Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2026.072633

ARTICLE

A Cooperative Hybrid Learning Framework for Automated Dandruff Severity Grading

Sin-Ye Jhong[1], Hui-Che Hsu[1,2], Hsin-Hua Huang[2], Chih-Hsien Hsia[3,4,*], Yulius Harjoseputro[2,5] and Yung-Yao Chen[2,*]

[1]Graduate Institute of Intelligent Manufacturing Technology, National Taiwan University of Science and Technology, Taipei, 106335, Taiwan

[2]Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei, 106335, Taiwan

[3]Department of Computer Science and Information Engineering, National Ilan University, Yilan, 26047, Taiwan

[4]Office of Research and Industry-Academia Development, Chaoyang University of Technology, Taichung City, 413310, Taiwan

[5]Department Informatics, Universitas Atma Jaya Yogyakarta, Yogyakarta, 55281, Indonesia

*Corresponding Authors: Chih-Hsien Hsia. Email: hsiach@niu.edu.tw; Yung-Yao Chen. Email: yungyaochen@gapps.ntust.edu.tw

Received: 31 August 2025; Accepted: 04 January 2026; Published: 10 February 2026

ABSTRACT: Automated grading of dandruff severity is a clinically significant but challenging task due to the inherent ordinal nature of severity levels and the high prevalence of label noise from subjective expert annotations. Standard classification methods fail to address these dual challenges, limiting their real-world performance. In this paper, a novel, three-phase training framework is proposed that learns a robust ordinal classifier directly from noisy labels. The approach synergistically combines a rank-based ordinal regression backbone with a cooperative, semi-supervised learning strategy to dynamically partition the data into clean and noisy subsets. A hybrid training objective is then employed, applying a supervised ordinal loss to the clean set. The noisy set is simultaneously trained using a dual-objective that combines a semi-supervised ordinal loss with a parallel, label-agnostic contrastive loss. This design allows the model to learn from the entire noisy subset while using contrastive learning to mitigate the risk of error propagation from potentially corrupt supervision. Extensive experiments on a new, large-scale, multi-site clinical dataset validate our approach. The method achieves state-of-the-art performance with 80.71% accuracy and a 76.86% F1-score, significantly outperforming existing approaches, including a 2.26% improvement over the strongest baseline method. This work provides not only a robust solution for a practical medical imaging problem but also a generalizable framework for other tasks plagued by noisy ordinal labels.

KEYWORDS: Dandruff severity grading; ordinal regression; noisy label learning; self-supervised learning; contrastive learning; medical image analysis

1 Introduction

The application of deep learning and artificial intelligence is rapidly transforming medical image analysis, offering tools that promise to enhance the speed, objectivity, and accessibility of clinical diagnosis [1]. Within this field, the automated analysis of dermatological conditions via computer vision is a significant and advancing area of research. Among common afflictions, scalp health issues such as excessive dandruff affect a large global population. Beyond physical discomfort, these conditions can significantly impact an individual's psychological well-being and daily life, while also serving as potential indicators for inflammatory disorders

like seborrheic dermatitis [2]. Although deep learning has achieved remarkable success in medical imaging, its application to the fine-grained task of grading dandruff severity from microscope images presents a unique and challenging set of problems that have not been fully addressed by prior work [3–5].

The primary challenges are rooted in the fundamental nature of the data and its annotation. First, severity grading is an ordinal classification task. The labels, ranging from "no dandruff" to "severe," possess a natural monotonic order. Standard classification methods that treat categories as independent and nominal ignore this crucial relationship, which can lead to clinically implausible errors, such as penalizing a misprediction from "severe" to "moderate" equally as one to "none". Second, collecting large-scale, accurately labeled medical datasets is notoriously difficult. The visual assessment of dandruff is highly subjective, leading to significant inter- and intra-rater variability. This issue is exacerbated in scalp analysis, where symptoms often manifest as subtle, fragmented visual cues. This results in minimal inter-class variance between adjacent severity levels, making consistent annotation exceptionally difficult even for trained experts [4]. Consequently, any realistic training dataset is inevitably corrupted with substantial label noise, which severely degrades the generalization performance of deep neural networks that are prone to memorizing incorrect labels.

This presents a significant dilemma for existing methods. While various approaches have been proposed to tackle either ordinal regression [6] or learning with noisy labels (LNL) [7] independently, they are insufficient for this problem. State-of-the-art LNL frameworks are class-agnostic, meaning they ignore the crucial monotonic order of severity and thus risk producing ordinally inconsistent results. Conversely, most ordinal regression methods inherently assume that the training labels are accurate and can be severely compromised by the high degree of annotation noise. This leaves a critical, unaddressed research gap for a framework that can simultaneously handle both the ordinal constraints and the endemic label noise in a unified manner.

To bridge this specific gap, our primary objective is to develop a novel framework for dandruff severity grading that learns robustly from noisy, ordered data. Our approach synergizes a self-supervised strategy for noisy label handling with an ordinal regression objective. Inspired by semi-supervised learning techniques, our method dynamically partitions the training data into presumed "clean" and "noisy" sets based on an ordinally-aware loss metric. For the noisy samples, where labels are deemed unreliable, a dual-objective is employed. This objective trains the noisy subset using a semi-supervised ordinal loss derived from pseudo-labels in parallel with a label-agnostic contrastive loss. This design allows the model to leverage the full dataset while the contrastive component acts as a robust regularizer, mitigating the risk of error propagation from potentially incorrect pseudo-labels. This entire process is built upon an ordinal regression backbone, which recasts the multi-class problem into a series of simpler, rank-consistent binary tasks. This ensures the model's predictions respect the inherent order of dandruff severity, leading to more reliable and clinically meaningful results.

In summary, our main contributions are as follows:

- We are the first to propose a deep learning framework that jointly addresses the coupled challenges of ordinal classification and label noise for automated dandruff severity grading.
- We introduce an ordinally-aware sample partitioning strategy that leverages the rank-based ordinal loss as its metric. This enables the partitioning process to be sensitive to the magnitude of ordinal errors, providing a more reliable separation of clean and noisy data than standard class-agnostic approaches.
- A hybrid training objective is proposed, which applies a supervised ordinal loss to the clean set while simultaneously training the noisy set with both a semi-supervised ordinal loss and a label-agnostic contrastive loss. This dual-objective design mitigates the risk of error propagation, as the contrastive

loss provides a robust, label-agnostic learning signal to counteract potential errors introduced by the pseudo-labels.
- An ordinal regression objective is incorporated throughout the learning process, enforcing the monotonic relationship between severity levels and significantly reducing the frequency of large, clinically implausible prediction errors.
- This study is validated through extensive experiments on a new, large-scale, multi-site dandruff severity grading dataset, demonstrating that the proposed method achieves state-of-the-art performance and significantly outperforms baseline approaches.

2 Related Works

Our research addresses the automated grading of dandruff severity, a task situated at the intersection of medical image analysis, ordinal classification, and robust learning under label noise. In this section, we review the literature from these three perspectives to contextualize our contribution.

2.1 Automated Dandruff Severity Grading

The application of deep learning to scalp analysis has gained significant traction. Early works focused on general scalp problem classification using CNN-based architectures [8,9] or robust representation learning in edge-cloud systems [10]. More recent work has shifted focus towards the more nuanced task of grading the severity of these conditions. Some approaches have repurposed object detection models, inferring severity from the density of detected problem areas [11,12]. However, this indirect method can be biased by non-scalp features or fail to capture fine-grained, fragmented symptoms. Consequently, a more direct approach based on established medical grading standards [13] has become prevalent. Jhong et al. enhanced a CNN with attention mechanisms [2], while Jin et al. [3] fine-tuned an EfficientNet for fine dandruff classification. Other works have explored ensemble models for robustness on limited data [5]. Or employed Vision Transformers (ViT) for grading multiple scalp conditions simultaneously [4]. While these studies demonstrate the potential of deep learning, they predominantly treat severity grading as a standard classification task. This overlooks two critical challenges: (1) the labels are ordinal in nature, and standard classification losses fail to capture this intrinsic order; (2) the labels are inherently noisy due to subjective expert assessment, which can degrade model performance. Our work explicitly addresses these two challenges in tandem.

2.2 Ordinal Classification

Ordinal classification (or regression) aims to solve tasks where labels possess a natural ranking. A prominent and modular strategy involves decomposing a K-class ordinal problem into K-1 independent binary classification sub-problems [14]. This rank-based approach is particularly adept at addressing the challenge of ambiguous instances, which are samples that fall near the blurred boundaries between adjacent categories and are a common issue in subjective grading. Foundational deep learning models like MO-CNN [14] first operationalized this concept by training multiple binary classifiers to distinguish adjacent ranks. More recent methods, such as CORAL [15], have refined this approach to enforce rank consistency with greater efficiency. For contrast, other families of methods, such as Distribution Ordering Learning, seek to model the ordinal relationship by learning the label distribution directly. Techniques in this category include using soft labels [16]. While effective in certain contexts, these distribution-based methods often rely heavily on the precise location of the ground-truth label to anchor the distribution. This can make them potentially more sensitive to the high degree of label noise present in our dataset, where the ground-truth label itself is often unreliable. Therefore, for our framework, we adopt the rank-based decomposition strategy.

Its proven effectiveness, modularity, and high compatibility with other learning paradigms make it an ideal backbone for integrating the noisy label learning methods we leverage.

2.3 Learning with Noisy Labels

To combat the memorization effect of deep networks on incorrect labels, various Learning with Noisy Labels (LNL) strategies have been proposed. These approaches can be broadly categorized into methods like Loss Adjustment and Sample Selection. Loss adjustment methods, such as Bootstrapping [17], attempt to correct the training signal for all samples, for example by using the model's own predictions as a soft target. However, these correction-based approaches risk error propagation, especially in high-noise environments; if the model's initial predictions are wrong, the corrected label will also be wrong, potentially reinforcing the error. In contrast, sample selection methods are particularly effective and relevant to our work. They operate on the principle that models learn from clean examples with small losses before fitting to noisy examples with large losses. Early methods in this domain, such as Co-teaching [18], utilized multiple networks to cross-filter clean samples for each other, mitigating confirmation bias. A leading approach, DivideMix [19], operationalizes this by modeling the per-sample loss distribution with a Gaussian Mixture Model (GMM) to dynamically separate the dataset into a likely clean set and a noisy set. Instead of discarding the noisy data, DivideMix reframes the task as a semi-supervised problem, using model predictions as pseudo-labels for the noisy subset. To further enhance feature learning on this subset without relying on potentially flawed pseudo-labels, contrastive learning has proven powerful for learning robust representations from the underlying images alone [20].

While these fields have advanced independently, a framework that synergistically integrates them is non-trivial and remains a critical gap. Ordinal regression methods typically assume that training labels are accurate [15,16]. Conversely, state-of-the-art LNL frameworks are class-agnostic [19] and do not enforce ordinal constraints. For instance, a standard class-agnostic loss metric, as used in many LNL methods, cannot differentiate minor from major ordinal errors, which is a vital distinction for robustly partitioning the data. Furthermore, reliance on pseudo-labeling for noisy data, a common LNL strategy, risks propagating ordinally-inconsistent errors. Our work bridges this gap by proposing the first unified framework built upon a rank-based ordinal backbone. This framework introduces two key innovations: it uses an ordinally-aware metric for robust sample partitioning and employs a dual-objective with contrastive learning to mitigate pseudo-label error propagation.

3 Methodology

3.1 Overall Framework

The task of grading dandruff severity presents dual challenges: the ordinal nature of severity levels and the prevalence of label noise. To address these, we propose a robust three-phase training framework that synergizes ordinal regression with a cooperative noisy label learning strategy. Our pipeline, illustrated in Fig. 1, progressively refines the model's capability to handle noisy ordinal data. The process begins with a crucial supervised pre-training phase. Two parallel networks are independently warmed-up on the entire noisy dataset using an ordinal regression loss. This crucial first step ensures the models develop an initial understanding of the ranking relationships between severity grades, which is a prerequisite for generating meaningful loss distributions for sample partitioning. Leveraging these pre-trained models, the framework then proceeds to a cooperative sample partitioning phase. At the start of each subsequent epoch, the two networks use each other's per-sample loss distributions to dynamically divide the dataset into a high-confidence clean subset and a low-confidence noisy subset. This cross-supervision or co-training design is critical for mitigating the confirmation bias inherent to self-training systems. With the data partitioned,

the framework enters its main hybrid training phase, which employs a dual-objective learning strategy tailored to these subsets. A semi-supervised ordinal loss is applied to data generated via MixUp from both subsets, allowing the model to learn from the entire dataset. Concurrently, a label-agnostic contrastive loss is applied exclusively to the noisy subset to learn robust, generalizable features without relying on their corrupt labels. The subsequent sections provide a detailed exposition of each component. The complete procedure is formalized in Algorithm 1.

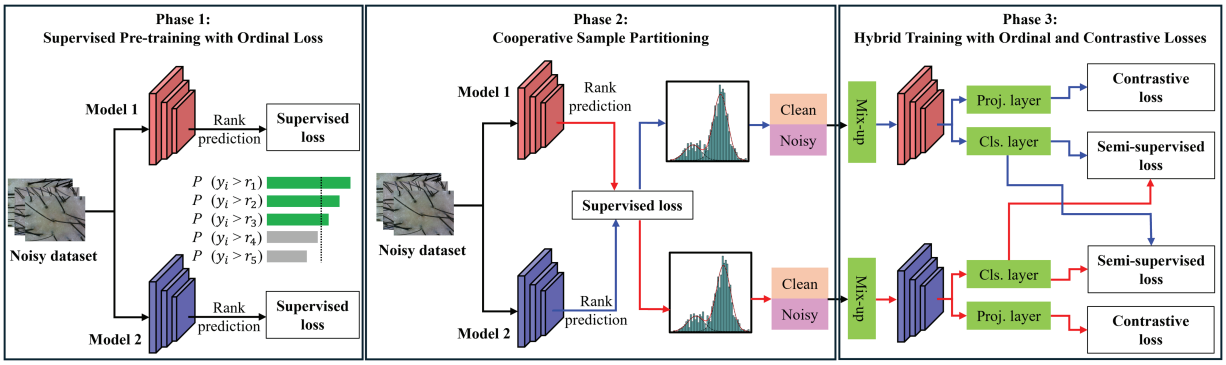

Figure 1: Overview of our proposed three-phase training framework for robust ordinal classification from noisy labels. (Phase 1) Supervised Pre-training: Two parallel networks are warmed-up on the full noisy dataset with an ordinal loss to establish a baseline understanding of the severity ranking. (Phase 2) Cooperative Sample Partitioning: Each network's loss distribution is used to partition the data into clean and noisy subsets for its peer, a cross-supervision strategy that mitigates confirmation bias. (Phase 3) Hybrid Training: A dual-objective strategy is employed. A semi-supervised ordinal loss is computed on Mixup-augmented data from both subsets, while a contrastive loss learns label-agnostic features exclusively from the noisy subset. The total loss combines these objectives for a comprehensive training update

Algorithm 1: Cooperative hybrid learning framework

Input: Labeled training set $D = \{(x_i, y_i)\}_{i=1}^{N}$, Two deep networks $f(\cdot;\theta_1) and f(\cdot;\theta_2)$, Warm-up epochs E_{warm}, Total epochs $E_{\max}$, MixUp Beta parameter α, Partition threshold τ, Loss weight λ_u, λ_r.
Output: Trained network parameters θ_1, θ_2
for $e = 1$ **to** E_{warm} **do**
    Train θ_1 and θ_2 independently using Supervised Ordinal Loss (Eq. (2)) on all samples in D.
end for
for $e = \mathrm{E}_{\text{warm}} + 1$ **to** $\mathrm{E}_{\max}$ **do**
    for $k = 1, 2$ **do**
        Calculate l_i for all samples in D using network θ_k.
        Fit GMM to the loss distribution.
        Calculate posterior probability $w_i = P(clean|l_i)$ for each sample.
        Partition D into Clean set X_c (where $w_i \geq \tau$) and Noisy set X_n (where $w_i < \tau$).
        Refine labels for X_c using Eq. (3): $\overline{y}_i^{'} = w_i \cdot y_i^{'} + (1 - w_i) \cdot p_i^{'}$.
    end for
    for $k = 1, 2$ **do**
        Let m be the peer network (if $k = 1$, them $m = 2$; if $k = 2$, them $m = 1$).
        Use the partition sets $x_c^{(m)}$ and $x_n^{(m)}$ generated by the peer network.
        for each mini-batch B sampled from $x_c^{(m)}$ and $x_n^{(m)}$ **do**

(Continued)

Algorithm 1 (continued)

Generate ensemble pseudo-labels $\overline{y}_n$ for noisy samples by averaging predictions from θ_1 and θ_2.
Construct training batch combining refined clean samples and pseudo-labeled noisy samples.
Apply MixUp augmentation to the batch inputs and targets using α.
Calculate Supervised Ordinal Loss L_s (Eq. (2)) on the mixed batch.
Calculate Contrastive Loss L_u (Eq. (5)) exclusively on noisy samples $x_n^{(m)}$.
Calculate Regularization Loss L_r (Eq. (4)) for distribution alignment.
Compute total loss: $L_{total} = L_s + \lambda_u L_u + \lambda_r L_r$.
Update network parameters θ_k via SGD to minimize L_{total}.
end for
end for
end for

3.2 Ordinal Regression for Severity Classification

Standard multi-class classification, which typically employs a softmax output with categorical cross-entropy loss, is suboptimal for severity grading as it treats labels as independent nominal entities. This approach ignores the critical ordinal relationship between grades (e.g., grade 4 is more severe than grade 2). To formally integrate this structure, we adopt a rank-based methodology inspired by [15]. We reframe the K-class ordinal problem into $K-1$ simpler binary classification tasks. Specifically, a ground-truth label $y \in \{0,1,\ldots,K-1\}$ is transformed into a binary vector $y' \in \{0,1\}^{K-1}$, where the k-th element indicates whether the severity grade is greater than rank k:

$$y'^{(k)} = \mathrm{I}(y \geq k) \ \ for\, k \in \{1,\ldots,K-1\} \tag{1}$$

To guarantee rank consistency in predictions (e.g., $P(y \geq 4) > P(y \geq 2)$), we constrain the network such that the final $K-1$ output neurons share the weights of the penultimate feature layer but each maintains an independent bias term. The model is then trained by minimizing the sum of binary cross-entropy (BCE) losses across all tasks. For a batch of B samples, this supervised loss L_s, is defined as:

$$L_s = \frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K-1}\mathrm{BCE}\left(\sigma(g(x_i,\theta)+b_k, y_i'^{(k)}\right) \tag{2}$$

where $g(x_i,\theta)$ denotes the feature vector from the penultimate layer, and b_k is the bias for the k-th task. Minimizing this objective ensures that the learned biases are monotonically non-increasing ($b_1 \geq b_2 \geq \cdots \geq b_{K-1}$), thereby guaranteeing rank-consistent probability estimates. This ordinal loss serves as the core objective for the pre-training (Phase 1) and the supervised component of the hybrid training (Phase 3).

3.3 Self-Supervised Learning via Cooperative Partitioning

A significant challenge in dandruff severity grading is the subtle visual distinction between adjacent levels, leading to inevitable label noise that can degrade model generalization. To address this, we employ a robust training strategy inspired by semi-supervised learning [19]. Instead of directly correcting labels, our approach dynamically partitions the data into a trusted clean set and an untrusted noisy set.

We posit that for a well-trained network, correctly labeled samples tend to exhibit lower losses than mislabeled ones. To formalize this, at the beginning of each epoch, we compute the per-sample ordinal loss for all training data. We then fit a two-component Gaussian Mixture Model (GMM) to the loss distribution to statistically separate clean and noisy samples. This process yields a posterior probability $w_i = P(clean|l_i)$ for each sample x_i. To prevent error accumulation from a single model (self-confirmation bias), we implement a cooperative "co-training" framework. Specifically, the partition derived from Model 1's loss distribution is used to train Model 2, and *vice-versa*. Samples with w_i exceeding a threshold τ form the clean set X_c while the remainder form the noisy set X_n for which the original labels are discarded.

For samples in X_c, we further mitigate potential noise by applying a label refinement mechanism, as shown in Fig. 2. We compute a "soft" target $\overline{y}'_i$ by linearly combining the original ground-truth label y'_i with the model's own prediction p'_i (averaged over augmentations):

$$\overline{y}'_i = w_i \cdot y'_i + (1 - w_i) \cdot p'_i \tag{3}$$

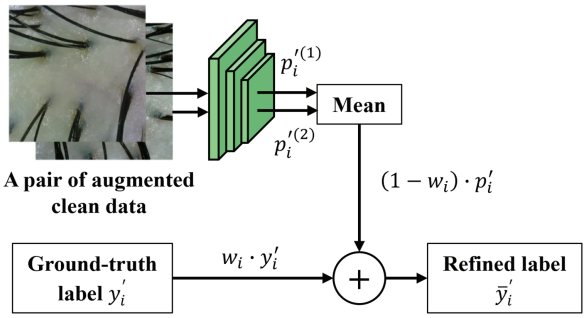

Figure 2: The label refinement mechanism for the clean set. A stabilized model prediction p'_i is obtained by averaging the outputs from two distinct augmentations of an input image. This prediction is then combined with the ground-truth label $\overline{y}'_i$ in a weighted sum, where the weight w_i is the sample's estimated probability of being clean. $\overline{y}'_i$ adaptively balances the influence of the original annotation and the model's own prediction

This mechanism adaptively balances supervision: it trusts the provided label for high-confidence samples (where $w_i \approx 1$) while relying more on the model's prediction for ambiguous cases.

3.4 Hybrid Training Strategy

With the data partitioned into a clean set X_c (using refined labels $\overline{y}'_i$) and an unlabeled noisy set X_n, we employ a hybrid training strategy to maximize feature learning. This strategy combines a semi-supervised ordinal objective with a label-agnostic contrastive objective. First, we generate targets for the noisy subset X_n. As illustrated in Fig. 3, we compute a robust ensemble prediction by averaging the outputs of both co-trained networks, serving as the pseudo-label for samples in X_n. We then apply MixUp augmentation [21] to construct interpolated training batches combining refined samples from X_c and pseudo-labeled samples from X_n. The supervised ordinal loss L_s is minimized on these mixed batches.

To prevent the model from collapsing to a trivial solution (e.g., predicting a single class for all inputs), we incorporate a regularization term L_r. This forces the moving average of the model's predictions $\overline{p_{model}}$ to align with the prior class distribution of the dataset p_{gt}:

$$L_r = KL(p_{gt} \,\|\, \overline{p_{model}}) \tag{4}$$

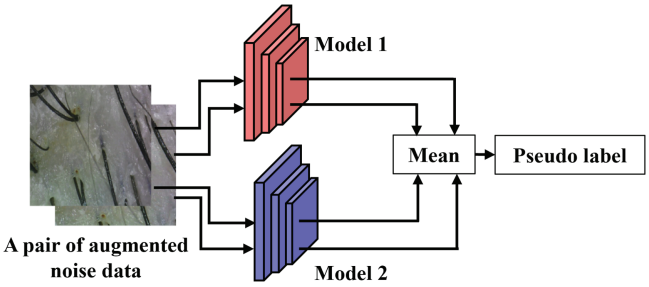

Figure 3: Ensemble pseudo-label generation for the noisy set. The process for generating pseudo-labels for samples in the noisy set X_n. We leverage the predictions from both co-trained networks (Model 1 and Model 2) to form a robust ensemble estimate. Specifically, a pair of augmented views of a noisy image is fed through both networks. The resulting output predictions are averaged to produce a single, high-confidence soft pseudo-label, which is then used as the target in the semi-supervised learning phase

Concurrently, to extract discriminative features from the noisy data without relying on potentially erroneous pseudo-labels, we apply a contrastive loss L_u exclusively toX_n. We utilize the InfoNCE objective to maximize the similarity between two augmented views (z_i, z_i^+) of the same image while minimizing similarity with other samples:

$$L_u = -\log \frac{e^{\frac{\text{sim}\left(z_i, z_i^+\right)}{K}}}{\sum_{j=1}^{N_n} I_{j\neq i} e^{\frac{\text{sim}\left(z_i, z_b\right)}{K}}} \tag{5}$$

where $sim(\cdot,\cdot)$ denotes cosine similarity, K is the temperature hyperparameter, and N_n is the batch size of the noisy subset. The total training objective is a weighted summation:

$$L_{total} = L_s + \lambda_u \cdot L_u + \lambda_{\mathrm{r}} \cdot L_{\mathrm{r}} \tag{6}$$

where λ_u and λ_{r} are balancing hyperparameters.

4 Experiments

In this section, we present a comprehensive empirical evaluation of our proposed framework. We begin by detailing our newly collected dataset and the evaluation protocol. We then provide a quantitative comparison against state-of-the-art (SOTA) methods, followed by a thorough ablation study to dissect the contribution of each core component of our methodology.

4.1 Dataset and Evaluation Metrics

To facilitate this research, we constructed a new, large-scale dataset for dandruff severity grading. The data was collected in collaboration with Taipei Hospital, Ministry of Health and Welfare and MacroHI Co., Ltd. in Taiwan, ensuring a diversity of clinical settings. This study was conducted in strict accordance with the Declaration of Helsinki and was approved by the Institutional Review Board (IRB) of Taipei Hospital, Ministry of Health and Welfare (Protocol No. TH-IRB-0022-0021). Using a digital microscope, clinicians and therapists captured images from eight distinct scalp regions, resulting in varied resolutions (768 × 576, 800 × 600, and 1024 × 768). A team of one dermatologist and three trained scalp therapists annotated a total of 22,537 images. To ensure a consistent and high-quality standard across this large-scale dataset, a rigorous adjudication protocol was employed. The trained therapists performed the initial annotations based on the clinically validated Adherent Scalp Flaking Score (ASFS) standard [13]. The dermatologist then served as the

final expert adjudicator, reviewing annotations and defining the ground-truth label for each image used in the study. Each image was assigned to one of six severity levels $\{0, 2, 4, 6, 8, 10\}$ based on the ASFS standard. The dataset exhibits a natural class imbalance representative of a clinical population: grade 0 (5817 images), grade 2 (9062), grade 4 (2966), grade 6 (2085), grade 8 (1893), and grade 10 (714). We randomly partitioned the data, allocating 18,871 images (~85%) for training and 3666 for testing, ensuring a consistent class distribution across splits. Fig. 4 shows example images for different severity grades. The creation of this multi-site, expert-annotated dataset is a key contribution to our work, providing a challenging and realistic benchmark for this task. Following standard practice, we report overall Accuracy. To provide a more robust assessment of our imbalanced dataset, we also report macro-averaged Precision, Recall, and F1-Score. The F1-Score is particularly important as it provides a balanced measure of performance across all six severity classes.

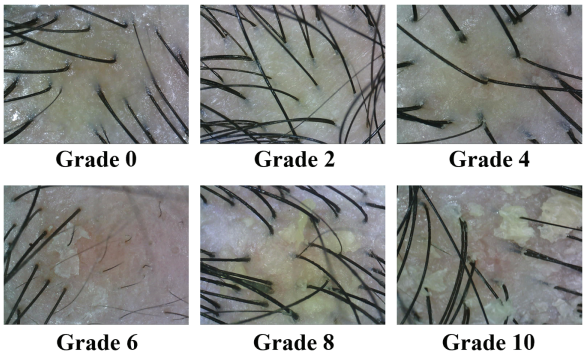

Figure 4: Representative images from our dandruff severity dataset. The six grades are annotated according to ASFS standard

4.2 Experiment Settings

All experiments were conducted on an NVIDIA GeForce RTX 4090 GPU using a ResNet-50 backbone. For our method, the backbone was modified in two ways: the final classification layer was changed to have 5 outputs to suit the ordinal regression task, and a 128-dimensional projection head was added in parallel to facilitate contrastive learning. For data preprocessing, all input images were resized and randomly cropped to 224×224. Across all experiments, we utilized the Adabelief optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, paired with a cosine annealing learning rate schedule. The training protocols were set as follows. For baseline and SOTA comparisons, models were trained for 450 epochs with a batch size of 64 and an initial learning rate of 1×10^{-4}. Our proposed method involved a 10-epoch warm-up phase, followed by 250 epochs of hybrid training with a batch size of 32, an initial learning rate of 1×10^{-3}, and a weight decay of 1×10^{-5}. The key hyperparameters for our framework were set as: GMM partition threshold $\tau = 0.5$, contrastive learning temperature $K = 0.05$, and regularization loss weight $\lambda_r = 0.01$. The contrastive loss weight λ_r was linearly ramped up from 0 to 0.06 over the first 100 epochs to stabilize initial training.

4.3 Comparison with State-of-the-Art Methods

We compare our framework against several SOTA methods, as shown in Table 1. As the literature on robust ordinal learning for dandruff grading is nascent, we benchmark against high-performing models from related scalp analysis tasks [2–5,8]. All competing methods were trained on our dataset using the parameter settings reported in their original papers. As shown in Table 1, our proposed method significantly outperforms all other approaches across all metrics. Our ResNet-50-based model achieves an Accuracy of 80.71% and an F1-Score of 76.86%. This represents a substantial improvement of 2.26% in accuracy and 1.25% in F1-score over the next best method, the Triple ensemble model proposed by Kim et al. [5]. Notably,

our single, principled framework surpasses not only this complex ensemble approach but also advanced architectures such as the Vision Transformer (ViT-B/16). These results validate that our framework's core principles of explicitly modeling ordinal relationships and robustly handling label noise provide a decisive advantage over methods relying solely on architectural sophistication or ensembling.

Table 1: Performance comparison with SOTA methods

Method	Backbone	Accuracy	Precision	Recall	F1-Score
Ha et al. [4]	ViT-B/16	69.91%	68.66%	65.67%	66.78%
Jin et al. [3]	EfficientNet-B0	74.71%	73.39%	70.48%	71.50%
Wang et al. [8]	VGG-16	77.85%	75.81%	74.28%	74.81%
Jhong et al. [2]	Enhanced DenseNet-121	78.18%	76.50%	73.67%	75.06%
Kim et al. [5]	Triple ensemble model	78.45%	77.16%	**75.10%**	75.61%
This work	**Resnet-50**	**80.71%**	**79.95%**	74.80%	**76.86%**

Note: The bold values indicate the best performance.

4.4 Ablation Studies

To verify the contribution of each component of our framework, we conducted a detailed ablation study. Starting with a standard supervised ResNet-50 baseline, we incrementally integrated Ordinal Regression (OR), our Self-Supervised noisy label learning strategy (SSL), and Contrastive Learning (CL). The quantitative results are summarized in Table 2. The baseline model achieves a respectable accuracy of 78.23%. Introducing the OR formulation provides immediate gains, confirming the importance of modeling the inherent ordinal structure of severity grades. Subsequently, incorporating the SSL strategy with cooperative partitioning yields the most significant performance leap. As detailed in the per-class F1-scores, this gain is largely driven by a substantial improvement in Class 0 (rising from 75.68% to 80.03%), strongly suggesting that the "no dandruff" category contains considerable label noise which our SSL mechanism successfully mitigates. Finally, the addition of the CL module provides a further performance boost, achieving the highest Accuracy (80.71%) and Macro F1-score (76.86%). To rigorously assess performance on our imbalanced dataset, we also evaluated the Weighted F1-score. The proposed method achieves a Weighted F1 of 80.56%, surpassing the baseline's 78.18%. This metric confirms that our framework maintains robust classification capabilities across all severity levels, rather than bias towards majority classes.

Table 2: Ablation study on each component for this study

Method			Per-class F1-score (%)						Overall performance (%)			Runtime
			0	2	4	6	8	10	Macro F1	Weighted F1	Accuracy	
Baseline			76.39	82.58	66.67	74.77	80.97	71.76	75.52	78.18%	78.23	0.00264 s
OR	SSL	CL	*Proposed method*									
✓			75.68	82.60	68.46	75.07	82.45	73.33	76.27	78.45%	78.45	0.00257 s
	✓		80.47	84.12	65.91	75.61	79.55	70.19	75.97	79.58%	79.60	0.00464 s
✓	✓		80.03	85.23	67.14	75.36	81.20	71.43	76.73	80.29%	80.36	0.00455 s
✓	✓	✓	80.28	85.80	67.17	73.78	81.88	72.27	76.86	80.56%	80.71	0.00476 s

To further investigate the sources of these performance gains, we visualized the confusion matrices for the key ablation stages in Fig. 5. As observed in Fig. 5a, the Baseline model exhibits a dispersed prediction pattern, with misclassifications frequently spanning across non-adjacent classes (e.g., confus-ing

Grade 2 with Grade 6). Incorporating the OR module (Fig. 5b) effectively constrains these predic-tions, resulting in a concentration of values along the diagonal and a significant reduction in large error margins. Furthermore, Fig. 5c,d illustrates the distinct contributions of the self-supervised strategies. The standalone SSL strategy (Fig. 5c) significantly corrects misclassifications in the noisy Grade 0 class. Subsequently, the intermediate OR + SSL stage (Fig. 5d) demonstrates how combining these approaches begins to align noise robustness with rank consistency. Finally, Fig. 5e presents the full Proposed Method (OR + SSL + CL). By integrating the contrastive learning objective, the model achieves the clearest diagonal structure with minimal off-axis errors. This visual evidence corroborates that our cooperative hybrid framework synergistically mitigates label noise while preserving the intrinsic ordinal structure.

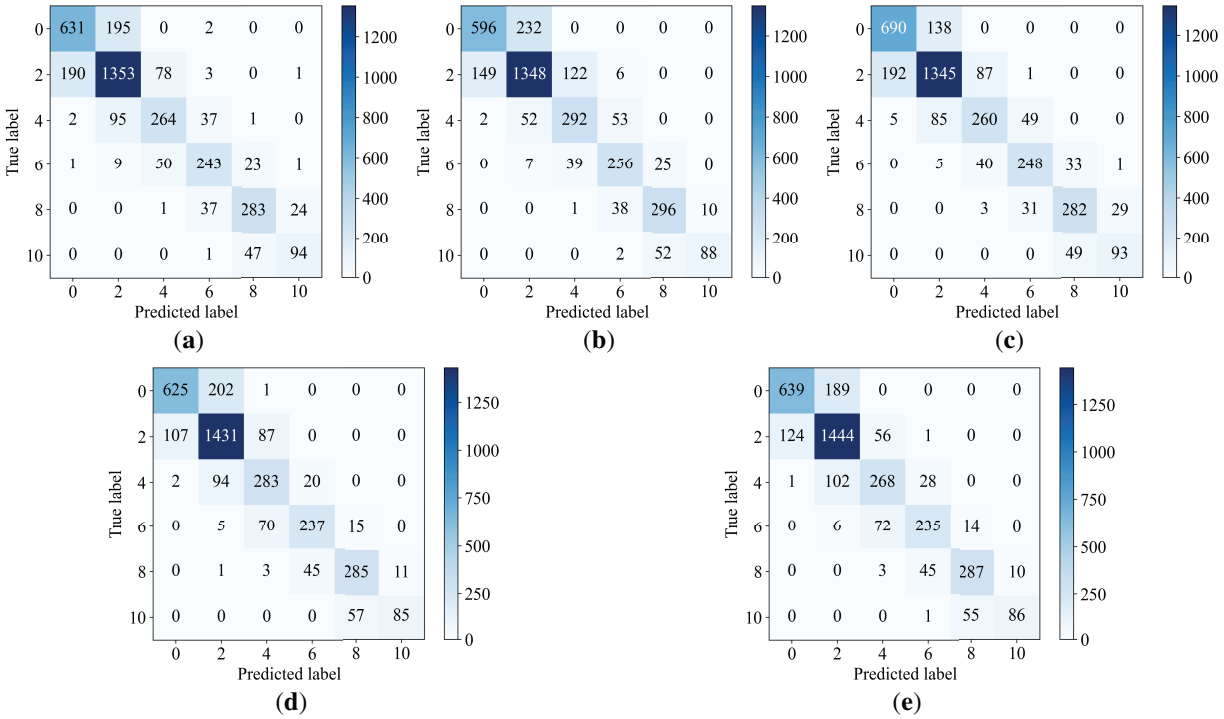

Figure 5: Visual comparison of confusion matrices across different ablation stages. (**a**) The Baseline model shows considerable confusion between adjacent severity grades. (**b**) The addition of OR concentrates predictions along the diagonal and reduces extreme outliers. (**c**) The SSL strategy significantly corrects misclassifications in the noisy Grade 0 class. (**d**) Combining OR + SSL further refines the diagonal structure. (**e**) The Proposed Method (OR + SSL + CL) achieves the most distinct diagonal with minimal off-axis misclassifications, demonstrating superior robustness against label noise

Beyond classification performance, we also addressed the practical requirements for clinical deployment by evaluating inference speed (Runtime) as reported in Table 2. While our hybrid framework introduces a marginal increase in computational cost compared to the baseline (0.00476 s vs. 0.00264 s per image) due to the architectural design, the system remains highly efficient. With an inference speed capable of processing over 200 frames per second, our method proves to be computationally lightweight and well-suited for real-time diagnostic applications.

5 Discussion

The experimental results presented in this study substantiate the efficacy of the proposed cooperative hybrid learning framework in addressing the dual challenges of ordinal severity grading and label noise.

A critical analysis of the performance metrics reveals that standard deep learning models often fail to capture the subtle inter-class variations characteristic of dandruff severity, leading to ordinally inconsistent predictions. By contrast, our method explicitly enforces rank consistency through the ordinal regression backbone, which significantly reduces clinically implausible errors. This is visually corroborated by the confusion matrices, where the predictions of our model are densely concentrated along the diagonal. Unlike the baseline model that exhibits dispersed errors across non-adjacent classes, our approach ensures that even when misclassification occurs, the predicted grade remains proximal to the ground truth. This characteristic is vital for clinical decision support systems, as it minimizes the risk of drastic misdiagnosis that could adversely affect treatment planning.

The ablation studies further highlight the specific contributions of the self-supervised partitioning and contrastive learning components. A notable finding is the substantial performance gain in identifying Grade 0 samples. In clinical practice, distinguishing between a healthy scalp (Grade 0) and mild dandruff (Grade 2) is notoriously difficult due to subjective interpretation, resulting in high label noise for these categories. The proposed cooperative partitioning strategy successfully identified and filtered these ambiguous samples, preventing the model from overfitting to inconsistent annotations. Furthermore, the integration of contrastive learning on the noisy subset proved essential. By pulling representations of the same image together while pushing others apart, the model learned robust and discriminative features directly from the visual data without relying on potentially corrupt labels. This synergy explains why our method outperforms complex architectures like Vision Transformers, which, despite their capacity, lack specific mechanisms to handle the ordinal and noisy nature of this specific medical task.

While the proposed framework demonstrates SOTA performance, we acknowledge certain limitations regarding the dataset demographics and imaging modality. The current study utilized a large-scale dataset collected from clinical centers in Taiwan, representing a specific ethnic group with predominantly dark hair and specific scalp characteristics. Consequently, the generalization of the model to other ethnic groups with different hair textures or scalp pigmentations has not yet been empirically verified. Additionally, the study focused exclusively on high-resolution digital microscope images. While this modality provides detailed visual information necessary for fine-grained grading, it limits the immediate applicability of the model to images captured by standard consumer devices or smartphones without distinct optical attachments. Future research will aim to address these limitations by incorporating multi-ethnic datasets and exploring domain adaptation techniques to extend the frameworks applicability across diverse populations and imaging devices.

6 Conclusions

This study presents a novel cooperative hybrid learning framework designed to overcome the persistent challenges of label noise and ordinality in automated dandruff severity grading. By synergizing a rank-consistent ordinal regression backbone with a self-supervised sample partitioning strategy, the proposed method effectively filters subjective annotation errors while preserving the intrinsic severity ranking. The incorporation of contrastive learning further enhances feature discrimination on noisy data, ensuring robust performance even when ground truth labels are unreliable. Extensive empirical validation on a large-scale clinical dataset confirms that our approach significantly outperforms state-of-the-art methods in both accuracy and stability. Beyond its theoretical contributions to noisy ordinal learning, this framework holds substantial practical value for dermatological healthcare. It offers a reliable, automated second opinion capable of standardizing diagnostic criteria and reducing inter-rater variability among clinicians. Furthermore, the high inference efficiency of the model supports its feasibility for deployment in real-time clinical decision support systems. To further extend the applicability of this framework, future research will prioritize

enhancing generalization capabilities by incorporating diverse multi-ethnic datasets and exploring domain adaptation techniques for smartphone-based image analysis. These advancements will ultimately aim to broaden the accessibility of professional scalp health diagnostics to a wider global population.

Acknowledgement: This work was partially supported by MacroHI Co., Ltd. through the provisioning of the scalp image dataset. We extend our sincere gratitude to the physicians at Taipei Hospital, Ministry of Health and Welfare and the professional scalp therapists at MacroHI Co., Ltd. for their expert contributions to data annotation and for providing invaluable clinical insights into scalp conditions.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Sin-Ye Jhong and Chih-Hsien Hsia; methodology, Sin-Ye Jhong, Chih-Hsien Hsia and Hsin-Hua Huang; software, Hsin-Hua Huang; validation, Hsin-Hua Huang; formal analysis, Sin-Ye Jhong; investigation, Sin-Ye Jhong; resources, Chih-Hsien Hsia and Yung-Yao Chen; data curation, Hui-Che Hsu and Hsin-Hua Huang; writing—original draft preparation, Sin-Ye Jhong and Yung-Yao Chen; writing—review and editing, Sin-Ye Jhong, Chih-Hsien Hsia and Hui-Che Hsu; visualization, Hui-Che Hsu and Hsin-Hua Huang; supervision, Yulius Harjoseputro and Yung-Yao Chen; project administration, Sin-Ye Jhong and Hui-Che Hsu; funding acquisition, Chih-Hsien Hsia and Yung-Yao Chen. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are subject to privacy and commercial restrictions as they were provided by a collaborating hospital and an industry partner. Therefore, the data are not publicly available.

Ethics Approval: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Taipei Hospital, Ministry of Health and Welfare (Protocol No. TH-IRB-0022-0021).

Informed Consent: Informed consent was obtained from all subjects involved in the study.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Mir AN, Rizvi DR. Advancements in deep learning and explainable artificial intelligence for enhanced medical image analysis: a comprehensive survey and future directions. Eng Appl Artif Intell. 2025;158:111413. doi:10.1016/j.engappai.2025.111413.
2. Jhong SY, Yang PY, Hsia CH. An expert smart scalp inspection system using deep learning. Sens Mater. 2022;34(4):1265. doi:10.18494/sam3462.
3. Jin YJ, Park YS, Kang SH, Kim DH, Lee JY. A study on the development of a web platform for scalp diagnosis using EfficientNet. Appl Sci. 2024;14(17):7574. doi:10.3390/app14177574.
4. Ha C, Go T, Choi W. Intelligent healthcare platform for diagnosis of scalp and hair disorders. Appl Sci. 2024;14(5):1734. doi:10.3390/app14051734.
5. Kim M, Gil Y, Kim Y, Kim J. Deep-learning-based scalp image analysis using limited data. Electronics. 2023;12(6):1380. doi:10.3390/electronics12061380.
6. Wang J, Chen J, Liu J, Tang D, Chen DZ, Wu J. A survey on ordinal regression: applications, advances and prospects. arXiv:2503.00952. 2025.
7. Song H, Kim M, Park D, Shin Y, Lee JG. Learning from noisy labels with deep neural networks: a survey. IEEE Trans Neural Netw Learn Syst. 2023;34(11):8135–53. doi:10.1109/TNNLS.2022.3152527.
8. Wang WC, Chen LB, Chang WJ. Development and experimental evaluation of machine-learning techniques for an intelligent hairy scalp detection system. Appl Sci. 2018;8(6):853. doi:10.3390/app8060853.
9. Roy M, Protity AT. Hair and scalp disease detection using machine learning and image processing. arXiv:2301.00122. 2023.

10. Jhong SY, Li GT, Hsia CH. An edge-cloud collaborative scalp inspection system based on robust representation learning. IEEE Trans Consum Electron. 2025;71(1):1551–62. doi:10.1109/TCE.2024.3474911.
11. Chang WJ, Chen LB, Chen MC, Chiu YC, Lin JY. ScalpEye: a deep learning-based scalp hair inspection and diagnosis system for scalp health. IEEE Access. 2020;8:134826–37. doi:10.1109/ACCESS.2020.3010847.
12. Chen LB, Chang WJ, Chiu YC, Huang XR. An efficient scalp inspection and diagnosis system using multiple deep learning-based modules. IEEE Can J Electr Comput Eng. 2024;47(1):22–35. doi:10.1109/icjece.2024.3354291.
13. Bacon RA, Mizoguchi H, Schwartz JR. Assessing therapeutic effectiveness of scalp treatments for dandruff and seborrheic dermatitis, part 1: a reliable and relevant method based on the adherent scalp flaking score (ASFS). J Dermatolog Treat. 2014;25(3):232–6. doi:10.3109/09546634.2012.687089.
14. Niu Z, Zhou M, Wang L, Gao X, Hua G. Ordinal regression with multiple output CNN for age estimation. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 4920–8. doi:10.1109/CVPR.2016.532.
15. Cao W, Mirjalili V, Raschka S. Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognit Lett. 2020;140(24):325–31. doi:10.1016/j.patrec.2020.11.008.
16. Díaz R, Marathe A. Soft labels for ordinal regression. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 4733–42. doi:10.1109/CVPR.2019.00487.
17. Reed S, Lee H, Anguelov D, Szegedy C, Erhan D, Rabinovich A. Training deep neural networks on noisy labels with bootstrapping. arXiv:1412.6596. 2015.
18. Han B, Yao Q, Yu X, Niu G, Xu M, Hu W, et al. Co-teaching: robust training of deep neural networks with extremely noisy labels. In: Neural Information Processing Systems; 2018 Dec 3–8; Montréal, QC, Canada. p. 8527–37.
19. Li J, Socher R, Hoi SCH. DivideMix: learning with noisy labels as semi-supervised learning. In: International Conference on Learning Representations; 2020 Apr 26–30; Addis Ababa, Ethiopia. p. 1–14.
20. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proceedings of the International Conference on Machine Learning; 2020 Jul 13–18; Vienna, Austria. p. 1597–607.
21. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. Mixup: beyond empirical risk minimization. In: Proceedings of the International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada. p. 1–13.