Adjacency-aware Fuzzy Label Learning for Skin Disease Diagnosis

Murong Zhou, Baifu Zuo, Guohua Wang, Gongning Luo, *Member, IEEE*, Fanding Li, Suyu Dong, *Member, IEEE*, Wei Wang, *Member, IEEE*, Kuanquan Wang, *Senior Member, IEEE*, Xiangyu Li, *Member, IEEE*, and Lifeng Xu

***Abstract*—Automatic acne severity grading is crucial for the accurate diagnosis and effective treatment of skin diseases. However, the acne severity grading process is often ambiguous due to the similar appearance of acne with close severity, making it challenging to achieve reliable acne severity grading. Following the idea of fuzzy logic for handling uncertainty in decision-making, we transforms the acne severity grading task into a fuzzy label learning problem, and propose a novel Adjacency-aware Fuzzy Label Learning (AFLL) framework to handle uncertainties in this task. The AFLL framework makes four significant contributions, each demonstrated to be highly effective in extensive experiments. First, we introduce a novel adjacency-aware decision sequence generation method that enhances sequence tree construction by reducing bias and improving discriminative power. Second, we present a consistency-guided decision sequence prediction method that mitigates error propagation in hierarchical decision-making through a novel selective masking decision strategy. Third, our proposed sequential conjoint distribution loss innovatively captures the differences for both high and low fuzzy memberships across the entire fuzzy label set while modeling the internal temporal order among different acne severity labels with a cumulative distribution, leading to substantial improvements in fuzzy label learning. Fourth, to the best of our knowledge, AFLL is the first approach to explicitly address the challenge of distinguishing adjacent categories in acne severity grading tasks. Experimental results on the public ACNE04 dataset demonstrate that AFLL significantly outperforms existing methods, establishing a new state-of-the-art in acne severity grading.**

This work was supported by the Key Science and Technology Research Projects of Quzhou under Grant 2022K50 and the Natural Science Foundation of Zhejiang Provincial under Grant LGF22G010009; the Key Research & Development Program of Heilongjiang Province under Grant 2023X01A08, the National Natural Science Foundation of China under Grants 62272135, 62372135; the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20242214; the China Postdoctoral Science Foundation under Grants 2024M754207.(Corresponding author: Lifeng Xu and Xiangyu Li)

Murong Zhou is with Quzhou Afffliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou, China; and also with the college of Computer and Control Engineering, Northeast Forestry University, Harbin, China (email: doriszmr@nefu.edu.cn).

Guohua Wang and Suyu Dong are with the College of computer and control engineering, Northeast Forestry University, Harbin, China(email: ghwang@nefu.edu.cn, dongsuyu@126.com).

Baifu Zuo, Gongning Luo, Fanding Li, Kuanquan Wang and Xiangyu Li are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China (email: zuobaifu@stu.hit.edu.cn, luogongning@hit.edu.cn, lifanding@stu.hit.edu.cn, wangkq@hit.edu.cn, lixiangyu@hit.edu.cn). Wei Wang is with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China (email: wangwei2019@hit.edu.cn).

Lifeng Xu is with the Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou, China (email: qz1109@wmu.edu.cn).

Code is available online: https://github.com/PerceptionComputingLab/AFLL

***Index Terms*—Acne severity grading, Fuzzy Label learning, Fuzzy set theory, Binary search method.**

I. Introduction

AUTOMATIC grading of skin disease severity, particularly acne severity grading(Fig.1), is of significant importance in medical image analysis. Acne vulgaris, commonly known as acne or pimples, is the most prevalent inflammatory skin disease linked to hair follicles and sebaceous glands, with the highest incidence among adolescents [1]. Globally, it affects approximately 650 million people, with over 85% of adolescents being impacted [2]. Due to its chronic nature, acne can persist for several years, adversely affecting patients' quality of life, self-esteem, and emotional well-being, while increasing the risk of anxiety, depression, and suicidal ideation [1], [3]. Thus, timely and effective diagnosis and treatment of acne are crucial. Meanwhile, accurate grading of acne severity is an essential step for acne diagnosis, and is significant for devising personalized treatment plans and monitoring patient's response over time [4]. Dermatologists commonly use the Hayashi criteria [4], which classifies acne severity into four levels: mild, moderate, severe, and very severe, based on lesion count. Traditional assessment methods rely on the observation and expertise of dermatologists, which are time-consuming and often struggle to deliver precise severity assessments [5]. Therefore, there is a pressing need for an accurate computer-aided acne severity grading method to support effective acne treatment planning.

Inspired by the fuzzy set theory in medical image analysis [6]–[8], the acne severity grading method based on fuzzy label learning (FLL) has demonstrated distinct advantages over other approaches. Shen et al. [5] extracted image features using a convolutional neural network, followed by applying an automated classifier for acne severity grading. The KIEGLFN method [9] enhanced critical details by first removing irrelevant information through image preprocessing techniques and then simulated a dermatologist's global estimation for severity grading. [10] predicted the lesion region using a segmentation-based method, employing it as diagnostic evidence, and then fused this evidence with the corresponding image to train a downstream diagnostic model. While these methods have made significant progress, they often overlook the inherent uncertainties in acne severity grading labels, which hinders further improvements in grading accuracy. Label uncertainty,

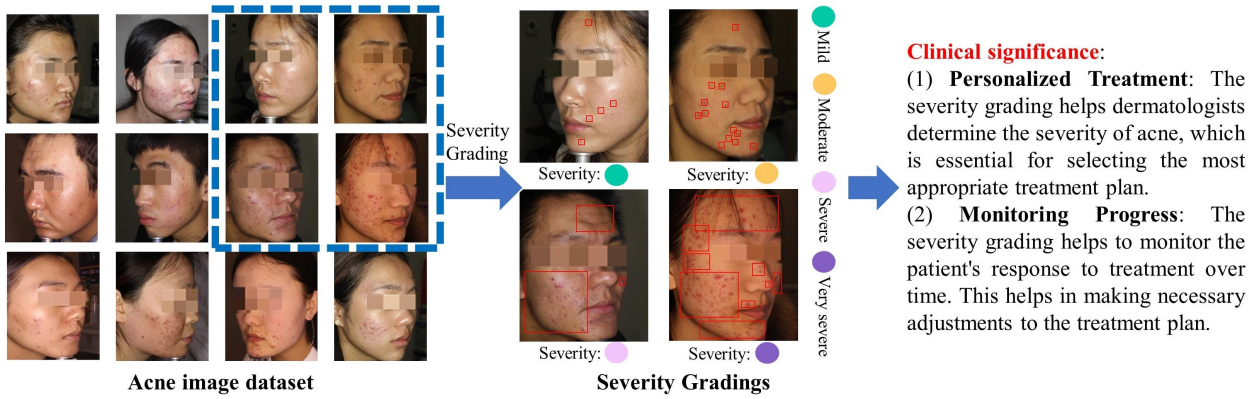


Fig. 1. The clinical significance of accurate acne severity grading for devising personalized treatment plans and monitoring patient's response over time.

reflecting the varying judgments by different individuals on the same acne image, contains valuable correlation information among inputs with adjacent labels, making it crucial for accurate acne severity grading. Following the ideas in fuzzy set theory, many researchers addressed this problem with fuzzy logic to effectively model uncertain information [4], [11]. They first construct a fuzzy label set with the possible acne severity grading values, and utilize the Gauss distribution as the fuzzy membership function to assign each label a degree of belonging to the given sample. They transform single-value labels into multiple fuzzy membership degree sets, and conduct fuzzy label learning upon them. Their methods efficiently leverage label uncertainty information, significantly enhancing acne severity grading performance.

Although existing FLL-based methods are highly competitive, they often reduce the discriminability of latent features between adjacent classes, making it challenging for models to distinguish between closely related severity gradings. During the construction of fuzzy label set (a probability distribution in our case), methods that create these label sets based on the proximity of labels result in very similar latent features for adjacent categories due to their sequential closeness. As a consequence, existing FLL-based approaches struggle to emphasize the boundaries between neighboring categories, leading to suboptimal performance in distinguishing them and limiting further improvements in the model's grading accuracy.

Binary search method [12]–[14], which recursively splits the search sequence to gradually find the target value, has the potential to enhance a model's ability to discriminate between adjacent categories. Although it has not yet been applied to the task of acne severity grading, it has proven to be highly effective in various computer vision tasks [13], [14]. The binary search approach transforms the image grading task into multiple binary classification tasks by first constructing a sequence tree and then iteratively refining the search interval using a hierarchical decision-making strategy (Fig.2). This coarse-to-fine strategy progressively narrows down the candidate classes, improving the discriminability between adjacent categories and ultimately achieving more accurate class predictions.

However, directly applying the binary search method in FLL-based approaches to address the challenge of adjacent severity grading presents several difficulties: **(1) Existing sequence tree construction methods often struggle to create a practical decision tree that enables the progressive discrimination of adjacent categories, leading to biased and less discriminative models.** As previously mentioned, the binary search method begins with constructing a sequence tree to iteratively refine the search interval through a hierarchical decision-making strategy. However, existing approaches [13], [14] often fail to account for variations in semantic similarity between neighboring categories and the issue of class imbalance within the dataset, which impairs the ability to differentiate between classes and results in biased model outcomes. **(2) The hierarchical decision-making strategies employed in current binary search methods fail to effectively mitigate error propagation, leading to significant performance degradation.** These strategies adopt a coarse-to-fine approach to iteratively narrow the search interval across multiple levels. At each level, existing methods [13], [14] typically rely on predictions made at the previous level, without adequately addressing the potential errors introduced at that stage. This oversight leads to substantial error propagation, ultimately compromising the accuracy of the final prediction outcomes. **(3) Existing FLL methods struggle to simultaneously capture the differences for both high and low fuzzy memberships in fuzzy label set and often overlook the sequential relationships among different acne severity labels, leading to unstable fuzzy label learning.** Existing fuzzy label learning (FLL) techniques typically use KL divergence or JS divergence for loss calculation. However, these approaches struggle to capture the differences between the low fuzzy memberships of the ground-truth fuzzy label set and the corresponding memberships in the predictions. Moreover, existing FLL methods neglect the sequential patterns inherent to the labels, making it difficult to differentiate between order-sensitive distributions. Order-sensitive distributions involve labels with a meaningful sequence, such as severity levels where the order represents increasing severity.

In this work, we propose a novel **A**djacency-aware **F**uzzy **L**abel **L**earning (**AFLL**) framework to address these challenges. Specifically, the proposed Adjacency-aware Decision Sequence Generation (ADSG) introduces a new Sequence Tree

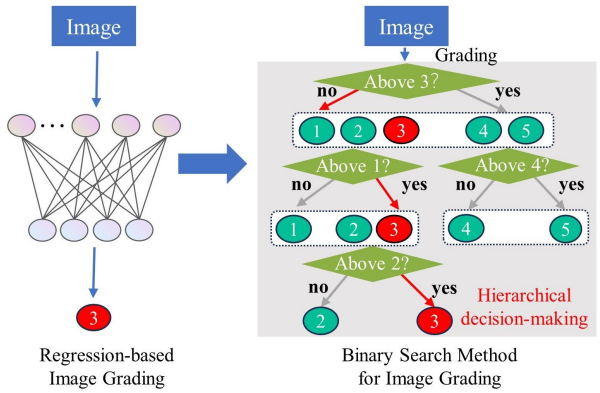


Fig. 2. The binary search method has the potential to enhance a model's ability to discriminate between adjacent categories by transforming image grading task into multiple binary classification tasks. (a) The regression-based image grading;(b) The binary search method for image grading with a hierarchical decision-making strategy.

Construction (STC) method, which enhances the sequence tree by reducing bias and improving discriminative ability through the effective incorporation of semantic similarity and class distributions. Furthermore, the Consistency-guided Decision Sequence Prediction (CDSP) introduces a novel selective masking decision strategy that alleviates error propagation in hierarchical decision-making by dynamically and selectively accessing previous predictions. Additionally, the proposed Sequential Conjoint Distribution Loss (SCDL) improves fuzzy label learning by simultaneously capturing the differences for both high and low fuzzy memberships across the entire fuzzy label set and modeling the internal temporal order among different acne severity labels with a cumulative distribution. The main contributions of our paper are four-fold:

- To the best of our knowledge, the proposed AFLL framework is the first to address the challenge of distinguishing adjacent categories in acne severity grading tasks.
- The ADSG proposes a new sequence tree construction method, which enhances the sequence tree with reduced bias and improved discriminative ability by effectively incorporating the semantic similarity between adjacent categories and their class distributions.
- The CDSP introduces a novel selective masking decision strategy, which effectively mitigates the impact of errors in preceding decisions on subsequent ones by dynamically and selectively accessing prior predictions.
- The SCDL introduces a novel sequential conjoint distribution loss, significantly improving fuzzy label learning by enabling a more comprehensive assessment of differences between distributions.

II. Related Work

A. Acne Severity Grading

Automatic acne severity grading has seen significant progress in recent years. Traditionally, the analysis of acne lesions has relied primarily on manually crafted features. For instance, Chantharaphaichi et al. [15] employed a binary thresholding method for feature extraction, and then the extracted features are utilized for acne detection, paving the way for subsequent severity grading. Maroni et al. [16] proposed an advanced feature engineering method by integrating various attributes, including color, texture and morphological characteristics. They also implemented a rigorous feature importance selection process to identify the most salient features, ultimately using these features to achieve acne severity grading. Compared to methods relying on handcrafted features, approaches based on Convolutional Neural Networks (CNNs) can effectively learn high-level feature representations, significantly improving severity grading performance [17]. Lin et al. [18] achieved effective acne severity grading by leveraging both local and global skin features within a deep CNN network. However, these methods often overlook the ambiguity in acne severity labels, which limits the potential for further refinement in grading capabilities. [4], [11] addressed this issue by modeling ambiguity information through the transformation of single-value labels into label distributions, effectively leveraging label ambiguity and achieving superior grading performance.

Despite the progress made by existing acne severity grading methods, they still face the challenge of simultaneously leveraging inherent label uncertainty and effectively distinguishing between adjacent categories, which limits further performance improvements. The proposed method addresses this challenge by introducing an iterative refinement strategy in FLL-based acne severity grading, enabling more effective discrimination between adjacent classes and achieving superior performance.

B. Fuzzy Label Learning

Fuzzy Label Learning (FLL) is a novel machine learning paradigm designed to address the issue of label uncertainty [19], [20]. In consistent with the fuzzy set theory, fuzzy label learning (FLL) assigns fuzzy memberships to each label for a given sample, and all the fuzzy labels constitute a probability distribution, which acts as a fuzzy set over possible labels. Each label has a membership degree, indicating the extent to which the sample belongs to that label. This enables FLL to handle ambiguity and uncertainty in the labeling process. Recently, FLL has been widely applied in various computer vision tasks, including age estimation, facial expression assessment and etc. In age estimation tasks, to tackle label ambiguity caused by multiple annotators, Yang et al. [21] transformed age values to multiple fuzzy labels and conducted an age distribution learning. Similarly, to address label ambiguity in bone age labels, Chen et al. [22] proposed a bone age prediction method that combines bone age distribution learning with regression, effectively leveraging correlations between adjacent bone age labels and achieving robust estimation. Furthermore, recognizing that different raters may assign varying scores to a facial image, Fan et al. [23] associated each image with a fuzzy set of score labels provided by multiple raters, thereby establishing a facial expression learning framework based on fuzzy label learning.

Although existing fuzzy label learning methods have make great progress in addressing label uncertainty, they still face

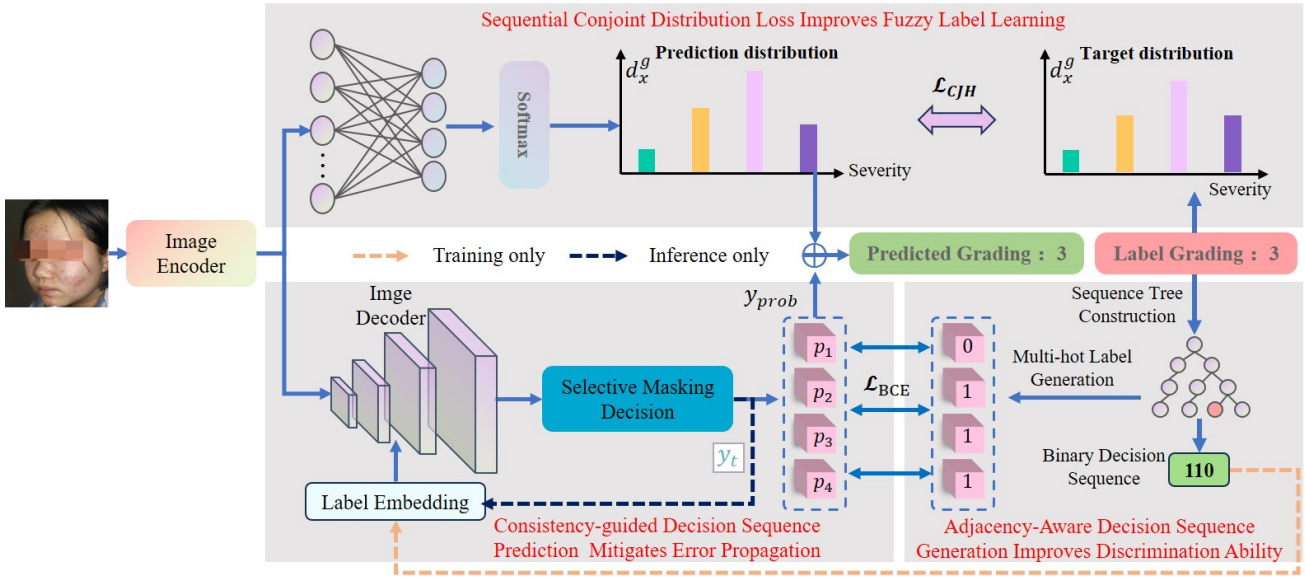


Fig. 3. The AFLL effectively addresses the challenge of distinguishing between adjacent categories with the proposed innovations. The ADSG enhances the sequence tree with reduced bias and improved discriminative ability by effectively incorporating the semantic similarity of adjacent categories and their class distributions. The CDSP mitigates the problem of error propagation in hierarchical decision-making by dynamically and selectively referencing previous predictions. The proposed SCDL significantly improves the performance of fuzzy label learning by offering a comprehensive loss function for measuring the discrepancy between distributions.

the challenge of the instability in fuzzy label learning due to the neglect of subtle differences for low fuzzy memberships and the intrinsic order of the labels. The proposed AFLL framework addresses these challenges by simultaneously capturing the differences for both high and low fuzzy memberships across the entire fuzzy label set. This approach significantly enhances the performance of fuzzy label learning by effectively addressing these problems.

III. Methods

The proposed AFLL framework (Fig.3) introduces an innovative approach to fuzzy label learning in the context of acne severity grading, effectively addressing the challenge of distinguishing between adjacent categories.

A. Fuzzy Label Learning with Gaussian Distribution

The fuzzy label learning (FLL) paradigm is an innovative machine learning method designed to address the issue of label uncertainty [19], [20], [24]. In this framework, given an input image $\boldsymbol{x}_i$ ($i \in 1, 2, ..., n$, where n represents the number of samples), the FLL first constructs a fuzzy set with all possible labels in the label space based on the fuzzy set theory. It then assigns a real value $d_{x_i}^{l_j}$ as the fuzzy membership degree to each possible label l_j in the fuzzy set, indicating the extent to which the input image $\boldsymbol{x}_i$ belongs to l_j, all the fuzzy membership degrees constitute a probability distribution. Typically, this label distribution is generated based on the ground-truth label associated with the input image, often using a Gaussian function. The degree to which a specific label l_j describes the input image $\boldsymbol{x}_i$ can be expressed as:

$$d_{x_i}^{l_j} = \frac{G(l_j \mid \mu, \sigma)}{\sum_{k=1}^{Z} G(l_k \mid \mu, \sigma)} = \frac{1}{\sqrt{2\pi}\sigma T} \exp(-\frac{(l_j - \mu)^2}{2\sigma^2}) \quad (1)$$

where T is a regularization factor and μ is the mean value of the Gaussian distribution. σ is a hyperparameter that controls the sharpness of the generated distribution. Z is the number of classes. Different from the vanilla fuzzy set theory [25], for image sample $\boldsymbol{x}_i$, the label distribution $\boldsymbol{d}_i = \{d_{x_i}^{l_1}, d_{x_i}^{l_2}, \ldots, d_{x_i}^{l_Z}\}$ needs to satisfy two rules: (1) $d_{x_i}^{l_j} \in [0, 1]$ and $\Sigma_{j=1}^{Z} d_{x_i}^{l_j} = 1$. (2) If l_j is the ground-truth label, then l_j has the highest probability value for the input image $\boldsymbol{x}_i$, and the labels further away from the ground-truth label have lower probabilities. For the first rule, the probability distribution $\boldsymbol{d_i}$ is constrained using a regularization factor T to ensure compliance with this rule, where T is defined by the following formula:

$$T = \sum_{k=1}^{Z} G(l_k \mid \mu, \sigma) = \sum_{k=1}^{Z} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(l_k - \mu)^2}{2\sigma^2}) \quad (2)$$

For the second rule, the requirement is fulfilled by setting the mean μ of the Gaussian distribution to the ground-truth label corresponding to the input image $\boldsymbol{x}_i$.

For the training of fuzzy label learning model, the FLL paradigm calculates the KL divergence [24], [26], [27] or JS divergence loss [28] between the ground truth and the predicted distributions. The model parameters are then optimized using the backpropagation algorithm.

B. Adjacency-aware Decision Sequence Generation Improves the Discrimination of Adjacent Categories

The Adjacency-aware Decision Sequence Generation (ADSG) introduces a novel sequence tree construction method that significantly enhances the sequence tree by reducing bias and improving discriminative ability. Leveraging the generated sequence tree, ADSG further converts acne severity level labels into binary decision sequences and generates multi-hot labels to serve as targets for training the binary search model.

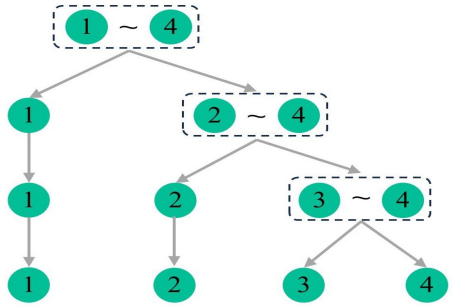


Fig. 4. An illustration of the generated sequence tree utilized in this paper, constructed using the proposed sequence tree construction algorithm.

1) Sequence Tree Construction: The proposed sequence tree construction method enables progressive discrimination of adjacent categories by effectively incorporating semantic similarities and class distributions. Specifically, using the binary search method, we design a sequence tree that transforms each acne severity category label into a binary decision sequence. An example of a constructed sequence tree is shown in Fig.4. During the tree construction process, a critical step involves splitting the class node at each layer of the tree. This challenge can be framed as identifying the division node within a set of nodes (i.e., $L = l_1, l_2, ..., l_Z$). Two key issues complicate the discrimination of adjacent categories when determining the division node: **(1) Homogeneity in Semantic Space:** Different acne severity labels represent distinct semantic categories, yet existing sequence tree construction methods often assume these severity labels shares the same representations in semantic space. This assumption of homogeneity makes it more challenging for the model to effectively differentiate between adjacent severity categories. **(2) Imbalanced Class Distribution:** In each layer of the sequence tree, class distributions are often imbalanced, making it challenging to learn an unbiased model.

To address these challenges, the proposed ADSG method incorporates both semantic similarity between adjacent categories and class distributions when determining the division node index. The primary criterion for splitting class nodes is to maximize the difference between the target division node and its neighboring nodes, thereby facilitating easier discrimination during binary classification. To achieve this, we first model the semantic similarity of adjacent class nodes by measuring the Quadratic Form Distance (QFD) [29] of their corresponding label distributions:

$$QFD(\boldsymbol{d}_i, \boldsymbol{d}_j) = \sqrt{(\boldsymbol{d}_i - \boldsymbol{d}_j)^T \cdot S \cdot (\boldsymbol{d}_i - \boldsymbol{d}_j)} \tag{3}$$

where $\boldsymbol{d}_i = \{d^{l_1}_{\boldsymbol{x_i}}, d^{l_2}_{\boldsymbol{x_i}}, \ldots, d^{l_Z}_{\boldsymbol{x_i}}\}$ and $\boldsymbol{d}_j = \{d^{l_1}_{\boldsymbol{x_j}}, d^{l_2}_{\boldsymbol{x_j}}, \ldots, d^{l_Z}_{\boldsymbol{x_j}}\}$ are two acne severity label distributions; $S \in R^{Z \times Z}$ is a correlation matrix, which captures the semantic relationships between different acne severity levels. The elements s_{ij} in the matrix S represent the semantic similarity between the i-th and j-th acne severity grade labels. It is denoted as follows:

$$s_{ij} = 1 - \frac{g_{ij}}{g_{max}} \tag{4}$$

where $g_{ij} = |i - j|$, $g_{max} = \max_{i,j} g_{ij}$. It is observed that adjacent acne severity labels exhibit higher semantic similarity, hence s_{ij} should approach 1. Conversely, if the i-th label is distant from the j-th label, s_{ij} should approach 0. Unlike common similarity measures such as Cosine [30] and Intersection [31], which only consider point-to-point differences in the distribution, the proposed similarity modeling method captures deeper semantic relationships between different acne severity levels. It achieves this by simultaneously modeling correlations in both the label distribution and the original label space.

Although it is possible to identify the target division node by directly searching for the index that maximizes the Quadratic Form Distance (QFD) in Eq.(3), this approach may introduce bias due to the imbalanced data distribution across different classes. To address this, we introduce a regularization term to ensure that the number of samples in the left and right subtrees after partitioning is balanced. Given the set of sample sizes for different categories, $C = \{c_1, c_2, ..., c_Z\}$, and assuming the i-th node is the partition node, the ratio of samples in the left and right subtrees after partitioning is calculated using the following formula:

$$\text{ratio}(i) = \frac{\min(\sum_{k=1}^{i} c_k, \sum_{k=i+1}^{Z} c_k)}{\max(\sum_{k=1}^{i} c_k, \sum_{k=i+1}^{Z} c_k)} \tag{5}$$

Considering all the aforementioned factors, the division node index is determined by maximizing the product of QFD similarities and the class ratio regularization term:

$$i^* = \underset{1 \le i < Z}{\text{argmax}}(ratio(i) * QFD(\boldsymbol{d}_i, \boldsymbol{d}_{i+1})) \tag{6}$$

After the partition operation, two new sets of nodes are generated: $L_{left} = \{l_1, \ldots, l_{i^*}\}$ and $L_{right} = \{l_{i^*+1}, ..., l_Z\}$. The new partition node index is then determined sequentially based on the node sets L_{left} and L_{right} using Eq.(6). This process is repeated until only one node remains in all generated node sets. The detailed construction process of the sequence tree is outlined in Algorithm 1.

Algorithm 1 STC

Input: The node set $L = \{l_s, l_{s+1}, \ldots, l_e\}$, starting index s=1, ending index $e = Z$ and the collection containing the set of nodes at each layer $V = \{L\}$

Output: The sequence tree generation series V

```
if s == e then
    return -1;
end if
Compute correlation matrix S by Eq.(4);
Get the index of the division node i* by Eq.(6);
L_left ← {l_s, ..., l_i*}, L_right ← {l_i*+1, ..., l_e};
V ← V + {L_left, L_right};
Call STC(L_left, s, i*, V);
Call STC(L_right, i* + 1, e, V);
```

2) Binary Decision Sequence Generation(BDSG): The proposed BDSG transforms single-value acne severity labels into binary decision sequences using the generated sequence tree. Specifically, each node in the sequence tree has two possible paths, represented by the left and right subtrees, denoted as

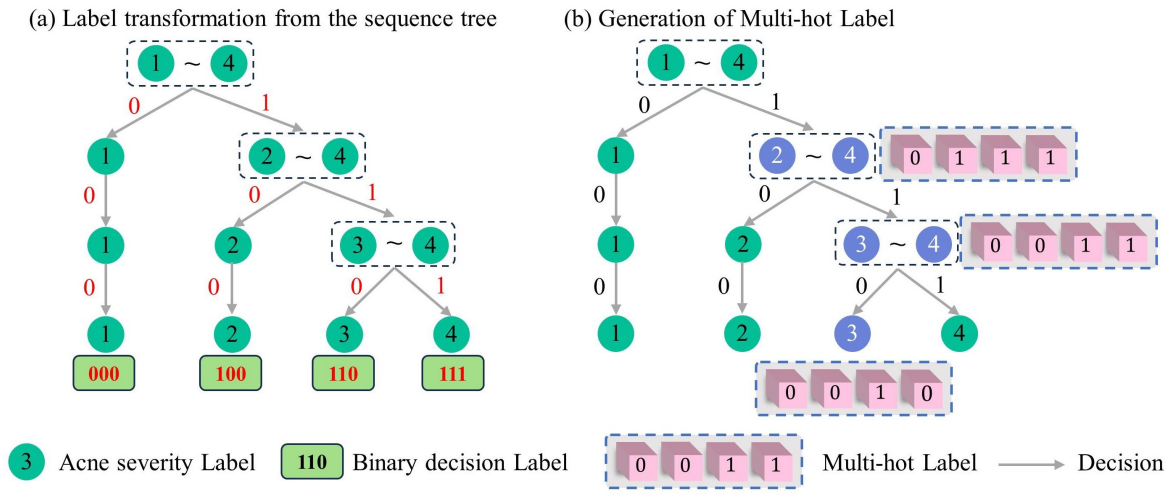


Fig. 5. An illustration of the label transformation in adjacency-aware decision sequence generation. (a) The process of binary decision sequence generation based on the sequence tree. (b) The process of the multi-hot label generation based on the sequence tree.

0 and 1, respectively. As illustrated in Fig.5(a), each acne severity label G is transformed into a corresponding binary decision sequence y_b using the sequence tree, which represents the path chosen for the acne severity level G from the root node to a leaf node. This transformation is represented by the following formula:

$$y_b = E(G) = [b_1, b_2, ..., b_h] \tag{7}$$

where $b_i \in \{0, 1\}$ indicates the two possible paths encoded by the left and right subtrees of the acne severity level G; E represents the label transformation function; and h is the depth of the sequence tree. Additionally, during the sequence prediction process, to ensure that each output at a given position depends only on the information from previous positions, the binary decision sequence is shifted one position to the right, and the first position is filled with a $start$ marker symbol:

$$y_{target} = [start, b_1, b_2, ..., b_{h-1}] \tag{8}$$

After the label transformation, the prediction target changes from acne severity level label to binary decision sequence.

3) Multi-hot Label Generation(MLG): Based on the binary decision sequence generated in Sec.III-B2, the most straightforward approach would be to use these sequences as targets and directly predict them with a neural network. However, this approach can degrade model performance due to the inherent ambiguity in the binary encoding of the decision sequence. For instance, the binary classifications at different layers of the sequence tree represent distinct meanings (e.g., the first decision differentiates among acne severity levels 1-4, while the second decision narrows it down to levels 2-4, as shown in Fig.5(a)). To address this ambiguity in binary encoding, we followed the method proposed by [13] and generated multi-hot label sequences based on the constructed sequence tree. An illustration of multi-hot label generation is presented in Fig. 5(b). In the sequence tree, each node corresponds to a multi-hot label. The acne severity level G is converted into a multi-hot label sequence as follows:

$$y_{mht} = M(G) = [m^{(1)}, m^{(2)}, ..., m^{(h)}] \tag{9}$$

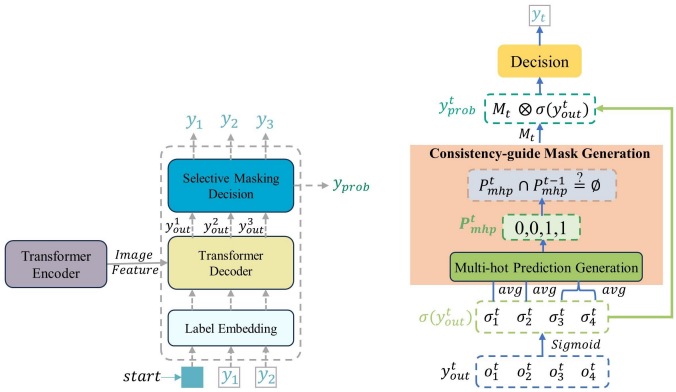
where $m^{(i)} = [m_1^{(i)}, m_2^{(i)}, ..., m_Z^{(i)}]$ represents a specific multi-hot label corresponding to the acne severity level G, $m_j^{(i)} \in \{0,1\}$. Ultimately, the model utilizes this multi-hot label sequence as the target for predicting probability sequences, which are then used to derive the binary decision sequence.

C. Consistency-guided Decision Sequence Prediction Effectively Mitigates the Error Propagation Problem

The proposed Consistency-Guided Decision Sequence Prediction (CDSP) introduces a novel selective masking decision strategy that effectively mitigates the error propagation problem in hierarchical decision-making and enhances the distinction between adjacent categories. Specifically, Fig.6(a) illustrates the comprehensive framework of CDSP, which comprises three key components: Label Embedding, Transformer Decoder, and Selective Masking Decision. Similar to the approaches in [32] and [13], the Label Embedding component converts the binary label sequence into a label embedding through the following formula:

$$y_{embedding} = E\left(y_{target}\right) \tag{10}$$

where E represents the embedding layer. The Transformer Decoder component operates according to the standard Transformer architecture. The Selective Masking Decision (SMD) strategy then converts the output y_{out} of the Transformer Decoder into the predicted probability sequence y_{prob} and the predicted binary decision sequence y. The predicted probability sequence y_{prob} is subsequently used for loss computation in binary decision sequence prediction. Typically, y_{prob} is obtained by computing $y_{prob} = \sigma(y_{out})$. However, observations suggest that predictions at the current time step t should be influenced by the forecast results from the previous time step



(a) The CDSP framework (b) The Selective Masking Decision

Fig. 6. The proposed CDSP effectively mitigates the error propagation problem and improves the prediction of adjacent categories. (a) An illustration of the CDSP model. (b) The proposed selective masking decision strategy.

$t-1$. While this approach is logically sound, it introduces a critical issue: error propagation. For instance, if a prediction error occurs at time step $t-1$, basing the prediction for the current time step t on this erroneous result will introduce interference, ultimately compromising the accuracy of the final sequence prediction outcomes.

To address the issue of error propagation, we introduce a novel decision-making strategy called Selective Masking Decision (SMD). The proposed SMD selectively masks nodes in the sequence tree by evaluating the consistency of predicted results between previous and current time steps. The selective masking decision process for a specific time step, $t=2$, is illustrated in Fig.6(b). Specifically, given an output y_{out}^t from the Transformer decoder, the SMD first applies a Sigmoid function to obtain the output probabilities $\sigma(y_{out}^t)$. Instead of directly using $\sigma(y_{out}^t)$ for decision-making at the current step, the SMD generates a node mask based on these output probabilities. The decision for the current step is then made by multiplying $\sigma(y_{out}^t)$ with the generated mask. To generate this node mask, the SMD introduces a consistency-guided mask generation (CMG) method. According to the structure of the sequence tree in Fig.5, the CMG first calculates the average probabilities for different nodes at layer t (e.g., for $t=2$, this would involve σ_1^t, σ_2^t, and the combination of σ_3^t and σ_4^t). The CMG then identifies the index of the maximum value among these averages and generates the corresponding multi-hot prediction P_{mhp}^t (e.g., if the index is 2, the multi-hot prediction would be $[0,0,1,1]$). Considering that the multi-hot predictions at two consecutive time steps are likely to have at least one matching position, the CMG assesses the consistency between the multi-hot predictions from the current and previous time steps:

$$C_t = P_{mhp}^t \cap P_{mhp}^{t-1} \tag{11}$$

According to the consistency, the CMG assign the node mask as:

$$M_t = \begin{cases} P_{mhp}^{t-1} & C_t \neq \emptyset, \\ \mathbf{1} & C_t = \emptyset, \end{cases} \tag{12}$$

The key insight is that it definitely exists error predictions in step t or $t-1$ if the consistency $C_t = \emptyset$. In such cases, the CMD skips the masking process to prevent error propagation. Otherwise, the CMD utilizes the multi-hot prediction from the previous step as a node mask to effectively reduce the interference caused by categories that were already eliminated in the prior time step:

$$y_{prob}^t = M_t \bigotimes \sigma(y_{out}^t) \tag{13}$$

where $\bigotimes$ represents element-wise multiplication.

Following the selective masking process, a decision strategy is applied to predict the binary decision label based on the output probability y_{prob}^t. At time step t, assume the categories in the left subtree fall within $[l,m]$, and those in the right subtree fall within $[m+1,r]$. Following the decision strategy outlined in [13], we compute the averages of all probability values in the left and right subtrees, respectively, and compare them. The binary decision label y_t is then derived from the comparison results, as follows:

$$\begin{aligned} P_{left}^t &= \frac{1}{m-l+1}\sum_{i=l}^{m} y_{prob}^{t,i}, \\ P_{right}^t &= \frac{1}{r-m}\sum_{i=m+1}^{r} y_{prob}^{t,i}, \end{aligned} \tag{14}$$

$$y_t = \begin{cases} 0 & P_{left}^t \geq P_{right}^t, \\ 1 & P_{left}^t < P_{right}^t, \end{cases} \tag{15}$$

where P_{left}^t and P_{right}^t denote the averages of all probability values corresponding to categories in the left and right subtrees, respectively. The prediction y_t will be used as the input for predicting the subsequent binary decision label in the inference process. As more binary decision labels are predicted, the number of remaining candidate categories gradually decreases, allowing adjacent categories to be distinguished with increasing confidence.

D. Sequential Conjoint Distribution Loss Improving the Fuzzy Label Learning

The proposed Sequential Conjoint Distribution Loss (SCDL) effectively complements existing distribution loss by simultaneously capturing the differences for both high and low fuzzy memberships across the entire fuzzy label set, while also leveraging the sequential relationships among different acne severity labels. Specifically, to effectively leverage the ambiguous information between acne severity levels, the proposed AFLL framework transforms the acne severity grading task into a fuzzy label learning problem, training the model by calculating the KL divergence or JS divergence between the ground truth distribution and the predicted distribution.

However, directly learning the acne severity distribution by minimizing the KL or JS divergence presents two key limitations: (1) KL and JS divergences are more sensitive to differences in high fuzzy memberships of the distributions, often overlooking subtle differences for low fuzzy memberships, which can result in model instability. (2) These divergences are typically applied on a per-label basis during optimization,

neglecting the sequential patterns inherent to the labels. This omission leads to a loss of crucial contextual and relational information between the labels, thereby limiting the prediction performance.

To address the first limitation, the proposed AFLL enhances the traditional KL and JS divergence methods by incorporating additional loss functions that focus on subtle differences in low fuzzy memberships. Specifically, we introduce the Hellinger distance [33] as a complementary objective. The Hellinger distance is particularly sensitive to small differences in regions where both distributions have low probabilities, offering a more comprehensive measurement of the overall difference between two distributions. The formula for the Hellinger distance loss is as follows:

$$H\left(p,q\right)=\sqrt{2\sum_{i=1}^{z}(\sqrt{p\left(i\right)}-\sqrt{q\left(i\right)})^{2}} \tag{16}$$

where p and q are predicted and label distributions, respectively. For the ordinary distribution loss, this paper chooses JS divergence because of its symmetry character. The JS divergence is given by:

$$D_{JS}(p||q)=\frac{1}{2}\sum_{j=1}^{Z}\left(p(j)\log\frac{p(j)}{M(j)}+q(j)\log\frac{q(j)}{M(j)}\right) \tag{17}$$

where $M(j)=(p(j)+q(j))/2$, Z represents the number of labels.

To address the second limitation, the proposed AFLL models the internal sequential patterns of labels by introducing an empirical cumulative distribution loss. In this framework, each label's value represents the cumulative sum of the values of all preceding labels, effectively incorporating temporal order information from previous labels. We apply this cumulative approach to a combination of JS divergence and Hellinger distance, with the loss function defined as follows:

$$\begin{aligned}\mathcal{L}_{CJH}=\beta*\sum_{n=1}^{Z}D_{JS}(CDF_{n}(p)||CDF_{n}(q))\\+(1-\beta)*\sum_{n=1}^{Z}H(CDF_{n}(p),CDF_{n}(q))\end{aligned} \tag{18}$$

where β is a hyperparameter, CDF$(\cdot)$ represents the cumulative density function, $\mathrm{CDF}_n(p)=\sum_{j=1}^{n}p(j)$.

E. Loss Function

The loss function of the model proposed in this paper comprises two major components: the loss $\mathcal{L}_{CJH}$ in the fuzzy label learning branch, as is denoted in Sec.III-D and the loss $\mathcal{L}_{BCE}$ in the sequence tree branch. The overall loss function of the model is formulated as follows:

$$\mathcal{L}=\mathcal{L}_{BCE}+\lambda\mathcal{L}_{CJH} \tag{19}$$

where λ is a hyper-parameter that determines the weighs of different loss functions. We utilize binary cross-entropy (BCE) loss in the sequence tree branch.

The primary motivation is that sequence prediction within the sequence tree branch can be viewed as a multi-label classification task using multi-hot labels. Therefore, treating this task as a series of binary classification problems and employing the Binary Cross-Entropy (BCE) loss is more effective than using Categorical Cross-Entropy (CE) loss. To compute the BCE loss for the model, the loss between y_{prob}^{t} and y_{mht}^{t} at each time step t is calculated and then summed, as shown below:

$$\begin{aligned}\mathcal{L}_{BCE}=\sum_{t=1}^{h}BCE(y_{prob}^{t},y_{mht}^{t})=-\frac{1}{n}\sum_{t=1}^{h}\sum_{i=1}^{C}\\(y_{mht}^{t,i}\log(y_{prob}^{t,i})+(1-y_{mht}^{t,i})\log(1-y_{prob}^{t,i}).\end{aligned} \tag{20}$$

where h is the length of the binary decision sequence.

Table I
THE DETAILED DISTRIBUTION OF THE TRAINING AND TESTING SETS OF THE ACNE04 DATASET AND THE MEDICAL CRITERION. THE CRITERION INDICATES THE RELATIONSHIP BETWEEN SEVERITY LEVEL AND NUMBER OF LESIONS.

Level of severity	Criterion	Number of images	
		Training	Testing
Mild	$1\sim5$	410	103
Moderate	$6\sim20$	506	127
Severe	$21\sim50$	146	36
Very severe	>50	103	26
Total	-	1,165	292

IV. Experimental Setup

A. Materials

To demonstrate the effectiveness of the proposed method, we utilized the publicly available ACNE04 dataset [4]. This dataset, which follows the Hayashi standard [34], provides facial acne diagnosis with severity graded as mild (label 0), moderate (label 1), severe (label 2), and very severe (label 3), corresponding to lesion counts of 1-5, 6-20, 21-50, and over 50, respectively. The ACNE04 dataset contains 1,457 images, with 80% allocated to the training set (1,165 images) and the remaining 20% to the testing set (292 images). A detailed description of the ACNE04 dataset is provided in Table I.

B. Evaluation Metrics

Following the implementation in [35], we also employed accuracy (**ACC**), precision(**PR**), sensitivity(**SE**), specificity(**SP**) and the Yorden Index(**YI**) as evaluation metrics. Accuracy and precision are standard metrics for classification problems, while sensitivity, specificity, and the Yorden Index are particularly relevant for medical diagnosis.

C. Implementation Details

Both branches of our model were trained simultaneously on an NVIDIA RTX 3090 GPU. We employ the PVTv2-b1 [36] as the backbone network, pretrained on the ImageNet dataset [37], and combine it with a vanilla Transformer encoder for

Table II
QUANTITATIVE RESULTS DEMONSTRATE THE SUPERIORITY OF OUR AFLL ON THE ACNE SEVERITY GRADING TASK. OUR AFLL OBTAINED THE BEST PERFORMANCE ON ALL METRICS BY CROSS-VALIDATION. BOLD TEXT DENOTES THE BEST RESULT FOR THAT METRIC. 'CLA' STANDS FOR THE CLASSIFICATION-BASED METHODS, WHILE 'FLL' REFERS TO THE FUZZY LABEL LEARNING-BASED METHODS; '-' MEANS THAT THE AUTHOR DIDN'T REPORT THAT METRIC. PR, SP, SE, YI, AND ACC: PRECISION, SENSITIVITY, SPECIFICITY, YORDEN INDEX AND ACCURACY.

Methods	Type	Metrics(%)				
		PR↑	SP↑	SE↑	YI↑	ACC↑
KIEGLFN [9]	CLA	83.58	94.11	81.95	76.06	84.52
Liu et al. [39]	CLA	85.56	94.55	83.71	78.26	85.82
EGPK [40]	CLA	84.01	94.40	84.62	79.01	85.27
Zhang et al. [10]	CLA	84.48	94.56	85.27	79.83	85.55
DED [35]	CLA	85.31	94.66	84.83	79.48	86.06
SLL [4]	CLA	75.81	-	-	67.21	78.42
JGC [4]	FLL	84.37	93.80	81.52	75.32	84.11
OLDL [11]	FLL	86.32	94.00	83.50	76.89	84.80
AFLL(Ours)	FLL	**87.38**	**95.69**	**86.41**	**82.11**	**88.56**

image encoding. Following [4], data preprocessing involves resizing the original images to 256×256, followed by random cropping to 224×224. For data augmentation, we follow the same approach as in [4], we apply random horizontal flipping and rotation. The data is then normalized by subtracting the mean and dividing by the standard deviation. We use the Adam optimizer [38] with a mini-batch size of 32. Training includes a warm-up iteration of 1 and a maximum of 1700 iterations, with an initial learning rate of 1.0×10^{-4} and no learning rate decay. Additionally, five-fold cross-validation is used to evaluate the stability of the results.

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present a series of experiments designed to demonstrate the advantages of the proposed AFLL method. We begin by comparing AFLL with current state-of-the-art acne severity grading methods to establish its superiority. Following this, we conduct several ablation studies to assess the impact of the innovations introduced in our method. Additionally, we analyze the network architecture under various hyperparameter settings. Finally, we evaluate the practical application of the diagnostic system by comparing its performance with that of different clinicians.

A. Quantitative Results Reveals Metric Superiority of the AFLL

The proposed AFLL model demonstrates substantial improvements in acne severity grading compared to current state-of-the-art methods (Table II). We conducted a comparative analysis with several leading approaches, including KIEGLFN [9], [39], EGPK [40], [10], DED [35], SLL [4], JGC [4], and OLDL [11]. The experimental results reveal that AFLL surpasses these methods across all metrics, especially in accuracy. Specifically, AFLL achieves a precision of 87.38%, specificity of 95.69%, sensitivity of 86.41%, Yorden Index of 82.11%, and accuracy of 88.56%. Notably, our approach enhances accuracy by 2.91% compared to the runner-up method. Moreover, direct acne severity classification using a single-label learning paradigm yielded inferior results across various metrics, with a precision of 75.81%, Yorden Index of 67.21%, and accuracy of 78.42%. By adopting a fuzzy label learning approach, JGC [4] and OLDL [11] significantly improved grading performance, with precision increasing by 11.29% and 13.86%, Yorden Index by 12.07% and 14.40%, and accuracy by 7.26% and 8.14%, respectively. These advancements can be attributed to the advantages of fuzzy label learning, which effectively captures the inherent ambiguity in acne severity labels. Furthermore, by addressing the challenge of distinguishing adjacent categories in existing fuzzy label learning techniques, AFLL improves grading performance by 1.23%, 6.79%, and 4.43% in precision, Yorden Index, and accuracy, respectively. These results further validate the effectiveness of the proposed innovations.

B. Ablation Studies Demonstrate Improvement of the Innovations

The ablation studies on acne severity grading demonstrate the significant improvements introduced by the proposed innovations. The baseline network, which utilizes fuzzy label learning with JS divergence loss and the PVTv2-b1 backbone (Table III, Row 1), achieves PR, SP, SE, YI, and ACC metrics of 79.38%, 92.30%, 76.86%, 69.17%, and 80.00%, respectively. When the vanilla binary search method is added (Table III, Row 3), the metrics improve by 6.48%, 2.20%, 8.57%, 10.31%, and 5.71% for PR, SP, SE, YI, and ACC, respectively. These improvements are primarily attributed to the binary search method, which enhances the ability of FLL methods to differentiate between adjacent categories. Introducing the Hellinger distance loss further boosts PR and ACC by 0.62% and 0.88% (Table III, Row 4), respectively, due to its more comprehensive measurement of distribution

Table III
ABLATION STUDIES ON ACNE SEVERITY GRADING TASK BY CROSS-VALIDATION DEMONSTRATE SIGNIFICANT IMPROVEMENTS OF THE PROPOSED INNOVATIONS. '*' DENOTES THE APPLICATION OF THE EMPIRICAL CUMULATIVE DISTRIBUTION. 'TREE' REPRESENTS THE UTILIZATION OF THE BINARY SEARCH METHOD; 'JS' MEANS JS DIVERGENCE LOSS;"H" DENOTES THE HELLINGER DISTANCE LOSS.

STC	SMD	SCDL		Tree	Metrics(%)				
		JS	H		PR↑	SP↑	SE↑	YI↑	ACC↑
		✓			79.38	92.30	76.86	69.17	80.00
				✓	83.65	93.86	82.62	75.52	83.27
		✓		✓	84.52	94.33	83.45	76.34	84.57
		✓	✓	✓	85.04	94.35	83.66	76.81	85.31
		✓*	✓*	✓	85.65	94.21	83.77	77.52	85.78
	✓	✓*	✓*	✓	86.06	95.14	84.69	79.27	87.12
✓	✓	✓*	✓*	✓	**87.38**	**95.69**	**86.41**	**82.11**	**88.56**

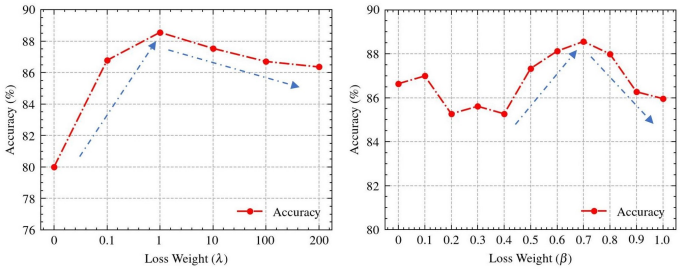


Fig. 7. Hyperparameter analysis of acne severity grading task. We analyzed two important hyperparameters in the framework, including the loss weight (λ) in the overall loss function and the loss weight (β) in SCDL.

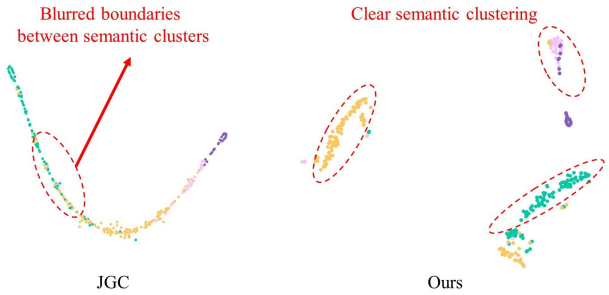


Fig. 8. Visualization of features representations of the JGC and our model using the t-SNE algorithm on the ACNE04 test set.

differences. The proposed empirical cumulative distribution effectively models the inherent temporal order among acne severity levels, resulting in significant improvements across all metrics, achieving 85.65%, 94.21%, 83.77%, 77.52%, and 85.78% for PR, SP, SE, YI, and ACC, respectively (Table III, Row 5). Moreover, the Selective Masking Decision (SMD) method effectively mitigates the error propagation problem in sequence prediction (Table III, Row 6), leading to further enhancements of 0.48%, 0.99%, 1.1%, 2.26%, and 1.56% in PR, SP, SE, YI, and ACC, respectively. Additionally, the proposed STC method improves sequence tree construction by reducing bias and enhancing discriminative ability (Table III, Row 7), resulting in increases of 1.53%, 0.58%, 2.04%, 3.58%, and 1.65% in PR, SP, SE, YI, and ACC, respectively. As a result, the final proposed AFLL achieves superior performance with PR, SP, SE, YI, and ACC metrics of 87.38%, 95.69%, 86.42%, 82.11%, and 88.56%, respectively, setting a new state-of-the-art in acne severity grading.

C. Analysis of different hyper-parameter settings

The analysis of different hyperparameters highlights the core innovations of the proposed approach (Fig. 7). We first examined the impact of the loss weight parameter, λ, on model performance. The parameter λ balances the tasks of predicting acne severity distribution and distinguishing between adjacent acne severity categories. We evaluated the model's accuracy across various λ settings $\{0, 0.1, 1, 10, 100, 200\}$. The results demonstrate that as λ increases, the model's performance initially improves but then gradually declines, with the highest accuracy achieved at $\lambda = 1$. This is because, at higher λ values, the model tends to overemphasize distinguishing between adjacent categories while neglecting the overall severity distribution, leading to decreased performance. Next, we explored different β settings within the Sequential Cumulative Distribution Loss (SCDL). Specifically, we uniformly sampled eleven β values within the range $[0, 1]$. The experimental results reveal that when β is small, the model's performance is suboptimal. Performance peaks at $\beta = 0.7$ and then gradually declines as β continues to increase. This pattern arises because, at lower β values, the model does not sufficiently focus on differences in high fuzzy memberships, leading to inadequate learning of primary modes, which negatively impacts overall performance. When $\beta = 0.7$, the model effectively balances accuracy in high fuzzy memberships without neglecting differences in low fuzzy memberships, resulting in superior performance. However, as β increases further, the model begins to overfit the high fuzzy memberships, neglecting differences in low fuzzy memberships, which diminishes generalization performance.

D. Validation of the Model's Effectiveness in Distinguishing Adjacent Categories

In this section, we validate the effectiveness of our model in distinguishing adjacent categories, which plays a significant

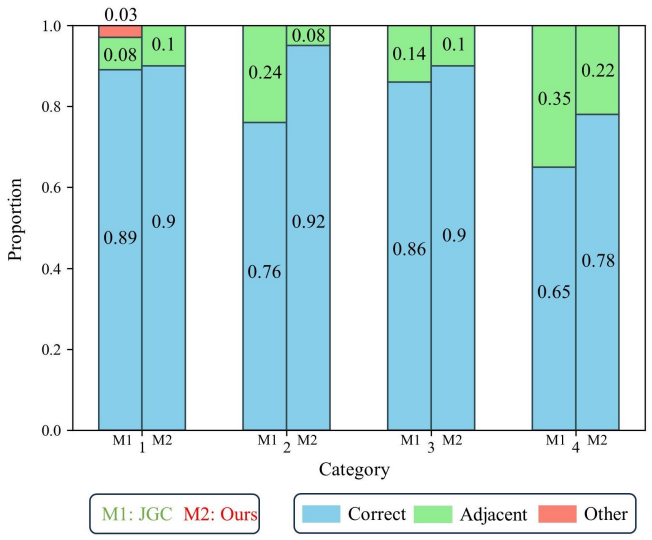


Fig. 9. Performances of the JGC and our model on each category within the ACNE04 dataset, demonstrating the proportion of samples predicted as the true class, adjacent classes, and other classes.

role in the overall performance improvement. This validation is approached from two perspectives: feature representation and classification results across different categories. First, we visualized the model features in a low-dimensional space by extracting the final layer features (4-dimensional) from both the JGC model [4] and our model on the test set. To obtain a 2-dimensional embedding of these features, we employed the popular dimensionality reduction algorithm t-SNE [41]. The resulting low-dimensional embeddings for JGC and our model on the ACNE04 test set are presented in Fig. 8, with points colored according to their corresponding acne severity levels. It can be observed that our model exhibits clear semantic clustering, whereas JGC shows blurred boundaries between different semantic clusters, thus validating the effectiveness of our model in distinguishing neighboring categories. Next, we visualized the performance of both JGC and our model on each category, highlighting the proportion of samples predicted as the true class, adjacent classes, and other classes. As shown in Fig. 9, our proposed AFLL achieves a higher overall correct prediction rate and a lower rate of incorrect predictions for adjacent categories. This indicates that AFLL enhances the model's ability to distinguish neighboring categories, thereby improving acne severity grading performance.

E. Discussion of the effectiveness of sequence tree construction algorithm

The performance of the models under all different sequence trees is tested (Table IV), thus demonstrating the effectiveness of our proposed sequence tree construction algorithm. As depicted in Fig. 10, all types of constructing sequence trees are presented. In the sequence tree constructed by our proposed algorithm (Table IV, first row), the precision, specificity, sensitivity, Yorden Index, and accuracy reach 87.38%, 95.69%, 86.41%, 82.11%, and 88.56%, respectively. The model's performance significantly exceeds that of capability under other types of sequence trees, thus demonstrating the effectiveness

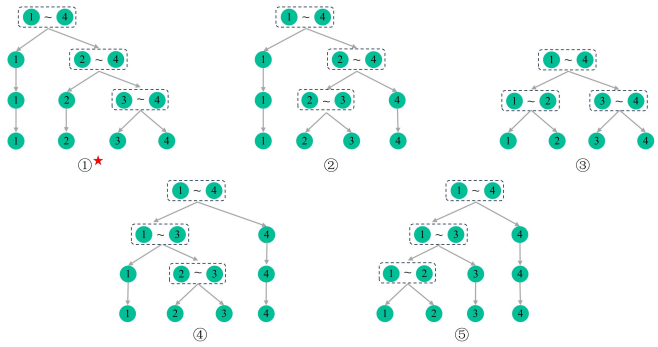


Fig. 10. Shows all the types of constructing sequence trees, where those with pentagrams are sequence trees constructed using our proposed sequence tree construction algorithm.

Table IV
SHOWS THE PERFORMANCE OF THE MODEL IN ALL DIFFERENT TREE-BUILDING PATTERNS, WHERE THOSE WITH RED PENTAGRAMS REPRESENT SEQUENCE TREES BUILT USING OUR PROPOSED ALGORITHM FOR BUILDING SEQUENCE TREES. THE BOLD TEXT INDICATES THE BEST RESULT FOR THE METRIC.

Sequence trees	Metrics				
	PR↑	SP↑	SE↑	YI↑	ACC↑
①★	**87.38**	**95.69**	**86.41**	**82.11**	**88.56**
②	70.71	88.06	65.40	53.44	70.55
③	83.90	93.66	80.91	74.56	83.77
④	81.07	92.97	76.16	69.12	80.96
⑤	85.47	94.24	83.60	77.84	85.21

of our proposed sequence tree construction algorithm. In other types to constructing sequence trees, either the imbalance between positive and negative samples arises from neglecting the ratio of left and right subtree samples during construction, or the semantic similarity between dividing nodes and adjacent nodes is disregarded. This leads to the model having difficulty in progressively classifying in binary-based sequence prediction, thereby impeding the model's performance in acne severity grading based on such sequence trees.

F. Comparison with different clinicians

To evaluate the practical utility of the proposed diagnostic system, we compared the performance of AFLL with that of two general clinicians and two dermatologists. As shown in Table V, the diagnostic outcomes for the doctors were reported in [4]. Given their extensive professional expertise, dermatologists consistently outperformed general clinicians across all metrics, highlighting the influence of specialized knowledge on diagnostic accuracy. The variations observed between the two dermatologists may be attributed to individual subjectivity. Notably, our experimental results demonstrate that the proposed method surpasses both dermatologists in all evaluated metrics. This indicates that our diagnostic model not only matches but potentially exceeds the diagnostic capabilities of professional dermatologists, offering significant promise as a valuable tool for clinicians and patients in real-world clinical settings.

Table V
COMPARISON OF PERFORMANCE WITH DIFFERENT CLINICIANS ON THE ACNE04 DATASET. GD: GENERAL DOCTOR, DERM: DERMATOLOGIST.

Methods	Metrics(%)				
	PR↑	SP↑	SE↑	YI↑	ACC↑
GD1	62.87	84.11	55.27	39.38	58.43
GD2	62.07	86.98	68.33	55.31	63.14
Derm1	77.33	90.60	72.56	63.22	75.29
Derm2	82.95	92.16	78.27	70.43	79.43
Ours	**87.38**	**95.69**	**86.41**	**82.11**	**88.56**

VI. CONCLUSION

The proposed AFLL framework provides a powerful benchmark for fuzzy label learning-based acne severity grading. It not only effectively leverages label uncertainty but also enhances the model's capability to differentiate between adjacent categories. The proposed ADSG effectively incorporates semantic similarity and class distributions into the sequence tree construction, resulting in a better sequence tree with reduced bias and improved discriminative power. Additionally, our CDSP effectively mitigates error propagation in hierarchical decision-making through the selective masking decision strategy, thereby increasing the overall robustness and accuracy of the decision-making process. Our SCDL further advances fuzzy label learning by capturing the differences for both high and low fuzzy memberships across the whole fuzzy label set while modeling internal temporal order with a cumulative distribution, leading to superior fuzzy label learning performance. Extensive experiments on the ACNE04 dataset demonstrate that our proposed AFLL model significantly outperforms existing state-of-the-art methods.

REFERENCES

[1] D. P. Krowchuk, "Managing acne in adolescents," *Pediatric Clinics of North America*, vol. 47, no. 4, pp. 841–857, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031395505702431

[2] B. Oulès, C. Philippeos, J. Segal, M. Tihy, M. Vietri Rudan, A.-M. Cujba, P. A. Grange, S. Quist, K. Natsuga, L. Deschamps *et al.*, "Contribution of gata6 to homeostasis of the human upper pilosebaceous unit and acne pathogenesis," *Nature communications*, vol. 11, no. 1, p. 5067, 2020.

[3] K. Bhate and H. Williams, "Epidemiology of acne vulgaris," *British Journal of Dermatology*, vol. 168, no. 3, pp. 474–485, 03 2013. [Online]. Available: https://doi.org/10.1111/bjd.12149

[4] X. Wu, N. Wen, J. Liang, Y.-K. Lai, D. She, M.-M. Cheng, and J. Yang, "Joint acne image grading and counting via label distribution learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 642–10 651.

[5] X. Shen, J. Zhang, C. Yan, and H. Zhou, "An automatic diagnosis method of facial acne vulgaris based on convolutional neural network," *Scientific reports*, vol. 8, no. 1, p. 5839, 2018.

[6] Y. Liu, B. Chen, S. Wang, G. Lu, and Z. Zhang, "Deep fuzzy multi-teacher distillation network for medical visual question answering," *IEEE Transactions on Fuzzy Systems*, 2024.

[7] D. Das and D. R. Nayak, "Fja-net: A fuzzy joint attention guided network for classification of glaucoma stages," *IEEE Transactions on Fuzzy Systems*, 2024.

[8] S. Zhang, M. Yin, F. Xiao, Z. Cao, and D. Pelusi, "A complex gaussian fuzzy numbers-based multisource information fusion for pattern classification," *IEEE Transactions on Fuzzy Systems*, 2024.

[9] Y. Lin, J. Jiang, Z. Ma, D. Chen, Y. Guan, H. You, X. Cheng, B. Liu, and G. Luo, "Kieglfn: A unified acne grading framework on face images," *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106911, 2022.

[10] Z. Zhang, Z. Liu, J. Jiang, C. Kong, Y. Guan, X. Liu, H. You, J. Yang, and Y. Lin, "Interpretable diagnosis of face acne via complementation learning of evidence localization and severity level grading," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2023, pp. 3414–3421.

[11] C. Wen, X. Zhang, X. Yao, and J. Yang, "Ordinal label distribution learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 481–23 491.

[12] L. F. Williams Jr, "A modification to the half-interval search (binary search) method," in *Proceedings of the 14th annual Southeast regional conference*, 1976, pp. 95–101.

[13] J. Wang, Y. Cheng, J. Chen, T. Chen, D. Chen, and J. Wu, "Ord2seq: Regarding ordinal regression as label sequence prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5865–5875.

[14] M.-S. Yang and J. B. Benjamin, "Sparse possibilistic c-means clustering with lasso," *Pattern Recognition*, vol. 138, p. 109348, 2023.

[15] T. Chantharaphaichi, B. Uyyanonvara, C. Sinthanayothin, and A. Nishihara, "Automatic acne detection for medical treatment," in *2015 6th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*. IEEE, 2015, pp. 1–6.

[16] G. Maroni, M. Ermidoro, F. Previdi, and G. Bigini, "Automated detection, extraction and counting of acne lesions for automatic evaluation and tracking of acne severity," in *2017 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2017, pp. 1–6.

[17] T. Rao, X. Li, and M. Xu, "Learning multi-level deep representations for image emotion classification," *Neural processing letters*, vol. 51, pp. 2043–2061, 2020.

[18] Y. Lin, Y. Guan, Z. Ma, H. You, X. Cheng, and J. Jiang, "An acne grading framework on face images via skin attention and sfnet," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 2407–2414.

[19] A. Campagner, "Learnability in "learning from fuzzy labels"," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2021, pp. 1–6.

[20] ——, "Learning from fuzzy labels: Theoretical issues and algorithmic solutions," *International Journal of Approximate Reasoning*, vol. 171, p. 108969, 2024.

[21] X. Yang, B.-B. Gao, C. Xing, Z.-W. Huo, X.-S. Wei, Y. Zhou, J. Wu, and X. Geng, "Deep label distribution learning for apparent age estimation," in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 102–108.

[22] C. Chen, Z. Chen, X. Jin, L. Li, W. Speier, and C. W. Arnold, "Attention-guided discriminative region localization and label distribution learning for bone age assessment," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1208–1218, 2021.

[23] Y.-Y. Fan, S. Liu, B. Li, Z. Guo, A. Samal, J. Wan, and S. Z. Li, "Label distribution-based facial attractiveness computation by deep residual learning," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2196–2208, 2017.

[24] X. Li, X. Liang, G. Luo, W. Wang, K. Wang, and S. Li, "Ultra: Uncertainty-aware label distribution learning for breast tumor cellularity assessment," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 303–312.

[25] W. Li, W. Zhang, Q. Zhang, X. Zhang, and X. Wang, "Weakly-supervised causal discovery based on fuzzy knowledge and complex data complementarity," *IEEE Transactions on Fuzzy Systems*, pp. 1–13, 2024.

[26] X. Li, X. Liang, G. Luo, W. Wang, K. Wang, and S. Li, "Ambiguity-aware breast tumor cellularity estimation via self-ensemble label distribution learning," *Medical Image Analysis*, vol. 90, p. 102944, 2023.

[27] X. Li, G. Luo, W. Wang, K. Wang, and S. Li, "Curriculum label distribution learning for imbalanced medical image segmentation," *Medical Image Analysis*, vol. 89, p. 102911, 2023.

[28] X. Zhao, L. Qi, Y. An, and X. Geng, "Generalizable label distribution learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8932–8941.

[29] C. Beecks, M. S. Uysal, and T. Seidl, "Earth mover's distance vs. quadratic form distance: an analytical and empirical comparison," in *2015 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2015, pp. 233–236.

[30] V. Monev, "Introduction to similarity searching in chemistry," *MATCH Commun. Math. Comput. Chem*, vol. 51, pp. 7–38, 2004.

[31] R. O. Duda, P. E. Hart *et al.*, *Pattern classification*. John Wiley & Sons, 2006.
[32] X. Ju, D. Zhang, J. Li, and G. Zhou, "Transformer-based label set generation for multi-modal multi-label emotion detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 512–520.
[33] V. González-Castro, R. Alaiz-Rodríguez, and E. Alegre, "Class distribution estimation based on the hellinger distance," *Information Sciences*, vol. 218, pp. 146–164, 2013.
[34] N. Hayashi, H. Akamatsu, M. Kawashima, and A. S. Group, "Establishment of grading criteria for acne severity," *The Journal of dermatology*, vol. 35, no. 5, pp. 255–260, 2008.
[35] Y. Lin, J. Jiang, D. Chen, Z. Ma, Y. Guan, X. Liu, H. You, and J. Yang, "Ded: Diagnostic evidence distillation for acne severity grading on face images," *Expert Systems with Applications*, vol. 228, p. 120312, 2023.
[36] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.
[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.
[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[39] S. Liu, Y. Fan, M. Duan, Y. Wang, G. Su, Y. Ren, L. Huang, and F. Zhou, "Acnegrader: An ensemble pruning of the deep learning base models to grade acne," *Skin Research and Technology*, vol. 28, no. 5, pp. 677–688, 2022.
[40] Y. Lin, J. Jiang, D. Chen, Z. Ma, Y. Guan, X. Liu, H. You, J. Yang, and X. Cheng, "Acne severity grading on face images via extraction and guidance of prior knowledge," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 1639–1643.
[41] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.