We incorporate VDMS into AT-based and relationship-based distillation loss to mitigate the impact of noise from the teacher. The results in Table 3 demonstrate that our proposed VAdaKD outperforms other state-of-the-art knowledge distillation methods across various teacher-student backbones, highlighting the effectiveness of our distillation framework.

Number of student model parameters and flops

The number of parameters and flops of all comparison methods and our proposed model during training and testing are detailed in Table 4. Auxiliary classifiers are incorporated during the training phase to form base classifiers and train the student using the AdaBoost method. Including auxiliary classifiers in our proposed model increases the number of parameters during training compared to testing. However, these auxiliary classifiers are not used during inference in the testing phase. Deploying our dermatologic diagnostic model to mobile devices will have more significant potential for clinical applications.

Variational difficulty mining strategy in skin lesion classification at different stages

In order to identify nuanced learning difficulties against noise interference from the teacher, we introduce a variational difficulty mining strategy (VDMS). VDMS first utilizes the GCN to capture nuanced higher-order information. It constructs a nearest-neighbor relationship matrix A_l to calculate the information of the current node's l th hop and then shares it with the student. The nearest-neighbor relationship matrix A_l is derived from the teacher model to ensure consistency in the relationship between the student and teacher. Next, VDMS eliminates noise interference from these nuanced difficulties by maximizing the mutual information between the teacher and student. Figure 3 demonstrates the efficacy of our proposed method to capture nuanced higher-order information, and therefore it performs better in difficult category classification performance. The results indicate

Table 1 Comparison of Accuracy (Acc), Recall, F1-score, and AUC of various distillation frameworks on the Dermnet dataset when using different teacher-student backbones.

Teacher Student	ResNet34				VGG13				ResNet50			
	ResNet18				MobileNetV2				VGG13			
	Acc	Recall	F1	AUC	Acc	Recall	F1	AUC	Acc	Recall	F1	AUC
Teacher	73.34	69.28	70.57	97.13	71.27	67.13	69.26	96.63	74.85	72.36	72.56	97.72
Student	71.41 ± 0.34	66.7 ± 0.44	67.82 ± 0.23	96.54 ± 0.12	66.87 ± 0.27	62.44 ± 0.46	66.17 ± 0.46	95.78 ± 0.13	69.92 ± 0.2	65.83 ± 0.32	66.08 ± 0.35	96.22 ± 0.15
DKD [8]	72.41 ± 0.31	68.21 ± 0.17	69.05 ± 0.14	96.73 ± 0.04	68.49 ± 0.23	64.21 ± 0.32	68.21 ± 0.24	96.58 ± 0.11	71.7 ± 0.34	67.53 ± 0.21	68.35 ± 0.31	96.52 ± 0.08
DIST [41]	72.58 ± 0.27	68.35 ± 0.16	69.28 ± 0.11	96.79 ± 0.03	68.63 ± 0.23	64.44 ± 0.13	68.37 ± 0.12	96.62 ± 0.13	71.76 ± 0.34	67.61 ± 0.26	68.52 ± 0.28	96.55 ± 0.07
LumiNet [40]	73.36 ± 0.16	68.76 ± 0.06	69.49 ± 0.28	96.87 ± 0.02	69.47 ± 0.26	64.87 ± 0.14	68.73 ± 0.27	96.65 ± 0.08	72.31 ± 0.27	67.88 ± 0.17	68.83 ± 0.27	96.57 ± 0.11
CAT-KD [35]	72.44 ± 0.03	68.23 ± 0.32	69 ± 0.21	96.8 ± 0.04	68.46 ± 0.26	64.82 ± 0.35	68.55 ± 0.33	96.54 ± 0.05	72.33 ± 0.13	68.35 ± 0.37	68.73 ± 0.12	96.8 ± 0.06
CWD [42]	72.4 ± 0.26	68.18 ± 0.35	68.86 ± 0.21	96.45 ± 0.03	68.79 ± 0.12	64.25 ± 0.27	68.76 ± 0.27	96.44 ± 0.12	72.24 ± 0.33	68.14 ± 0.32	68.38 ± 0.41	96.46 ± 0.09
SP [16]	72 ± 0.26	67.52 ± 0.42	68.24 ± 0.21	96.36 ± 0.08	68.22 ± 0.12	64.03 ± 0.21	68.78 ± 0.35	96.49 ± 0.07	70.68 ± 0.33	66.52 ± 0.45	67.26 ± 0.36	96.24 ± 0.13
WSDL [32]	72.61 ± 0.15	68.2 ± 0.2	68.55 ± 0.33	96.68 ± 0.03	68.65 ± 0.24	64.45 ± 0.28	68.92 ± 0.21	96.66 ± 0.04	71.3 ± 0.17	66.91 ± 0.33	68.05 ± 0.28	96.41 ± 0.05
AdaKD	73.28 ± 0.46	68.4 ± 0.28	68.96 ± 0.17	96.73 ± 0.02	69.27 ± 0.22	65.19 ± 0.34	68.83 ± 0.26	96.68 ± 0.04	72.17 ± 0.15	68.31 ± 0.37	68.76 ± 0.37	96.67 ± 0.04
VAdaKD	73.61 ± 0.24	68.81 ± 0.28	70.08 ± 0.27	96.64 ± 0.04	69.75 ± 0.24	65.57 ± 0.26	69.13 ± 0.21	96.75 ± 0.04	72.63 ± 0.27	68.65 ± 0.53	69.14 ± 0.11	96.84 ± 0.07

The best results are shown in bold, and the second-best result is underlined. AdaKD represents our proposed AdaBoost-based KD with GCN, and VAdaKD represents our proposed AdaBoost-based KD with VDMS

Table 2 Comparison of Accuracy (Acc), Recall, F1-score, and AUC of various distillation frameworks on the ISIC2019 dataset when using different teacher-student backbones

Teacher	ResNet34				VGG13				ResNet50			
	ResNet18				MobileNetV2				VGG13			
	Acc	Recall	F1	AUC	Acc	Recall	F1	AUC	Acc	Recall	F1	AUC
Teacher	91.16	83.78	84.85	97.88	87.24	79.7	80.82	96.83	91.5	85.68	86.74	98.26
Student	89.93 ± 0.18	82.35 ± 0.3	83.57 ± 0.17	96.95 ± 0.12	79.67 ± 0.42	67.76 ± 0.73	69.04 ± 0.32	96.08 ± 0.17	86.3 ± 0.54	78.57 ± 1.29	79.47 ± 0.35	96.24 ± 0.17
DKD [8]	90.54 ± 0.11	83.21 ± 0.14	84.42 ± 0.17	97.18 ± 0.12	81.28 ± 0.22	69.46 ± 0.48	71.53 ± 0.59	96.17 ± 0.25	87.85 ± 0.12	80.89 ± 0.36	81.85 ± 0.18	96.72 ± 0.18
DIST [41]	90.58 ± 0.21	83.33 ± 0.15	84.47 ± 0.18	97.14 ± 0.11	82.14 ± 0.23	69.96 ± 0.32	72.82 ± 0.59	96.35 ± 0.21	87.65 ± 0.12	80.52 ± 0.27	81.47 ± 0.24	96.63 ± 0.14
LumiNet [40]	90.74 ± 0.23	83.41 ± 0.22	84.25 ± 0.35	97.2 ± 0.08	84.25 ± 0.35	75.16 ± 0.44	75.88 ± 0.34	96.34 ± 0.13	88.12 ± 0.25	81.14 ± 0.42	81.43 ± 0.26	96.75 ± 0.12
CAT-KD [35]	90.55 ± 0.27	83.27 ± 0.22	84.36 ± 0.21	97.21 ± 0.15	84.22 ± 0.31	74.94 ± 0.45	75.46 ± 0.27	96.94 ± 0.12	87.76 ± 0.37	81.44 ± 0.25	82.51 ± 0.14	96.85 ± 0.21
CWD [42]	90.38 ± 0.15	83.54 ± 0.25	84.43 ± 0.23	97.15 ± 0.19	82.53 ± 0.41	71.06 ± 0.46	72.14 ± 0.31	96.35 ± 0.14	87.56 ± 0.21	80.75 ± 0.53	81.74 ± 0.21	96.73 ± 0.23
SP [16]	90.3 ± 0.15	83.46 ± 0.25	84.31 ± 0.23	97.12 ± 0.17	82.61 ± 0.41	71.23 ± 0.37	72.37 ± 0.28	96.37 ± 0.16	87.44 ± 0.21	80.46 ± 0.68	81.56 ± 0.32	96.71 ± 0.14
WSDL [32]	90.68 ± 0.17	83.26 ± 0.27	84.56 ± 0.32	97.28 ± 0.23	84.14 ± 0.55	74.87 ± 0.96	76.92 ± 0.37	96.65 ± 0.14	88.03 ± 0.16	81.16 ± 0.34	82.22 ± 0.15	96.68 ± 0.13
AdaKD	90.87 ± 0.16	83.56 ± 0.21	84.53 ± 0.26	97.25 ± 0.14	84.43 ± 0.25	75.03 ± 0.58	75.96 ± 0.15	96.63 ± 0.08	88.17 ± 0.23	81.33 ± 0.38	82.37 ± 0.22	96.73 ± 0.16
VAdaKD	91.04 ± 0.13	83.67 ± 0.17	84.72 ± 0.15	97.28 ± 0.11	84.86 ± 0.43	75.23 ± 0.47	76.35 ± 0.34	96.88 ± 0.13	88.36 ± 0.25	81.52 ± 0.23	82.64 ± 0.13	97.25 ± 0.24

The best results are shown in bold, and the second-best result is underlined. AdaKD represents our proposed AdaBoost-based KD with GCN, and VAdaKD represents our proposed AdaBoost-based KD with VDMS

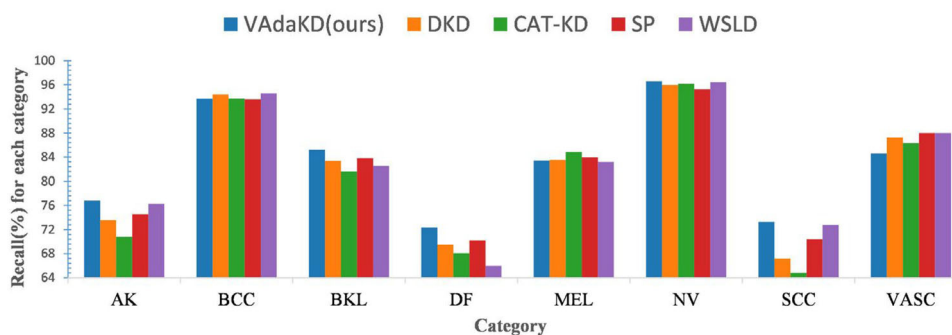
Table 3 Comparison of Accuracy (Acc), Recall, F1-score, and AUC of various distillation frameworks on the HAM10000 dataset when using different teacher-student backbones

Teacher	ResNet34				VGG13				ResNet50			
	ResNet18				MobileNetV2				VGG13			
	Acc	Recall	F1	AUC	Acc	Recall	F1	AUC	Acc	Recall	F1	AUC
Teacher	93.99	74.42	81.28	96.47	89.68	65.05	72.34	95.26	94.45	76.17	83.24	97.53
Student	93.11 ± 0.19	70.37 ± 0.32	77.58 ± 0.15	94.92 ± 0.21	82.65 ± 0.37	51.73 ± 0.46	57.48 ± 0.28	94.25 ± 0.16	89.18 ± 0.46	64.67 ± 0.84	71.46 ± 0.28	95.23 ± 0.16
DKD [8]	93.21 ± 0.14	70.65 ± 0.13	77.41 ± 0.17	95.43 ± 0.15	84.37 ± 0.24	56.18 ± 0.34	62.32 ± 0.18	94.85 ± 0.23	90.13 ± 0.37	65.56 ± 0.23	72.73 ± 0.17	95.74 ± 0.13
DIST [41]	93.19 ± 0.23	70.58 ± 0.14	77.62 ± 0.2	96.18 ± 0.12	84.73 ± 0.31	56.46 ± 0.41	62.43 ± 0.54	95.22 ± 0.31	90.21 ± 0.16	65.34 ± 0.24	72.62 ± 0.25	95.68 ± 0.11
LumiNet [40]	93.43 ± 0.11	71.36 ± 0.26	78.57 ± 0.16	96.58 ± 0.13	84.28 ± 0.22	56.46 ± 0.44	62.47 ± 0.47	95.36 ± 0.22	90.36 ± 0.23	65.52 ± 0.19	72.64 ± 0.13	95.65 ± 0.16
CAT-KD [35]	93.15 ± 0.21	70.27 ± 0.22	77.36 ± 0.21	96.21 ± 0.15	85.22 ± 0.33	57.67 ± 0.52	62.38 ± 0.25	95.78 ± 0.16	89.76 ± 0.18	64.89 ± 0.27	72.57 ± 0.14	95.83 ± 0.17
CWD [42]	93.2 ± 0.13	70.46 ± 0.14	77.31 ± 0.23	96.12 ± 0.17	84.61 ± 0.47	56.23 ± 0.36	62.61 ± 0.21	95.35 ± 0.27	90.05 ± 0.25	65.43 ± 0.43	72.68 ± 0.26	95.49 ± 0.23
SP [16]	93.23 ± 0.15	70.57 ± 0.25	77.31 ± 0.23	96.12 ± 0.17	83.61 ± 0.43	55.23 ± 0.47	61.37 ± 0.23	95.43 ± 0.15	90.23 ± 0.21	65.51 ± 0.37	72.57 ± 0.18	95.62 ± 0.24
WSDL [32]	93.24 ± 0.17	70.79 ± 0.28	77.56 ± 0.32	96.28 ± 0.23	84.14 ± 0.34	56.87 ± 0.69	62.92 ± 0.26	95.62 ± 0.13	90.37 ± 0.13	65.22 ± 0.18	72.81 ± 0.14	95.66 ± 0.11
AdaKD	93.28 ± 0.12	70.85 ± 0.47	77.98 ± 0.23	96.07 ± 0.15	84.46 ± 0.25	56.53 ± 0.35	62.96 ± 0.18	95.63 ± 0.12	90.17 ± 0.18	65.38 ± 0.16	72.37 ± 0.26	95.74 ± 0.18
VAdaKD	93.51 ± 0.14	71.91 ± 0.88	78.75 ± 0.17	<u>96.33 ± 0.11</u>	85.53 ± 0.32	57.82 ± 0.29	63.35 ± 0.27	95.88 ± 0.24	90.44 ± 0.23	65.64 ± 0.25	73.14 ± 0.15	96.17 ± 0.21

The best results are shown in bold, and the second-best result is underlined. AdaKD represents our proposed AdaBoost-based KD with GCN and VAdaKD represents our proposed AdaBoost-based KD with VDMs

Table 4 Number of parameters and flops during training and testing phases of all comparison methods and our proposed method

Teacher	ResNet34				VGG13				ResNet50			
Student	ResNet18				MobileNetV2				VGG13			
Comp. burden	Training		Testing		Training		Testing		Training		Testing	
	Params	Flops	Params	Flops	Params	Flops	Params	Flops	Params	Flops	Params	Flops
DKD [8]	103.81 M	11.68G	11.69 M	1.82G	47.2 M	50.63G	0.69 M	1.42G	117.37 M	25.23G	9.92 M	14.01G
DIST [41]	103.81 M	11.68G	11.69 M	1.82G	47.2 M	50.63G	0.69 M	1.42G	117.37 M	25.23G	9.92 M	14.01G
LumiNet [40]	103.81 M	11.68G	11.69 M	1.82G	47.2 M	50.63G	0.69 M	1.42G	117.37 M	25.23G	9.92 M	14.01G
CAT-KD [35]	103.81 M	11.71G	11.69 M	1.85G	47.21 M	50.68G	0.7 M	1.47G	117.41 M	25.27G	9.96 M	14.05G
CWD [42]	103.81 M	11.68G	11.69 M	1.82G	47.2 M	50.63G	0.69 M	1.42G	117.37 M	25.23G	9.92 M	14.01G
SP [16]	103.81 M	11.68G	11.69 M	1.82G	47.2 M	50.63G	0.69 M	1.42G	117.37 M	25.23G	9.92 M	14.01G
WSLD [32]	103.81 M	11.68G	11.69 M	1.82G	47.2 M	50.63G	0.69 M	1.42G	117.37 M	25.23G	9.92 M	14.01G
AdaKD	145.25 M	14.62G	11.69 M	1.82G	49.86 M	53.72G	0.69 M	1.42G	153.96 M	60.43G	9.92 M	14.01G
VAdaKD	145.25 M	14.62G	11.69 M	1.82G	49.86 M	53.72G	0.69 M	1.42G	153.96 M	60.43G	9.92 M	14.01G

**Fig. 3** Recall performance of our proposed VAdaKD against the comparison methods on the ISIC2019 dataset. The comparison methods include logits-based (e.g., DKD), intermediate feature-based (e.g., CAT-KD), relationship-based (e.g., SP) knowledge distillation, and active knowledge distillation (e.g., WSLD), respectively. ResNet34 acts as

the teacher, and ad ResNet18 acts as the student. It is worth noting that our VAdaKD performs better in hard classes like DF and SCC, which highlights the effectiveness of actively identifying learning difficulties against noise interference from the teacher

that our proposed method outperforms others in acquiring challenging knowledge and enhancing classification performance on complex samples. To validate the effectiveness of our proposed VDMS strategy, we conducted the corresponding experimental study, and the detailed results are presented in Table 5. We employ BoostNet as a baseline on the Dermnet dataset. 'Stu + AdaBoost' denotes that the student model is trained using AdaBoost but without distillation; 'Stu + AdaBoost + GCN' denotes that the student model is trained using AdaBoost along with GCN but without distillation. Meanwhile, AdaKD represents our proposed AdaBoost-based KD with GCN, and VAdaKD represents our proposed AdaBoost-based KD with VDMS, respectively.

Table 5 shows that our proposed VAdaKD achieves satisfactory results. The results demonstrate that VAdaKD performs better than AdaKD, indicating the effectiveness of the VDMS strategy for identifying increasing learning difficulties to some extent at different stages.

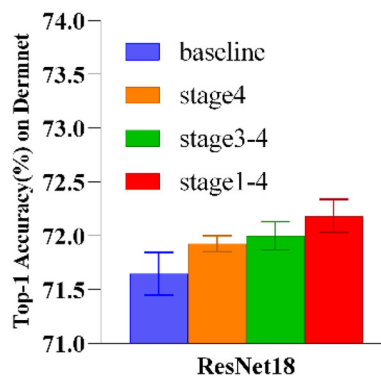
Ablation Study

Effect of the number on auxiliary classifiers. We investigate the impact of appending auxiliary classifiers on the performance of the distillation framework in skin lesion

Table 5 Comparison of the effectiveness of the VDMS strategy in skin lesion classification at different stages

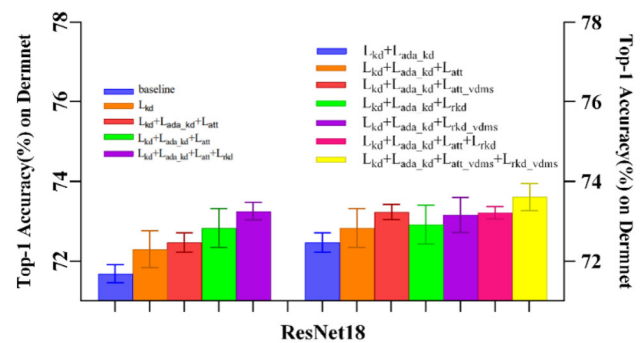
BackBone	ResNet18		VGG13		MobileNetv2	
	Acc	Recall	Acc	Recall	Acc	Recall
BoostNet* [33]	72.05 ± 0.24	67.59 ± 0.14	70.88 ± 0.27	66.71 ± 0.12	67.32 ± 0.21	63.12 ± 0.18
Stu + AdaBoost	72.18 ± 0.12	67.82 ± 0.18	71.09 ± 0.13	66.85 ± 0.28	67.63 ± 0.24	63.14 ± 0.32
Stu + AdaBoost	72.42 ± 0.16	68.03 ± 0.12	71.33 ± 0.11	67.25 ± 0.17	68.17 ± 0.16	64.22 ± 0.27
+ GCN	73.28 ± 0.46	68.4 ± 0.28	72.17 ± 0.15	68.31 ± 0.37	69.27 ± 0.22	65.19 ± 0.34
AdaKD						
AdaKD + VDMS	73.61 ± 0.24	68.81 ± 0.28	72.63 ± 0.27	68.65 ± 0.53	69.75 ± 0.24	65.57 ± 0.26

BoostNet* refers to the modified network by incorporating ResNet18, VGG13, and MobileNetv2 using BoostResNet as a prototype

**Fig. 4** Ablation experiments on the Dermnet dataset. We progressively increase the number of auxiliary classifiers for the student model ResNet18 along with vanilla knowledge distillation KD as the baseline. The best results are obtained when auxiliary classifiers are added across all stages

classification. We gradually increase the number of auxiliary classifiers and evaluate their influence on the Dermnet dataset. We use the student model along with vanilla knowledge distillation, KD as the benchmark. From the Fig. 4, it is evident that the best results are obtained when auxiliary classifiers are added for all stages. This result demonstrates that AdaBoost boosting theory can effectively train auxiliary classifiers across different stages sequentially, thereby improving the performance of the student model in skin lesion classification.

Effect of distillation loss on each component. We conduct an ablation study to investigate the contribution of each component for the weighted distillation loss with sample-level weights, L_{mimic} during the distillation process on the Dermnet dataset. The student model (without KD) is utilized as the baseline while maintaining the same task loss of true-label supervision L_{ce} and AdaBoost loss of true-label

**Fig. 5** Ablation experiments on the Dermnet dataset. The left panel shows ablation experiments of weighted distillation loss with sample-level weights L_{mimic} , which examines each component's role in the distillation process based on the student ResNet18. The right panel illustrates ablation experiments of the AT-based and relationship-based distillation loss with the VDMS strategy, which examines VDMS's role in the distillation process based on the student ResNet18

supervision L_{ada_ce} . The experimental results are presented in the left panel of Fig. 5. We take the student baseline as the benchmark and gradually append the logits-based distillation loss. This loss includes two forms: the KL loss L_{kd} between soft outputs with temperature for the last layer of the student and teacher, and the distillation loss L_{ada_kd} of the soft outputs with temperature for the last layer of the student–teacher correspondence auxiliary classifiers. Additionally, we introduce the intermediate feature-based distillation loss L_{att} and the relationship-based distillation loss L_{rkd} . Our experiments show that each component positively contributes to the final recognition performance, highlighting their effectiveness in leveraging different forms of knowledge embedded in the teacher's model.

Effect of the variational difficulty mining strategy. We introduce the VDMS to AT-based distillation loss and get the AT-based distillation loss with VDMS, L_{att_vdms} according

Table 6 Effect of different balancing factors on student model accuracy (%) on the HAM10000 dataset

λ_1	0.1	0.2	0.2	0.3	0.3	0.4	0.5	0.6	0.6	0.7	0.8
λ_2	0.1	0.1	0.2	0.2	0.3	0.3	0.3	0.3	0.4	0.4	0.4
Accuracy	93.19	93.22	93.28	93.34	93.38	93.44	93.51	93.36	93.42	93.22	93.17

The teacher model is ResNet34. This table shows the experimental results of taking different λ_1 and λ_2 when fixing η to be 0.25

Table 7 Different balancing factors affect the accuracy (%) of the ResNet18 student model on the HAM10000 dataset.

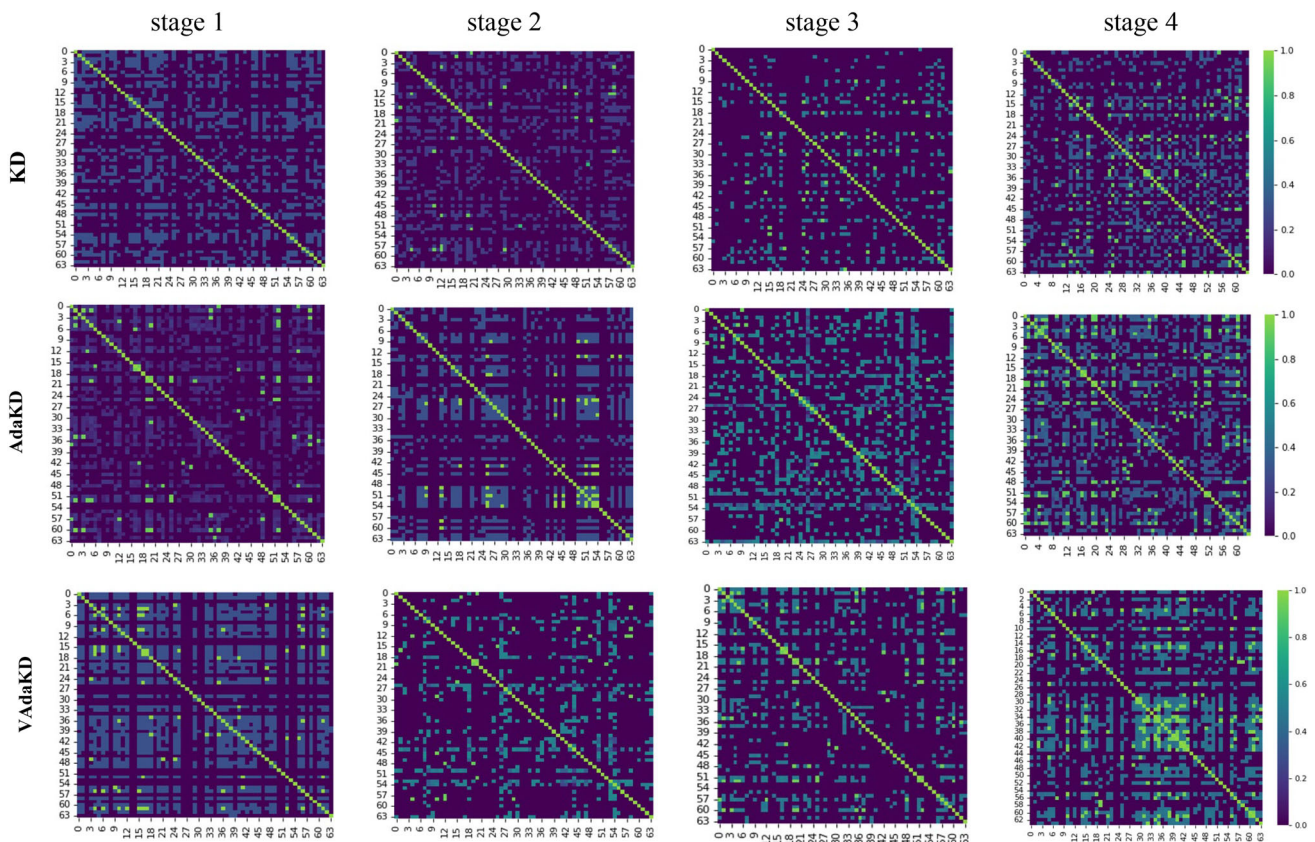
η	0.1	0.15	0.2	0.25	0.3	0.35	0.4
Accuracy	93.24	93.37	93.43	93.51	93.29	93.26	93.19

The teacher model is ResNet34. This table shows the experimental results of taking different η when fixing λ_1 and λ_2 to be 0.5 and 0.3, respectively

to Eq. (8). Similarly, we introduce the VDMS to relationship-based distillation loss and obtain L_{rkd_vdms} according to Eq. (10). Our experiment results show that each component positively contributes to the final recognition performance, indicating that VDMS can identify nuanced learning difficulties against noise interference from the teacher.

Effect of different hyperparameters. The effect of various hyperparameters is investigated through experiments on the

loss balancing factors outlined in Eq. (11) and (12) using the HAM10000 dataset. The results of these experiments for different factor values are detailed in Tables 6 and 7. It should be mentioned that the hyperparameters γ and β in Eq. (12) are set empirically with reference to prior works AT [10] and SP [16] and thus are not included in the experimental analysis. Ultimately, the values of λ_1 and λ_2 are set to 0.5 and 0.3, and the value of η is set to 0.25 in our proposed method.

**Fig. 6** Comparison of the visualization results of the correlation between KD, AdaKD, and VAdaKD at different stages using the same batch of dermatology samples as input. The visualization results of the correlation from stage 1 to stage 4 are represented in columns 1 to 4, respectively

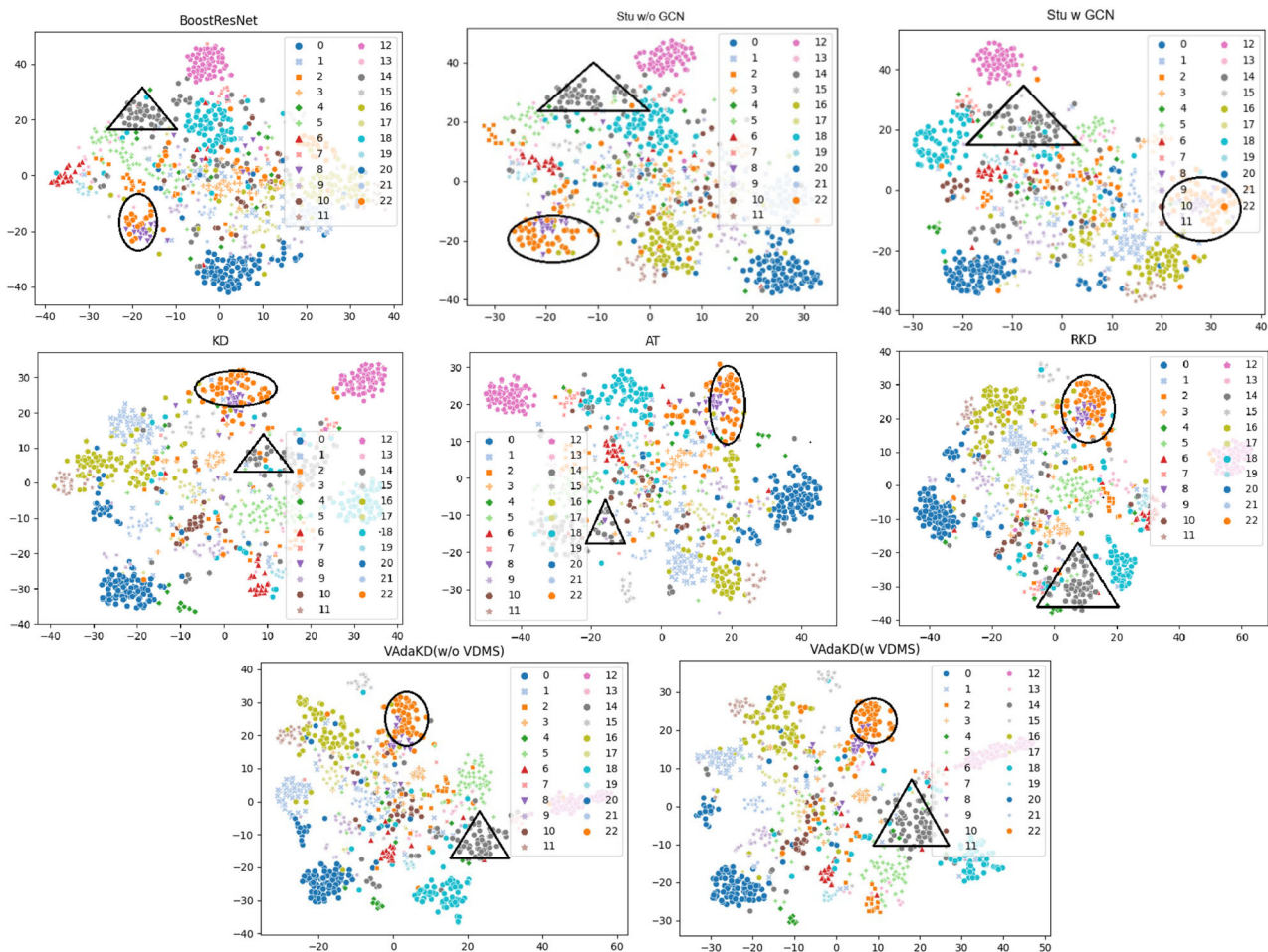


Fig. 7 Comparison of our proposed VAdaKD along with GCN strategy and three knowledge distillation methods in skin lesion classification. VAdaKD achieves a more ‘high cohesion-low coupling’ effect of classification on the Dermnet dataset. However, the dispersion between multiple categories is slightly weaker in the other methods. VAdaKD outperforms the other distillation frameworks for

categories 14, denoted by \triangle , and 22, denoted by \bigcirc , indicating its superior performance. ‘Stu w/o GCN’ denotes that the student model is trained using AdaBoost but without distillation, and GCN strategy is not adopted; ‘Stu w GCN’ denotes that the student model is trained using AdaBoost along with GCN but without distillation

Visualization

The visualization performance of our proposed VAdaKD is presented in this subsection. We examine VAdaKD’s ability to identify nuanced learning difficulties in the presence of noise interference from the teacher across different stages. Correlation visualization results are showcased using the same batch of dermatology samples as input, with vanilla KD and AdaKD serving as benchmarks (refer to Fig. 6). The results reveal that VAdaKD mines more relevant information at stage 1 and stage 2 in comparison to vanilla KD. It suggests that VAdaKD can effectively utilize inter-sample information in these two stages. At stage 3, VAdaKD has a weaker inter-sample correlation than AdaKD, but at stage 4, it perceives richer inter-sample details. This suggests that at stage

4, VAdaKD identifies more nuanced classification difficulties by mining multi-hop information between dermatology samples against noise interference from the teacher.

To further demonstrate the discriminative representation of VAdaKD, we utilize t-SNE to visualize the multi-classification effect of our proposed VAdaKD in skin lesion classification. We illustrate the effectiveness of VDMS in skin lesion classification along with logits-based distillation, intermediate features-based distillation, and relationship-based distillation. In addition, we also visualize the results of BoostResnet, Stu w/o GCN, and Stu w GCN, as presented in Fig. 7. From the figure, it is evident that VAdaKD achieves a more pronounced ‘high cohesion-low coupling’ classification effectiveness on the Dermnet dataset. For example, VAdaKD outperforms other distillation frameworks significantly in the 14th and 22nd categories.

Conclusion

This research proposes a Variational AdaBoost Knowledge Distillation framework, VAdaKD, which utilizes an advanced knowledge distillation paradigm. This paradigm enables the student to actively acquire and extract valuable knowledge for classifying dermatology samples at the current stage. The VDMS strategy helps reduce noise interference from the teacher by maximizing the mutual information between the teacher and student. Our proposed VAdaKD framework demonstrates superior performance on Dermnet, ISIC2019, and HAM10000 datasets compared to other knowledge distillation methods. Experimental results underscore the efficacy of the Active Knowledge Distillation framework in the classification of dermatology images.

Globally, 1.9 billion people are impacted by skin diseases. With a scarcity of dermatologists, many turn to general practitioners for dermatologic care, leading to less precise diagnoses. This research focuses on developing lightweight dermatology diagnostic models for mobile devices. Implementing these models could significantly enhance the practicality of dermatologic diagnostic systems, facilitating early screening and enhancing diagnosis and treatment in rural regions, ultimately leading to substantial economic advantages.

Acknowledgements We express our gratitude to the High-Performance Laboratory of the School of Information Engineering at Jiangxi University of Science and Technology for providing the computational resources. We would also like to thank all the teachers and students in this academic group for their valuable suggestions. Additionally, we appreciate the contributions of other researchers who helped adjust the ideas, designs, and experiments in this paper.

Funding This research is supported in part by the Jiangxi Provincial Natural Science Foundation under grants 20224BAB212013, 20224BAB212008, and 20224BAB202002, in part by the National Natural Science Foundation of China under grants 62266020 and 62261027, and in part by the Science and Technology Research Project of Jiangxi Provincial Department of Education under grants GJJ2200830 and GJJ190467.

Data availability The data that support the findings of this study are available on request from the corresponding author upon reasonable request.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in

the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Wang Y, Wang Y, Cai J, Lee TK, Miao C, Wang ZJ (2023) Ssd-kd: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images. *Med Image Anal* 84:102693
2. Khan MS, Alam KN, Dhruva AR, Zunair H, Mohammed N (2022) Knowledge distillation approach towards melanoma detection. *Comput Biol Med* 146:105581
3. Elbatel M, Martí R, Li X (2024) FoPro-KD: Fourier Prompted Effective Knowledge Distillation for Long-Tailed Medical Image Recognition. *IEEE Trans Med Imaging* 43(3):954–965
4. Adepu AK, Sahayam S, Jayaraman U, Arramraju R (2023) Melanoma classification from dermoscopy images using knowledge distillation for highly imbalanced data. *Comput Biol Med* 154:106571
5. Liu Q, Yu L, Luo L, Dou Q, Heng PA (2020) Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE Trans Med Imaging* 39(11):3429–3440
6. Gou J, Yu B, Maybank SJ, Tao D (2021) Knowledge distillation: a survey. *Int J Computer Vis* 129(6):1789–1819
7. Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. *Comput Sci* 14(7):38–39
8. Zhao B, Cui Q, Song R, Qiu Y, Liang J (2022) Decoupled knowledge distillation. In: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11953–11962
9. Adriana R, Nicolas B, Ebrahimi KS, Antoine C, Carlo G, Yoshua B (2015) Fitnets: Hints for thin deep nets. *Proc Int Conf Learn Represent* 2(3):1
10. Komodakis N, Sergey Z (2017) Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In: *Proceedings of International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.1612.03928>
11. Kim J, Park S, Kwak N (2018) Paraphrasing complex network: network compression via factor transfer. *Adv Neural Inform Process Syst*. <https://doi.org/10.48550/arXiv.1802.04977>
12. Yang C, An Z, Cai L, Xu Y (2021) Hierarchical self-supervised augmented knowledge distillation. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. <https://doi.org/10.24963/ijcai.2021/168>
13. Jin X, Peng B, Wu Y, Liu Y, Liu J, Liang D, Yan J, Hu X (2019) Knowledge distillation via route constrained optimization. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1345–1354
14. Passalis N, Anastasios T (2018) Learning deep representations with probabilistic knowledge transfer. In: *Proceedings of the European Conference on Computer Vision*, pp. 268–284
15. Park W, Kim D, Lu Y, Cho M (2019) Relational knowledge distillation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3967–3976
16. Tung F, Mori G (2019) Similarity-preserving knowledge distillation. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1365–1374
17. Yim J, Joo D, Bae J, Kim J (2017) A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4133–4141

18. Lee SH, Kim DH, Song BC (2018) Self-supervised knowledge distillation using singular value decomposition. In: Proceedings of the European conference on computer vision, pp. 335–350
19. Liu Y, Cao J, Li B, Yuan C, Hu W, Li Y, Duan Y (2019) Knowledge distillation via instance relationship graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7096–7104
20. Passalis N, Tzelepi M, Tefas A (2020) Heterogeneous knowledge distillation using information flow modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2339–2348
21. Peng B, Jin X, Liu J, Li D, Wu Y, Liu Y, Zhang Z (2019) Correlation congruence for knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5007–5016
22. Zhang R, Yu Y, Shen J, Cui X, Zhang C (2023) Local boosting for weakly-supervised learning. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3364–3375
23. Zhang R, Yu Y, Shetty P, Song L, Zhang C (2022) Prboost: prompt-based rule discovery and boosting for interactive weakly-supervised learning. <https://doi.org/10.18653/v1/2022.acl-long.55>. arXiv preprint [arXiv:2203.09735](https://arxiv.org/abs/2203.09735)
24. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. *Int Conf Mach Learn* 96:148–156
25. Baig MM, Awais MM, El-Alfy ESM (2017) AdaBoost-based artificial neural network learning. *Neurocomputing* 248:120–126
26. Gao Y, Rong W, Shen Y, Xiong Z (2016) Convolutional neural network based sentiment analysis using Adaboost combination. In: International Joint Conference on Neural Networks, pp. 1333–1338
27. Taherkhani A, Cosma G, McGinnity TM (2020) AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. *Neurocomputing* 404:351–366
28. Yang S, Chen LF, Yan T, Zhao YH, Fan YJ (2017) An ensemble classification algorithm for convolutional neural network based on AdaBoost. In: International Conference on Computer and Information Science, pp. 401–406
29. Shakeel PM, Tolba A, Al-Makhadmeh Z, Jaber MM (2020) Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks. *Neural Comput Appl* 32(3):777–790
30. Sun K, Zhu Z, Lin Z (2019) Adagcn: Adaboosting graph convolutional networks into deep models. In: Proceedings of International Conference on Learning Representations. <https://doi.org/10.48550/arXiv.1908.05081>
31. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. In: Proceedings of International Conference on Learning Representations. <https://doi.org/10.48550/arXiv.1609.02907>
32. Zhou H, Song L, Chen J, Zhou Y, Wang G, Yuan J, Zhang Q (2021) Rethinking soft labels for knowledge distillation: a bias-variance tradeoff perspective. *Proc Int Conf Learn Represent*. <https://doi.org/10.48550/arXiv.2102.00650>
33. Li Z, Li X, Yang L, Zhao B, Song R, Luo L, Yang J (2023) Curriculum temperature for knowledge distillation. *Proc AAAI Conf Artif Intell* 37(2):1504–1512
34. Huang F, Ash J, Langford J, Schapire R (2018) Learning deep resnet blocks sequentially using boosting theory. In: International Conference on Machine Learning, pp. 2058–2067
35. Guo Z, Yan H, Li H, Lin X (2023) Class attention transfer based knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11868–11877
36. Oakley, A. DermNet New Zealand. Topical formulations. Updated February.
37. Tschandl P, Rosendahl C, Kittler H (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data* 5(1):1–9
38. Codella NC, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, Halpern A (2018) 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: International Symposium on Biomedical Imaging, 168–172
39. Combalia M, Codella NC, Rotemberg V, Helba B, Vilaplana V, Reiter O, Malvey J (2019) Bcn20000: dermoscopic lesions in the wild. <https://doi.org/10.48550/arXiv.1908.02288> preprint [arXiv:1908.02288](https://arxiv.org/abs/1908.02288)
40. Hossain MI, Elahi MM, Ramasinghe S, Cheraghian A, Rahman F, Mohammed N, Rahman S (2023) LumiNet: the bright side of perceptual knowledge distillation. <https://doi.org/10.48550/arXiv.2310.03669>. arXiv preprint [arXiv:2310.03669](https://arxiv.org/abs/2310.03669)
41. Huang T, You S, Wang F, Qian C, Xu C (2022) Knowledge distillation from a stronger teacher. *Adv Neural Inf Process Syst* 35(33716):33727
42. Shu C, Liu Y, Gao J, Yan Z, Shen C (2021) Channel-wise knowledge distillation for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5311–5320
43. Haberman HF, Norwich KH, Diehl DL, Evans SJ, Harvey B, Landau J, Zingg W (1985) DIAG: a computer-assisted dermatologic diagnostic system—clinical experience and insight. *J Am Acad Dermatol* 12(1):132–143
44. Brooks GJ, Ashton RE, Pethybridge RJ (1992) DERMIS: a computer system for assisting primary-care physicians with dermatological diagnosis. *Br J Dermatol* 127(6):614–619
45. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, Coz D (2020) A deep learning system for differential diagnosis of skin diseases. *Nat Med* 26(6):900–908
46. Hameed SA, Haddad A, Nirabi A (2019) Dermatological diagnosis by mobile application. *Bull Elect Eng Inform* 8(3):847–854
47. Joshi SK (2023) Chaos embedded opposition based learning for gravitational search algorithm. *Appl Intell* 53(5):5567–5586
48. Wang Y, Yu Y, Gao S, Pan H, Yang G (2019) A hierarchical gravitational search algorithm with an effective gravitational constant. *Swarm Evolut Comput* 46:118–139
49. Wang Y, Gao S, Yu Y, Cai Z, Wang Z (2021) A gravitational search algorithm with hierarchy and distributed framework. *Knowl Based Syst* 218:106877
50. Mohammadi A, Sheikholeslam F, Mirjalili S (2023) Nature-inspired metaheuristic search algorithms for optimizing benchmark problems: inclined planes system optimization to state-of-the-art methods. *Arch Comput Methods Eng* 30(1):331–389
51. Bolotnik N, Figurina T (2023) Controllability of a two-body crawling system on an inclined plane. *Meccanica* 58(2):321–336
52. Song X, Song Y, Stojanovic V, Song S (2023) Improved dynamic event-triggered security control for T-S fuzzy LPV-PDE systems via pointwise measurements and point control. *Int J Fuzzy Syst* 25(8):3177–3192
53. Zhang Z, Song X, Sun X, Stojanovic V (2023) Hybrid-driven-based fuzzy secure filtering for nonlinear parabolic partial differential equation systems with cyber attacks. *Int J Adap Control Signal Process* 37(2):380–398
54. Zhang X, He S, Stojanovic V, Luan X, Liu F (2021) Finite-time asynchronous dissipative filtering of conic-type nonlinear Markov jump systems. *Sci China Inform Sci* 64(5):152206