

Article

A Study of SimCLR-Based Self-Supervised Learning for Acne Severity Grading Under Label-Scarce Conditions

Krittakom Srijiranon, Nanmanat Varisthanist and Tanatorn Tanantong *

Thammasat University Research Unit in Data Innovation and Artificial Intelligence, Department of Computer Science, Faculty of Science and Technology, Thammasat University, Pathum Thani 12121, Thailand; krittakom_s@sci.tu.ac.th (K.S.); nanmanat.v@gmail.com (N.V.)

* Correspondence: tanatorn@sci.tu.ac.th

Abstract

Acne severity grading is an important dermatological task that supports clinical diagnosis, treatment planning, and disease monitoring. Although self-supervised learning (SSL) has gained interest as a means to reduce reliance on large annotated datasets, its effectiveness for fine-grained and ordinal dermatological tasks remains unclear. This research systematically evaluates contrastive SSL for acne severity grading by comparing SimCLR-based pretraining with a diverse set of supervised deep learning models, including Convolutional Neural Networks and Vision Transformers, under controlled experimental conditions. The evaluation considers full-data training, label-scarce scenarios, and temperature tuning of the contrastive loss. The results consistently demonstrate the superiority of supervised transfer learning, which achieves Quadratic Weighted Kappa (QWK) scores ranging from 0.7616 to 0.8533. In contrast, SimCLR-based models exhibit substantially lower performance, with QWK values between 0.2343 and 0.4548 after fine-tuning. Although temperature adjustment achieved modest performance gains, it does not close this gap, with the best configuration attaining a QWK of 0.4548 using a ResNet18 backbone. Qualitative analysis using Grad-CAM further reveals that SimCLR-based contrastive SSL tends to exhibit diffuse attention patterns and limited localization of clinically relevant acne regions. Overall, these findings indicate that generic contrastive SSL objectives are poorly aligned with the subtle and localized visual cues required for acne severity grading. The results highlight the need for domain-aware representation learning in fine-grained dermatological image analysis.

Keywords: acne severity grading; deep learning; dermatological image analysis; self-supervised learning; SimCLR

Academic Editors: Chih-Yu Hsu and Shuo-Tsung Chen

Received: 31 December 2025

Revised: 8 February 2026

Accepted: 10 February 2026

Published: 12 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

1.1. Motivation

Acne is one of the most common skin diseases, affecting nearly 90% of the global population [1], and it occurs more frequently in adolescents and adults. Beyond its high prevalence, acne is associated with substantial physical and psychosocial burden, including permanent scarring, impaired quality of life, and increased risks of anxiety and depression, as consistently reported in clinical and population-based studies [2,3]. Acne presents in various forms, including whiteheads (closed plugged pores), blackheads (open plugged pores), small red tender bumps (papules), pimples (pustules), large solid painful

lumps under the skin (nodules), and painful pus-filled lumps under the skin (cystic lesions). These lesions commonly appear on the face, chest, shoulders, and upper back [4]. Acne can cause several negative effects, including severe skin damage and adverse mental health impacts, which often lead to reduced self-esteem, especially among adolescents [5]. Diagnosing acne typically involves grading severity by counting the number of lesions based on established criteria such as the Hayashi Criteria [6] and the Global Acne Grading System (GAGS) [7]. Currently, dermatologists diagnose acne severity by manually counting lesions. However, this method is time-consuming and prone to human errors, which may result in misdiagnosis.

In recent years, substantial advances in Artificial Intelligence (AI) have led to its application across various domains, particularly in healthcare. Within dermatology, deep learning-based approaches have been increasingly adopted for skin disease classification, including mobile-based diagnostic systems and population-specific applications under data-limited conditions [8,9]. Building upon these classification capabilities, more advanced AI techniques have been developed for object detection and localization, which are fundamentally rooted in machine learning, particularly deep learning. Such approaches have also been applied to a wide range of medical diagnostic tasks, including tumor detection using Raman spectroscopy [10], glaucoma identification from medical images [11], and prognosis or mortality prediction of COVID-19 using chest X-rays and electronic health record data [12]. Extending these developments to acne analysis, several studies have applied deep learning models for acne detection and severity grading to support individuals with limited access to dermatologists and to streamline clinical workflows. For example, regression models and Faster Region-based Convolutional Neural Networks (Faster-RCNN) have been used to grade acne severity on the ACNE04 dataset [13]. The Modified Pyramid Scene Parsing Network (PSPNet) [14] has also been applied to grade acne severity using a private dataset collected from Xiangya Hospital. However, most existing studies focus on training models on a single dataset and fine-tuning them to achieve optimal accuracy.

Despite these advances, most existing acne severity grading systems rely on fully supervised learning, which requires large amounts of expert-annotated data. In dermatology, such annotations are costly, time-consuming, and prone to inter-observer variability, limiting scalability and generalization across datasets and clinical settings. Self-supervised learning (SSL) [15] has emerged as a potential alternative by enabling representation learning from unlabeled data. However, its effectiveness for fine-grained acne severity grading remains unclear. In particular, it is uncertain whether generic contrastive SSL methods can capture the subtle, localized features and ordinal relationships required for clinical assessment. This gap motivates a systematic evaluation of SSL in comparison with established supervised approaches, and an analysis of its limitations and suitability for dermatological image analysis, which this study aims to address.

1.2. Background

Deep learning is a subset of machine learning and is commonly implemented as neural networks with three or more layers that simulate the functionality of the human brain. This structure enables deep learning models to learn meaningful patterns from large datasets. Deep learning has been widely applied in tasks such as voice recognition, text generation, and object detection. Convolutional Neural Networks (CNNs) are among the most widely used deep learning architectures, particularly for image recognition and computer vision tasks. CNNs eliminate the need for manual feature extraction by automatically learning features directly from images. This capability makes CNN-based models highly effective for computer vision applications and contributes to their strong performance in image classification. For example:

- VGG [16] is a 16-layer CNN architecture characterized by the use of multiple sequential 3×3 convolutional filters. This design emphasizes depth and uniformity in convolutional operations, enabling effective hierarchical feature extraction while maintaining a simple architecture.
- ResNet [17] introduces residual connections to address the increasing error rate associated with deeper networks, particularly the vanishing gradient problem. These skip connections enable more stable training of deeper CNNs and significantly improve both accuracy and convergence.
- EfficientNet [18] incorporates compound scaling, which uniformly scales network depth, width, and resolution using a single coefficient. This approach achieves a strong accuracy-efficiency tradeoff, allowing the model to deliver high performance while maintaining low computational cost.
- ConvNeXt [19] is a modernized CNN architecture inspired by Vision Transformer design principles. It employs simplified convolutional blocks, larger kernel sizes, and improved normalization strategies, enabling CNNs to achieve competitive performance with Transformer-based models while preserving convolutional efficiency.
- DenseNet121 [20] is a densely connected architecture in which each layer receives feature maps from all preceding layers. This connectivity pattern enhances feature reuse, strengthens gradient flow, reduces parameter redundancy, and improves overall learning efficiency.
- Regnet [21] is developed through design-space exploration to identify simple and regular network patterns that scale predictably. It focuses on constructing architectures with consistent structural rules, achieving strong performance and efficiency across various computational budgets.
- MobileNet [22] is a lightweight CNN model optimized for mobile and embedded devices. It utilizes depthwise separable convolutions to drastically reduce computational cost and parameter count while maintaining competitive accuracy, making it suitable for real-time and resource-constrained environments.

Beyond CNNs, the emergence of Vision Transformers (ViT) [23] represents a newer approach in computer vision. ViT applies the Transformer architecture, originally designed for natural language processing, to image understanding by dividing images into fixed-size patches and processing them through self-attention mechanisms. While CNNs inherently emphasize local spatial patterns through convolutional filters, ViTs learn long-range dependencies globally from the outset, enabling them to capture richer contextual relationships within an image. This global modeling capability, combined with reduced reliance on handcrafted inductive biases, allows ViTs to scale effectively with large datasets and often surpass CNN performance when trained on a large amount of data. As deep learning continues to advance, ViTs have become a powerful alternative to traditional CNN architectures. For example:

- Swin Transformer [24] is a hierarchical Transformer architecture that processes images through shifted windows. This approach improves computational efficiency, preserves spatial locality, and scales effectively to high-resolution inputs.
- Data-efficient Image Transformers (DeiT) [25] is a training-optimized version of ViT that incorporates knowledge distillation and improved optimization strategies, enabling Transformer models to achieve high accuracy without requiring extremely large datasets.

SSL has emerged as a powerful alternative to traditional supervised learning, particularly in domains where labeled data is limited or expensive to obtain. Instead of relying on manual annotations, SSL trains models by designing pretext tasks that allow the network to learn meaningful representations directly from unlabeled data. In computer

vision, these tasks typically involve predicting transformations, contrasting image views, or reconstructing missing information, allowing the model to capture semantic features without human supervision. By learning robust and generalizable representations, SSL models can outperform or match supervised counterparts when fine-tuned on downstream tasks, especially in scenarios with scarce labeled datasets. As a result, SSL has become an increasingly important technique in modern deep learning pipelines, bridging the gap between large-scale pretraining and real-world applications with limited data availability. Specifically, Simple Framework for Contrastive Learning of Visual Representations (SimCLR) [26] is a contrastive learning framework that trains models by maximizing the agreement between different augmented views of the same image. It relies on strong data augmentations and a projection head to learn meaningful visual representations without requiring labeled data.

1.3. Prior Studies on Acne Detection and Severity Grading

From a previous study, deep learning approaches for acne analysis can generally be categorized into three groups. The first group focuses on fine-tuning existing deep learning architectures [27–33]. For example, Ref. [30] adapted Inception V3 for acne detection and achieved an accuracy of 89.39% through parameter fine-tuning. The second group introduces newly proposed models that use established architectures as backbones [14,34–43]. In ref. [41], VGG16 served as the backbone with an additional skin-patch module, yielding accuracies of 84.52% on the ACNE04 dataset and 52.85% on the PLSBRACNE01 dataset. The last group employs ensemble or hybrid models that combine multiple deep learning architectures [13,36,40,43–45]. In ref. [45], ResNet-50 and YOLOv5 were employed to create a hybrid deep learning model, and a fine-tuning process was applied to both networks. ResNet-50 was used for the classification module, while YOLOv5 was used for the localization module, resulting in an accuracy of 99.31% on the ACNE04 dataset.

The conclusions drawn from the 20 selected articles are summarized in Table 1, which includes the dataset source, severity grading criteria, image processing techniques, deep learning models used, statistical indicators, and model outcomes. The key points are as follows:

- The collective findings indicate that 15 studies utilized open datasets, with 80% of them employing the ACNE04 dataset. This observation highlights ACNE04 as the most commonly adopted open dataset for acne severity grading.
- Notably, only four articles incorporated more than one dataset, and their outcomes clearly demonstrate a significant accuracy drop on the second dataset. For example, in ref. [34], Diagnostic Evidence Distillation (DED) was proposed using CNNs as the backbone for acne severity grading. The precision and accuracy achieved were 86.06% and 85.31% on the ACNE04 dataset, respectively, but these values decreased to 69.16% and 67.56% on the PLSBRACNE01 dataset. Similarly, in ref. [37], Prior Knowledge Guided (PKG) modeling was used with CNN as a backbone, resulting in an accuracy of 85.27% on the ACNE04 dataset, which dropped to 65.85% on the PLSBRACNE01 dataset.
- All deep learning models reviewed in the literature rely heavily on labeled data, which is particularly challenging to obtain in dermatology and medical imaging. Acne severity annotation requires expert dermatologists to manually grade each image, a process that is time-consuming, costly, and prone to inter-observer variability. Moreover, publicly available acne datasets remain limited in size and diversity, making it difficult for supervised models to generalize effectively across variations in skin tones, lighting conditions, and imaging devices.

Table 1. Summary of previous studies on acne detection and severity grading.

Authors	Dataset	SGC	Image Processing	Deep Learning Model	Statistical Indicators	Outcomes
H. Wen et al. [44]	ACNE04	Hayashi	-	R-CNN and YOLOv4	mAP	ASGC
H. Zein et al. [33]	ACNE04	Hayashi	-	InceptionResNetv2	Acc	ASGC
H. Zhang et al. [45]	ACNE04	Hayashi	-	ResNet50 and YOLOv5	Acc	ASGC
J. Wang et al. [14]	Private dataset	GAGS	Background removal	Localization-DL based on PSPNet and Class-Seg based on HRNet	Acc	ASGC
J. Zhang et al. [43]	Private Dataset and ACNE04	-	-	SADH with Mask R-CNN and ResNet50-based	AP	AD
K. Min et al. [39]	ACNE04	-	-	ACNet with CNN-based	mAP	AD
K. Naidu et al. [30]	DermNet	-	Background removal	Inception V3	Acc	AD
K. Rashataprucksa et al. [31]	Private Dataset	-	-	R-FCN	Acc, Prec, AP	AD
M. S. Junayed et al. [38]	Private dataset	-	Contrast Enhancement and Color Conversion	ScarNet with CNN-based	Acc, Spec, KS	AD
N. S. Suriani et al. [32]	DermNet	-	Augmentation	MobileNetV2	Acc, F1	-
N. Yadav et al. [27]	DermNet	-	k-means, Gaussian method, and HSV segmentation	CNN	Acc	AS
Q. T. Huynh et al. [40]	ACNE04 and PLSBRACNE01	Hayashi	Information Fold, Key area segmentation	GLFN with VGG16-based	Acc, Sen, Spec, YI	ASGC
S. Alzahrani et al. [13]	ACNE04	Hayashi	-	UNet dense regressor and Faster R-CNN	MAE, MSE, Prec, Spec, Sens, Acc	ASGC
T. Zhao et al. [28]	Private dataset	Hayashi	Extracted and rolling skin patches	ResNet	RMSE	ASGC
X. Wei et al. [42]	Private Dataset and ACNE04	-	-	DSDH with Mask R-CNN-based	AP	AD
X. Wu et al. [36]	ACNE04	Hayashi	-	LDL and ResNet-50-based	Acc	ASGC
Y. Lin et al. [34]	ACNE04 and PLSBRACNE01	Hayashi	-	DED with EfficientNet-based	Acc, Prec, Sens, Spec, YI	ASGC
Y. Lin et al. [35]	ACNE04 and PLSBRACNE01	Hayashi	AES	SFNet with VGG16-based	Acc	ASGC
Y. Lin et al. [37]	ACNE04 and PLSBRACNE01	Hayashi	-	PKG with CNN-based	Acc	ASGC
Y. Lin et al. [41]	ACNE04 and PLSBRACNE01	Hayashi	Information Fold, Key area segmentation, Augmentation	VGG16	Acc, Prec, Sens, Spec, YI	ASGC

Acc denotes accuracy, AD denotes acne detection, AES denotes Attention Extraction Segmentation, AP denotes average precision, AS denotes Acne Segmentation, ASGC denotes acne severity grading classification, DSDH denotes Decoupled Sequential Detection Head, F1 denotes F1-score, GLFN denotes global local fusion network, KS denotes Kappa Score, LDL denotes Label Distribution

Learning, MAE denotes mean absolute error, mAP denotes mean average precision, MSE denotes mean squared error, Prec denotes precision, SADH denotes Spatial Aware Double Head, Sens denotes sensitivity, SGC denotes severity grading criteria, Spec denotes specificity, YI denotes Youden Index.

1.4. Our Research

This research systematically investigates the applicability of SSL for acne severity grading by leveraging contrastive pretraining with SimCLR and comparing its effectiveness against a broad range of supervised deep learning models. The study evaluates multiple CNN and ViT backbones under consistent experimental settings, including full-data and label-scarce scenarios, to assess whether SimCLR-based contrastive SSL representations can provide robust and clinically meaningful performance. In addition, this research examines the effect of contrastive loss temperature on downstream acne severity classification performance and representation behavior, highlighting its influence on both classification accuracy and spatial attention characteristics. The main contribution of this research is a controlled benchmark of standard contrastive SSL against established supervised approaches for acne severity grading with subtle differences between levels, together with an analysis of temperature sensitivity and representation behavior using PCA, t-SNE, and Grad-CAM. Through these analyses, this research systematically reveals persistent performance gaps and clarifies the strengths and limitations of SimCLR for dermatological image analysis, motivating the need for domain-adapted self-supervised representation learning strategies for medical image classification tasks with subtle visual differences.

2. Methodology

2.1. Datasets

The datasets used in this research were obtained through a systematic search of publicly available repositories, including widely used data-sharing platforms such as Kaggle and Roboflow. The preliminary open dataset comprises two widely used acne image datasets collected from different research sources [46]. The ACNE04 dataset [36] contains 1419 images consisting exclusively of acne images with corresponding labels. Similarly, the AcneSCU dataset [47] includes 276 images with high-resolution images, precise annotations, and fine-grained lesion categories. These datasets form the initial foundation for model development and evaluation in this research, where both datasets are combined and subsequently split into training, validation, and test sets using a random seed of 42.

Severity grading criteria are commonly used to quantify acne severity, yet their definitions and grading scales vary across studies. Two widely adopted standards are the Hayashi Grading Criteria and GAGS. The Hayashi criteria classify acne severity based solely on the total number of acne lesions, ranging from no acne (0–5 lesions) to moderate (6–20 lesions), severe (21–50 lesions), and very severe cases with more than 50 lesions. In contrast, GAGS determines acne severity using a composite score that accounts for both the distribution of acne across different facial regions and lesion types, categorizing cases as none (score 0), mild (1–18), moderate (19–30), severe (31–38), and very severe when the score exceeds 38. Although GAGS provides a more comprehensive clinical assessment, this study adopts the Hayashi Grading Criteria due to their simplicity and direct applicability to image-based lesion counting, which facilitates consistent annotation and model evaluation and has been applied to the AcneSCU dataset for further model training and evaluation.

2.2. System Architecture

SimCLR is an SSL framework designed to learn discriminative and transferable visual representations without relying on labeled data. Its architecture is intentionally simple, yet effective, consisting of a backbone encoder, a projection head, and a contrastive learning objective that together enable representation learning from large collections of unlabeled images. At the core of SimCLR is a deep convolutional encoder $f(\cdot)$, which serves as the feature extraction backbone. Common choices for this encoder include ResNet-based architectures, where the final classification layer is removed, and the output of the global average pooling layer is used as the representation h . This encoder is responsible for capturing high-level semantic features such as texture, shape, and structural patterns, which are critical for visual understanding tasks. In practice, the encoder architecture can be adapted to different backbone depths or variants depending on computational constraints and task complexity.

Following the encoder, SimCLR introduces a projection head $g(\cdot)$, which is typically implemented as a small multilayer perceptron (MLP) composed of two or three fully connected layers with non-linear activation functions. The projection head maps the encoder output h into a lower-dimensional latent space z , where the contrastive loss is applied. Although the projection head is used only during self-supervised pretraining, it plays a crucial role in improving the quality of learned representations. It allows the encoder to retain task-relevant information while the contrastive objective is optimized in a separate representation space.

For each input image x , two stochastic data augmentations are applied to generate correlated views $\tilde{x}_i$ and $\tilde{x}_j$. These augmented images are independently processed by the shared encoder and projection head, producing representations z_i and z_j . The similarity between these representations is measured using cosine similarity, and training is driven by the normalized temperature-scaled cross-entropy (NT-Xent) loss [26]. For a positive pair (i, j) in a batch of $2N$ augmented samples, the loss is defined in

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\mathcal{T})}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_j)/\mathcal{T})} \quad (1)$$

where $\text{sim}(\cdot)$ denotes cosine similarity, $\mathcal{T}$ is the temperature parameter, and the denominator includes all negative samples within the batch. The overall loss is obtained by averaging over all positive pairs.

After self-supervised pretraining, the projection head $g(\cdot)$ is discarded, and the pre-trained encoder $f(\cdot)$ is reused for downstream tasks. A lightweight classifier is then attached to the encoder for tasks such as acne severity grading, either by freezing the encoder and training a linear classifier or by fine-tuning the entire network. This architectural design enables SimCLR to learn robust and transferable representations that are particularly effective in label-scarce medical imaging settings.

While SimCLR was originally designed for generic natural image representation learning, applying it to acne severity grading introduces additional challenges. As illustrated in Figure 1, SimCLR is used as a standard self-supervised baseline to examine how instance-level contrastive objectives interact with dermatological image characteristics, where discriminative features are often localized, and severity labels exhibit ordinal structure. No task-specific modifications are introduced to the SimCLR objective, allowing the observed performance and representation behavior to be attributed to the intrinsic properties of the contrastive formulation rather than architectural customization. This design enables a controlled assessment of the suitability and limitations of standard contrastive SSL for medical image classification with subtle visual differences between severity levels.

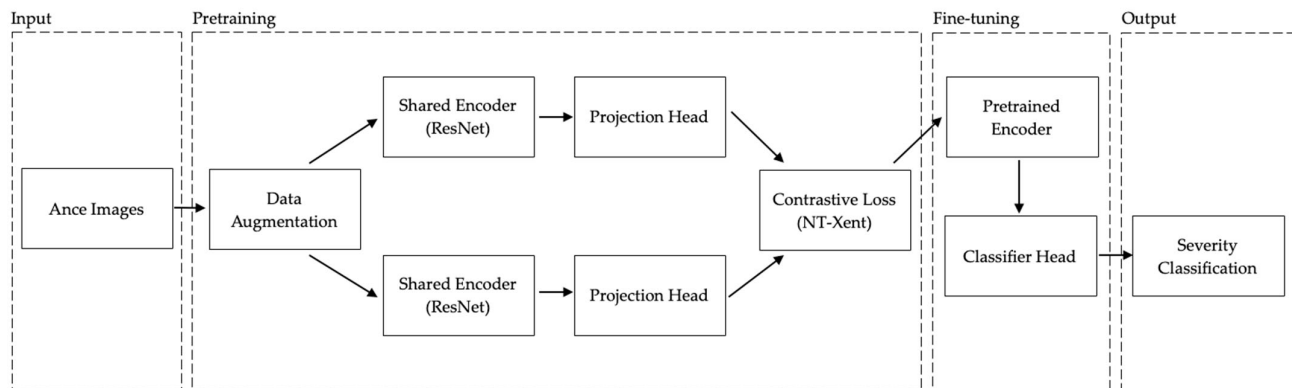


Figure 1. System architecture for SimCLR-based self-supervised pretraining and acne severity classification.

2.3. Evaluation Metrics

To provide a comprehensive and clinically meaningful evaluation of acne severity grading performance, we report metrics that capture ordinal agreement, class-balanced accuracy, discriminative ability, and probabilistic calibration. Acne severity is an ordinal task; misclassifying “Severe” as “Moderate” is less harmful than misclassifying “Severe” as “Clear.” Therefore, accuracy alone is insufficient to quantify model reliability.

Quadratic Weighted Kappa (QWK) [48], originally introduced for assessing human rater agreement, quantifies ordinal consistency between the prediction models and the ground-truth acne severity labels. QWK is particularly appropriate when the target variable exhibits an ordered progression, as is typical in acne grading. QWK (κ) is calculated from

$$\kappa = 1 - \frac{\sum_{i=1}^N \sum_{j=1}^N \omega_{ij} O_{ij}}{\sum_{i=1}^N \sum_{j=1}^N \omega_{ij} E_{ij}} \quad (2)$$

where N is the number of classes. There are three related variables. Firstly, O_{ij} is the observed agreement matrix representing the proportion of examples with true class i and predicted class j . Secondly, E_{ij} is the expected agreement matrix under chance agreement, calculated from

$$E_{ij} = \frac{r_i \times c_j}{T} \quad (3)$$

where r_i is the number of samples in row i of the confusion matrix, c_j is the number of samples in column j of the confusion matrix, T is the number of samples in the confusion matrix. Finally, ω_{ij} is the quadratic weight calculated by

$$\omega_{ij} = \frac{(i-j)^2}{(N-1)^2} \quad (4)$$

Balanced Accuracy (BalAcc) [49] mitigates class imbalance by computing the mean recall across all classes, giving equal weight to underrepresented categories, calculated by

$$BalAcc = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FN_i} \quad (5)$$

where TP_i is a true positive for class i and FN_i is a false negative for class i . Acne severity datasets are naturally imbalanced; most real-world samples fall into “Mild” and “Moderate” categories, while “Severe” cases are far less common.

Macro-Averaged F1 Score (Macro-F1) [50] emphasizes per-class performance, summarizing the harmonic mean of precision and recall per class and averaging them equally. Macro-F1 is sensitive to both incorrect predictions and missed detections. Macro-F1 is calculated from

$$Macro - F1 = \frac{1}{N} \sum_{i=1}^N \frac{2P_i R_i}{P_i + R_i} \quad (6)$$

where FP_i is a false positive for class i , P_i is a precision for class i calculated from

$$P_i = \frac{TP_i}{TP_i + FP_i} \quad (7)$$

and R_i is a recall for class i calculated from

$$R_i = \frac{TP_i}{TP_i + FN_i} \quad (8)$$

This metric is particularly important for acne severity grading, where misclassifying “Severe” as “Mild” and vice versa carries different clinical implications.

Unlike accuracy or F1-score, Macro-Averaged AUROC (Macro-AUROC) [51] is calculated from

$$Macro - AUROC = \frac{1}{N} \sum_{i=1}^N AUC_i \quad (9)$$

where AUC_i is an area under the curve of class i , evaluates discrimination independent of classification threshold, measuring how well the model can rank samples by severity likelihood.

3. Experiments

This research conducted a comprehensive series of experiments to evaluate the effectiveness and limitations of SimCLR-based SSL for acne severity grading under varying annotation budgets. The experimental design investigates three main experiments, including whether learned representations improve ordinal classification performance compared with supervised transfer learning, how performance scales under label scarcity, and how fine-tuning affects the SimCLR model. This research benchmarks SimCLR models against strong supervised baselines, evaluates performance using linear probing, and assesses calibration, robustness, and subgroup fairness. All experiments are conducted on identical train-validation-test splits with subject-level separation. The subsections below describe the experimental setup, baselines, ablation configurations, implementation details, and evaluation protocols.

3.1. Experiment Setup

All models are trained using the combination of ACNE04 and AcneSCU dataset split with 70%, 15%, and 15% allocated for training, validation, and testing, respectively. This partitioning strategy has also been used in prior dermatology-related medical image analysis studies to support robust model selection and unbiased performance evaluation [52,53]. To evaluate label efficiency, this research creates four additional training subsets containing 1%, 5%, 10%, and 25% of the labeled training samples. These subsets are sampled using a fixed random seed to ensure reproducibility. All experiments are conducted on a single NVIDIA A4000 GPU with 16 GB VRAM using PyTorch with version 2.6.0 and automatic mixed precision (AMP). All reported metrics, including QWK, Balanced Accuracy, Macro-F1, and Macro-AUROC, are computed on the testing dataset.

SimCLR is trained on the entire unlabeled image pool using two augmented views per image. Unless otherwise specified, this research uses ResNet-18 as the backbone, an input image size of 224×224 , a temperature value of 0.1, an MLP-2 projection head with a 128-dimensional output, and LARS (Layer-wise Adaptive Rate Scaling) as the optimizer. Training is conducted for 100 epochs using the default SimCLR augmentations, including random crops, flip, color jitter, Gaussian blur.

To provide evidence of training stability, Figure 2 presents a representative SimCLR pretraining loss curve using ResNet18 with a temperature value of 1.5 over 8f00 epochs.

The figure shows both training and validation contrastive loss curves, which follow the same decreasing trajectory and converge with a consistently small gap across epochs, suggesting stable optimization with no clear signs of severe overfitting. The absence of divergence between the curves supports the interpretation that the downstream performance limitations are more likely related to the contrastive objective than to training instability. The contrastive loss decreases rapidly in the early phase and transitions into a slow-converging regime with diminishing improvements well before 100 epochs. Due to computational resource and training time constraints, 100 epochs were therefore selected as the default SimCLR pretraining length across the experimental grid, providing a practical balance between training stability and computational feasibility. Although longer pretraining can further reduce the contrastive loss, extending all configurations in the full grid to 600–800 epochs was not feasible given the scale of the experiments. Moreover, no downstream performance trends were observed to suggest that substantially longer pretraining would materially change the study's conclusions. Under extended pretraining, the achieved macro precision, macro recall, Macro-F1, QWK, balanced accuracy, and Macro-AUROC were 0.4494, 0.4365, 0.4419, 0.4752, 0.4365, and 0.6655, respectively, which are consistent with the reported results.

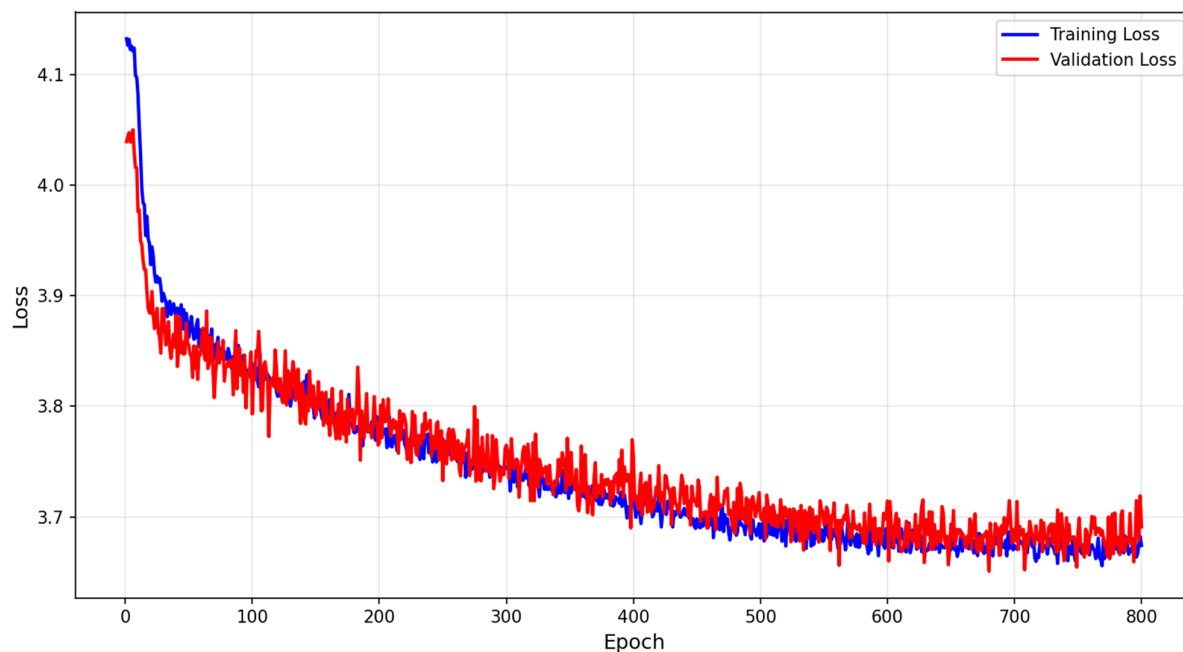


Figure 2. SimCLR pretraining loss curve for ResNet18 at temperature = 1.5 over 800 epochs.

3.2. Comparison Between SimCLR-Based Contrastive Self-Supervised and Supervised Models

The first experiment evaluates whether SimCLR-based self-supervised pretraining provides any measurable advantage over conventional supervised transfer learning for acne severity grading. This experiment compares a SimCLR-pretrained ResNet-50 encoder, evaluated using linear probing, against a diverse suite of ten supervised baselines. In addition, cross-dataset generalization is assessed by training on ACNE04 (80% for training and 20% for validation) and testing on AcneSCU, and vice versa, as summarized in Table 2. The supervised baselines include seven CNN architectures, namely VGG16, ResNet50, EfficientNetV2-S, ConvNeXt-Tiny, DenseNet121, RegNetY-8GF, MobileNetV3-Large, and three ViT models, including ViT-Small, Swin-Tiny, and DeiT-Small. All models share identical data splits, input resolutions, and training hyperparameters to ensure

comparability. The comparative performance of self-supervised and supervised models is summarized in Table 3.

Table 2 reports cross-dataset performance when training on ACNE04 and testing on AcneSCU, and vice versa. When trained on ACNE04, ConvNeXt-Tiny achieves the highest performance among CNNs with a QWK of 0.5439, Balanced Accuracy of 0.5303, Macro-F1 of 0.3927, and Macro-AUROC of 0.7389, while most other CNNs remain below a QWK of 0.27 and a Macro-F1 of 0.31. When trained on AcneSCU and tested on ACNE04, transformer-based models outperform CNNs, with Swin-Tiny achieving a QWK of 0.5483, Balanced Accuracy of 0.4554, Macro-F1 of 0.4650, and Macro-AUROC of 0.7892, and DeiT-Small achieving a QWK of 0.5589, Balanced Accuracy of 0.4584, Macro-F1 of 0.4711, and Macro-AUROC of 0.7564. Overall, the consistently low performance of most models across both directions highlights limited cross-dataset generalization when trained on a single dataset, likely due to domain shifts in acquisition conditions and annotation characteristics between datasets.

Table 2. Comparative performance of models under cross-dataset testing.

Model	Training on ACNE04 and Testing on AnceSCU				Training on AnceSCU and Testing on ACNE04			
	QWK	BalAcc	Macro-F1	Macro-AUROC	QWK	BalAcc	Macro-F1	Macro-AUROC
CNN								
VGG16	0.2669	0.3352	0.1426	0.6659	0.2970	0.3106	0.3072	0.6539
ResNet50	0.0116	0.3750	0.1607	0.5672	0.0070	0.2618	0.2032	0.6065
EfficientNetV2-S	0.0443	0.3750	0.1655	0.6008	0.0538	0.2665	0.1968	0.5676
ConvNeXt-Tiny	0.5439	0.5303	0.3927	0.7389	0.1853	0.2881	0.2814	0.6173
DenseNet121	0.1797	0.4640	0.3130	0.5367	0.0733	0.2745	0.2092	0.6138
RegNetY-8GF	0.0000	0.2500	0.1622	0.6675	0.1640	0.3129	0.2655	0.6485
MobileNetV3-Large	0.0690	0.3390	0.1776	0.5032	0.1627	0.3026	0.2906	0.6695
ViTs								
Swin-Tiny	0.3553	0.3769	0.1958	0.6547	0.5483	0.4554	0.4650	0.7892
DeiT-Small	0.3284	0.4413	0.2916	0.7076	0.5589	0.4584	0.4711	0.7564
ViT-Small	0.1115	0.3163	0.1341	0.6733	0.3024	0.3776	0.3142	0.6753

Table 3 shows that SimCLR representations are substantially weaker than those of all supervised baselines. Linear probing shows limited feature separability, and full fine-tuning fails to match the performance of even mid-sized supervised CNNs. Using QWK as the primary metric, SimCLR underperforms across all evaluation criteria, including Balanced Accuracy, Macro-F1, and Macro-AUROC. These results indicate that contrastive pretraining does not yield clinically meaningful ordinal representations for this task. Although negative, this finding is scientifically valuable because it highlights the limitations of standard contrastive self-supervision for dermatological image analysis and motivates the need for domain-adapted SSL strategies or alternative pretraining frameworks better suited to subtle and localized skin-lesion features.

Table 3. Comparative performance of SimCLR models and supervised CNN and ViT models.

Models	QWK	BalAcc	Macro-F1	Macro-AUROC	Inference (ms)
CNN					
VGG16	0.8115	0.7202	0.7272	0.9178	2.61
ResNet50	0.8011	0.7067	0.7033	0.9102	2.00
EfficientNetV2-S	0.8533	0.7535	0.7396	0.9158	1.85
ConvNeXt-Tiny	0.7990	0.7162	0.7157	0.9025	1.79
DenseNet121	0.8066	0.7211	0.7174	0.9257	2.27
RegNetY-8GF	0.8169	0.6839	0.7179	0.9292	0.90
MobileNetV3-Large	0.7616	0.7095	0.6906	0.8994	1.10
ViTs					
Swin-Tiny	0.8714	0.7880	0.8038	0.9445	0.31
DeiT-Small	0.8411	0.7682	0.7566	0.9382	1.57
ViT-Small	0.8694	0.7904	0.7960	0.9432	0.13
SSL					
SimCLR + ResNet18	0.3501	0.3721	0.3902	0.6310	3.49
SimCLR + ResNet50	0.2773	0.3443	0.3567	0.6348	9.86

The comparative results demonstrate a significant performance gap between self-supervised SimCLR pretraining and conventional supervised transfer learning for acne severity grading. All supervised models, across both CNN and ViT families, achieve consistently high agreement with dermatologist labels, with QWK scores ranging from 0.7616 to 0.8533, indicating strong ordinal consistency and clinical relevance. ViT models, particularly Swin Transformer, achieved 0.8714 in QWK, exhibit the best overall performance, suggesting an advantage in modeling global context and fine-grained visual patterns present in dermatological images.

In contrast, SimCLR-based models perform substantially worse across all evaluation metrics. Both SimCLR + ResNet18 and SimCLR + ResNet50 yield very low QWK scores of 0.3501 and 0.2773, respectively, along with near-random Balanced Accuracy and Macro-F1 values. These results indicate poor feature separability and an inability of contrastive pretraining to capture the subtle and localized cues required for ordinal acne severity assessment. Notably, increasing backbone capacity from ResNet18 to ResNet50 does not improve performance, suggesting that the primary limitation lies in the learned representation rather than model size or inference capability.

The uniformly low Macro-AUROC values observed for SimCLR further imply weak class discrimination, reinforcing that generic instance-level contrastive objectives are not well suited to dermatology tasks, where intra-class visual variation is limited and clinically meaningful differences are often subtle and localized. Overall, these findings suggest that standard SimCLR pretraining does not provide clinically meaningful representations for acne severity grading. While negative, this result is informative and underscores the need for domain-adapted self-supervised objectives, lesion-aware augmentation, or alternative pretraining paradigms that better align with the visual characteristics of medical skin images.

3.3. Evaluation Under Label Scarcity

The second experiment investigates how the performance of SimCLR changes when the amount of labeled data is progressively reduced, focusing on how each model behaves relative to its fully supervised performance at 100% labels. In theory, self-supervised pretraining should mitigate the impact of label scarcity by providing strong initialization, allowing downstream fine-tuning to remain stable even when only a small fraction of labels is available. However, our results show that SimCLR exhibits only minimal

degradation between the 100% labels setting and the 1–25% label settings, suggesting that model performance is relatively insensitive to annotation volume. Importantly, this stability does not arise from particularly strong representations but rather, overall performance of SimCLR remains consistently low across all label budgets. The small performance gap among the 1%, 5%, 10%, 25%, and 100% labels settings, therefore, indicates that the limiting factor is representation quality rather than label quantity. This finding implies that if a stronger or domain-adapted self-supervised feature extractor were available, a similar pattern of label-insensitive fine-tuning could allow the model to achieve clinically competitive accuracy, even with limited labeled data. Thus, although SimCLR does not yield high performance for this task, the observed robustness across label budgets highlights the potential value of improved SSL methods specifically designed for dermatological imagery.

Across all label budgets, from 1% to 25%, and up to 100%, SimCLR shows only marginal differences in performance. Metrics such as QWK, balanced accuracy, and Macro-F1 remain clustered within a narrow range, indicating that fine-tuning is surprisingly stable even when labels are extremely limited. The detailed quantitative results across different label budgets are summarized in Table 4. At first glance, such invariance might appear beneficial, but the underlying cause is more concerning. The overall performance of SimCLR remains consistently low, regardless of label availability. In other words, the primary bottleneck is not labeled quantity but the quality of the learned representations.

Table 4. Comparative performance of SimCLR models across label budget conditions.

Model	Label	QWK	Accuracy	Macro-F1	Macro-AUROC	Inference (ms)
SimCLR + ResNet18	0%	0.3225	0.3061	0.3028	0.5669	3.49
	1%	0.3008	0.3095	0.3257	0.5577	3.49
	5%	0.3519	0.3226	0.3317	0.5683	3.49
	10%	0.2888	0.3053	0.3094	0.6124	3.49
	25%	0.4116	0.3756	0.3892	0.6279	3.49
	100%	0.3501	0.3721	0.3902	0.6310	3.49
SimCLR + ResNet50	0%	0.0964	0.2546	0.2213	0.4900	9.86
	1%	0.0772	0.2418	0.2152	0.5253	9.86
	5%	0.2142	0.3010	0.2961	0.6376	9.86
	10%	0.2751	0.3170	0.3280	0.6037	9.86
	25%	0.2186	0.3578	0.3770	0.6346	9.86
	100%	0.2773	0.3443	0.3567	0.6348	9.86

The PCA visualizations of SimCLR embeddings using both ResNet18 and ResNet50 backbones reveal limited severity-aware structure in the learned feature space. Figure 3a and Figure 3b show the 2D PCA projections of SimCLR feature representations using ResNet18 and ResNet50 backbones, respectively, where Levels 0–3 correspond to mild, moderate, severe, and very severe acne. For SimCLR + ResNet18, the 2D PCA projection explains 87.0% of the total variance (PC1: 83.0%, PC2: 4.0%), while the 3D projection increases this slightly to 88.4%, as shown in Figure 4a. The dominance of PC1 indicates that the learned representations capture a strong underlying feature direction aligned with acne severity. As shown in Figure 3a, Level 0 and Level 3 samples form relatively distinct regions along PC1, suggesting clearer separation between the extreme severity cases. In contrast, Level 1 and Level 2 samples show substantial overlap, which is consistent with the visual and clinical similarity of intermediate severity cases. Notably, the samples exhibit an ordinal progression from Level 0 to Level 3 along the principal axis, consistent with clinical severity ordering.

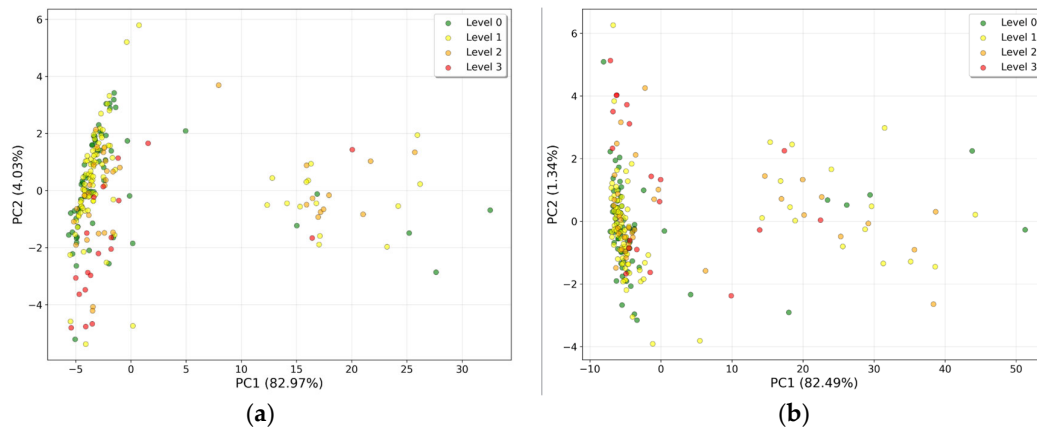


Figure 3. 2D PCA visualization of SimCLR feature representations: (a) ResNet-18 and (b) ResNet-50.

For SimCLR + ResNet50, the variance explained by the 2D PCA projection is slightly lower at 83.8% (PC1: 82.5%, PC2: 1.3%), increasing to 84.6% in 3D, as shown in Figure 4b. The feature space exhibits an even stronger concentration of variance along PC1, with higher components contributing only marginally. As shown in Figure 3b, although this indicates highly aligned representations, the severity levels, particularly Levels 1 and 2, still overlap substantially. Compared to ResNet18, ResNet50 shows a more compressed variance distribution across principal components, implying that increased representational capacity does not directly yield improved low-dimensional separability. Overall, both backbones capture clinically meaningful severity ordering, with clearer separation at severity extremes and overlap among moderate cases, while adding a third principal component achieved only marginal representational gains.

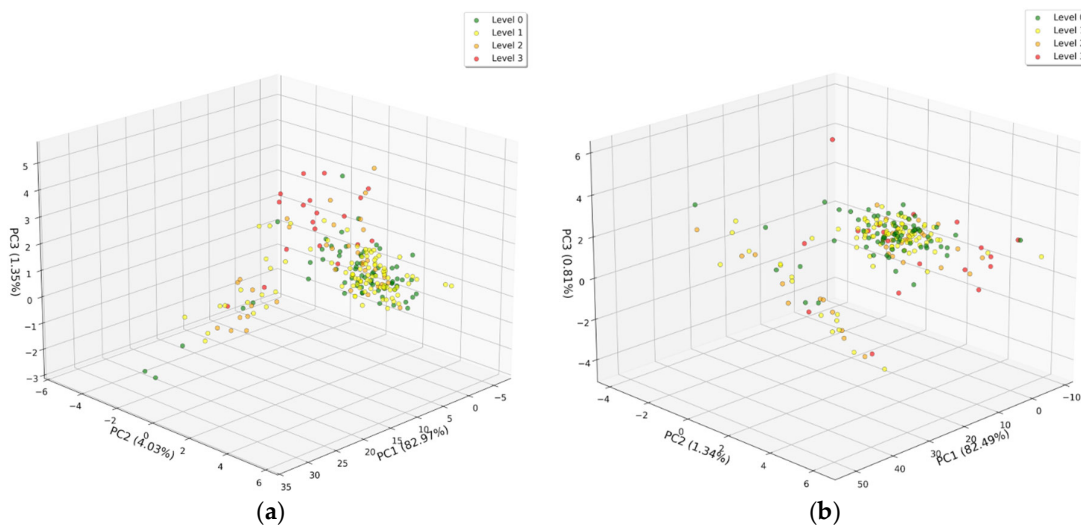


Figure 4. 3D PCA visualization of SimCLR feature representations: (a) ResNet-18 and (b) ResNet-50.

The t-SNE visualizations provide complementary insights into the local structure of the learned representations and further support the severity-aware trends observed in the PCA analysis. As shown in Figure 5a,b, the 2D t-SNE projections of SimCLR embeddings obtained with ResNet18 and ResNet50 backbones reveal the emergence of localized, severity-consistent neighborhoods. For SimCLR + ResNet18, the embeddings form discernible clusters corresponding to the four acne severity levels. Level 0 and Level 3 samples

exhibit the most compact and well-separated clusters, suggesting strong discriminative features for extreme severity cases. In contrast, Level 1 and Level 2 samples show partial overlap and transitional regions between clusters, reflecting the visual ambiguity and gradual progression typical of moderate acne severity.

SimCLR + ResNet50 exhibits a similar clustering structure, with slightly tighter and smoother local neighborhoods, suggesting more aligned feature representations due to the deeper backbone. However, overlap between the intermediate severity levels persists, and the global arrangement of clusters follows an approximate ordinal progression from Level 0 to Level 3 rather than forming strictly isolated groups. This behavior indicates that the learned representations encode acne severity as a continuum rather than as independent categorical classes. Overall, the t-SNE results highlight clear separation at the severity extremes and clinically meaningful transition zones among moderate cases, reinforcing the conclusion that SimCLR captures relevant disease progression patterns while still struggling to distinguish fine-grained boundaries between adjacent severity levels.

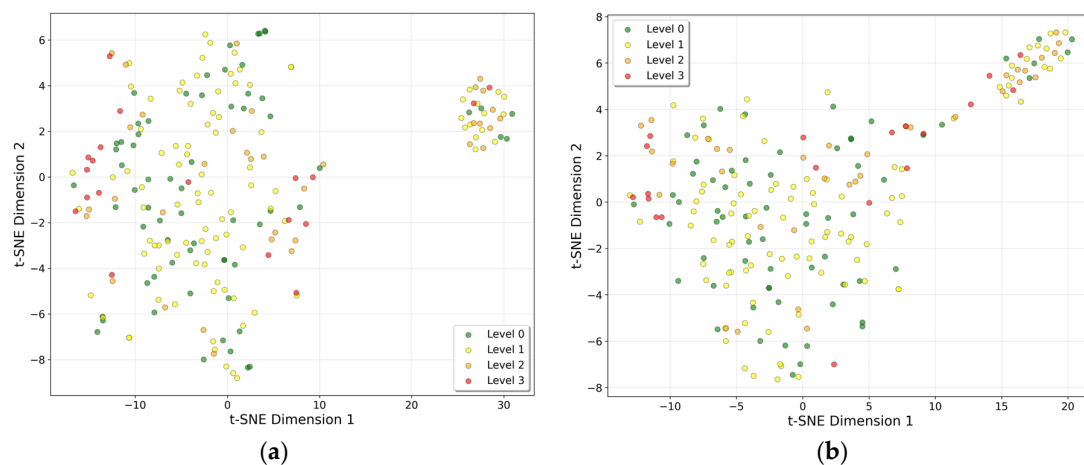


Figure 5. 2D t-SNE Visualization of SimCLR with (a) ResNet-18 and (b) ResNet-50.

3.4. Fine-Tuning the SimCLR Model

The final experiment investigates whether refining the contrastive learning formulation, specifically by adjusting the temperature parameter in the SimCLR loss, can improve representation quality and downstream acne severity grading performance. The temperature parameter controls the softness of the contrastive softmax distribution and directly affects the degree of separation between positive and negative sample pairs during self-supervised pretraining. To evaluate its impact, SimCLR models are pretrained using multiple temperature values ($\tau = 0.5, 0.7, 1.0, 1.2$, and 1.5) and subsequently fine-tuned end-to-end on the supervised acne severity classification task. This experiment aims to assess whether suboptimal temperature selection contributes to the poor performance observed in previous SimCLR evaluations and whether careful tuning can narrow the gap with supervised transfer learning baselines.

The comparative results across different temperature settings are summarized in Table 5. The results indicate that temperature tuning has a measurable but limited effect on SimCLR performance. For both ResNet18 and ResNet50 backbones, moderate temperature values, including $\tau = 0.5, 0.7$ and $\tau = 1.2$, consistently outperform the default setting of $\tau = 1.0$, achieving improvements in QWK, Macro-F1, and Macro-AUROC. The best-performing configurations achieve QWK scores of 0.4548 for SimCLR + ResNet18 at $\tau = 1.5$ and 0.3881 for SimCLR + ResNet50 $\tau = 0.5$. Although these results represent relative gains over the baseline configuration, they remain substantially lower than those achieved by supervised models.

Table 5. Comparative performance of SimCLR models under different temperature settings.

Model	τ	QWK	Accuracy	Macro-F1	Macro-AUROC	Inference (ms)
SimCLR + ResNet18	0.5	0.4055	0.4018	0.4012	0.6657	3.23
	0.7	0.4258	0.3835	0.3901	0.6389	3.30
	1.0	0.3501	0.3721	0.3902	0.6310	3.49
	1.2	0.3520	0.3952	0.4060	0.6681	3.29
	1.5	0.4548	0.4223	0.4299	0.6418	3.30
SimCLR + ResNet50	0.5	0.3881	0.4254	0.4337	0.6794	9.66
	0.7	0.2343	0.3661	0.3727	0.6447	9.86
	1.0	0.2773	0.3443	0.3567	0.6348	9.86
	1.2	0.3772	0.3725	0.3893	0.6435	9.96
	1.5	0.3255	0.3554	0.3687	0.6464	11.02

From a class-balanced perspective, the Macro-F1 scores remain in the range of 0.35–0.43, indicating limited per-class consistency even after temperature tuning. Similarly, the Macro-AUROC values range from 0.63 to 0.68 across all settings, reflecting only moderate discriminative ability across acne severity levels. These results suggest that although adjusting the temperature can modestly improve representation quality, it does not substantially enhance the ability of model to separate fine-grained and ordinal severity grades in a clinically meaningful feature.

Notably, higher temperature values (e.g., $\tau = 1.5$) do not yield uniform gains across backbones, with improvements observed for ResNet18 but not consistently for ResNet50, indicating sensitivity to backbone capacity. Conversely, very low temperature settings fail to produce robust improvements in Macro-F1 or Macro-AUROC, potentially due to over-sharpening effects that amplify noise and intra-class variability in dermatological images where visual differences are subtle. Overall, despite modest gains over the baseline configuration, SimCLR-based models remain substantially inferior to supervised counterparts in both class-balanced performance and discriminative ability.

Figure 6 further contextualizes these findings by visualizing the class-wise error distribution of the best-performing SimCLR configuration under the full-label setting. The confusion matrix reveals substantial misclassification between adjacent acne severity levels, particularly between mild and moderate classes. Only 42.42% of mild cases are correctly predicted, while 50.0% are misclassified as moderate. Similarly, 45.3% of moderate cases are correctly predicted, while 40.00% are misclassified as mild. Although the very severe class exhibits a higher recall at 52.63%, it is still confused with the neighboring severe class at 21.05%. In contrast, the severe class shows the lowest recall at 34.29% and is frequently confused with moderate at 28.57% and very severe at 17.14%, indicating limited ordinal discrimination. This pattern helps explain why improvements in aggregate metrics from temperature tuning do not translate into clinically reliable severity separation. Moreover, the confusion matrix exposes a tendency toward majority-class predictions and reduced sensitivity for intermediate severity levels, which are both clinically important and visually subtle.

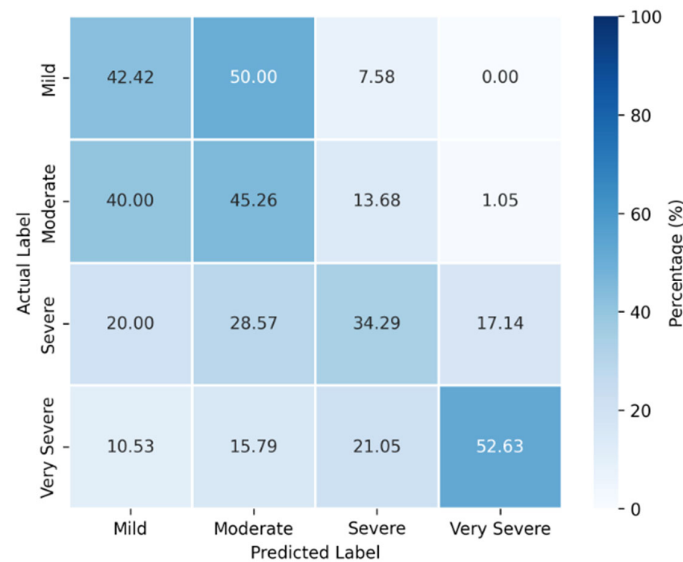
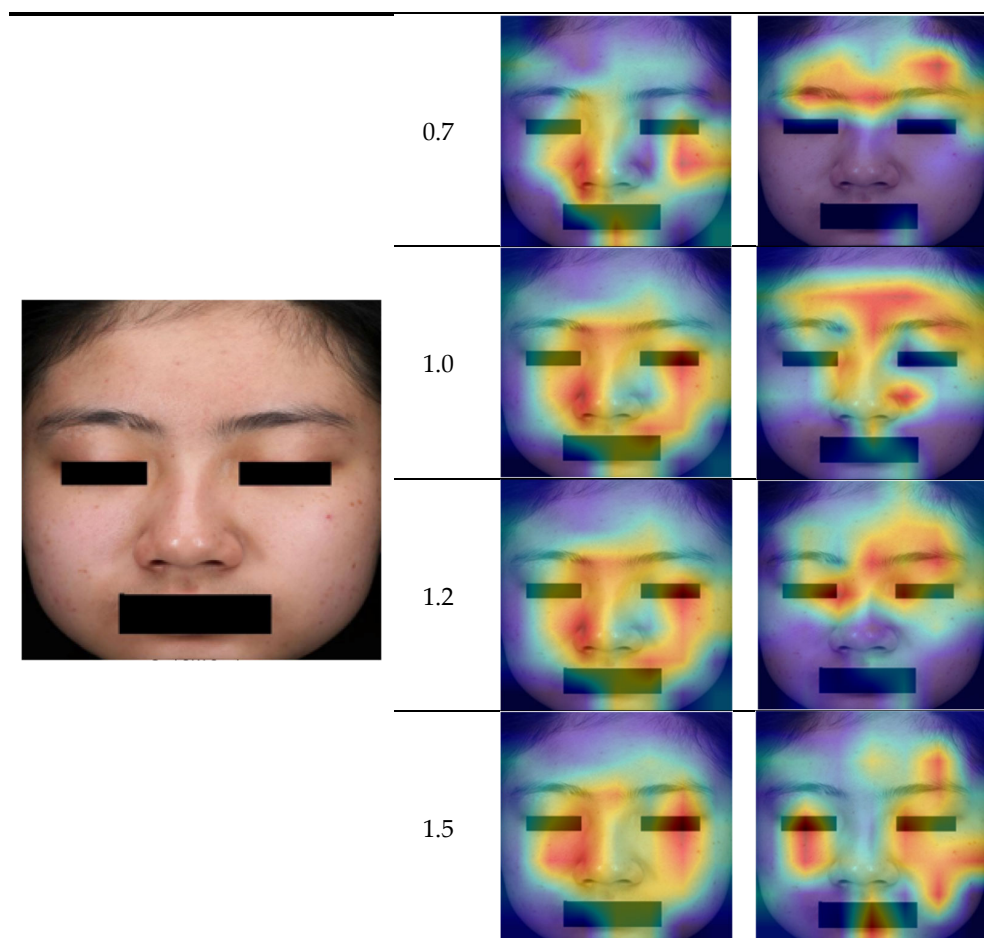


Figure 6. Row-normalized confusion matrix of the best-performing SimCLR model.

As shown in Table 6, moderate temperature settings result in Grad-CAM visualizations that are relatively more concentrated over acne-prone facial regions, such as the cheek and jawline. In contrast, higher temperature values result in more diffuse attention patterns that extend to non-diagnostic regions, including hair and background areas. Clinically, acne lesions may be distributed across multiple facial regions. However, the Grad-CAM results indicate that the model often concentrates its attention on a subset of representative acne-prone areas rather than uniformly covering all affected regions. This behavior suggests that the model may rely on salient regions that are sufficient for severity discrimination, although it does not provide quantitative validation against dermatologist-annotated lesion regions. Quantitative attention evaluation is an important direction for future work. This qualitative trend is consistently observed across both ResNet18 and ResNet50 backbones, indicating limited improvement in spatial feature localization with increased model capacity. Notably, increasing backbone capacity from ResNet18 to ResNet50 does not yield systematic improvements, reinforcing the conclusion that representation quality is constrained primarily by the contrastive objective rather than model expressiveness.

Table 6. Grad-CAM visualization comparison of ResNet-18 and ResNet-50 models under different Temperature settings.

Original Image	τ	Grad-Cam ResNet-18	Grad-Cam ResNet-50
	0.5		
	0.7		



Overall, while temperature is confirmed to be a meaningful hyperparameter for dermatology-focused SSL, its optimization alone is insufficient to achieve clinically relevant ordinal representations. These findings suggest that closing the performance gap will likely require more fundamental changes, such as lesion-aware data augmentations, spatially localized contrastive objectives, or hybrid self-supervised and supervised frameworks tailored to fine-grained medical image analysis.

4. Discussion

This research presents a comprehensive evaluation of self-supervised learning for acne severity grading and provides critical insights into its applicability and limitations within dermatological image analysis. Through a series of controlled experiments, this research examined whether contrastive self-supervised pretraining using SimCLR can provide meaningful advantages over supervised transfer learning, particularly under label-scarce conditions, and whether modifications to the contrastive formulation could improve downstream performance. Across all experimental settings, supervised models consistently outperformed SimCLR-based contrastive SSL approaches by a substantial margin, highlighting the continued effectiveness of supervised transfer learning and the current limitations of generic contrastive self-supervision for medical image classification with subtle visual differences. In addition, although no explicit class imbalance mitigation techniques were applied, the impact of class imbalance was examined using class-balanced evaluation metrics and confusion matrices. As shown in Figure 6, all severity classes, including the minority severe class, were detected and were not collapsed into majority categories. However, the confusion matrix reveals class-dependent performance

variations, including reduced recall for intermediate severity levels and a tendency toward majority-class predictions. These observations suggest that class imbalance influences the error distribution even when balanced metrics are reported. While this analysis improves transparency, explicitly addressing class imbalance through data rebalancing or cost-sensitive learning remains important future work.

The results demonstrate that supervised CNN and ViT models achieve strong and consistent performance across all evaluation metrics, including QWK, Balanced Accuracy, Macro-F1, and Macro-AUROC. Vision transformer-based architectures, such as Swin Transformer and DeiT, exhibit particularly high ordinal agreement with expert annotations. Future work should further investigate Transformer-based architectures for acne severity grading, as their self-attention mechanisms can model global contextual relationships and lesion distribution patterns that are clinically relevant for dermatological assessment. Although not evaluated in this study, such models may be better aligned with the spatial and ordinal nature of acne severity. This observation suggests that the ability to model global context while preserving detailed spatial information is beneficial for acne severity grading. These findings also indicate that supervised pretraining on large-scale natural image datasets provides transferable representations that align well with dermatological severity assessment, despite domain differences between natural and medical images.

In contrast, SimCLR-based self-supervised models consistently exhibit weak performance, even after fine-tuning. Low ordinal agreement and near-random class discrimination indicate that the learned representations lack clinically meaningful structure. Increasing backbone capacity from ResNet18 to ResNet50 does not lead to systematic improvements, suggesting that representation quality is constrained primarily by the contrastive learning objective rather than model expressiveness. The instance-level invariance encouraged by SimCLR may suppress subtle but clinically important variations in lesion appearance, density, and spatial distribution, which are critical for accurate severity grading. Importantly, this performance gap likely reflects a mismatch between the generic SimCLR-style contrastive objective and dermatological visual characteristics, rather than an inherent limitation of self-supervised learning in general.

The label-scarcity experiments further reinforce these observations. Although self-supervised learning is often proposed as a solution for reducing reliance on annotated data, SimCLR does not demonstrate a clear advantage over supervised transfer learning when the amount of labeled data is reduced. Supervised models degrade more gracefully under limited supervision, indicating stronger inductive biases for dermatological image interpretation. This finding challenges the assumption that generic self-supervised pretraining alone can compensate for annotation scarcity in medical imaging tasks characterized by subtle inter-class differences and ordinal relationships.

Adjusting the temperature parameter of the SimCLR contrastive loss yields measurable but limited improvements in downstream performance, confirming that representation quality is sensitive to contrastive formulation. Moderate temperature values consistently outperform the default setting, while higher temperatures degrade performance due to insufficient separation between feature clusters. Overall, the observed gains are modest and do not change the overall performance trend. However, even the best-performing temperature configurations fail to approach supervised baselines, indicating that hyperparameter tuning alone is insufficient to address the fundamental mismatch between generic contrastive objectives and the requirements of acne severity grading.

Recent self-supervised frameworks such as DINO, which rely on self-distillation and teacher-student training paradigms, have demonstrated promising representation learning capabilities, particularly when paired with vision transformer architectures. Such approaches may better capture global structural relationships without relying on explicit

negative pairs. However, due to computational resource limitations and the substantial training cost associated with large-scale transformer-based self-supervised pretraining, a direct evaluation of DINO was not feasible within the scope of this research. Therefore, the limitations observed here should be interpreted as specific to the evaluated SimCLR-based contrastive self-supervised setting rather than representative of self-supervised learning methods in general (e.g., DINO or other teacher–student frameworks), which were not evaluated in this research. Consequently, while DINO is a promising alternative, its effectiveness for fine-grained acne severity grading, where localized lesion features and ordinal severity relationships are critical, remains to be systematically validated.

Overall, the findings of this research underscore the importance of domain-specific adaptation in self-supervised learning for medical image analysis. Dermatological images present unique challenges, including high intra-class similarity, subtle inter-class variation, and strong dependence on localized visual cues. Generic self-supervised objectives designed for natural images may not capture these characteristics. Effective representation learning for acne severity grading is likely to require lesion-aware augmentations, spatially localized learning objectives, or hybrid frameworks that integrate self-supervision with weak or partial supervision and ordinal constraints. While the self-supervised approaches evaluated in this research do not yet achieve clinically meaningful performance, the systematic negative findings reported here provide valuable guidance for future research and highlight key directions for developing more effective self-supervised models in dermatology.

5. Conclusions

This research presents a systematic evaluation of SSL for acne severity grading by benchmarking SimCLR against a diverse set of supervised deep learning models under controlled experimental conditions. Across full-data training, label-scarce settings, and temperature-tuned contrastive learning experiments, supervised transfer learning consistently demonstrated superior performance. These results indicate that supervised models are more effective at capturing the subtle visual cues required for clinically meaningful acne assessment. In contrast, standard SimCLR pretraining yielded only limited performance gains and remains below supervised baselines, even after fine-tuning.

Future research should therefore explore domain-aware representation learning strategies that explicitly incorporate dermatological knowledge into the training process, including Transformer-based architectures such as ViT and Swin Transformer. This may include pretraining on large-scale skin-disease datasets; designing contrastive objectives that emphasize lesion boundaries and inflammatory patterns; integrating clinical metadata such as lesion type, anatomical location, or severity information; or adopting weakly supervised and multi-task learning frameworks that jointly predict acne severity alongside auxiliary dermatological attributes. Such domain-aligned objectives may help bridge the performance gap observed with generic contrastive self-supervised learning and provide deeper insights into how self-supervised methods can be effectively adapted for fine-grained dermatological image analysis.

In addition, qualitative analysis suggests that the current models focus on a limited subset of salient acne-prone regions rather than comprehensively attending to all affected facial areas. While such behavior may be sufficient for coarse severity discrimination, it limits interpretability and clinical trust. Future work should incorporate quantitative attention evaluation, ideally using dermatologist-annotated lesion regions, to assess whether the learned attention patterns align with clinical assessment criteria and to guide the development of more spatially faithful representation learning strategies.

Although SSL offers a promising solution for reducing reliance on annotated data, these findings suggest that generic contrastive learning objectives are not well aligned

with the subtle, localized, and severity-dependent patterns in dermatological imagery. While temperature adjustment influenced SimCLR performance, the observed improvements were modest and insufficient to close the performance gap. Collectively, these results highlight the limitations of generic contrastive SSL for acne severity grading and motivate further investigation into domain-adapted self-supervised strategies. Moreover, this research provides negative evidence that can inform future methodological development and help prevent overgeneralization of SSL effectiveness in clinical image analysis.

Author Contributions: Conceptualization, K.S., N.V. and T.T.; methodology, K.S. and N.V.; software, N.V.; validation, K.S. and T.T.; formal analysis, K.S. and N.V.; investigation, K.S.; resources, K.S. and T.T.; data curation, N.V.; writing—original draft preparation, K.S. and N.V.; project administration, T.T.; writing—review and editing, K.S. and T.T.; visualization, K.S.; supervision, T.T. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by Thammasat University Research Fund, Contract No. TUFT 50/2567.

Institutional Review Board Statement: The study was approved by the Human Research Ethics Committee of Thammasat University (Science), Thailand (COA No. 057/2567, Project No. 67SC057, approved on 4 June 2024).

Informed Consent Statement: Not applicable. This study used publicly available and no direct patient recruitment or interaction was conducted by the authors.

Data Availability Statement: The raw data supporting the findings of this study are available online. Further inquiries should be directed to the corresponding author.

Acknowledgments: This research was supported by the Thammasat University Research Unit in Data Innovation and Artificial Intelligence. During the preparation of this study, the authors used ChatGPT 5.2 to check the grammar. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ghodsi, S.Z.; Orawa, H.; Zouboulis, C.C. Prevalence, Severity, and Severity Risk Factors of Acne in High School Pupils: A Community-Based Study. *J. Investig. Dermatol.* **2009**, *129*, 2136–2141. <https://doi.org/10.1038/jid.2009.47>.
2. Pagliarello, C.; Di Pietro, C.; Tabolli, S. A Comprehensive Health Impact Assessment and Determinants of Quality of Life, Health and Psychological Status in Acne Patients. *G. Ital. Dermatol. Venereol.* **2015**, *150*, 303–308.
3. Dumont, S.; Lorthé, E.; Loizeau, A.; Richard, V.; Nehme, M.; Posfay-Barbe, K.M.; Barbe, R.P.; Trellu, L.T.; Stringhini, S.; Guessous, I.; et al. Acne-Related Quality of Life and Mental Health among Adolescents: A Cross-Sectional Analysis. *Clin. Exp. Dermatol.* **2025**, *50*, 795–803. <https://doi.org/10.1093/ced/llae453>.
4. Adityan, B.; Kumari, R.; Thappa, D.M. Scoring Systems in Acne Vulgaris. *Indian J. Dermatol. Venereol. Leprol.* **2009**, *75*, 323–326. <https://doi.org/10.4103/0378-6323.51258>.
5. Koo, J. The Psychosocial Impact of Acne: Patients' Perceptions. *J. Am. Acad. Dermatol.* **1995**, *32*, S26–S30. [https://doi.org/10.1016/0190-9622\(95\)90417-4](https://doi.org/10.1016/0190-9622(95)90417-4).
6. Hayashi, N.; Akamatsu, H.; Kawashima, M. Establishment of Grading Criteria for Acne Severity. *J. Dermatol.* **2008**, *35*, 255–260. <https://doi.org/10.1111/j.1346-8138.2008.00462.x>.
7. Doshi, A.; Zaheer, A.; Stiller, M.J. A Comparison of Current Acne Grading Systems and Proposal of a Novel System. *Int. J. Dermatol.* **1997**, *36*, 416–418. <https://doi.org/10.1046/j.1365-4362.1997.00099.x>.
8. Tanantong, T.; Chalarak, N.; Pandecha, P.; Tanantong, K.; Srihiranon, K. Mobile-Based Deep Learning Framework for Classifying Common Skin Diseases in Thailand. *ICIC Express Lett. Part B Appl.* **2024**, *15*, 495–503. <https://doi.org/10.24507/icicelb.15.05.495>.

9. Tanantong, T.; Chalarak, N.; Jirattisak, S.; Tanantong, K.; Srijaranon, K. A Study on Area Assessment of Psoriasis Lesions Using Image Augmentation and Deep Learning: Addressing the Lack of Thai Skin Disease Data. *J. Curr. Sci. Technol.* **2025**, *15*, 119. <https://doi.org/10.59796/jcst.V15N3.2025.119>.
10. Qi, Y.; Liu, Y.; Luo, J. Recent Application of Raman Spectroscopy in Tumor Diagnosis: From Conventional Methods to Artificial Intelligence Fusion. *Photonix* **2023**, *4*, 1. <https://doi.org/10.1186/s43074-023-00098-0>.
11. Hemelings, R.; Elen, B.; Schuster, A.K.; Blaschko, M.B.; Barbosa-Breda, J.; Hujanen, P.; Junglas, A.; Nickels, S.; White, A.; Pfeiffer, N.; et al. A Generalizable Deep Learning Regression Model for Automated Glaucoma Screening from Fundus Images. *NPJ Digit. Med.* **2023**, *6*, 112. <https://doi.org/10.1038/s41746-023-00857-0>.
12. Baik, S.M.; Hong, K.S.; Park, D.J. Deep Learning Approach for Early Prediction of COVID-19 Mortality Using Chest X-ray and Electronic Health Records. *BMC Bioinform.* **2023**, *24*, 190. <https://doi.org/10.1186/s12859-023-05321-0>.
13. Alzahrani, S.; Al-Bander, B.; Al-Nuaimy, W. Attention Mechanism Guided Deep Regression Model for Acne Severity Grading. *Computers* **2022**, *11*, 3. <https://doi.org/10.3390/computers11030031>.
14. Wang, J.; Wang, C.; Wang, Z.; Hounye, A.H.; Li, Z.; Kong, M.; Hou, M.; Zhang, J.; Qi, M. A Novel Automatic Acne Detection and Severity Quantification Scheme Using Deep Learning. *Biomed. Signal Process. Control* **2023**, *84*, 104803. <https://doi.org/10.1016/j.bspc.2023.104803>.
15. Ericsson, L.; Gouk, H.; Loy, C.C.; Hospedales, T.M. Self-Supervised Representation Learning: Introduction, Advances, and Challenges. *IEEE Signal Process. Mag.* **2022**, *39*, 42–62. <https://doi.org/10.1109/MSP.2021.3134634>.
16. Liu, S.; Deng, W. Very Deep Convolutional Neural Network Based Image Classification Using Small Training Sample Size. In Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 730–734. <https://doi.org/10.1109/ACPR.2015.7486599>.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
18. Tan, M.; Le, Q.V. EfficientNetV2: Smaller Models and Faster Training. *arXiv* **2021**, arXiv:2104.00298. <https://doi.org/10.48550/arXiv.2104.00298>.
19. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. ConvNeXt V2: Co-Designing and Scaling ConvNets with Masked Autoencoders. *arXiv* **2023**, arXiv:2301.00808. <https://doi.org/10.48550/arXiv.2301.00808>.
20. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2016**, arXiv:1608.06993. <https://doi.org/10.48550/arXiv.1608.06993>.
21. Xu, J.; Pan, Y.; Pan, X.; Hoi, S.; Yi, Z.; Xu, Z. RegNet: Self-Regulated Network for Image Classification. *arXiv* **2021**, arXiv:2101.00590. <https://doi.org/10.48550/arXiv.2101.00590>.
22. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *arXiv* **2018**, arXiv:1801.04381. <https://doi.org/10.48550/arXiv.1801.04381>.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. <https://doi.org/10.48550/arXiv.2010.11929>.
24. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *arXiv* **2021**, arXiv:2103.14030. <https://doi.org/10.48550/arXiv.2103.14030>.
25. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training Data-Efficient Image Transformers and Distillation through Attention. *arXiv* **2020**, arXiv:2012.12877. <https://doi.org/10.48550/arXiv.2012.12877>.
26. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv* **2020**, arXiv:2002.05709. <https://doi.org/10.48550/arXiv.2002.05709>.
27. Neha, Y.; Alfayeed, S.M.; Khamparia, A.; Pandey, B.; Thanh, D.N.H.; Pande, S. HSV Model-Based Segmentation Driven Facial Acne Detection Using Deep Learning. *Expert Syst.* **2022**, *39*, e12760. <https://doi.org/10.1111/exsy.12760>.
28. Zhao, T.; Zhang, H.; Spoelstra, J. A Computer Vision Application for Assessing Facial Acne Severity from Selfie Images. *arXiv* **2019**, arXiv:1907.07901. <https://doi.org/10.48550/arXiv.1907.07901>.
29. Zein, H.; Chantaf, S.; Fournier, R.; Nait-Ali, A. GANs for Anonymous Acneic Face Dataset Generation. *arXiv* **2022**, arXiv:2211.04214. <https://doi.org/10.48550/arXiv.2211.04214>.
30. Naidu, K.; Kareppa, O.; Menon, S.; Bhole, C.; Poojary, S. Dermato: A Deep Learning Based Application for Acne Subtype and Severity Detection. In Proceedings of the 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 14–16 March 2023; pp. 569–574. <https://doi.org/10.1109/ICIDCA56705.2023.10100165>.

31. Rashataprucksa, K.; Chuangchaichatchavarn, C.; Triukose, S.; Nitinawarat, S.; Pongprutthipan, M.; Piromsopa, K. Acne Detection with Deep Neural Networks. In Proceedings of the 2020 2nd International Conference on Image Processing and Machine Vision, Bangkok, Thailand, 5–7 August 2020; pp. 53–56. <https://doi.org/10.1145/3421558.3421566>.
32. Suriani, N.S.; Tarmizi, S.S.A.; Mohd, M.N.H.; Shah, S.M. Acne Severity Classification on Mobile Devices Using Lightweight Deep Learning Approach. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 680–687. <https://doi.org/10.14569/IJACSA.2024.0150668>.
33. Zein, H.; Chantaf, S.; Fournier, R.; Nait-Ali, A. Generative Adversarial Networks for Anonymous Acneic Face Dataset Generation. *PLoS ONE* **2024**, *19*, e0297958. <https://doi.org/10.1371/journal.pone.0297958>.
34. Lin, Y.; Jiang, J.; Chen, D.; Ma, Z.; Guan, Y.; Liu, X.; You, H.; Yang, J. DED: Diagnostic Evidence Distillation for Acne Severity Grading on Face Images. *Expert Syst. Appl.* **2023**, *228*, 120312. <https://doi.org/10.1016/j.eswa.2023.120312>.
35. Lin, Y.; Guan, Y.; Ma, Z.; You, H.; Cheng, X.; Jiang, J. An Acne Grading Framework on Face Images via Skin Attention and SFNet. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 9–12 December 2021; pp. 2407–2414. <https://doi.org/10.1109/BIBM52615.2021.9669431>.
36. Wu, X.; Wen, N.; Liang, J.; Lai, Y.-K.; She, D.; Cheng, M.-M.; Yang, J. Joint Acne Image Grading and Counting via Label Distribution Learning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 10641–10650. <https://doi.org/10.1109/ICCV.2019.01074>.
37. Lin, Y.; Jiang, J.; Chen, D.; Ma, Z.; Guan, Y.; Liu, X.; You, H.; Yang, J.; Cheng, X. Acne Severity Grading on Face Images via Extraction and Guidance of Prior Knowledge. In Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 6–8 December 2022; pp. 1639–1643. <https://doi.org/10.1109/BIBM55620.2022.9995101>.
38. Junayed, M.S.; Islam, B.; Jeny, A.A.; Sadeghzadeh, A.; Biswas, T.; Shah, A.F.M.S. ScarNet: Development and Validation of a Novel Deep CNN Model for Acne Scar Classification. *IEEE Access* **2022**, *10*, 1245–1258. <https://doi.org/10.1109/ACCESS.2021.3138021>.
39. Min, K.; Lee, G.H.; Lee, S.W. ACNet: Mask-Aware Attention with Dynamic Context Enhancement for Robust Acne Detection. In Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, 17–20 October 2021; pp. 2724–2729. <https://doi.org/10.1109/SMC52423.2021.9659243>.
40. Huynh, Q.T.; Nguyen, P.H.; Le, H.X.; Ngo, L.T.; Trinh, N.-T.; Tran, M.T.-T.; Nguyen, H.T.; Vu, N.T.; Nguyen, A.T.; Suda, K.; et al. Automatic Acne Object Detection and Acne Severity Grading Using Smartphone Images and Artificial Intelligence. *Diagnostics* **2022**, *12*, 1879. <https://doi.org/10.3390/diagnostics12081879>.
41. Lin, Y.; Jiang, J.; Ma, Z.; Chen, D.; Guan, Y.; You, H.; Cheng, X.; Liu, B.; Luo, G. KIEGLFN: A Unified Acne Grading Framework on Face Images. *Comput. Methods Programs Biomed.* **2022**, *221*, 106911. <https://doi.org/10.1016/j.cmpb.2022.106911>.
42. Wei, X.; Zhang, L.; Zhang, J.; Wang, J.; Liu, W.; Li, J.; Jiang, X. Decoupled Sequential Detection Head for Accurate Acne Detection. *Knowl.-Based Syst.* **2024**, *284*, 111305. <https://doi.org/10.1016/j.knosys.2023.111305>.
43. Zhang, J.; Zhang, L.; Wang, J.; Wei, X.; Li, J.; Jiang, X.; Du, D. Learning High-Quality Proposals for Acne Detection. *arXiv* **2022**, arXiv:2207.03674. <https://doi.org/10.48550/arXiv.2207.03674>.
44. Wen, H.; Yu, W.; Wu, Y.; Zhao, J.; Liu, X.; Kuang, Z.; Fan, R. Acne Detection and Severity Evaluation with Interpretable Convolutional Neural Network Models. *Technol. Health Care* **2022**, *30*, S143–S153. <https://doi.org/10.3233/THC-228014>.
45. Zhang, H.; Ma, T. Acne Detection by Ensemble Neural Networks. *Sensors* **2022**, *22*, 6828. <https://doi.org/10.3390/s22186828>.
46. Traini, D.O.; Palmisano, G.; Guerriero, C.; Peris, K. Artificial Intelligence in the Assessment and Grading of Acne Vulgaris: A Systematic Review. *J. Pers. Med.* **2025**, *15*, 238. <https://doi.org/10.3390/jpm15060238>.
47. Zhang, J.; Zhang, L.; Wang, J.; Wei, X.; Li, J.; Jiang, X.; Du, D. SA-RPN: A Spatial Aware Region Proposal Network for Acne Detection. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 5439–5448. <https://doi.org/10.1109/JBHI.2023.3304727>.
48. Fleiss, J.L.; Cohen, J. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educ. Psychol. Meas.* **1973**, *33*, 613–619. <https://doi.org/10.1177/001316447303300309>.
49. Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. The Balanced Accuracy and Its Posterior Distribution. In Proceedings of the 20th International Conference on Pattern Recognition (ICPR 2010), Istanbul, Turkey, 23–26 August 2010; pp. 3121–3124. <https://doi.org/10.1109/ICPR.2010.764>.
50. Sokolova, M.; Lapalme, G. A Systematic Analysis of Performance Measures for Classification Tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>.
51. Hand, D.J.; Till, R.J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Mach. Learn.* **2001**, *45*, 171–186. <https://doi.org/10.1023/A:1010920819831>.

52. Sarwar, N.; Irshad, A.; Naith, Q.H.; DAlsufiani, K.; Almalki, F.A. Skin Lesion Segmentation Using Deep Learning Algorithm with Ant Colony Optimization. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 265. <https://doi.org/10.1186/s12911-024-02686-x>.
53. Abu Owida, H.; Abd El-Fattah, I.; Abuowaida, S.; Alshdaifat, N.; Mashagba, H.A.; Abd Aziz, A.B.; Alzoubi, A.; Larguech, S.; Al-Bawri, S.S. A Deep Learning-Based Dual-Branch Framework for Automated Skin Lesion Segmentation and Classification via Dermoscopic Images. *Sci. Rep.* **2025**, *15*, 37823. <https://doi.org/10.1038/s41598-025-21783-z>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.