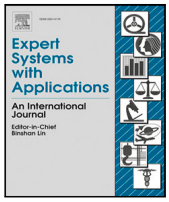
Expert Systems With Applications 228 (2023) 120312

Contents lists available at ScienceDirect

Expert Systems With Applications

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



# DED: Diagnostic Evidence Distillation for acne severity grading on face images

Yi Lin [a,1], Jingchi Jiang [a,1], Dongxin Chen [a], Zhaoyang Ma [a], Yi Guan [a,*], Xiguang Liu [b,*], Haiyan You [b], Jing Yang [b]

[a] Harbin Institute of Technology, Harbin, 150001, Heilongjiang, China
[b] Heilongjiang Provincial Hospital, Harbin, 150001, Heilongjiang, China

ARTICLE INFO

*Keywords:*
Skin disease
Acne grading
Medical image processing
Convolutional neural networks
Knowledge distillation

ABSTRACT

Acne seriously affects people's daily life. Acne severity level grading plays a decisive role in the cure. However, the acne criterion is not unified in the medical field. Most of the current studies explore the application of advanced visual models on acne severity grading but lack the adaptation to the characteristics of acne diagnosis. Meanwhile, some studies propose specifically adapted methods for respectively used acne criteria but cannot be used on other acne criteria. In this study, we propose an acne diagnosis method, Diagnostic Evidence Distillation (DED), that suitably adapts the characteristics of acne diagnosis and can be applied to diagnose under different acne criteria. Firstly, we fully investigate and analyze the commonality of various acne criteria and condense the face acne diagnosis into an unconventional image classification problem based on the diagnostic evidence of type and number of fine-grained lesions distributed on a whole face. Next, we propose the DED framework to adapt the characteristics of acne diagnosis. This framework uses the teacher–student structure of knowledge distillation methods to bring the diagnostic evidence, which is not available for new patients but only in training data, into the diagnosis model. To break the limitation of different criteria, the framework uses convolutional neural networks (CNNs) as backbones to imitate the global estimation of dermatology. We also propose a subtask joint learning for the teacher network to enhance its guidance to the student. The DED is applied to diagnose acne on two datasets ACNE04 and PLSBRACNE01 based on different mainstream acne criteria. The experimental results demonstrate that on both datasets, the DED effectively improves the diagnosis performance, exceeds the state-of-the-art and reaches the diagnostic level of dermatologists. The precision, sensitivity, specificity, Youden Index and accuracy reach 85.31%, 84.83%, 94.66%, 79.48% and 86.06% on the ACNE04 dataset and 69.16%, 65.62%, 88.93%, 54.54% and 67.56% on the PLSBRACNE01 dataset, respectively.

## 1. Introduction

Acne seriously interferes with people's physical and mental health. As one of the most common diseases (Dréno & Poli, 2003), more than 90% of people suffer from acne in their teenage years, because of the hormonal fluctuation of puberty (Ghodsi et al., 2009). Due to the unequal distribution of medical resources, dermatologists are scarce in some remote areas, while the waiting time for medical appointments is too long in prosperous areas (Creadore et al., 2021). It causes many patients to miss the best time to see a doctor, resulting in more than 10% of patients remaining uncured even in adulthood (Ghodsi et al., 2009). More seriously, pigments and scars always remain on the faces of patients without being cured in time (Williams et al., 2012). It leads many patients to feel unhappy, lose confidence, fall into depression, and even avoid social interactions (Koo, 1995). In order to alleviate the physical and mental suffering of so many acne patients, an accurate diagnosis is a vitally important step before establishing an appropriate treatment plan (Hayashi et al., 2008).

The diagnosis of acne, for dermatologists, is grading its severity level based on a specific acne criterion by global estimation or lesion counting (Witkowski & Parish, 2004), but the acne criteria in the medical community have not been unified yet (Ramli et al., 2012). That is, dermatologists in different areas use different criteria to diagnose. The diagnosis rules for different acne criteria will be introduced in

* Corresponding authors.
*E-mail addresses:* linyi@stu.hit.edu.cn (Y. Lin), jiangjingchi@hit.edu.cn (J. Jiang), 21S103150@stu.hit.edu.cn (D. Chen), mazhaoyang2021@163.com (Z. Ma), guanyi@hit.edu.cn (Y. Guan), liuxiguang5714@163.com (X. Liu), youhaiyan0524@qq.com (H. You), 2118yangjing@163.com (J. Yang).
[1] First Author and Second Author contribute equally to this work.

https://doi.org/10.1016/j.eswa.2023.120312
Received 21 December 2022; Received in revised form 11 April 2023; Accepted 28 April 2023
Available online 3 May 2023


**Fig. 1.** Some samples of four severity levels (mild, moderate, severe and very severe from left to right) under the Hayashi criterion.

Section 3. Generally, the prevailing criteria mostly grade acne into four severity levels, *i.e.* mild, moderate, severe and very severe (Cho et al., 2020; Pillsbury et al., 1962; Zhao, 2010) but follow different rules regarding the type and amount of five types of lesions, *i.e.* comedo, papules, pustules, cysts and nodules (Lin et al., 2021; Lin, Jiang, Ma, Chen, Guan, You et al., 2022; Ramli et al., 2012). That is to say, the number and type of various lesions distributed on the face form two aspects of diagnostic evidence.

The diagnostic evidence for grading the severity is almost imperceptible. Some samples of the four severity levels (under Hayashi criterion) and five types of lesions are displayed in Figs. 1 and 2. From the figures, as the disease progresses, the number of facial lesions increases, and more types of lesions occur. In these mild cases, only a few scattered lesions are on the face. While in very severe cases, the lesions form scattered or continuous areas and are darker. At the same time, the lesions are significantly tiny compared to the face. And some lesions of different types are much too similar. The average person without dermatologic knowledge can hardly tell the difference between some of these five lesion types. For example, some of these comedones and pustules show white raised dots. Some of these papules, pustules and nodules all form red irregular bumps. Their differences are almost impossible to capture by ordinary people.

The non-uniform acne criteria and imperceptible diagnostic evidence pose a severe challenge for developing intelligent acne diagnostic methods. Generally, for computers, acne severity grading is an image classification issue. But acne severity grading is somewhat different from conventional classification. On the one hand, for the varying acne criteria, conventional image classification methods learn to recognize the class in a certain task, that is, the rule for recognizing is unchanging. Although the conventional image classification methods


**Fig. 2.** Some samples of five types of lesions (comedo, papule, cyst, pustule and nodule from top to bottom).

can diagnose under different criteria by retraining the models on specific criterion-based image data. And simultaneously, various excellent image classification methods have achieved amazing results surpassing humans on general tasks (He et al., 2015). But these general classification methods without specific adaptation to acne are difficult to obtain satisfactory diagnostic accuracy. Meanwhile, once a diagnostic model is developed for adapting a specific criterion, it will not be able to diagnose under other criteria, such as the method proposed by Wu et al. (2019). On the other hand, for the imperceptible diagnostic evidence, the conventional classification methods recognize the pattern of different classes almost based on the edge, texture, color, and some other characteristics of one whole large object (Lin, Jiang, Ma, Chen, Guan, You et al., 2022). But the acne diagnosis is based on the type and amount of many significantly tiny lesions distributed on a whole face. Many researchers also have proposed excellent skin diagnosis methods using the rule-based or deep learning-based classification paradigm with the development of various methods and models (Alenezi et al., 2022; Alizadeh et al., 2020; Camacho-Gutiérrez et al., 2022; Ghalejoogh et al., 2020; Hameed et al., 2020; Lin, Jiang, Ma, Chen, Guan, Liu et al., 2022; ul Ain et al., 2022). However, these skin diseases mostly do not have acne-like characteristics.

Current intelligent acne diagnosis studies mostly develop their systems using existing excellent visual classification models that can be adopted on different acne criteria but lack the adaption to the characteristics of acne diagnosis or propose novel models well-adapted but over-adapted to a specific criterion that cannot be used on other acne criteria (Lim et al., 2019; Malgina & Kurochkina, 2021; Seité et al., 2019; Wang et al., 2022). For example, the research of Lin et al. (2021) and Lin, Jiang, Ma, Chen, Guan, You et al. (2022) proposed a unified framework by imitating the global estimation of dermatology diagnosis. Their framework is the first one that can diagnose under different acne criteria. But their framework lacks the adaption to acne criteria. Wu et al. (2019) proposed a novel method that jointly learns the severity level grading and lesion counting. This strategy is well-adapted to the Hayashi criterion (Hayashi et al., 2008) but cannot be used on other criteria. Wang et al. (2022) proposed a procedure that first segment lesions, then classify the lesions and grade the severity finally. This method is also well-adapted to some of the acne criteria that consider the type of lesions but cannot be used on the acne criteria regardless of the type. Therefore, there are few methods that adapt to the characteristics of acne diagnosis as well as can be used on different criteria. We believe that reasonable integration of diagnostic evidence into the classification method to adapt to the commonality of various acne criteria will improve the diagnostic performance as well as be used to diagnose under different acne criteria. Regardless of the acne criteria, the diagnosis is based on lesion cases, only with different grading rules considering the number and type of lesions. The lesion areas can be considered as diagnostic evidence to be incorporated into the diagnostic model. On the one hand, the lesion areas can indicate the key areas of the model observation and enhance its recognition of the lesion type which is one aspect of diagnostic evidence. On the other hand, the areas of lesions also contain the number of lesions, which is the other aspect of diagnostic evidence.

In fact, fusing diagnostic evidence of lesion areas into the diagnostic process is difficult to realize. For the classification model, to fuse the lesion areas into the diagnosis, the lesion areas should be one piece of input of the model. However, if lesion areas can be obtained, the patient has been diagnosed. For a patient in need of medical attention, the lesion areas are unknown. Similarly, for a diagnostic model that fuses diagnostic evidence, the lesion areas can only be obtained in the dataset well-annotated by dermatologists. When a new image sample needs to be diagnosed, the lesion areas are unknown. In the research of Lin, Jiang, Chen et al. (2022), they try to use segmentation methods to extract the lesion areas first and then fuse the area map with its corresponding face image to the downstream diagnosis model. But this process will cause errors to accumulate. If an upstream segmentation error occurs, then the downstream diagnosis receives the wrong area of lesions.

In this study, we propose an acne diagnosis method for acne severity grading on face images named Diagnostic Evidence Distillation (DED) which suitably adapts the characteristics of acne diagnosis and can be applied to diagnose under different acne criteria. Firstly, we fully investigate and analyze the commonality of various acne criteria and condense the face acne diagnosis into an unconventional image classification problem based on the diagnostic evidence of type and number of fine-grained lesions distributed on a whole face. Next, we propose the DED framework to adapt the characteristics of acne diagnosis. The DED uses the teacher–student structure of knowledge distillation

methods to bring the diagnostic evidence, which is not available for new patients but only in training data, into the diagnosis model. The teacher model learns the diagnosis process from an image with its real diagnostic evidence which is only available in training data. A student model learns to grade the severity level from only face images and simultaneously receives guidance from the evidence-fused teacher during the training stage. And the student model is the final diagnosis model, diagnosing new samples only with face images. To break the limitation of different criteria, the framework uses convolutional neural networks (CNNs) as backbones to imitate the global estimation of dermatology diagnosis. In addition, a joint learning subtask that learns lesion counting on the teacher network is also proposed to enhance the mining of tiny lesions and the guidance to the student.

The main contributions of this work are as follows:

- This work condenses face acne diagnosis of ununified criteria into an unconventional image classification problem based on a specific rule of the type and amount of fine-grained lesions distributed on a whole face.
- This work proposes an acne diagnosis method for acne severity grading on face images which suitably adapts the characteristics of acne diagnosis and can be applied to diagnose under different acne criteria.
- The proposed DED framework is a new knowledge distillation paradigm that can bring meaningful features, not available in new samples but in training data, into downstream inference.
- We propose a joint learning subtask for the teacher model to enhance its recognition of key areas and its guidance to the student model.

We conduct comprehensive experiments using DED to diagnose acne on two datasets ACNE04 and PLSBRACNE01 based on different mainstream acne criteria. The experimental results demonstrate that on both datasets, the DED effectively improves the diagnosis performance. The whole framework exceeds the state-of-the-art and reaches the diagnostic level of dermatologists. The precision, sensitivity, specificity, Youden Index and accuracy reach 85.31%, 84.83%, 94.66%, 79.48% and 86.06% on the ACNE04 dataset and 69.16%, 65.62%, 88.93%, 54.54% and 67.56% on the PLSBRACNE01 dataset, respectively. The code for this work is opened on GitHub.[2]

## 2. Related works

### 2.1. Intelligent acne diagnosis method

Recently, intelligent acne diagnosis studies mostly explore existing classification models on acne diagnosis or propose novel models well-adapted to a specific criterion (Lim et al., 2019; Malgina & Kurochkina, 2021; Seité et al., 2019; Wang et al., 2022). The existing models can be used on different criteria but lack adaption to the characteristics of acne diagnosis. Meanwhile, the models well-adapted to specific criteria cannot be used on different criteria.

The existing acne intelligent diagnosis methods are basically developed according to specific criteria or by imitating dermatologists. Dermatologists usually use two separate methods to diagnose acne: criterion-based lesion counting and experience-based global estimation (Wu et al., 2019). Lesion counting is usually used by junior dermatologists. The type and number of lesions are counted first, and then the diagnosis is made according to a specific criterion. Instead, experienced dermatologists often use global estimation. Observing the entire face and key lesions gives a diagnosis based on the experience from a specific criterion (Lin et al., 2021). Some criteria determine severity only by the number of lesions, such as Hayashi et al. (2008), while others consider both types and amounts, such as Pillsbury et al. (1962). Although the criteria vary, rich diagnostic experience under one criterion can still support global estimation for diagnosis. So many deep learning methods learn to diagnose by observing the whole image, which is mining the experience of global estimation.

Some studies use existing models for global estimation under only one criterion due to data limitations such as the works of Lim et al. (2019) and Malgina and Kurochkina (2021). These works explored the application of using existing deep learning methods to diagnose acne on their own collected datasets based on different criteria. The results show that deep learning can achieve a good diagnosis under the criteria they use, and provide a good basis for other research.

Some studies consider global estimation and lesion counting, but cannot be used universally among different criteria. Seité et al. (2019) developed an intelligence algorithm for acne grading from smartphone photographs. Their acne grading method uses the Global Acne Severity Scale (GEA) widely used in Europe. The algorithm contains two stages, acne segmentation, and severity classification. The acne segmentation stage contains three tasks, detecting and segmenting the retentional lesions, inflammatory lesions, and postinflammatory hyperpigmentation (PIHI). Then the severity classification stage uses the results of retentional lesions and inflammatory lesions to grade the severity level based on the GEA scale. This algorithm first extracts some specific lesions guided from their criterion as diagnostic evidence. Then mine the diagnosis rule with the evidence to grade the final severity level. Their algorithm reaches a high accuracy with a performance close to that of dermatologists. However, for the two-stage approach, errors should occur in evidence mining. Then, the error-contained evidence will misguide the downstream model to learn the diagnosis rule with deviation. In addition, this algorithm uses many models to recognize each type of lesion. This approach can be modified by adding or reducing the detection models to adapt to other criteria. But this modification will take a lot of time. Wu et al. (2019) proposed a framework jointly learning the severity level grading and lesion counting. This diagnosis framework is developed based on the Hayashi criterion (Hayashi et al., 2008). The Hayashi criterion considers only the number of lesions to determine the severity. Their method poses lesion counting as a joint task that shares the same model parameters with the grading task that perfectly adapts the acne criterion. And the performance of their framework also achieves the diagnostic level of dermatologists. But this adaption makes the framework difficult to learn and diagnose under other criteria. Wang et al. (2022) in their research develop a cell phone app for facial acne severity assessment. The diagnosis procedure firstly segments all lesion areas, then classifies these lesions and grades the severity finally. This method performs significantly on the criteria considering both the type and the number of lesions but cannot be used on the criteria regardless of the type.

Some studies diagnose under different criteria only by global estimation but ignore lesion counting. Lin et al. (2021) and Lin, Jiang, Ma, Chen, Guan, You et al. (2022) developed a unified acne grading framework. This framework imitates the global estimation of dermatologists containing two stages, key information enhancement and global local fusion grading. The key information enhancement is an image preprocessing stage. It detects and cuts the face into a box and blackens the background area around the face. Then the global local fusion grading uses a CNN as the backbone to fuse the global observation of the image and local feature from selection to jointly predict the severity. Due to the imitation of global estimation, this framework can be applied to diagnose under different criteria. Their modification of image preprocessing and global local fusion also improve their framework effectively. The performance reaches the state-of-the-art and diagnostic level of dermatologists. However, their method adapted the acne diagnosis with only the approach of global estimation but ignored some more valuable diagnostic evidence.

Overall, there are few methods adapted to the characteristics of acne diagnosis and can be applied to different acne criteria. Lin, Jiang, Chen et al. (2022) proposes a framework that adapts characteristics of acne

[2] https://github.com/linyi0604/DED.

diagnosis as well as can be used on different criteria. They first segment all lesion areas, then fuse the area map with its corresponding face image to train the downstream diagnosis model. But this method may lead to error accumulation. When data is insufficient, the segmentation model will not be trained well, and the downstream diagnosis model will receive the incorrect lesion areas.

We believe that a suitable way to adapt to the characteristics of acne diagnosis will improve the diagnostic performance as well as be able to diagnose under different criteria. To break the limitation of different criteria, our method first uses a CNN as the image feature extractor to realize the global estimation diagnosis. Then, to complement the global estimation, the diagnostic evidence is condensed by analyzing the commonalities of various acne criteria. To avoid the accumulation of errors in fusing diagnostic evidence, our method does not fuse the diagnostic evidence by predicting it for sending downstream. Instead, our method uses the knowledge distillation approach to transform the diagnostic evidence from a teacher to a student. The ground truth diagnostic evidence from the annotation is integrated into a teacher network, and a student network learns to grade the severity from only images as well as get guidance from the teacher network. To cover various acne criteria, the diagnostic evidence is extracted as the lesion areas. On the one hand, the lesion areas indicate the key areas to be carefully observed for the teacher model to enhance its recognition of different types of lesions which is one aspect of diagnostic evidence. On the other hand, the lesion areas directly contain the number of lesions which is the other aspect of diagnostic evidence. This way, our method not only fuses the diagnostic evidence of both global estimation and lesion counting in various criteria but also can be applied to diagnose under various criteria.

### 2.2. Image classification

Acne diagnosis is a kind of unconventional image classification problem, which is difficult to be modeled effectively by general classification methods.

Image classification tasks shine in the field of computer vision, almost surpassing humans (He et al., 2015). Since the sudden intrusion of deep learning, a large number of state-of-the-art methods have been continuously refreshed every year. Various CNNs bear the brunt of the image classification list for many years (He et al., 2016; Hu et al., 2020; Simonyan & Zisserman, 2014; Tan & Le, 2019). This series of methods benefit from the combination of convolutional and fully connected networks. Trainable filters identify edge and texture feature more efficiently, and fully connected networks combine features to analyze image categories.

Later, the transformer for modeling time series sequences in natural language processing was introduced into computer vision as a replacement for CNNs, which once became a hotspot (Dosovitskiy et al., 2021). Subsequently, a large number of Vision Transformer-based methods emerged in image classification (Touvron et al., 2021; Wang et al., 2021; Yang et al., 2021; Yuan et al., 2021). This series of methods cuts the image into patches in order, first extracts the features of the patches, and then calculates the attention among them. This simplifies the pressure of deep convolution to extract features layer by layer and models the complementarity of various features for target recognition. Recently, some scholars have even combined convolution and attention to achieve better performance (Dai, Liu, Le et al., 2021). All of these methods maximize the recognition and classification of one large object in an image.

However, in the diagnosis of acne images, the overall classification is often determined according to the number and types of small objects distributed on a whole area. Almost none of the existing methods can exploit such features. In this situation, bringing more meaningful features to CNN models suitable in this case will be helpful. We propose incorporating diagnostic evidence into CNNs to improve the identification ability of multi-specific region combinations. But the diagnostic evidence is unavailable in new patients but only in training data. So we use the teacher–student structure of knowledge distillation to fuse the diagnostic evidence from the teacher model into the student model. The teacher model leans from diagnostic evidence in the training data and the student learns only from face images as the diagnosis model.

### 2.3. Knowledge distillation

Knowledge distillation was proposed and improved for deep learning model compression and achieved impressive results (Alkhulaifi et al., 2021; Blakeney et al., 2021; Cui et al., 2021). Later on, it was applied to solve practical compression problems in various fields quickly (Dai, Liu, Li et al., 2021; Fan et al., 2021; Fu et al., 2021). These works all obtained smaller models whose performances were close to the teachers' through various distillation methods and achieved excellent results. At present, knowledge distillation is widely used to obtain a smaller, faster and well-performing model for practical scenarios in the case, where the sample volume is large enough to support a very large and perfect-performing model with an accuracy close to 100% but the large model is not suitable for application.

The core of knowledge distillation is to use the "genius" teacher network whose size is huge to guide a lighter "clumsy" student network, letting the intermediate features extracted from the student approximate the teacher ones. This prevents the student from being unable to capture the efficient pattern and improves the performance of the student network. Indeed, expanding the teacher size improves its performance but only when the data volume is large enough. Instead, when the data volume is insufficient, a large model is hard to obtain excellent enough performance, the practical application is not available, compressing models becomes meaningless, and knowledge distillation is not widely used.

Inspired by knowledge distillation transforming the extracted knowledge from a large model to a small model, we incorporate helpful features which are only available in training data to train a teacher network and then transfer the feature-fused experience to a student network when larger models cannot achieve better performance. In this work, the transformed features are diagnostic evidence, that is, the lesion cases distributed on the whole face.

## 3. Acne criteria analysis

The acne criteria are not unified. In this section, we fully investigate and analyze the commonality of various acne criteria and condense the face acne diagnosis into an unconventional image classification problem based on the diagnostic evidence of type and number of fine-grained lesions distributed on a whole face.

From the perspective of diagnosis approaches, acne criteria are divided into two categories: global estimation and lesion counting. From the perspective of diagnostic evidence, acne criteria have two mainstream separately considering only the number of lesions and both the type and number of lesions. The criteria using the global estimation approach to diagnose always simultaneously consider both the type and number of lesions but without the precise number of each type of lesion. And the criteria using the lesion counting approach mostly consider only the number of lesions regardless of the type of lesions. In addition, experienced dermatologists generally use global estimation diagnosis although some criteria use the lesion counting approach, while junior dermatologists or general practitioners generally use lesion counting (Lin, Jiang, Ma, Chen, Guan, You et al., 2022; Wu et al., 2019). Various acne criteria and their measurements have been introduced in some previous studies (Cho et al., 2020; Lin, Jiang, Ma, Chen, Guan, You et al., 2022; Ramli et al., 2012; Witkowski & Parish, 2004). Here we have a brief review.

For the criteria of the global estimation approach, the Pillsbury criterion (Pillsbury et al., 1962) is one of the most popular diagnostic criteria in modern medicine. It is the first known acne criterion

**Table 1**
The criterion description and sample quantity of each severity level for two datasets: Hayashi-based ANCE04 and Pillsbury-based PLSBRACNE01.

| Dataset (criterion) | ACNE04 (Hayashi) | | PLSBRACNE01 (Pillsbury) | |
|---|---|---|---|---|
| Severity level | Criterion description (Number of lesions) | Sample quantity | Criterion description | Sample quantity |
| Mild | 1~5 | 513 | Comedones and occasional small cysts confined to the face | 58 |
| Moderate | 6~20 | 633 | Comedones with occasional pustules and small cysts confined to the face | 47 |
| Severe | 21~50 | 182 | Many comedones and small and large inflammatory papules and pustules, more extensive but confined to the face | 65 |
| Very Severe | >50 | 129 | Many comedones and deep lesions tending to coalesce and canalize, and involving the face and the upper aspects of the trunk | 30 |
| Total | – | 1457 | – | 200 |

The Hayashi criterion determines the four levels according to the number of lesions regardless of lesion type.
The Pillsbury criterion determines the four levels according to the number and type combination of lesions.

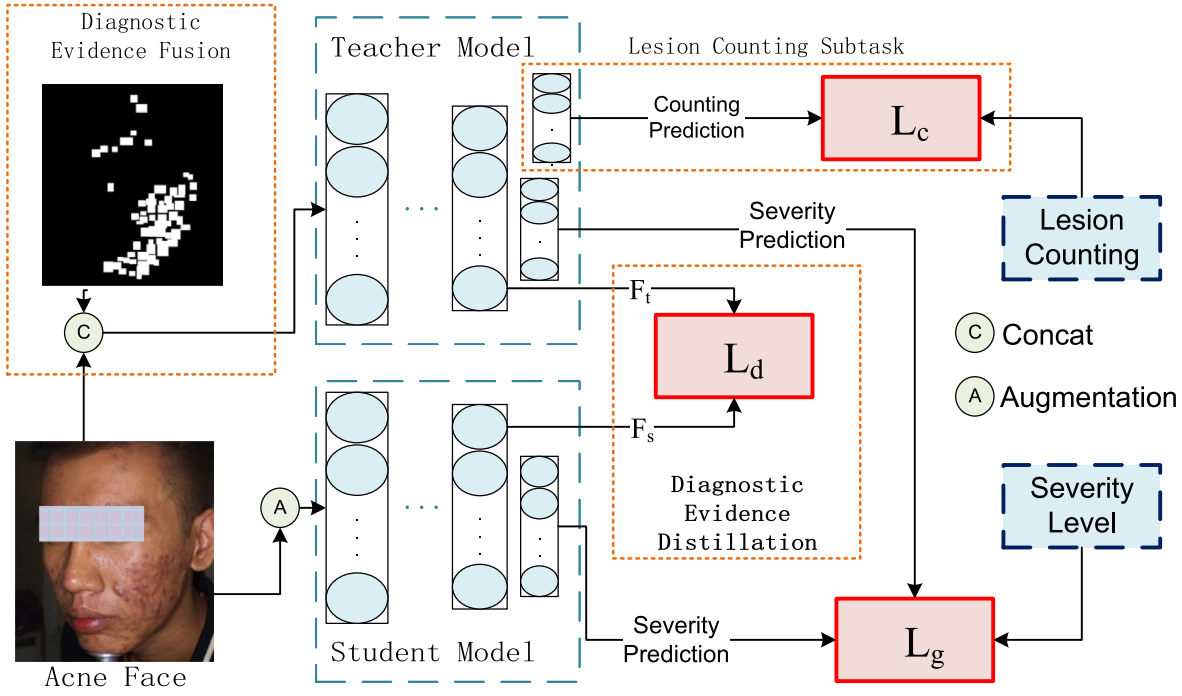



**Fig. 3.** An overview of our diagnostic evidence distillation framework for face acne severity grading. $F_t$: Teacher feature, $F_s$: Student feature, $L_d$: Distillation loss, $L_g$: Grading loss and $L_c$: Counting loss.

proposed by Pillsbury et al. (1962) in their published dermatological textbook. This criterion grades the acne severity level based on a global estimation of the type and number of lesions from observation but does not provides the precise number of each type of lesion. Its description can be found in the right half of Table 1. This criterion uses the global estimation approach to diagnose and simultaneously considers both the type and number of lesions. There are also other alternative grading schemes, such as the one proposed by James and Tisserand (1958). But these criteria are significantly similar to the Pillsbury criterion and not as influential as the Pillsbury criterion.

The lesion counting approach to grading acne severity was first initiated by Witkowski and Parish (1999). Subsequently, various diagnostic criteria using lesion counting have been proposed such as those from researches (Doshi et al., 1997; Dréno et al., 1999; Frank, 1971; Hayashi et al., 2008; Lucky et al., 1996). The grading rules of these acne criteria based on lesion counting are similar. A description of the Hayashi criterion can be found in the left half of Table 1. These criteria grade the acne severity level by counting a precise number of lesions distributed on the whole face (some also contain chest and back) regardless of any type of lesions. There are also some rare grading criteria using lesion counting but different from the above ones. For example, the criterion proposed by Michaelson et al. (1977) uses a type-weighted number summation to calculate the final severity level.

In summary, the acne criteria are not unified. Dermatologists in different areas use different acne criteria to diagnose. The acne criteria form two mainstreams. The one stream of criteria using global estimation to diagnose simultaneously considers both the number and type of lesions but not the precise number. The other stream using lesion counting to diagnose mostly considers only the number of lesions regardless of any type of lesions. No matter which diagnosis approach,

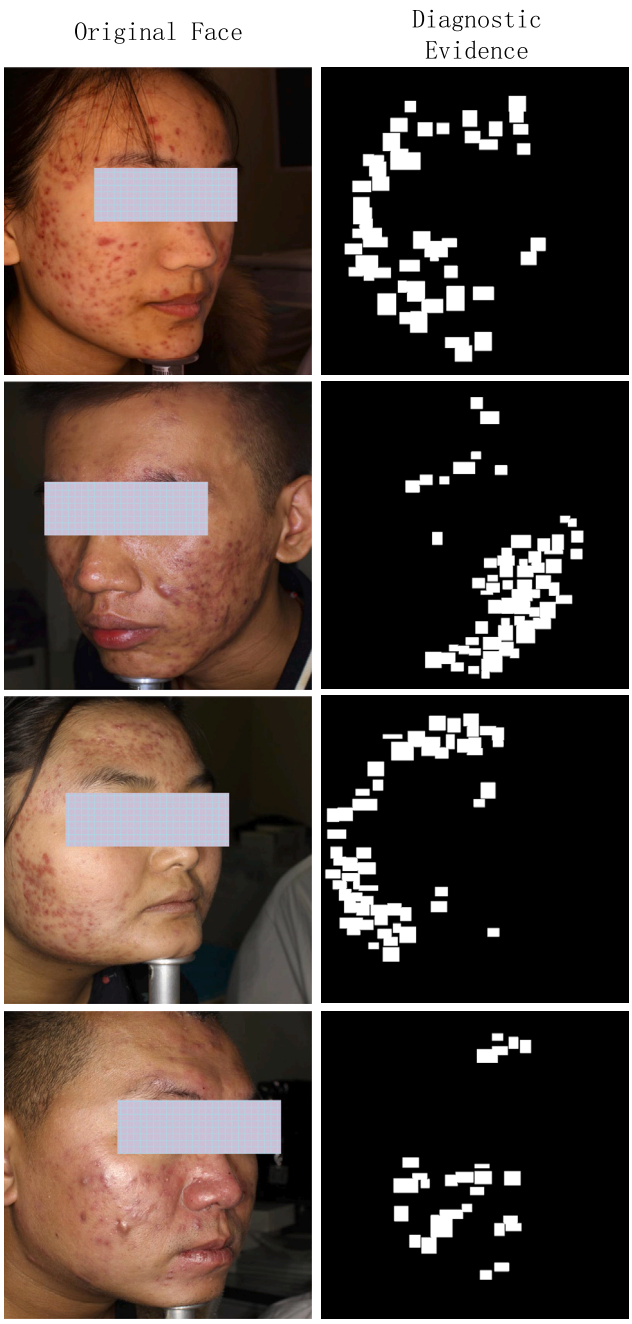



**Fig. 4.** Some examples of original face images with their corresponding constructed diagnostic evidence of lesion areas.

most of the acne criteria set four levels: mild, moderate, severe and very severe.

This work aims to develop a unified acne diagnosis framework that can be applied to diagnose under both two mainstream acne criteria. Guided by dermatologists, our framework considers both diagnosis approaches of global estimation and lesion counting. Furthermore, we propose to fuse diagnostic evidence into the diagnosis process. To break the limitation among different acne criteria, the diagnostic evidence covers both two mainstream criteria and does not over-adapt to any specific criteria.

## 4. Method

An overview of the diagnostic evidence distillation framework is displayed in Fig. 3. This framework uses the teacher–student structure of knowledge distillation. The core components of this framework are the lesion location fusion, lesion counting subtask and diagnostic evidence distillation. The diagnostic evidence fusion combines the image and lesion areas into the teacher network. This strategy indicates the key areas to be observed for the teacher net to recognize the patterns of different types of lesions which is one aspect of diagnostic evidence. The lesion areas also contain the number of lesions which is the other aspect of lesions. The lesion counting subtask further enhances the teacher model to recognize tiny lesions. The diagnostic evidence distillation transforms evidenced-fused knowledge into the student model. The student model is the final diagnosis model. During the training stage, it learns to grade the severity from input images as well as obtains guidance from the evidence-fused teacher network. During the diagnosis stage, it predicts the severity level only from input images.

### 4.1. Teacher network

The teacher network aims to learn the process of acne diagnosis by fusing the diagnostic evidence. Two strategies are proposed to enhance the teacher network, diagnostic evidence fusion and lesion counting subtask.

#### 4.1.1. Diagnostic evidence construction

As described above, the diagnosis of acne is an image classification problem. But due to its characteristics, the diagnosis of acne is somewhat different from the conventional image classification problem based on the edges, texture, color and some other features of one whole target. Instead, the problem of acne diagnosis is based on the type and number of very tiny lesions distributed on a face according to different criteria. Some criteria consider the number of lesions and others consider both the number and type of lesions as evidence. At the same time, excessive adaptation to one acne criterion will cause the method being not able to diagnose under other criteria. To cover various mainstream acne criteria without over-adapting, the lesion areas are considered as the diagnostic evidence for our framework. On the one hand, lesion areas can indicate key areas for the diagnosis model to recognize the patterns of various lesions. On the other hand, the lesion areas directly contain the number of lesions. Therefore, constructing a rational structure to incorporate the lesion areas as diagnostic evidence into the diagnosis of the teacher network is a vital issue.

Inspired by the fields of image inpainting and image segmentation, we construct a mask matrix as the diagnostic evidence for every face image where the pixel values of lesions are set as 1 and the rest pixel values are set as 0. All lesions were identified and annotated by dermatologists. An example of the diagnostic evidence and its corresponding image is shown on the left side of Fig. 4. This way, the white areas with pixel values of 1 are the area that the diagnostic process should focus on. Separate white areas contain the type of different lesions. Whole white areas contain the number of all lesions. The constructed diagnostic evidence is a one-channel gray image.

#### 4.1.2. Diagnostic evidence fusion

To fuse the diagnostic evidence into the diagnosis process of the teacher network, the diagnostic evidence will be concatenated with its corresponding original face image together to form a four-channel image as the input sent into the teacher network. To ensure that the four-channel images can be processed as inputs, we add a channel adaption layer in front of the selected backbone networks as the new input layer of the teacher model. The channel adaption layer is a convolutional layer with 3 filters of $1 \times 1$ size. Processed by the channel adaption layer, the four-channel input images are mapped into three-channel feature maps. Then, the processed three-channel feature maps

will be input into the backbone network. As a part of the teacher model, the parameters of the channel adaption layer will be trained together with the backbone network.

Indeed, this approach of fusing lesion areas as diagnostic evidence does not directly integrate the diagnostic basis into the teacher network. But the implicit fusion ensures that our framework is applicable to diagnose under different acne criteria.

#### 4.1.3. Lesion counting subtask

The key to acne diagnosis is to closely observe the lesion areas. The diagnostic evidence fuses the lesion areas as evidence into the diagnosis process of the teacher network. At the same time, the lesion counting subtask further strengthens the teacher network's focus on diagnostic evidence areas. Specifically, the teacher network learns to predict the lesion counting results as a branching task while diagnosing acne.

In fact, the lesion counting subtask is predicting the number of lesions regardless of the type of lesions. But the lesion counting learning is slightly different from the conventional regression. On the one hand, the results of lesion counting are discrete positive integers, while the results of conventional regression are continuous values. On the other hand, even for patients with the same severity level, the difference in the number of facial lesions is also dramatic. To adapt these special characteristics in acne diagnosis, one-hot distribution learning is adopted instead of continuous regression learning. In particular, all counting ground truths are converted to one-hot distribution. For a counting label, each dimension refers to a specific lesion number. To learn the lesion counting of one-hot distribution, a fully connected counting layer is added to the teacher network parallel with the grading layer and shares all the previous network parameters with the grading layer. We believe that the use of one-hot distribution learning and shared weights can enhance the teacher network's attention to minor lesions and avoid it falling into local observation. This way, the lesion counting of one-hot distribution learning optimizes the counting loss $L_c$ as:

$$L_c = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^{(c)} \log(p_i^{(c)}) \tag{1}$$

where $N$ denotes the number of the total sample and $C$ refers to the maximum counting number. $y_i^{(c)}$ is 0 or 1 indicates the counting of sample $i$ is $c$ or not. $p_i^{(c)}$ is the logit softmax output corresponding to the counting $c$ for sample $i$.

#### 4.1.4. Grading learning

Acne diagnosis of severity grading falls under the classification. So the teacher network optimizes the commonly used cross-entropy as grading loss $L_{g,t}$ to learn the acne diagnosis as:

$$L_{g,t} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L} y_i^{(l)} \log(p_i^{(l)}) \tag{2}$$

where $N$ denotes the total sample and $L$ refers to the number of severity levels. $y_i^{(l)}$ is the ground truth of 0 or 1 indicating the severity level of the sample $i$ belongs to $l$ or not. $p_i^{(l)}$ is the logit softmax output corresponding to the severity $l$.

#### 4.1.5. Network design

The teacher network uses an advanced and commonly used pre-trained visual network as its backbone. And the backbone selection will be discussed in the experiment. To process the evidence-fused four-channel input, a channel adaption layer is added in front of the original input layer as the new input layer. The channel adaption layer uses a $1 \times 1$ convolutional layer of three filters. To adapt the severity grading and lesion counting tasks, the original output layer is removed and replaced by dual parallel fully connected layers. One of the fully connected layers is the grading layer with the same number of neural units to the severity levels. The other one is the counting layer with the same number of neural units to the maximum lesion number.

This way, the input of the teacher network is the fusion of an original face image and its corresponding diagnostic evidence. Because the grading layer and the counting layer are a pair of parallel output layers at the end of the teacher network, when samples are input into the teacher model, the grading prediction and the counting prediction will output simultaneously after the network forward computing.

#### 4.1.6. Overall optimizing of the teacher model

During training, the teacher network is first trained to achieve an optimal acne grading performance. The optimization goal of the teacher network is:

$$L_t = \lambda_g L_{g,t} + \lambda_c L_c \tag{3}$$

where $\lambda_g$ and $\lambda_c$ are two hyperparameters used to reconcile the weight of severity grading and lesion counting.

### 4.2. Student network

The student network is the final diagnosis model. During the training stage, the student network learns to grade the acne severity level from original face images as well as obtains guidance from the diagnostic evidence-fused teacher network by distillation learning. During diagnosis, the student network predicts the severity level from only the original face images.

#### 4.2.1. Distillation learning

Distilling learning is a strategy that effectively guides the learning process of the student network with the knowledge from the diagnostic evidence-fused teacher model. Since the diagnostic evidence of the new samples to be diagnosed, including the number, type and area of lesions, is unknown, the diagnostic evidence cannot be directly fused into the diagnosis model. By using knowledge distillation, the diagnostic evidence can be obtained indirectly from the knowledge learned by the teacher network without any other modification of the diagnosis model.

There are various distillation methods proposed in recent years. They use different approaches but all minimize the difference in hidden output distributions between the student and the teacher using a certain distance metric. In this work, our framework employs an effective and direct feature distillation strategy where features from the first previous layer of the output layer for both teacher and student networks are organized to distill using L1 distance. This simple feature distillation forcing the student features to approximate the teacher features as:

$$L_d = \frac{1}{K}\sum_{k=1}^{K} \|f_s(k) - f_t(k)\|_1 \tag{4}$$

where $L_d$ denotes the distillation loss, $f_s$ is the hidden feature from the student network, $f_t$ is the hidden feature from the teacher network and $k$ refers to the dimension of the features.

#### 4.2.2. Grading learning

Different from the teacher network, the student network only learns to grade the severity level without any other subtask as the final diagnosis model. But also similar to the teacher network, the student network learns to grade the severity level by optimizing the commonly used cross-entropy loss $L_{g,s}$ as:

$$L_{g,s} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L} y_i^{(l)} \log(p_i^{(l)}) \tag{5}$$

where $N$ denotes the number of the total sample and $L$ refers to the maximum severity level. $y_i^{(l)}$ is the ground truth of 0 or 1 indicating the severity level of the sample $i$ belongs to $l$ or not. $p_i^{(l)}$ is the logit softmax output corresponding to the severity $l$.

#### 4.2.3. Network design

Similar to the teacher network, the student network also uses a pretrained visual classification network as its backbone. The selection of the backbone will be discussed in the experiment section. To adapt the severity grading, the original output layer of the backbone is removed and replaced by a grading layer. The grading layer is a fully connected layer with the same number of neural units as the severity numbers. Due to no modifications on the student network, it can be easily generalized to apply under different acne criteria.

#### 4.2.4. Overall optimizing

During training, the teacher network is first trained to achieve an optimal acne grading performance. Then, train the student network under the guidance of the teacher network whose weight parameters are fixed.

The optimization goal of the student network is:

$$L_s = \delta_g L_{g,s} + \delta_d L_d \tag{6}$$

where $\delta_g$ and $\delta_d$ are the other two hyperparameters to reconcile the weight of severity grading and distillation from the teacher network.

Note that the grading process for both the student and teacher models is classification optimization, so the same loss function is used. But in the optimization process are two different network optimization goals, so two same items are given for the student network and teacher network separately.

### 4.3. Augmentation

Image augmentation is a common strategy in image methods. In this work, image augmentation technology is also used to improve the generalization ability of the model. In most previous studies, the image augmentation strategy is only used on a single classification model. This study involves two image classification networks, the teacher model and the student model.

In this work, we perform the image augmentation only on the inputs of the student network, specifically including random resize, cropping, rotation and flipping but not on the teacher inputs. This approach forces the student network to mine similar features when observing different variations of the same sample and improves the generalization ability of the student network.

## 5. Experiment settings

### 5.1. Datasets

Two face acne datasets, ACNE04 and PLSBRACNE01, are adopted to validate our framework and conduct all the experiments. The two datasets are annotated by dermatologists under different acne criteria. A detailed description of the two datasets and their based acne criteria can be found in Table 1.

ACNE04: This dataset is the first publicly available face acne dataset collected and annotated based on Hayashi Criterion (Hayashi et al., 2008) by Wu et al. (2019). The Hayashi Criterion (Hayashi et al., 2008) determines the severity level by only considering the number of lesions regardless of the type. The annotation results for this dataset include the locations of all skin lesions, the total number of lesions and the severity level of each image. It contains 1457 face images with 513, 633, 182 and 129 samples for each of the four severity levels, respectively. The dataset has been split into 80% training set and 20% test set. The numbers of training samples for the four levels are: 410, 506, 146 and 103, and the numbers of test samples are 103, 127, 36, and 26. A detailed description of the ACNE04 dataset can be found in the left half of Table 1.

PLSBRACNE01: This dataset was collected and manually annotated with the help of dermatologists by Lin et al. (2021). The annotation for this dataset is based on the Pillsbury Criterion (Pillsbury et al., 1962) which is widely used clinically in China. The Pillsbury Criterion (Pillsbury et al., 1962) determines the severity level by considering both the number and the type of lesions. The annotation results include the locations and types of all skin lesions and the severity level of each sample. The dataset totally contains 200 samples with 58, 47, 65 and 30 samples for each of the four severity levels, respectively. It has been split into 80% training set and 20% test set. A detailed description of the PLSBRACNE01 dataset can be found in the right half of Table 1.

### 5.2. Evaluation metrics

Some widely used evaluation metrics are adopted in this work including accuracy, precision, sensitivity, specificity and Youden Index. Accuracy and precision are two popular evaluation metrics for the classification problem. Sensitivity, specificity and Youden Index are three common evaluation metrics for diagnosis.

The accuracy is calculated as:

$$ACC = \frac{N_c}{N} \times 100\% \tag{7}$$

where $N_c$ is the number of samples that are correctly predicted by the framework. $N$ is the total number of all samples that are predicted by the framework.

The precision is calculated as:

$$PRE = \frac{TP}{TP + FP} \times 100\% \tag{8}$$

where $TP$ is the number of true positive samples, that is, the number of positive samples that are correctly predicted by the framework. $FP$ is the number of false positive samples, that is, the number of samples that are wrongly predicted into positive by the framework.

The sensitivity is calculated as:

$$SEN = \frac{TP}{TP + FN} \times 100\% \tag{9}$$

where $FN$ is the number of false negative samples, that is, the number of samples that are wrongly predicted as negative by the framework.

The specificity is calculated as:

$$SPE = \frac{TN}{TN + FP} \times 100\% \tag{10}$$

where $TN$ is the number of true negative samples, that is, the number of negative samples that are correctly predicted by the framework.

The Youden Index is calculated as:

$$YI = SEN + SPE - 1 \tag{11}$$

The Yoden Index is a harmonic assessment of sensitivity and specificity.

### 5.3. Implementation details

By comparing the performances of various pre-trained models on two datasets, pre-trained VGG16 (Simonyan & Zisserman, 2014) is selected as the backbones for both the teacher network and the student network due to its best performance on the ACNE04 dataset while VGG11 (Simonyan & Zisserman, 2014) is chosen as the backbones for the teacher and student networks on the PLSBRACNE01 dataset. For the loss function, the $\lambda_g$ and $\lambda_c$ are both set as 0.5 and the $\delta_g$ and $\delta_d$ are set with 0.7 and 0.3 on both datasets. The teacher network, the student network and all other comparative networks are initialized with pretrained weights on ImageNet (Deng et al., 2009). For all models in the training stage, the learning rate is set as 0.01, the batch size is set as 32 and the maximum training epoch is 120. Adam is chosen as the optimizer employed for training all models.

All programs are coded using Python with PyTorch framework. The experimental results obtained by the analysis are calculated using the python language. All the experiments are conducted on a Linux Ubuntu with 32G RAM and an NVIDIA Geforce RTX 3090 GPU of 24G VRAM. All schematic figures are drawn using Microsoft Visio, and statistical figures are drawn using Microsoft Excel in this paper. The final results of our proposed method are the average of the evaluation results with 10 independent training times.

**Table 2**
Performances of some advanced CNN-based methods and ViT-based methods on ANCE04 (left) and PLSBRACNE01 (right) datasets.

| Dataset | ACNE04 | | | | | PLSBRACNE01 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | PR↑ | SE↑ | SP↑ | YI↑ | ACC↑ | PR↑ | SE↑ | SP↑ | YI↑ | ACC↑ |
| CNN-based networks | | | | | | | | | | |
| EfficientNet0 (Tan & Le, 2019) | 74.72 | 72.62 | 91.69 | 64.31 | 77.74 | **66.39** | 50.90 | 85.15 | 36.05 | 58.54 |
| EfficientNet1 (Tan & Le, 2019) | 75.45 | 75.03 | 92.24 | 67.27 | 79.11 | 51.48 | 50.74 | 85.36 | 36.09 | 58.54 |
| EfficientNet2 (Tan & Le, 2019) | 73.73 | 72.94 | 91.80 | 64.74 | 77.74 | 51.64 | 50.90 | 85.33 | 36.22 | 58.54 |
| EfficientNet3 (Tan & Le, 2019) | 73.77 | 71.85 | 91.26 | 63.11 | 76.71 | 59.98 | 48.81 | 84.49 | 33.30 | 56.10 |
| EfficientNet4 (Tan & Le, 2019) | 67.70 | 67.28 | 89.73 | 57.01 | 72.95 | 40.03 | 42.40 | 82.56 | 24.97 | 51.22 |
| EfficientNet5 (Tan & Le, 2019) | 70.00 | 71.35 | 90.71 | 62.07 | 75.00 | 50.83 | 49.55 | 84.64 | 34.19 | 56.10 |
| EfficientNet6 (Tan & Le, 2019) | 58.65 | 61.26 | 88.23 | 49.50 | 66.78 | 62.50 | 47.88 | 83.79 | 31.68 | 53.66 |
| EfficientNet7 (Tan & Le, 2019) | 66.91 | 71.37 | 90.20 | 61.58 | 72.60 | 37.61 | 43.30 | 83.30 | 26.6 | 51.22 |
| Res18 (He et al., 2016) | 67.91 | 67.01 | 88.89 | 55.90 | 71.23 | 63.62 | 50.74 | 85.15 | 35.89 | 58.54 |
| Res34 (He et al., 2016) | 71.58 | 66.38 | 88.92 | 55.29 | 72.26 | 63.75 | 49.39 | 84.46 | 33.86 | 56.10 |
| Res50 (He et al., 2016) | 72.28 | 69.17 | 90.99 | 60.17 | 76.71 | 51.07 | 48.81 | 84.40 | 33.22 | 56.10 |
| Res101 (He et al., 2016) | 74.30 | 70.38 | 91.16 | 61.54 | 77.40 | 42.05 | 46.73 | 84.35 | 31.08 | 56.10 |
| Res152 (He et al., 2016) | 68.58 | 68.34 | 90.41 | 58.75 | 74.32 | 36.97 | 44.71 | 82.29 | 27.00 | 48.78 |
| VGG11 (Simonyan & Zisserman, 2014) | 78.48 | 73.69 | 91.15 | 64.84 | 77.74 | 62.70 | **59.65** | **87.32** | **46.97** | **63.41** |
| VGG13 (Simonyan & Zisserman, 2014) | 73.81 | 72.01 | 91.40 | 63.41 | 77.40 | 63.77 | 57.15 | 86.04 | 43.19 | 60.98 |
| VGG16 (Simonyan & Zisserman, 2014) | **77.51** | **75.36** | **92.26** | **67.36** | **79.79** | 56.97 | 52.66 | 86.10 | 38.76 | 60.98 |
| VGG19 (Simonyan & Zisserman, 2014) | 71.56 | 69.68 | 90.48 | 60.16 | 75.00 | 44.2 | 47.15 | 84.45 | 31.60 | 56.10 |
| ViT-based networks | | | | | | | | | | |
| ViT (Dosovitskiy et al., 2021) | 70.38 | 63.95 | 87.63 | 51.58 | 68.49 | 38.03 | 44.65 | 83.51 | 28.16 | 53.66 |
| T2T-ViT$_t$14 (Yuan et al., 2021) | 71.16 | 71.39 | 90.96 | 62.34 | 75.68 | 63.43 | 51.06 | 85.27 | 36.33 | 56.10 |
| PVT (Wang et al., 2021) | 59.78 | 58.06 | 85.46 | 43.52 | 62.67 | 42.50 | 49.04 | 84.79 | 33.82 | 56.10 |
| DeiT (Touvron et al., 2021) | 71.61 | 69.53 | 89.40 | 58.93 | 72.95 | 38.97 | 44.71 | 82.05 | 26.76 | 48.78 |

"PR", "SE", "SP", "YI" and "ACC": Precision, Sensitivity, Specificity, Youden Index and Accuracy, respectively.

## 6. Experiment results and discussion

### 6.1. Backbone selection

In order to select a suitable network as the backbone of our framework, we implement some advanced visual classification networks on the two datasets. These advanced methods include three classical CNN-based networks and four recent ViT-based networks. The three CNN-based networks are EfficientNet (Tan & Le, 2019), ResNet (He et al., 2016) and VGGNet (Simonyan & Zisserman, 2014). The four ViT-based methods are ViT (Dosovitskiy et al., 2021), T2T-ViT (Yuan et al., 2021), PVT (Wang et al., 2021) and DeiT (Touvron et al., 2021). For the three CNN-based methods, we implement all versions of the proposed networks from shallow to deep to analyze the performance of deepening their network structure. The performances of these advanced methods on the ACNE04 and PLSBRACNE01 datasets are shown in Table 2.

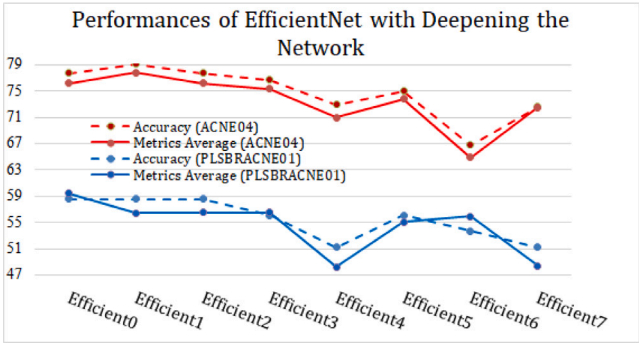
For the EfficientNet, the best performing model on the ACNE04 dataset is Efficient1 which has an accuracy of 79.11%, a precision of 75.45%, a sensitivity of 75.03%, a specificity of 92.24%, and a Youden Index of 67.27%. Scores on all indicators exceed other EfficientNet models. On the PLSBRACNE01 dataset, the best performing model is EfficientNet0, with an accuracy of 58.54%, a precision of 66.39%, a sensitivity of 50.90%, a specificity of 85.15% and a Youden Index of 36.05%. Most indicators outperform other EfficientNet models. For the ResNet, Res101 performs the best on the ACNE04 dataset with an accuracy of 77.40%, a precision of 74.30%, a sensitivity of 70.38%, a specificity of 91.16% and a Youden Index of 61.54%. All metrics outperform other ResNet models. On the PLSBRACNE01 dataset, Res18 achieves the best performance with an accuracy of 58.54%, a precision of 63.62%, a sensitivity of 50.74%, a specificity of 85.15% and a Youden Index of 35.89%. Except the precision is slightly lower than Res34, all other indicators exceed other models. For the VGGNet, on the ACNE04 dataset, VGG16 performs the best with an accuracy of 79.79%, a precision of 77.51%, a sensitivity of 75.36%, a specificity of 92.26%, and a Youden Index of 67.36%. On the PLSBRACNE01 dataset, VGG11 performs the best with an accuracy of 63.41%, a precision of 62.70%, a sensitivity of 59.65%, a specificity of 87.32% and a Youden Index of 46.97%. Except the precision is slightly lower than that of VGG13, other indicators exceeded other VGGNets.

For the ViT-based methods, the basic ViT (Dosovitskiy et al., 2021) and PVT (Wang et al., 2021) perform poorly on this task, with an accuracy rate of only not over 70.00% on the ACNE04 dataset. And other transformer-based methods perform better with accuracies of around 75.00%. On the PLSBRACNE01 dataset, the DeiT (Touvron et al., 2021) performs poorly, with an accuracy of only 48.78%. The performances of other transformer-based methods achieve accuracies of about 55%.

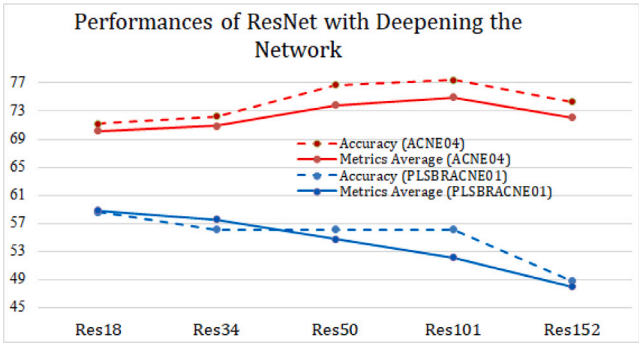
We can find that the best-performing method on the ACNE04 dataset is VGG16 and that on the PLSBRACNE01 dataset is VGG11. So the VGG16 is selected as the backbones of both the teacher network and student network for the ACNE04 dataset. Similarly, the VGG11 is selected as the backbones of both the teacher network and student network for the PLSBRACNE01 dataset.

From the experiment results, all these methods perform significantly better on the ACNE04 dataset than the PLSBRACNE01 dataset. The reason may be two aspects. On the one hand, the data volume of the ACNE04 dataset is larger than that of the PLSBRACNE01 dataset. It is well established that machine learning methods tend to be more effective at mining task patterns with more data. With insufficient data, it is difficult for these methods to achieve very high performance. On the other hand, the Pillsbury criterion adopted by the PLSBRACNE01 dataset is also more complex than the Hayashi criterion adopted by the ACNE04 dataset. The Pillsbury criterion considers both the type and number of various lesions to grade the severity level while the Hayashi criterion considers only the number of lesions regardless of the type of lesions. So these deep learning methods are easier to achieve good performance on the ACNE04 dataset.

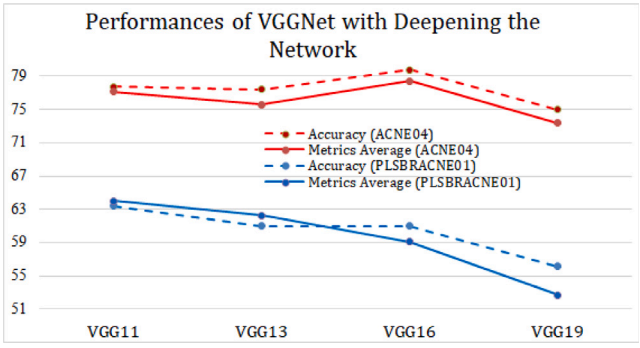
The performance of the four ViT-based methods is not significantly better than that of the CNN-based methods. This result is unexpected because the ViT-based methods achieve have been proven to have better performance than the CNN-based methods on many public image tasks. Some previous studies of acne diagnosis methods such as Lin, Jiang, Ma, Chen, Guan, You et al. (2022) and Lin et al. (2021) also found the same results as ours and discussed the phenomenon. The ViT-based networks cut the original input image into many small patches. And calculate attention scores among these patches to mine the patterns of different kinds of targets. This attention computes the relevance of the different components of a larger target and is a great help in distinguishing the type of the target. But for acne support for its diagnosis




(a) Performances of EfficientNet with Deepening the Network




(b) Performances of ResNet with Deepening the Network




(c) Performances of VGGNet with Deepening the Network

**Fig. 5.** Performances of EfficientNet, ResNet and VGGNet with Deepening the Networks. The red broken lines show the performance on the ACNE04 dataset, and the blue broken lines show the performance on the PLSBRACNE01 dataset. The solid lines display the average score of all evaluation indicators, and the dotted lines show the accuracy.

lies in the cases of many tiny lesions distributed on the whole face. Such cutting of the original image may lead to the large lesions being cut into multiple patches, and at the same time, a patch may contain multiple different smaller lesions. This time, it becomes meaningless to calculate the correlation among different patches. So these state-of-the-art ViT-based methods failed to significantly outperform the CNN-based methods.

Different deep versions of the same model perform very differently. On the ACNE04 dataset, the medium-depth versions of these models perform slightly better than others. In terms of PLSBRACNE01 data, the shallow versions of these models perform better than the others. Researchers generally believe that deeper network models provide better performance. But only if the amount of data is sufficient enough. Both datasets in this work are relatively few compared to other common image classification tasks. So the largest models are not the best performers in either dataset.

### 6.2. Performance changes with deepening the network structure

The knowledge distillation methods usually use a large network as the teacher model to achieve good enough performance. But in the previous comparison experiment, we found that the largest networks of various CNNs do not perform best and different versions of the same model performed significantly differently on the two datasets. So this section analyzes the performances of the three CNN methods of their versions from shallow to deep. Performances of EfficientNet, ResNet and VGGNet with Deepening the Networks are shown in Fig. 5. The red broken lines show the performance on the ACNE04 dataset, and the blue broken lines show the performance on the PLSBRACNE01 dataset. The solid lines display the average score of all evaluation indicators, and the dotted lines show the accuracy.

For the EfficientNet on the ACNE04 dataset, the performance of the model increases first and then decreases as the network gradually becomes larger, and the model achieves the best performance at EfficientNet1. Although the performance fluctuates after EfficientNet3, the overall trend remained downward. On the PLSBRACNE01 dataset, EfficientNet0 achieves the best performance. With the gradual deepening of the network, the overall performance shows a downward trend. Similarly to the phenomenon on the ACNE04 dataset, the performance fluctuates after EfficientNet3.

For the ResNet on the ACNE04 dataset, the performance of various versions increases first and then decreases as the network gradually becomes deeper. The model achieves the best performance at Res101. The performance of ResNet models is relatively flat, without fluctuation. On the PLSBRACNE01 dataset, Res18 achieves the best performance. With the gradual deepening of the network, the overall performance shows a downward trend.

For the VGGNet, the performance is similar to that of ResNet. On the ACNE04 dataset, the performance of various versions increases first and then decreases as the network gradually becomes deeper. The model achieves the best performance at VGG16. On the PLSBRACNE01 dataset, VGG11 achieves the best performance. With the gradual deepening of the network, the overall performance shows a downward trend.

In general, the three network models all show the same trend. As the network gradually grows larger, the performance on the ACNE04 dataset rises first and then falls, while that on the PLSBRACNE01 dataset declines gradually. This phenomenon shows that larger networks do not necessarily achieve better performance. Under the premise of limited data volume, optimal performance can be obtained by selecting a model with an appropriate size.

### 6.3. Effectiveness of diagnostic evidence

In previous experiments, we compared the performance of various visual classification models in acne diagnosis and selected the optimal model as the backbone of our framework. This section analyzes the effectiveness of fusing the diagnostic evidence into the optimal models and constructs the teacher network for our framework. The VGG16 and VGG11 are the optimal models for the ACNE04 dataset and PLSBRACNE01 dataset respectively. So the constructed diagnostic evidence is fused into these two models for both datasets. The experiment results are shown in Table 3.

For the ACNE04 dataset, after fusing the diagnostic evidence, the VGG16 obtains an accuracy of 98.63%, a precision of 97.90%, a sensitivity of 98.72%, a specificity of 99.54%, and a Youden Index of 98.26%. The scores of all evaluation indexes have been improved significantly. All the evaluation results surpass 90%. Most of the scores are improved by over 20%. Similarly, for the PLSBRACNE01 dataset, the diagnostic evidence improves the VGG11 to obtain an accuracy of 73.17%, a precision of 73.62%, a sensitivity of 69.23%, a specificity of 90.65%, and a Youden Index of 59.88%. The scores of all evaluation

**Table 3**
Performance of the optimal networks and optimal networks with fusing diagnostic evidence on the ACNE04 dataset and PLSBRACNE01 dataset.

| Dataset | ACNE04 | | | | | PLSBRACNE01 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | PR↑ | SE↑ | SP↑ | YI↑ | ACC↑ | PR↑ | SE↑ | SP↑ | YI↑ | ACC↑ |
| VGG11 (Simonyan & Zisserman, 2014) | 78.48 | 73.69 | 91.15 | 64.84 | 77.74 | 62.70 | 59.65 | 87.32 | 46.97 | 63.41 |
| VGG16 (Simonyan & Zisserman, 2014) | 77.51 | 75.36 | 92.26 | 67.36 | 79.79 | 56.97 | 52.66 | 86.10 | 38.76 | 60.98 |
| VGG16+Diagnostic Evidence | **97.90** | **98.72** | **99.54** | **98.26** | **98.63** | – | – | – | – | – |
| VGG11+Diagnostic Evidence | – | – | – | – | – | **73.62** | **69.23** | **90.65** | **59.88** | **73.17** |

"PR", "SE", "SP", "YI" and "ACC": Precision, Sensitivity, Specificity, Youden Index and Accuracy, respectively.

**Table 4**
Experiment results of the diagnostic evidence distillation and ablation of different components on the ACNE04 dataset and PLSBRACNE01 dataset.

| Dataset (Backbone) | ACNE04 (VGG16) | | | | | PLSBRACNE01 (VGG11) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Distillation strategy | PR↑ | SE↑ | SP↑ | YI↑ | ACC↑ | PR↑ | SE↑ | SP↑ | YI↑ | ACC↑ |
| B | 77.51 | 75.36 | 92.26 | 67.36 | 79.79 | 62.70 | 59.65 | 87.32 | 46.97 | 63.41 |
| S distilled by T(B+DE) | 84.55 | 82.02 | 93.71 | 75.73 | 83.90 | 65.12 | 61.73 | 88.07 | 49.80 | 65.85 |
| S distilled by T(B+DE+ST) | 84.88 | 82.63 | 93.82 | 76.45 | 84.25 | 65.01 | 63.97 | 88.21 | 52.19 | 65.85 |
| S+A distilled by T(B+DE+ST) | **85.31** | **84.83** | **94.66** | **79.48** | **86.06** | **69.16** | **65.62** | **88.93** | **54.54** | **67.56** |

"B", "S" and "T": Backbone network, Student network and Teacher network.
"DE", "ST" and "A": Diagnostic Evidence, SubTask and Augmentation.
"PR", "SE", "SP", "YI" and "ACC": Precision, Sensitivity, Specificity, Youden Index and Accuracy, respectively.

metrics also increase significantly. Most of these evaluation scores are improved by around 10%.

We can find that after fusing the diagnostic evidence, the two networks are improved significantly, demonstrating that the integration of diagnostic evidence is meaningful. We believe that the diagnostic evidence provides vital guidance for mining the pattern of acne diagnosis. The diagnostic evidence indirectly fuses both the type and the number of all lesions distributed on the faces. This operation relieves the pressure on deep learning models to learn acne diagnosis. So the performance on both datasets increases significantly.

Next, we will use these two diagnostic evidence-fused models as the teacher networks to conduct final diagnosis models.

### 6.4. Diagnostic evidence distillation and ablation study

In the previous experiments, we obtained the optimal diagnostic evidence-fused teacher networks and found the backbone networks for the two datasets. But the constructed diagnostic evidence in this work is not available when a new sample needs to be diagnosed. That is, the excellent-performing teacher models cannot be the final diagnosis models. So this section transforms the knowledge of diagnostic evidence-fused teacher networks into the student backbone networks as the final diagnosis models using knowledge distillation. The experiment results of gradually adding the components of our framework to obtain the optimal student networks as the final diagnostic models are shown in Table 4.

The first row shows the performances of selected backbones for the two datasets. Next, we fuse diagnostic evidence with the corresponding image into the backbones to train the teacher networks and fix the parameters of the teachers to guide the student networks. Under the guidance of the teacher networks, the distilled student networks are significantly improved in all metrics, as shown in the second row. On the ACNE04 dataset, the student network reaches a precision of 84.55%, a sensitivity of 82.02% and an accuracy of 83.90%. On the PLSBRACNE01 dataset, the student network reaches a precision of 65.12%, a sensitivity of 61.73% and an accuracy of 65.85%. Next, we bring the subtask of learning the lesion counting along with grading the severity to retrain the evidence-fused teacher network. The performances of the student networks are improved again as shown in the third row. On the ACNE04 dataset, the student network reaches a precision of 84.88%, a sensitivity of 82.63% and an accuracy of 84.25%. On the PLSBRACNE01 dataset, the student network reaches a precision of 65.01%, a sensitivity of 63.97% and an accuracy of 65.85%. Most of the evaluation results increase slightly again on both datasets. Finally, we fix the structure of the teacher networks and introduce the data augmentation on the input of student networks including size scaling, random cropping, random flipping and random rotation as shown in the last row The performances are further improved on most of the metrics. On the ACNE04 dataset, the student network reaches a precision of 85.31%, a sensitivity of 84.83%, a specificity of 94.66%, a Youden Index of 79.48% and an accuracy of 86.06%. On the PLSBRACNE01 dataset, the student network reaches a precision of 69.16%, a sensitivity of 65.62%, a specificity of 88.93%, a Youden Index of 54.54% and an accuracy of 67.56%. After all strategies are introduced, the performances of the student networks reach the best. Compared with the backbones, the accuracy is improved by more than 6%, the precision is enhanced by over 6%, and the Youden Index increases by more than 10% on the ACNE04 dataset. And on the PLSBRACNE01 dataset, the precision increases about 6% and the Youden Index is improved around 8%.

Experiments show that all components in the framework are helpful for improving performance. Firstly, the distillation from the teacher networks significantly improves the performance of student networks on two datasets. The teacher networks have fused diagnostic evidence which directly guides the diagnosis. So the distillation which guides the student networks and transforms the knowledge from teachers into students improves the students significantly. Next, the subtask conducted on the teacher networks also improves the performance of students. The subtask uses the teachers to learn lesion counting jointly with the severity grading together. This strategy coerces the teacher networks to pay more attention to the lesion areas that are directly related to the lesion counting. It enhances the teachers themselves and also enhances the guidance to the student networks. So The subtask on the teacher networks improves student performance. The augmentation conducted on the student input also improves the students slightly. This strategy compels the student to extract the same hidden features when observing the varieties of the same sample. It effectively improves the generalization of the student models. So the performance of the two students increases again.

However, we also found that the student network eventually falls short of the teacher network. We believe that the experience of fusing diagnostic evidence is more difficult to learn than conventional distillation. Meanwhile, the improvement for the framework on the ACNE04 is more obvious than that on the PLSBRACNE01. We believe that the data volume of ACNE04 is relatively large, and it is easier for the model to mine diagnostic experience to obtain significant effects.

**Table 5**
Comparison with state-of-the-arts on ANCE04 (left) and PLSBRACNE01 (right) datasets separately.

| Dataset | ACNE04 | | | | | PLSBRACNE01 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | PR↑ | SE↑ | SP↑ | YI↑ | ACC↑ | PR↑ | SE↑ | SP↑ | YI↑ | ACC↑ |
| Wu et al. (2019) | 84.37 | 81.52 | 93.80 | 75.32 | 84.11 | – | – | – | – | – |
| Lin et al. (2021) | 83.58 | 81.95 | 94.11 | 76.06 | 84.52 | 48.43 | 50.48 | 83.50 | 33.98 | 52.85 |
| KIEGLFN (Lin, Jiang, Ma, Chen, Guan, You et al., 2022) | 83.58 | 81.95 | 94.11 | 76.06 | 84.52 | 62.67 | 55.77 | 85.71 | 41.84 | 59.35 |
| EGPK (Lin, Jiang, Chen et al., 2022) | 84.01 | 84.62 | 94.40 | 79.01 | 85.27 | 65.00 | 61.79 | 88.15 | 49.94 | 65.85 |
| DED | **85.31** | **84.83** | **94.66** | **79.48** | **86.06** | **69.16** | **65.62** | **88.93** | **54.54** | **67.56** |

"PR", "SE", "SP", "YI" and "ACC": Precision, Sensitivity, Specificity, Youden Index and Accuracy.

### 6.5. Comparison with state-of-the-arts

To explore the effect of the proposed framework, we conduct comparative experiments with some state-of-the-art methods including four previous proposed methods by Wu et al. (2019) and Lin et al. (2021), KIEGLFN (Lin, Jiang, Ma, Chen, Guan, You et al., 2022) and EGPK (Lin, Jiang, Chen et al., 2022) as shown in Table 5. Because previous studies (Lin, Jiang, Ma, Chen, Guan, You et al., 2022) have shown that their methods outperform state-of-the-art CNNs and ViT-based methods (He et al., 2016; Hu et al., 2020; Simonyan & Zisserman, 2014; Tan & Le, 2019), we do not include them in our comparative experiments.

For the four methods for this task, most of them are superior to the various CNN and transformer-based methods discussed above. Since the method proposed by Wu et al. is based on Hayashi Criterion and is not applicable to the PLSBRACNE01 dataset, we only show its performance on the ACNE04 dataset. Ultimately, the performance of the student network obtained under our framework is shown in the last row of the table. Our student network slightly outperforms state-of-the-art methods on all metrics. On the ACNE04 dataset, the performance reaches an accuracy of 86.06%, a precision of 85.31%, a sensitivity of 84.83%, a specificity of 94.66% and a Youden Index of 79.48%. On the PLSBRACNE01 dataset, the performance reaches an accuracy of 67.56%, a precision of 69.16%, a sensitivity of 65.62%, a specificity of 88.93% and a Youden Index of 54.54%. The experiment results demonstrate that the final diagnosis models obtained from our framework surpass the state-of-the-art methods.

In the first row, the key strategy of the method proposed by Wu et al. (2019) is joint learning lesion counting and severity grading which is an adaption to the Hayashi Criterion. So it can only be adopted on the ACNE04 dataset. In the second row, the key idea of Lin et al. (2021) is imitating the global estimation of dermatologists which makes the methods can be adopted on different criteria. Then, in the third row, Lin, Jiang, Ma, Chen, Guan, You et al. (2022) proposes a transfer learning strategy based on the method of Lin et al. (2021) that transfers the experience on a large dataset into another small dataset of different criteria. So its performance on the PLSBRACNE01 dataset is improved significantly. But both two methods lack adaptation to the characteristics of acne diagnosis. In the fourth row, the EGPK proposed by Lin, Jiang, Chen et al. (2022) has a similar key idea to our proposed DED. They believe the lesion areas are the prior knowledge for acne diagnosis. So firstly, the EGPK uses a segmentation model to identify the lesion areas. Then, they use the segmented lesion map with its corresponding face image to train the downstream diagnosis methods. Indeed, this is an explicit fusion of diagnostic evidence that uses a model to predict the diagnostic evidence for downstream diagnosis. So this method achieves good performance with all of the metrics slightly higher than the above two. But if the predicted diagnostic evidence has errors, the errors will accumulate downstream. So its performance on PLSBRACNE01 is slightly lower than the DED. Compared to these methods, DED is an implicit way to fuse diagnostic evidence. This fusion ensures that the guidance from the teacher model to the downstream diagnosis model contains the real diagnostic evidence but not the predicted one.

### 6.6. Comparison with doctors

To explore the practice of our framework, we compare the performance of our framework with that of some doctors. Table 6 shows the comparative results of doctors and our proposed framework. For the ACNE04 dataset, the diagnostic results of doctors are published by Wu et al. (2019). The results of the two general doctors underperforming the dermatologists are reasonable. Dermatologists with specialized knowledge certainly have better diagnostic performance than general doctors. Meanwhile, the two dermatologists performed extremely well, beating most of the comparison models. Fortunately, our proposed framework can reach the level of professional dermatology diagnosis. For the PLSBRACNE01 dataset, two dermatologists are invited to perform the diagnosis (Lin, Jiang, Ma, Chen, Guan, You et al., 2022). Between them, the performance of Derm1 is better than that of Derm2 slightly, but the overall difference is not significant. Finally, our framework also reaches the diagnosis level of dermatologists.

The experiment results demonstrate that our diagnosis models reach the diagnostic level of dermatologists on both two datasets.

### 6.7. Diagnostic stability

To evaluate the diagnostic stability of our proposed method, we independently train the method ten times on both datasets. The evaluation results are sorted by the average of precision, sensitivity, specificity, Youden Index and accuracy as shown in Table 7. The statistical analysis results including mean, standard deviation, median, maximum, minimum and range are shown in Table 8 and the boxplot is shown in Fig. 6.

From the Statistical results, on the ACNE04 dataset, the standard deviations of the five scores are all 0.00 or 0.01. The average of the standard deviation on the five scores is 0.01. The mean and the median of all evaluation scores are very close. The range of sensitivity and Youden Index is around 3%, slightly greater than the other evaluation metrics. The range of precision and specificity is lower than 1%. This result demonstrates that the method is stable. The performance of ten independent evaluations has little difference. For the PLSBRACNE01 dataset, the standard deviations of the five scores are all between 0.01 to 0.04. The average of the standard deviation on the five scores is 0.02. The mean and the median of all evaluation scores are also very close. The range of sensitivity and Youden Index is over 10%, slightly greater than the other evaluation metrics. The range of specificity is 2.94% and the range of other metrics is around 7%. Compared to the performance on the ACNE04 dataset, the stability on the PLSBRACNE01 dataset performs slightly less well but within the acceptable range.

The boxplot shows similar results to the statistical results. On the ANCE04 dataset, there is no outlier for each evaluation metric of the model, while on the PLSBRACNE01 dataset, there is an outlier better than common values for accuracy. The boxes of most of the evaluation metrics for the two datasets are small, indicating that they are relatively stable, such as precision, specificity, and accuracy. The box for the average of all metrics is also small, meaning that the overall performance of the model is stable. The box of Youden Index is obviously larger than that of other metrics, and the box of sensitivity is slightly larger than that of other metrics. It means that the stability of the two metrics is

**Table 6**
Comparison with Doctor performances on ANCE04 (top) and PLSBRACNE01 (bottom) datasets separately.

| Method | PR↑ | SE↑ | SP↑ | YI↑ | ACC↑ |
|---|---|---|---|---|---|
| ACNE04 dataset | | | | | |
| GD1 (Wu et al., 2019) | 62.87 | 55.27 | 84.11 | 39.38 | 58.43 |
| GD2 (Wu et al., 2019) | 62.07 | 68.33 | 86.98 | 55.31 | 63.14 |
| Derm1 (Wu et al., 2019) | 77.33 | 72.56 | 90.60 | 63.22 | 75.29 |
| Derm2 (Wu et al., 2019) | 82.95 | 78.27 | 92.16 | 70.43 | 79.43 |
| DED | **85.31** | **84.83** | **94.66** | **79.48** | **86.06** |
| PLSBRACNE01 dataset | | | | | |
| Derm1 (Lin et al., 2021) | 56.75 | 41.07 | 80.24 | 21.30 | 41.00 |
| Derm2 (Lin et al., 2021) | 50.27 | 36.41 | 78.77 | 15.18 | 37.00 |
| DED | **69.16** | **65.62** | **88.93** | **54.54** | **67.56** |

"PR", "SE", "SP", "YI" and "ACC": Precision, Sensitivity, Specificity, Youden Index and Accuracy, respectively.
"GD" and "Derm": general doctor and dermatologist.

**Table 7**
Experiment results for diagnostic stability of evaluation with ten independent training times on the ACNE04 dataset (left) and PLSBRACNE01 dataset (right).

| Dataset | ACNE04 | | | | | | PLSBRACNE01 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ranking | PR↑ | SE↑ | SP↑ | YI↑ | ACC↑ | AVG↑ | PR↑ | SE↑ | SP↑ | YI↑ | ACC↑ | AVG↑ |
| 1 | 86.26 | 86.10 | 95.01 | 81.11 | 86.99 | 87.09 | 73.62 | 69.23 | 90.65 | 59.88 | 73.17 | 73.31 |
| 2 | 84.82 | 85.73 | 95.06 | 80.79 | 86.64 | 86.61 | 67.36 | 68.21 | 89.46 | 57.66 | 68.29 | 70.20 |
| 3 | 85.15 | 86.18 | 94.65 | 80.83 | 85.96 | 86.55 | 66.67 | 67.05 | 89.07 | 56.12 | 68.29 | 70.20 |
| 4 | 85.42 | 85.36 | 94.90 | 80.26 | 86.64 | 86.52 | 72.03 | 64.81 | 89.14 | 53.95 | 68.29 | 69.64 |
| 5 | 85.40 | 84.61 | 94.61 | 79.22 | 85.96 | 85.96 | 65.79 | 67.63 | 89.40 | 57.03 | 68.29 | 69.63 |
| 6 | 85.40 | 84.61 | 94.61 | 79.22 | 85.96 | 85.96 | 68.45 | 67.85 | 88.84 | 56.70 | 65.85 | 69.54 |
| 7 | 85.45 | 84.47 | 94.56 | 79.02 | 85.96 | 85.89 | 68.45 | 67.85 | 88.84 | 56.70 | 65.85 | 69.54 |
| 8 | 84.66 | 84.39 | 94.68 | 79.06 | 85.96 | 85.75 | 69.33 | 62.31 | 88.09 | 50.39 | 65.85 | 67.19 |
| 9 | 84.91 | 83.58 | 94.27 | 77.85 | 85.27 | 85.18 | 69.33 | 62.31 | 88.09 | 50.39 | 65.85 | 67.19 |
| 10 | 85.61 | 83.22 | 94.20 | 77.41 | 85.27 | 85.14 | 70.54 | 58.91 | 87.71 | 46.62 | 65.85 | 65.93 |

"PR", "SE", "SP", "YI", "ACC" and "AVG": Precision, Sensitivity, Specificity, Youden Index, Accuracy and Average of the Precision, Sensitivity, Specificity, Youden Index and Accuracy.

**Table 8**
Statistical results for diagnostic stability of evaluation with ten independent training times on the ACNE04 dataset (left) and PLSBRACNE01 dataset (right).

| Dataset | ACNE04 | | | | | | PLSBRACNE01 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | PR↑ | SE↑ | SP↑ | YI↑ | ACC↑ | AVG↑ | PR↑ | SE↑ | SP↑ | YI↑ | ACC↑ | AVG↑ |
| Mean | 85.31 | 84.83 | 94.66 | 79.48 | 86.06 | 86.07 | 69.16 | 65.62 | 88.93 | 54.54 | 67.56 | 69.24 |
| Standard Deviation | 0.00 | 0.01 | 0.00 | 0.01 | 0.01 | 0.01 | 0.02 | 0.03 | 0.01 | 0.04 | 0.02 | 0.02 |
| Median | 85.40 | 84.61 | 94.63 | 79.22 | 85.96 | 85.96 | 68.89 | 67.34 | 88.96 | 56.41 | 67.07 | 69.58 |
| Maximum | 86.26 | 86.18 | 95.06 | 81.11 | 86.99 | 87.09 | 73.62 | 69.23 | 90.65 | 59.88 | 73.17 | 73.31 |
| Minimum | 84.66 | 83.22 | 94.20 | 77.41 | 85.27 | 85.14 | 65.79 | 58.91 | 87.71 | 46.62 | 65.85 | 65.93 |
| Range | 0.60 | 2.96 | 0.86 | 3.70 | 1.72 | 1.95 | 7.83 | 10.32 | 2.94 | 13.26 | 7.32 | 7.38 |

"PR", "SE", "SP", "YI", "ACC" and "AVG": Precision, Sensitivity, Specificity, Youden Index, Accuracy and Average of the Precision, Sensitivity, Specificity, Youden Index and Accuracy.


**Fig. 6.** Boxplot for evaluating diagnostic stability of evaluation with ten independent training times on the ACNE04 dataset (left) and PLSBRACNE01 dataset (right).

not better than the others. The box size of specificity is obviously small compared with other metrics, indicating the performance of specificity is very stable.

The stability of the PLSBRACNE01 dataset is slightly poorer than that of the ACNE04. The reason may be due to the data insufficiency in this dataset, leading the model prone to be overfitted. In addition, the criterion used by PLSBRACNE01 is more complex than that used by ACNE04, which makes it more difficult for models to learn diagnostic patterns on PLSBRACNE01. On both datasets, the fluctuation of Youden Index score is significantly higher than that of other indexes. Youden Index is a reconciliation of sensitivity and specificity, which are positively correlated, so the changes of the two indicators will be superimposed on the Youden Index, resulting in a greater fluctuation. Of course, it cannot be denied that the fluctuation of sensitivity of the model is also

relatively large. Despite this, the minimum values of Youden Index and sensitivity can still reach the state-of-the-art level. So this fluctuation is acceptable.

Overall, the experiment results demonstrate that the performance of the method is acceptable. The performance of ten independent evaluations has little difference.

## 7. Conclusion

In this paper, we propose a Diagnostic Evidence Distillation framework to diagnose face acne under different acne criteria with suitably adapting the characteristics of acne diagnosis. This framework is a novel distillation paradigm that uses the teacher–student structure of knowledge distillation methods to bring the diagnostic evidence, which is not available for new patients but only in training data, into the diagnosis model. To break the limitation of different acne criteria, we use CNNs as the backbone for the diagnosis model to imitate the global estimation diagnosis of dermatologists. We also propose a joint learning subtask to enhance the mining of tiny lesions and the guidance to the student. We conduct comprehensive experiments on the framework of diagnosing face acne on two datasets based on different mainstream acne criteria. The experimental results show that the framework effectively improves the diagnosis performance. The performance reaches the state-of-the-art and diagnostic level of dermatologists. The precision, sensitivity, specificity, Youden Index and accuracy reach 85.31%, 84.83%, 94.66%, 79.48% and 86.06% on the ACNE04 dataset and 69.16%, 65.62%, 88.93%, 54.54% and 67.56% on the PLSBRACNE01 dataset, respectively.

Compared to the previous studies, the main contribution of this work is proposing a novel face acne diagnosis method that can be adopted on different mainstream acne criteria as well as suitably adapts to the characteristics of acne diagnosis. Some studies exploring the diagnosis of advanced visual models can diagnose acne on different criteria but lack adaptation to the characteristics of acne diagnosis. Some other studies proposed models over-adapting a specific criterion, rendering these methods unusable on other criteria. We analyze and summarize the commonalities of various mainstream acne criteria and reduce the over-adaptation. Also, one previous study adapts the characteristics of acne diagnosis and can be used on different criteria but has the problem of error accumulation. We propose a more suitable adaption method to avoid error accumulation.

For future studies related to this study, maybe three folds as follows: Firstly, construct more face acne datasets based on various commonly used acne criteria to validate the generalization of our method. Secondly, explore the generalization of our method in other classification tasks of different fields. Finally, research about more accurate acne diagnosis methods to help patients without medical conditions and alleviate their suffering.

## CRediT authorship contribution statement

**Yi Lin:** Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Project administration. **Jingchi Jiang:** Methodology, Writing – review & editing. **Dongxin Chen:** Software, Validation, Formal analysis, Writing – review & editing. **Zhaoyang Ma:** Software, Validation, Formal analysis. **Yi Guan:** Investigation, Resources, Supervision. **Xiguang Liu:** Investigation, Resources, Supervision. **Haiyan You:** Investigation, Resources. **Jing Yang:** Investigation, Resources.

## Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript.

## Data availability

Two datasets are used in this study. One of them is ACNE04 which is an open dataset. For the other dataset, we do not have the permission to share

## References


Alenezi, F. S., Armghan, A., & Polat, K. (2022). Wavelet transform based deep residual neural network and ReLU based extreme learning machine for skin lesion classification. *Expert Systems with Applications*.

Alizadeh, R., Allen, J. K., & Mistree, F. (2020). Managing computational complexity using surrogate models: a critical review. *Research in Engineering Design*, (5).

Alkhulaifi, A., Alsahli, F. S., & Ahmad, I. (2021). Knowledge distillation in deep learning and its applications. *PeerJ Computer Science*, *7*.

Blakeney, C., Li, X., Yan, Y., & Zong, Z. (2021). Parallel blockwise knowledge distillation for deep neural network compression. *IEEE Transactions on Parallel and Distributed Systems*, *32*, 1765–1776.

Camacho-Gutiérrez, J. A., Solorza-Calderón, S., & Álvarez-Borrego, J. (2022). Multi-class skin lesion classification using prism- and segmentation-based fractal signatures. *Expert Systems with Applications*, *197*, Article 116671.

Cho, S. I., Yang, J., & Suh, D. K. (2020). Analysis of trends and status of physician-based evaluation methods in acne vulgaris from 2000 to 2019. *The Journal of Dermatology*, *48*.

Creadore, A., Desai, S., Li, S. J., Lee, K. J., Bui, A.-T. N., Villa-Ruiz, C., Lo, K., Zhou, G., Joyce, C. J., Resneck, J. S., Seiger, K., & Mostaghimi, A. (2021). Insurance acceptance, appointment wait time, and dermatologist access across practice types in the US. *JAMA Dermatology*.

Cui, B., Li, Y., & Zhang, Z. (2021). Joint structured pruning and dense knowledge distillation for efficient transformer model compression. *Neurocomputing*, *458*, 56–69.

Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). CoAtNet: Marrying convolution and attention for all data sizes. ArXiv, abs/2106.04803.

Dai, C., Liu, X., Li, Z., & Chen, M.-Y. (2021). A tucker decomposition based knowledge distillation for intelligent edge applications. *Applied Soft Computing*, *101*, Article 107051.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Doshi, A., Zaheer, A., & Stiller, M. J. (1997). A comparison of current acne grading systems and proposal of a novel system. *International Journal of Dermatology*, *36*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Dréno, B., Bodokh, I., Chivot, M., Daniel, F., Humbert, P., Poli, F., Clerson, P., & Berrou, J. P. (1999). [ECLA grading: a system of acne classification for every day dermatological practice]. *Annales de Dermatologie et de Venereologie*, *126*(2), 136–141.

Dréno, B., & Poli, F. (2003). Epidemiology of acne. *Dermatology*, *206*, 7–10.

Fan, S., Zhang, X., & Song, Z. (2021). Reinforced knowledge distillation: Multi-class imbalanced classifier based on policy gradient reinforcement learning. *Neurocomputing*, *463*, 422–436.

Frank, S. B. (1971). Acne vulgaris. *Journal of Investigative Dermatology*.

Fu, S., Li, Z., Liu, Z., & Yang, X. (2021). Interactive knowledge distillation for image classification. *Neurocomputing*, *449*, 411–421.

Ghalejoogh, G. S., Kordy, H. M., & Ebrahimi, F. (2020). A hierarchical structure based on stacking approach for skin lesion classification. *Expert Systems with Applications*, *145*, Article 113127.

Ghodsi, S. Z., Orawa, H., & Zouboulis, C. C. (2009). Prevalence, severity, and severity risk factors of acne in high school pupils: a community-based study. *The Journal of Investigative Dermatology*, *129*(9), 2136–2141.

Hameed, N., Shabut, A. M., Ghosh, M. K., & Hossain, M. A. (2020). Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques. *Expert Systems with Applications*, *141*.

Hayashi, N., Akamatsu, H., & Kawashima, M. (2008). Establishment of grading criteria for acne severity. *The Journal of Dermatology*, *35*(5), 255–260.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *2015 IEEE international conference on computer vision (ICCV)* (pp. 1026–1034).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778).

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2020). Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*, 2011–2023.

James, K., & Tisserand, J. (1958). Treatment of acne vulgaris. *GP*, *18*(3), 130–139.

Koo, J. (1995). The psychosocial impact of acne: patients' perceptions. *Journal of the American Academy of Dermatology*, *32*(5 Pt 3), S26–30.

Lim, Z. V., Akram, F., Ngo, C. P., Winarto, A. A., Lee, W. Q., Liang, K., Oon, H. H., Thng, S. T. G., & Lee, H. K. (2019). Automated grading of acne vulgaris by deep learning with convolutional neural networks. *Skin Research and Technology*, *26*, 187–192.

Lin, Y., Guan, Y., Ma, Z., You, H., Cheng, X., & Jiang, J. (2021). An acne grading framework on face images via skin attention and SFNet. In *2021 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 2407–2414).

Lin, Y., Jiang, J., Chen, D., Ma, Z., Guan, Y., Liu, X., You, H., Yang, J., & Cheng, X. (2022). Acne severity grading on face images via extraction and guidance of prior knowledge. In *2022 IEEE international conference on bioinformatics and biomedicine* BIBM, (pp. 1639–1643).

Lin, Y., Jiang, J., Ma, Z., Chen, D., Guan, Y., Liu, X., You, H., Yang, J., & Cheng, X. (2022). CGPG-GAN: An acne lesion inpainting model for boosting downstream diagnosis. In *2022 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 1634–1638).

Lin, Y., Jiang, J., Ma, Z., Chen, D., Guan, Y., You, H., Cheng, X., Liu, B., & Luo, G. (2022). KIEGLFN: A unified acne grading framework on face images. *Computer Methods and Programs in Biomedicine*, *221*, Article 106911.

Lucky, A. W., Barber, B. L., Girman, C. J., Williams, J., Ratterman, J., & Waldstreicher, J. (1996). A multirater validation study to assess the reliability of acne lesion counting. *Journal of the American Academy of Dermatology*, *35*(4), 559–565.

Malgina, E., & Kurochkina, M. (2021). Development of the mobile application for assessing facial acne severity from photos. In *2021 IEEE conference of russian young researchers in electrical and electronic engineering (ElConRus)* (pp. 1790–1793).

Michaelson, G., Juhlin, L., & Vahlquist, A. (1977). Oral zinc sulphate therapy for acne vulgaris. *Acta Dermato-Venereologica*, *57*(4), 372.

Pillsbury, D. M., Shelley, W., & Kligman, A. M. (1962). A manual of cutaneous medicine. *The American Journal of the Medical Sciences*, *243*, 131.

Ramli, R., Malik, A. S., Hani, A. F. M., & Jamil, A. (2012). Acne analysis, grading and computational assessment methods: an overview. *Skin Research and Technology*, *18*.

Seité, S., Khammari, A., Benzaquen, M., Moyal, D. D., & Dréno, B. (2019). Development and accuracy of an artificial intelligence algorithm for acne grading from smartphone photographs. *Experimental Dermatology*, *28*, 1252–1257.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556.

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. ArXiv, abs/1905.11946.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & J'egou, H. (2021). Training data-efficient image transformers & distillation through attention. In *ICML*.

ul Ain, Q., Al-Sahaf, H., Xue, B., & Zhang, M. (2022). Genetic programming for automatic skin cancer image classification. *Expert Systems with Applications*, *197*, Article 116680.

Wang, J., Luo, Y., Wang, Z., Hounye, A. H., Cao, C., Hou, M., & Zhang, J. (2022). A cell phone app for facial acne severity assessment. *Applied Intelligence*, 1–20.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *2021 IEEE/CVF international conference on computer vision (ICCV)* (pp. 548–558).

Williams, H., Dellavalle, R., & Garner, S. (2012). Acne vulgaris. *The Lancet*, *379*, 361–372.

Witkowski, J. A., & Parish, L. C. (1999). From other ghosts of the past: acne lesion counting. *Journal of the American Academy of Dermatology*, *40*(1), 131.

Witkowski, J. A., & Parish, L. (2004). The assessment of acne: an evaluation of and lesion counting in the measurement of acne. *Clinics in Dermatology*, *22*(5), 394–397.

Wu, X., Wen, N., Liang, J., Lai, Y.-K., She, D., Cheng, M.-M., & Yang, J. (2019). Joint acne image grading and counting via label distribution learning. In *2019 IEEE/CVF international conference on computer vision (ICCV)* (pp. 10641–10650).

Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., & Gao, J. (2021). Focal attention for long-range interactions in vision transformers. *Advances in Neural Information Processing Systems*, *34*.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Tay, F. E. H., Feng, J., & Yan, S. (2021). Tokens-to-token ViT: Training vision transformers from scratch on ImageNet. In *2021 IEEE/CVF international conference on computer vision (ICCV)* (pp. 538–547).

Zhao, B. (2010). *Chinese clinical dermatology (a Chinese dermatology book of authority)*. Jiangsu Science and Technology Press.