cosmetics

MDPI

Article

Systematic Optimization of Transfer Learning for Acne Severity Classification Using Restricted, Imbalanced and Non-Public Facial Images: An Experimental Study

Taradon Khonsiri [1], Woottichai Nachaiwieng [1,2], Anon Paichitrojjana [3] and Pattaramon Vuttipittayamongkol [4,5,*]

[1] School of Health Science, Mae Fah Luang University, 333, Thasud, Chiang Rai 57100, Thailand; 6451811501@lamduan.mfu.ac.th (T.K.); woottichai.nac@mfu.ac.th (W.N.)
[2] Biomedical Technology Research Group for Vulnerable Populations, Mae Fah Luang University, 333, Thasud, Chiang Rai 57100, Thailand
[3] School of Anti-Aging and Regenerative Medicine, PS Tower 25Fl, Mae Fah Luang University, 36/87-88, Bangkok 10110, Thailand; anon.pai@mfu.ac.th
[4] Center of Excellence in AI and Emerging Technologies, Mae Fah Luang University, 333, Thasud, Chiang Rai 57100, Thailand
[5] School of Applied Digital Technology, Mae Fah Luang University, 333, Thasud, Chiang Rai 57100, Thailand
* Correspondence: pattaramon.vut@mfu.ac.th

Abstract

Acne vulgaris is a prevalent inflammatory skin condition that requires accurate severity assessment for effective management. As a step toward more objective and reproducible severity assessment, this study developed an optimized transfer learning-based convolutional neural network (CNN) framework for facial acne severity classification using a restricted, imbalanced, non-public facial image dataset. A total of 442 frontal facial images were collected under natural lighting conditions. Acne severity was graded by a board-certified dermatologist using the Investigator's Global Assessment (IGA) scale and categorized into three levels. The study systematically investigated model architecture selection, hyperparameter tuning, optimizer comparison, cross-validation, and class-imbalance handling through random oversampling, Synthetic Minority Over-sampling Technique (SMOTE), and Generative Adversarial Networks (GANs). The 5-fold cross-validation experiment supported the reliability of model performance across different data partitions, achieving an accuracy of 0.51. Resampling methods showed limited overall benefit, although some methods altered class-wise prediction patterns. To further examine model behavior, Gradient-weighted Class Activation Mapping (Grad-CAM) visualization was used to provide qualitative insight into the facial regions contributing to model predictions. Although the findings are limited by dataset size and diversity, the proposed framework suggests exploratory feasibility for automated acne severity assessment. Rather than serving as an immediately deployable clinical tool, this pipeline provides a preliminary baseline framework that requires further validation using larger, more diverse datasets, particularly to address subtle visual differences between acne severity classes.

Keywords: acne severity classification; natural-setting images; transfer learning; imbalanced data; model optimization

Check for updates

Academic Editor: Steven Paul Nisticò

Received: 18 May 2026
Revised: 23 June 2026
Accepted: 25 June 2026
Published: 29 June 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) license.

1. Introduction

Acne vulgaris is a prevalent chronic inflammatory skin disorder affecting mainly adolescents and young adults, with severity ranging from comedones to inflammatory lesions that may cause scarring and post-inflammatory hyperpigmentation [1]. Cutaneous manifestations may significantly affect psychological well-being and overall quality of life, underscoring the need for timely and appropriate clinical management [2,3]. Treatment is guided by severity and includes topical agents, systemic antibiotics, and oral isotretinoin [4]. However, limited access to dermatological care and the financial burden associated with medications and systemic therapies remain significant challenges [5,6]. These limitations have prompted growing interest in accessible, cost-effective technological solutions to support early assessment and treatment guidance [7]. Advances in machine learning, particularly convolutional neural networks (CNNs), have enabled automated dermatological image analysis [8]. In acne management, CNN-based systems can enhance diagnostic consistency and support personalized care by learning complex visual patterns directly from images, outperforming traditional handcrafted approaches [9,10].

Several studies have investigated deep learning approaches for automated acne severity assessment. CNN architectures, including VGG-based models, MobileNet, EfficientNet, and hybrid attention-based networks, have shown promising classification performance under controlled experimental conditions [11–13]. However, these studies often emphasize specific model designs or performance gains, while the combined effects of model configuration choices, hyperparameter settings, and class imbalance remain less explored [14,15]. This is important because acne datasets collected in real-world settings are often limited in size and unevenly distributed across severity levels, which may affect the reliability of model evaluation.

Some studies have incorporated attention mechanisms or capsule-based architectures to enhance feature discrimination [15]. Beyond model architecture, recent dermatological artificial intelligence (AI) research has also explored enriched image representations beyond standard color photographs. For example, Huang et al. (2025) [16] showed that spectrum-enhanced image conversion can improve dermatological lesion detection by revealing visual cues that may be less apparent in ordinary RGB images. These findings suggest that image representation remains an important challenge in dermatological AI, especially when subtle skin features must be identified from conventional facial photographs.

Class imbalance is another common challenge in acne datasets, as moderate-to-severe cases are typically underrepresented [11,12], potentially biasing model training and reducing sensitivity for clinically important classes. To mitigate this issue, previous studies have applied resampling strategies such as random oversampling (ROS) and the Synthetic Minority Over-sampling Technique (SMOTE) [17]. More recently, Generative Adversarial Networks (GANs) have been proposed to generate realistic synthetic medical images, demonstrating potential benefits in enhancing minority-class performance [18,19]. However, systematic comparisons between traditional resampling approaches and GAN-based augmentation in multi-class acne severity classification remain limited.

Furthermore, an important consideration for clinical adoption is model robustness and interpretability. In contrast, many studies emphasize that overall accuracy, balanced performance across precision, recall, specificity, and F1-score are essential for clinical relevance, particularly for moderate-to-severe acne, where misclassification may directly affect treatment decisions [20]. Specifically, the lack of explainability in deep learning models remains a major barrier to clinical trust and adoption [21]. Explainable AI techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM), have been increasingly used to visualize model attention and support clinical plausibility [22]; however, their application in acne severity classification has been relatively limited.

Despite these advances, evidence remains limited on acne severity classification under restricted, imbalanced, and real-world facial-image conditions. Existing studies have often focused on individual architectures or performance improvements in relatively controlled settings, whereas less attention has been given to how modelling decisions, including architecture selection, hyperparameter configuration, optimizer choice, and resampling strategy, interact within a unified experimental setting. This gap is clinically relevant because unreliable severity classification may affect treatment guidance and reduce confidence in automated acne assessment systems. Therefore, this study aimed to evaluate a transfer learning-based CNN framework that combines model optimization, class-imbalance assessment, and qualitative interpretability for multiclass acne severity classification using participant-submitted facial images.

The main contributions of this study are:

1. Application of transfer learning for multiclass acne severity classification using participant-submitted images acquired under real-world conditions.
2. Systematic comparison of CNN architectures, optimizers, and hyperparameter configurations.
3. Comparative evaluation of ROS, SMOTE, and GAN-based augmentation to examine the effect of resampling on class-imbalance handling under identical experimental settings.
4. Integration of Grad-CAM to support qualitative model interpretability.

2. Materials and Methods

This study collected smartphone facial images and obtained dermatologist annotations of acne severity based on the Investigator's Global Assessment (IGA) scale [14]. The images were preprocessed using face parsing, stratified data splitting, normalization, and data augmentation. Transfer learning-based convolutional neural networks were subsequently employed for acne severity classification. Model optimization included hyperparameter tuning, optimizer comparison, K-fold cross-validation, and class imbalance handling through multiple oversampling methods, including random oversampling, SMOTE, and GAN-based resampling. Finally, Grad-CAM was incorporated to improve model interpretability and provide visual explanations for acne severity prediction, as illustrated in Figure 1.

2.1. Tools

Experiments were conducted on an Apple Silicon M1-based system (8-core CPU, 8-core GPU, 16-core Neural Engine, 8 GB unified memory). Data preprocessing was performed in Python 3.10.14 using Google Colab, while model training and transfer learning were implemented in MATLAB (R2023a).

2.2. Study Design, Data Collection and Management

Frontal and bilateral side-view facial images were collected from each of the 524 participants using smartphones under guided conditions, with an approximate capture distance of 20 cm and sufficient lighting. Images had a resolution of approximately 1538 × 2051 pixels and were obtained either by researchers or self-uploaded by participants via a Google Form (Figure 1). Data collection was approved by the Human Research Ethics Committee of Mae Fah Luang University.

For data annotation, a board-certified dermatologist assessed acne severity using the frontal and bilateral side-view facial images of each participant according to the Investigator's Global Assessment (IGA) scale. The IGA scale classifies acne severity into five grades (0–4): 0 = clear skin/no acne, 1 = non-inflammatory lesions, 2 = mild, 3 = moderate, and 4 = severe, which detailed description provided in Table 1.

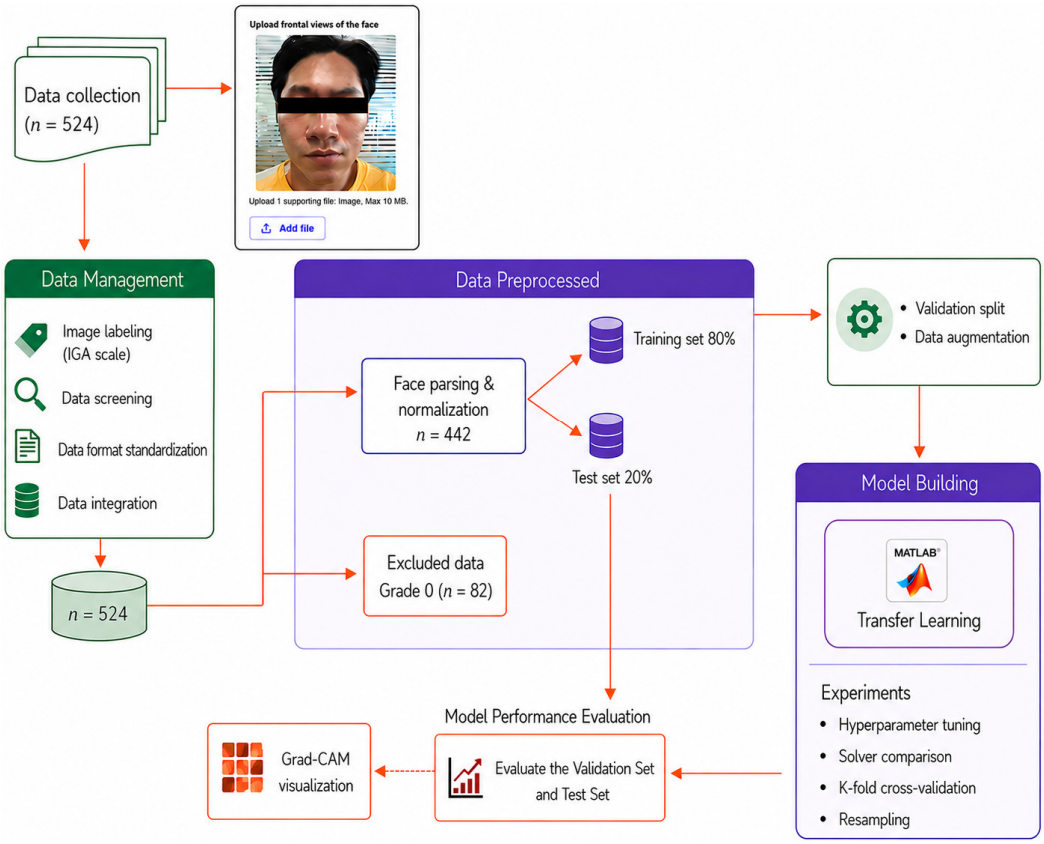


Figure 1. Overview of the proposed acne severity classification framework and data flow. *n* denotes the number of samples at each stage of the pipeline. To protect participant privacy, the eye region was masked during data management and was not used during model training.

Table 1. Acne Severity Grading System and Treatment-Based Class Label Definition [4].

Grade	Description	Treatment	Class Label
0	Clear skin with no inflammatory or non-inflammatory lesions	-	(Excluded *n* = 82)
1	Almost clear: rare non-inflammatory lesions with no more than one small inflammatory lesion	Topical product monotherapy: Retinoid (0.01–0.1%), Benzoyl peroxide (2.5–10%), Azelaic acid (20%)	1 (*n* = 106)
2	Mild severity: some non-inflammatory lesions with no more than a few inflammatory lesions (papules/pustules only, no nodular lesions)	Topical product combination therapy: Benzoyl peroxide (2.5–10%) + Clindamycin (1%) or Erythromycin (1–4%), Benzoyl peroxide (2.5%) + Adapalene (0.1%).	2 (*n* = 141)
3	Moderate severity: up to many non-inflammatory lesions and may have some inflammatory lesions, but no more than one small nodular lesion	Topical product with an oral antibiotic or isotretinoin. Additional, topical combined oral contraceptive or oral spironolactone (female).	3 (*n* = 195)
4	Severe: up to many non-inflammatory lesions and may have some inflammatory lesions, but no more than a few nodular lesions		

For model development, images classified as Grade 0 (clear skin without acne lesions; $n = 82$) were excluded because they do not require therapeutic intervention and are therefore outside the clinical scope of this study. In addition, Grades 3 and 4 were merged into a single moderate–severe category. The consolidation of Grades 3 and 4 was justified on both methodological and clinical grounds. Methodologically, the number of Grade 4 samples was insufficient to support stable independent model training. Clinically, Grades 3 and 4 generally require similar treatment approaches according to established clinical guidelines from the Dermatological Society of Thailand (DST) [23], the American Academy of Dermatology (AAD) [4], the South-East Asia Study Alliance Guidelines (SEA) [24], and the European Dermatological Forum (EDF) [25]. The final dataset comprised three classes: Class 1 (non-inflammatory; $n = 106$), Class 2 (mild; $n = 141$), and Class 3 (moderate–severe; $n = 195$).

As part of quality control, samples were excluded if their images exhibited poor resolution, non-standard facial orientation, occlusions, inadequate illumination, or incomplete facial regions. Following quality control and class consolidation, 442 frontal facial images were retained for subsequent model training. The retained frontal images were converted to a standardized JPG format and used for model training to provide a consistent input representation and reduce variability associated with pose differences across samples.

2.3. Data Preprocessing

Face parsing was applied as a preprocessing step to standardize image inputs by removing background variability and non-relevant facial regions, as illustrated in Figure 2. Semantic segmentation was performed using a pretrained U-Net-based model trained on the CelebAMask-HQ dataset, capable of pixel-wise classification of facial components, including background, skin, eyes, nose, mouth, hair, and clothing. The implementation followed the face parsing framework proposed by Jonathan Dinu [26] and was executed using Python-based processing to generate segmentation masks. The final input images retained only the facial skin regions, while all non-skin pixels were masked, thereby reducing feature noise and improving the robustness and efficiency of downstream CNN-based acne severity classification.

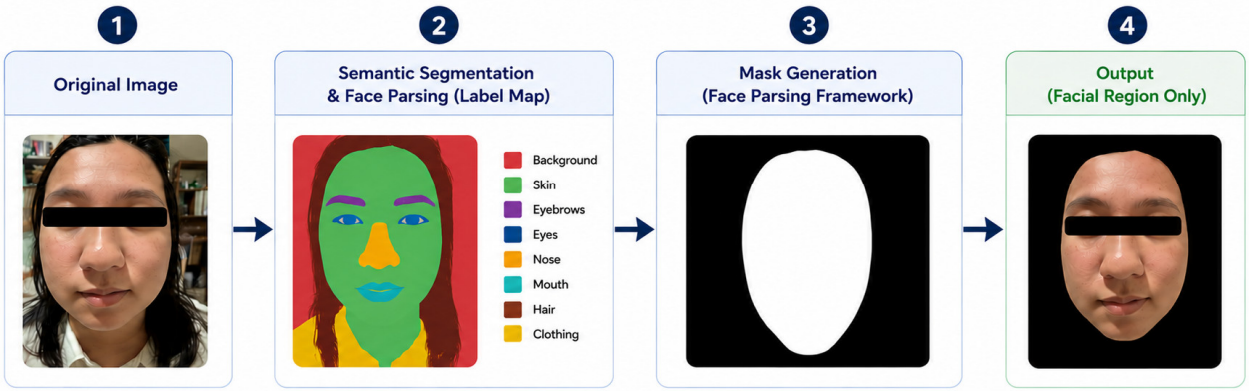


Figure 2. Face parsing preprocessing pipeline.

After segmentation, the dataset was partitioned into training and test sets using an 80:20 split, resulting in 354 images for training and 88 images for independent testing. To support model optimization and performance monitoring, the training set was further subdivided into a training subset ($n = 283$) and a validation subset ($n = 71$). The validation subset was used exclusively for hyperparameter tuning and early stopping, while the test set remained completely unseen and was reserved for final performance evaluation.

To improve model generalization and reduce overfitting, data augmentation was applied only to the training set. The random transformations included rotation (−30° to 30°), horizontal and vertical translation, scaling (0.8–1.2), and horizontal flipping. For experiments involving class-imbalance handling, resampling was subsequently applied only to the training subset, while the validation and test sets were kept unchanged to avoid data leakage and ensure unbiased performance evaluation.

2.4. Transfer Learning Using a Pre-Trained Network

Transfer learning is an effective technique for creating models for specialized tasks by using pre-trained models developed on large datasets. This approach improves training efficiency and accuracy, especially when data is limited [27]. Transfer learning involves two main methods: feature extraction and fine-tuning. Feature extraction uses the pre-trained model's learned features as input for related tasks, while fine-tuning adapts the model by adjusting certain layers to suit the new task [28].

The transfer learning architecture used in this study is illustrated in Figure 3. Preprocessed frontal facial images with an input size of 224 × 224 × 3 were fed into pretrained CNN base models, including ResNet18, ResNet50, GoogLeNet, and InceptionV3. These models were initialized with ImageNet pretrained weights and fine-tuned for the acne severity classification task. A task-specific classification head was then added, consisting of global average pooling, a fully connected layer, and a softmax layer. The final output classified each image into one of three acne severity classes: non-inflammatory, mild, or moderate–severe. In the baseline transfer learning setting, no class-imbalance handling was applied.

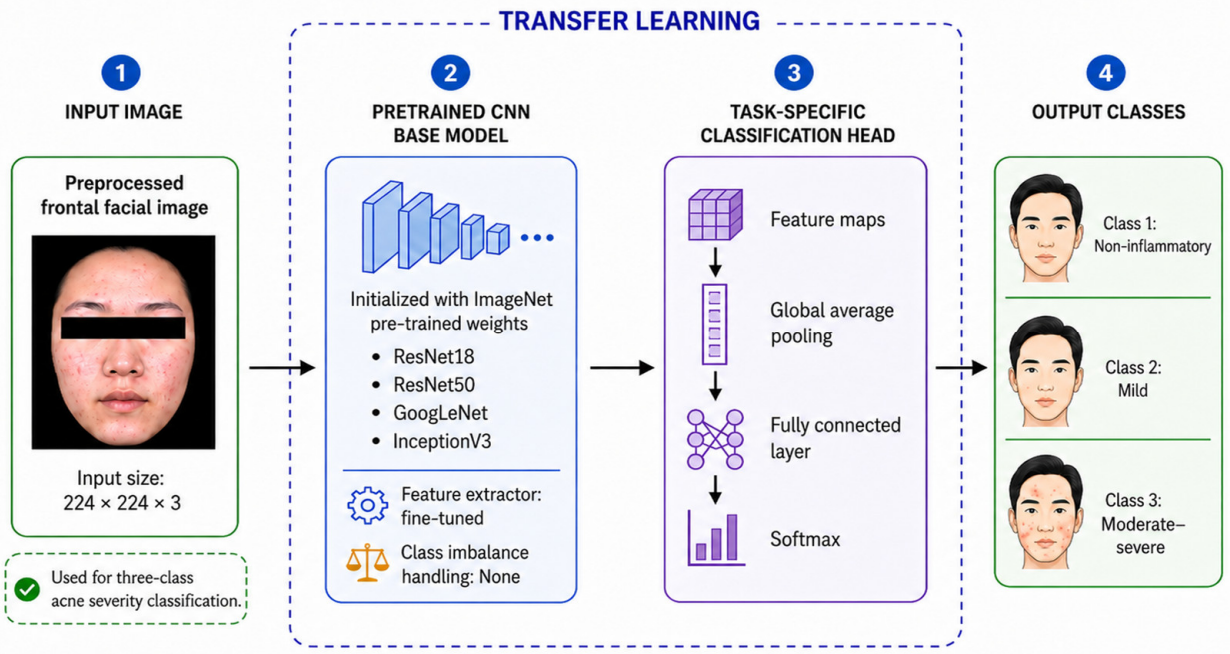


Figure 3. Transfer learning architecture for three-class acne severity classification using ImageNet-pretrained CNN base models, and a task-specific classification head.

The experimental workflow was organized into sequential stages, as illustrated in Figure 4, to ensure systematic model development and optimization. The methodology comprised two main parts: (1) model selection and hyperparameter tuning, and (2) model robustness assessment and class imbalance handling.

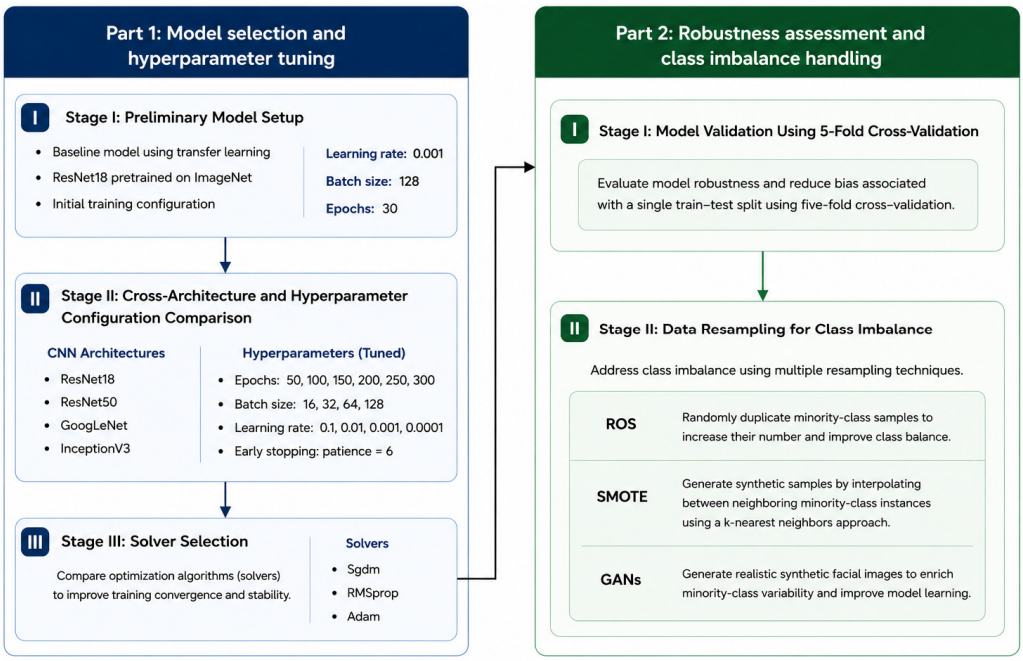


Figure 4. Overview of the experimental workflow (Sgdm: Stochastic Gradient Descent with Momentum, RMSprop: Root Mean Square Propagation, Adam: Adaptive Moment Estimation, ROS: Random Oversampling, SMOTE: the Synthetic Minority Over-sampling Technique, and GANs: Generative Adversarial Networks).

Part 1: Model selection and hyperparameter tuning

Stage I: Preliminary Model Setup. In the first stage, a baseline deep learning model was constructed using transfer learning. ResNet18, pretrained on ImageNet, was selected as the initial base model due to its moderate network depth and computational efficiency. Model training was initialized with a learning rate of 0.001, a batch size of 128, and 30 training epochs. This baseline configuration served as a reference for subsequent model comparison across different CNN architectures and for systematic hyperparameter tuning aimed at optimizing classification performance.

Stage II: Cross-Architecture and Hyperparameter Configuration Comparison. Following baseline construction, multiple convolutional neural network architectures including ResNet18, ResNet50, GoogLeNet, and InceptionV3. The models were systematically evaluated under identical experimental conditions. To optimize learning performance, key hyperparameters were tuned by varying the number of training epochs (50, 100, 150, 200, 250, and 300), batch size (16, 32, 64, and 128), learning rate (0.1, 0.01, 0.001, and 0.0001), and Early stopping (patience = 6). This systematic experiment aimed to identify the most suitable combination of network architecture and hyperparameter settings for accurate and robust acne severity classification. Given the dataset's restricted and imbalanced nature, the optimal configuration identified at this stage was established as the baseline model. Subsequently, different optimization algorithms were evaluated to enhance training convergence and overall model performance further.

Stage III: Solver Selection. After determining the optimal model configuration, different optimization algorithms (solvers) were compared to further enhance training convergence and stability. The solvers evaluated included Stochastic Gradient Descent with Momentum (Sgdm), Root Mean Square Propagation (RMSprop), and Adaptive Moment Estimation (Adam). Each solver was tested using the selected hyperparameters, and performance was assessed to determine the most effective optimization strategy.

Part 2: Robustness assessment and class imbalance handling

Stage I: Model Validation Using 5-Fold Cross-Validation. To evaluate model robustness and reduce bias associated with a single train–test split, five-fold cross-validation was performed on the training dataset.

Stage II: Data Resampling for Class Imbalance. In the final experiment, class imbalance was addressed using multiple resampling techniques: ROS, SMOTE [17], and GAN-based oversampling [29]. ROS increases the number of minority-class samples through random duplication of existing instances, thereby improving class balance without altering the original data distribution. In contrast, SMOTE generates new synthetic samples by interpolating between neighboring minority-class instances based on a k-nearest neighbors approach, enabling a more diverse representation of the minority class. GAN-based oversampling generates realistic synthetic facial images with the aim of enriching minority-class variability and improving model learning.

In this study, SMOTE was applied with k = 5, following the original formulation [17]. For GAN-based oversampling, Cycle-Consistent Generative Adversarial Network (CycleGAN) [29], was employed for unpaired image-to-image translation between acne severity classes. Both the generator and discriminator were trained with a learning rate of 0.0002, a gradient decay factor of 0.5, and a squared gradient decay factor of 0.999. The generator used a ResNet-based architecture with nine residual blocks, while the discriminator used a 70 × 70 PatchGAN architecture with Instance Normalization and LeakyReLU activation, where alpha = 0.2. Images were resized to 512 × 512 pixels before training. A higher input resolution of 1024 × 1024 pixels was also explored during preliminary experimentation; however, this setting resulted in lower classification performance and was therefore not used in the final oversampling experiment. The training objective combined least-squares adversarial loss, cycle-consistency loss (λ_{cyc} = 5.0), and identity loss (λ_{id} = 0.25). Training was performed for 300 epochs using the Adam optimizer, and the best model checkpoint was selected based on the lowest Kernel Inception Distance score [30]. Synthetic images were generated by translating images from a source class to a target acne severity class. Only the highest-quality synthetic images were retained through discriminator-based filtering, and the generated images inherited the labels of the corresponding target class. Representative CycleGAN-generated images are provided in Figure S1 to illustrate the visual reference of the augmented samples.

For all oversampling methods, the training data were resampled to obtain 155, 300, and 500 samples per class. The target of 155 samples per class was selected as a near-balanced baseline approximating the original majority-class size, while 300 and 500 were used to examine the effect of more intensive oversampling on classification performance and model generalization.

Resampling was performed only on the training set, while the validation and test sets were strictly preserved to ensure an unbiased assessment of model performance. These approaches were then comparatively evaluated to examine their effects on classification performance and model generalization.

2.5. Model Evaluation

The models were evaluated on the held-out test set (20% of the dataset) to ensure an unbiased assessment of generalization performance. Model performance was assessed using a confusion matrix, which provides detailed insight into class-wise predictions and identifies misclassifications, including false positives and false negatives. Based on the confusion matrix, key evaluation metrics, accuracy, precision, recall, F1-score, and specificity were computed. For the multiclass classification task, a one-vs-rest strategy was used to compute class-wise metrics, and overall performance was summarized as a weighted average.

2.6. Model Explainability

To improve transparency and support clinical validation, Grad-CAM was applied to the trained convolutional neural network models [22]. Grad-CAM heatmaps generated from the test images across all acne severity levels were used to confirm that model predictions were driven by clinically relevant facial regions associated with acne lesions.

3. Results

This section reports the performance of the proposed transfer learning-based model and compares its effectiveness with alternative data enhancement strategies applied to the acne image dataset. The impact of different data resampling methods on classification accuracy and overall model performance is also evaluated.

3.1. Hyperparameter Configuration for Base Models in Part 1 of Stage I & II

The baseline performance of the transfer learning models is summarized in Table 2, with the highest macro average values presented in bold. When ResNet18 was trained on raw images, the validation accuracy reached 0.54, whereas the test accuracy decreased markedly to 0.28, indicating poor generalization. The macro-average precision, recall, and F1-score were low (0.27, 0.26, and 0.26, respectively), with particularly low recall for class 2 (0.14). After image preprocessing, the test accuracy improved to 0.37 despite a lower validation accuracy (0.40), suggesting improved generalization. Macro-average precision, recall, F1-score, and specificity also increased to 0.42, 0.38, 0.39, and 0.70, respectively. The greatest improvement was observed in class 3, where the F1-score increased from 0.34 to 0.50, indicating improved detection of moderate-to-severe acne.

Table 2. Baseline performance of transfer learning models using raw and preprocessed data.

Model	ValAcc	TestAcc (95% CI)	Class	Precision	Recall	F1	Specificity
ResNet18 (Base) Raw data	0.54	0.28 (0.20–0.38)	1	0.21	0.28	0.24	0.68
			2	0.27	0.14	0.19	0.82
			3	0.32	0.36	0.34	0.33
			MC(AVG)	0.27	0.26	0.26	0.61
ResNet18 Preprocessed data	0.40	0.37 (0.28–0.48)	1	0.25	0.39	0.31	0.60
			2	0.29	0.19	0.23	0.80
			3	0.52	0.48	0.5	0.65
			MC(AVG)	0.42	0.38	0.39	0.70
ResNet50 (Tuning) Preprocessed data	0.52	**0.52** (0.42–0.62)	1	0.40	0.43	0.42	0.77
			2	0.60	0.12	0.20	0.96
			3	0.56	0.85	0.68	0.49
			MC(AVG)	**0.52**	**0.46**	**0.43**	**0.74**

ValAcc = Validation accuracy, TesAcc = Test accuracy, MC(AVG) = Macro Average.

For ResNet50 with hyperparameter tuning, both validation and test accuracy were 0.52, indicating consistent performance across datasets. The macro-average precision, recall, and F1-score were 0.52, 0.46, and 0.43, respectively, with a specificity of 0.74. The model showed strong performance in class 3, achieving a recall of 0.85 and an F1-score of 0.68, reflecting effective detection of moderate-to-severe acne. In contrast, class 2 performance remained limited, with low recall (0.12) despite relatively high precision (0.60), suggesting under-detection. Overall, ResNet50 outperformed ResNet18 and provided more balanced performance, particularly in terms of generalization and detection of more severe acne.

The comparative evaluation of convolutional neural network architectures and hyperparameter settings demonstrated that model performance was strongly influenced by

both network design and training configuration (Figure 5). Among the evaluated architectures, ResNet50 consistently outperformed ResNet18, GoogLeNet, and InceptionV3 across multiple parameter combinations.

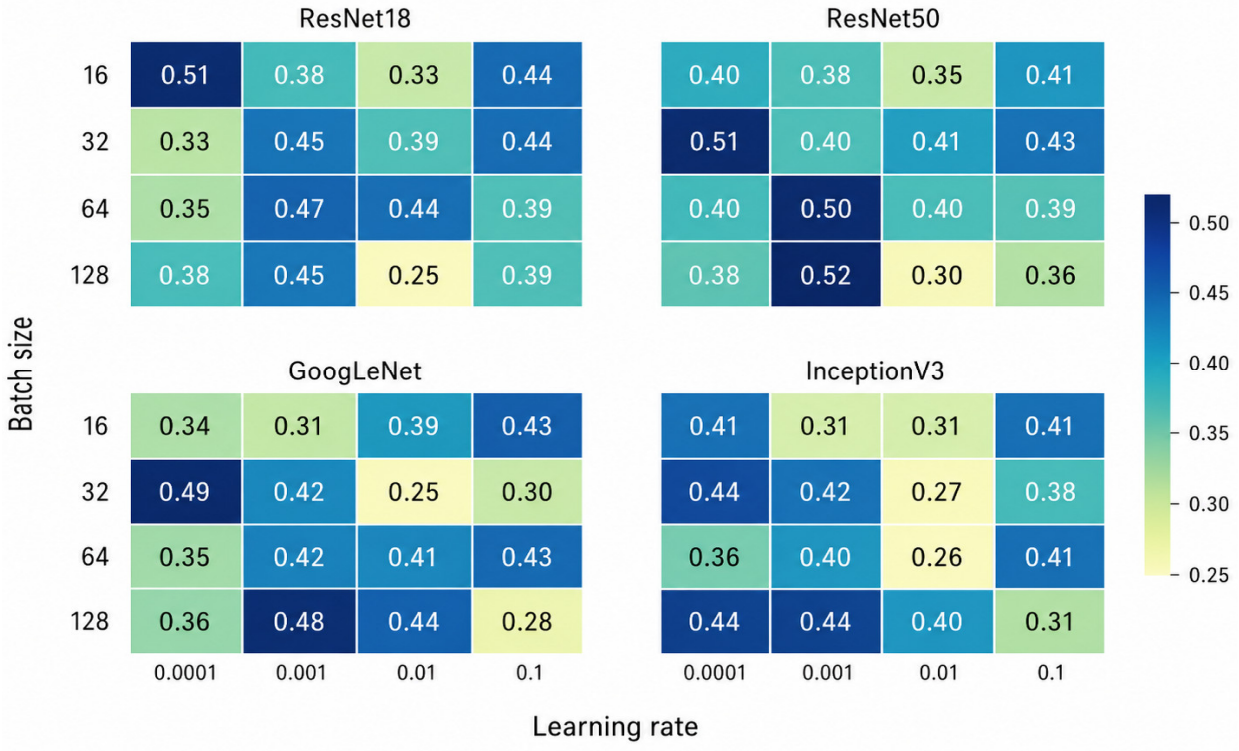


Figure 5. Heatmap of test accuracy for different CNN architectures under varying learning rates and batch sizes.

The optimal performance was achieved using ResNet50 with a learning rate of 0.001 and a batch size of 128, yielding a test accuracy of 0.52. Compared with the preprocessed ResNet18 baseline (0.37), this represents a clear improvement in predictive capacity. Class-wise evaluation further showed enhanced discrimination for class 3 (recall = 0.85, F1 = 0.68), indicating improved sensitivity for moderate-to-severe acne cases. However, class 2 remained challenging, with low recall (0.12), reflecting persistent class imbalance effects. This configuration was therefore selected as the final model for subsequent experiments, including optimizer comparison, cross-validation, and data resampling analyses.

3.2. Optimization Algorithm Comparison in Part 1 of Stage III

The comparison of performance with different optimization algorithms was summarized in Table 3, with the highest macro average values presented in bold. Both Sgdm and RMSprop produced identical test accuracies (0.41) with similar macro-average precision, recall, and F1-score (0.41). In contrast, Adam achieved the highest test accuracy (0.52) and improved macro-average performance, including precision (0.52), recall (0.46), F1-score (0.43), and specificity (0.73). At the class level, Adam showed the strongest performance for Class 3 (recall = 0.84, F1-score = 0.68). Therefore, Adam was selected as the optimizer for subsequent experiments.

Table 3. Performance evaluation of different optimization algorithms and cross-validation strategies.

Optimizer	TestAcc 95% CI	Class	Precision	Recall	F1	Specificity
Sgdm	0.41 (0.31–0.52)	1	0.35	0.43	0.39	0.72
		2	0.29	0.3	0.3	0.69
		3	0.6	0.51	0.55	0.73
		MC(AVG)	0.41	0.41	0.41	0.71
RMSprop	0.41 (0.31–0.52)	1	0.33	0.35	0.34	0.75
		2	0.4	0.35	0.37	0.79
		3	0.5	0.54	0.51	0.57
		MC(AVG)	0.41	0.41	0.41	0.70
Adam	**0.52** (0.42–0.62)	1	0.4	0.43	0.41	0.76
		2	0.6	0.11	0.19	0.96
		3	0.56	0.84	0.68	0.48
		MC(AVG)	**0.52**	0.46	0.43	0.73
Adam (5-fold CV)	0.51 (0.41–0.61)	1	0.39	0.39	0.39	0.78
		2	0.56	0.53	0.54	0.82
		3	0.55	0.56	0.55	0.63
		MC(AVG)	0.50	**0.49**	**0.49**	**0.74**

3.3. Cross-Validation Performance in Part 2 of Stage I

The 5-fold cross-validation experiment was conducted to assess the reliability of model performance across different data partitions. As shown in Table 3, standard Adam training achieved performance comparable to that of the 5-fold cross-validation model. These findings support the reliability of the model by showing that the results were not solely dependent on a single train–validation split and that the model maintained comparable class-level performance across repeated data partitions.

3.4. Effects of Data Resampling for Class Imbalance in Part 2 of Stage II

The effects of different resampling techniques and the number of images generated per class are presented in Table 4, with the highest macro-average values shown in bold. Overall, the resampling methods did not clearly improve overall model performance compared with the baseline 5-fold cross-validation result reported in Table 3. Although some resampling settings improved recognition of certain classes, these gains were not consistent across all acne severity levels. In particular, improvements in Class 1 or Class 3 were often accompanied by reduced performance for Class 2, suggesting a shift in class-wise prediction behavior rather than a uniform improvement in classification performance.

Across the evaluated settings, ROS showed limited benefit, with test accuracy decreasing as the number of duplicated samples increased. SMOTE achieved the highest test accuracy at 300 images per class, but this improvement was not balanced across classes, as Class 2 performance remained weak. GAN-based resampling at 300 images per class produced overall performance comparable to the baseline 5-fold cross-validation model. Although it improved class-wise accuracy for Class 1 and Class 3, this was accompanied by reduced performance for Class 2. Therefore, GAN-based resampling did not provide a clear overall improvement but instead redistributed model performance across classes.

Table 4. Performance comparison of data resampling techniques (ROS, SMOTE, and GANs) at different sample sizes per class, evaluated using test accuracy and average classification metrics.

Resampling Method	Images per Class	TestAcc (95% CI)	Class	Precision	Recall	F1	Specificity
ROS	155	0.44 (0.34–0.55)	1	0.42	0.47	0.44	0.76
			2	0.23	0.11	0.15	0.83
			3	0.51	0.64	0.56	51
			MC(AVG)	0.39	0.41	0.38	0.70
	300	0.43 (0.33–0.54)	1	0.20	0.13	0.15	0.81
			2	0.34	0.46	0.39	0.62
			3	0.60	0.58	0.59	0.69
			MC(AVG)	0.38	0.39	0.38	0.71
	500	0.38 (0.28–0.49)	1	0.25	0.30	0.28	0.69
			2	0.26	0.23	0.24	0.72
			3	0.55	0.53	0.54	0.65
			MC(AVG)	0.35	0.35	0.35	0.69
SMOTE	155	0.37 (0.28–0.48)	1	0.33	0.10	0.16	0.92
			2	0.26	0.19	0.22	0.78
			3	0.38	0.61	0.48	0.22
			MC(AVG)	0.32	0.30	0.28	0.64
	300	**0.52** (0.42–0.62)	1	0.40	0.43	0.41	0.76
			2	0.60	0.23	0.33	0.93
			3	0.56	0.76	0.65	0.53
			MC(AVG)	**0.52**	0.47	0.46	0.74
	500	0.42 (0.32–0.53)	1	0.20	0.10	0.10	0.84
			2	0.40	0.09	0.15	0.94
			3	0.39	0.70	0.50	0.15
			MC(AVG)	0.33	0.30	0.25	0.64
CycleGAN	155	0.48 (0.42–0.62)	1	0.33	0.44	0.38	0.76
			2	0.40	0.14	0.21	0.89
			3	0.56	0.74	0.64	0.52
			MC(AVG)	0.43	0.44	0.41	0.72
	300	0.51 (0.42–0.62)	1	0.35	0.44	0.39	0.78
			2	0.47	0.32	0.38	0.82
			3	0.63	0.69	0.66	0.65
			MC(AVG)	0.48	**0.49**	**0.48**	**0.75**
	500	0.37 (0.41–0.61)	1	0.19	0.33	0.24	0.62
			2	0.41	0.25	0.31	0.82
			3	0.51	0.48	0.50	0.60
			MC(AVG)	0.37	0.35	0.35	0.68

Across all methods, increasing the resampling target to 500 images per class reduced performance, suggesting that excessive resampling may have introduced redundancy or synthetic variability that did not support generalization. Overall, these findings indicate that resampling alone was insufficient to substantially improve acne severity classification performance in this dataset.

3.5. Statistical Tests

To statistically examine whether resampling-based models produced different prediction outcomes from the baseline model, McNemar's test was applied using paired predictions on the same test samples. The analysis was based on discordant cases, where one model produced a correct prediction and the other did not. Exact McNemar's tests

were used due to the relatively small number of discordant pairs, and Holm–Bonferroni correction was applied to control for multiple comparisons at a significance level of $\alpha = 0.05$.

Table 5 summarizes the pairwise statistical comparison between the baseline model and the resampling-based models. The table reports the discordant case counts, exact *p*-values (Exact *p*), and Holm–Bonferroni-adjusted *p*-values (Holm *p*). The adjusted *p*-values did not provide sufficient evidence to conclude that ROS, SMOTE, or GAN-based resampling improved prediction outcomes relative to the baseline model. These findings support the interpretation that resampling alone was insufficient to substantially improve acne severity classification performance in this dataset.

Table 5. Pairwise statistical comparison between the baseline model and resampling-based models using McNemar's test.

Comparison	n_{01}	n_{10}	Exact *p*	Holm *p*
Base Model vs. ROS_300	10	12	0.832	1.000
Base Model vs. SMOTE_300	7	8	1.000	1.000
Base Model vs. GANs_300	20	16	0.618	1.000

n_{01} = cases incorrectly predicted by the baseline model but correctly predicted by the comparator model, and n_{10} = cases correctly predicted by the baseline model but incorrectly predicted by the comparator model.

3.6. Explainability Analysis via Grad-CAM Visualization

Grad-CAM visualization [22] was used to qualitatively assess the facial regions contributing to model predictions across acne severity levels. In the representative examples shown in Figure 6, the activation maps suggested that the model attended to some acne-relevant facial skin regions. The non-inflammatory case showed relatively limited and diffuse activation, which may reflect the less visually prominent lesion patterns in this class. The mild case showed more localized activation, whereas the moderate–severe case showed broader activation across facial regions where acne lesions were visible.

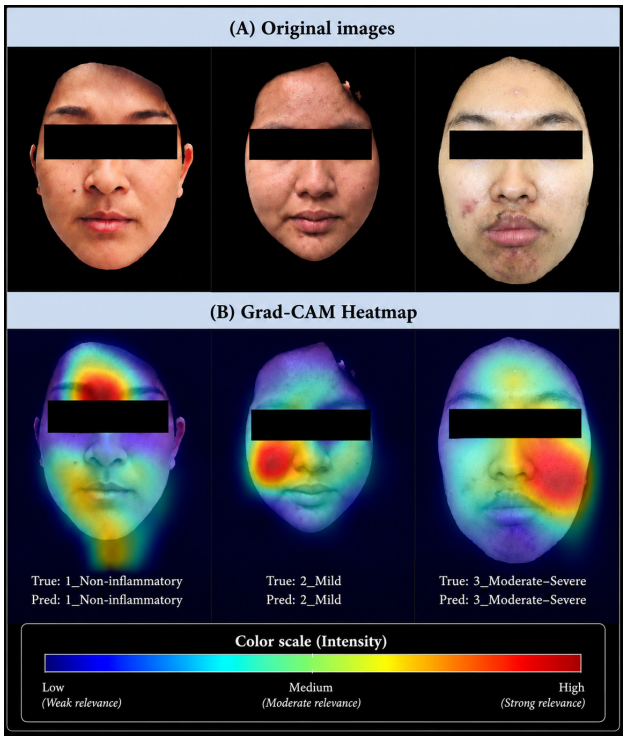


Figure 6. Grad-CAM heatmaps highlighting lesion-related facial regions used for acne severity classification.

However, some activations were also observed around the lips and adjacent perioral areas. This suggests that the model may have attended not only to acne-related skin patterns but also to visually salient facial structures. To examine this issue further, an additional preprocessing experiment was conducted in which the lip region was excluded from the model input. However, this modification did not lead to a clear improvement in classification performance, suggesting that the observed lip-related activation may reflect broader challenges in learning acne-relevant visual patterns from limited facial-image data rather than a simple masking issue.

4. Discussion

This study shows that model performance appeared to be influenced by the facial parsing approach, model architecture, optimization strategy, and class-imbalance handling method. The baseline results suggested limited generalization when raw facial images were used, as reflected by the discrepancy between validation and test performance. After image preprocessing, model performance improved, suggesting that background removal and facial region standardization enhanced feature consistency and supported better generalization. This finding is consistent with prior studies emphasizing the importance of preprocessing in medical image analysis [31].

Among the evaluated convolutional neural network architectures, ResNet50 achieved stronger overall performance than ResNet18, GoogLeNet, and InceptionV3 across a wide range of hyperparameter configurations. This finding aligns with previous studies reporting that deeper residual networks can provide improved representational capacity and training stability for complex visual recognition tasks [32]. The results also underscore the importance of hyperparameter tuning, particularly learning-rate selection, for achieving stable and efficient convergence in deep neural networks [33]. The comparison of optimization algorithms showed that Adam performed better than Sgdm and RMSprop, particularly in terms of recall and F1-score for moderate–severe acne cases. This result is consistent with prior work demonstrating that Adam's adaptive learning-rate mechanism can be effective for handling heterogeneous data distributions and class imbalance in deep learning models [34]. The 5-fold cross-validation experiment was conducted to evaluate the reliability and stability of model performance across different data partitions. Although it achieved slightly lower test accuracy than the single-split model, it produced more balanced precision, recall, F1-score, and specificity. This suggests that the model maintained comparable class-level performance across repeated partitions, supporting the reliability of the performance estimates under limited-data conditions [35].

The limited improvement observed after resampling suggests that class imbalance was not the only factor affecting model performance. Given the relatively small dataset size, resampling may have increased sample quantity without sufficiently increasing clinically meaningful visual diversity. In this study, higher resampling levels, particularly 500 images per class, appeared to reduce performance, suggesting possible overfitting to duplicated or lower-quality synthetic patterns [36,37].

The overall test accuracy of approximately 0.51 further reflects the difficulty of automated acne severity classification under data-constrained conditions. One possible explanation is that the visual boundaries between severity classes are often subtle, especially between adjacent categories such as non-inflammatory and mild acne or mild and moderate–severe acne. In real-world facial images, lesion visibility may also be affected by lighting, skin texture, facial pose, and uneven lesion distribution. These factors may make it difficult for an image-level CNN model to learn distinct severity patterns, even when class imbalance is addressed through resampling.

From a clinical perspective, this limitation is important because misclassification, particularly underestimation of severity, may lead to delayed or inadequate treatment [4,20]. Therefore, the proposed framework should not be interpreted as an autonomous diagnostic tool at this stage. Instead, it should be viewed as an exploratory benchmark for evaluating transfer learning, resampling-based class-imbalance handling, and interpretability under restricted, imbalanced, and non-public facial-image conditions. Future studies are needed to further evaluate the framework using larger and more diverse datasets, multi-view facial images, lesion-level annotation, and clinical evaluation to better capture fine-grained differences among acne severity levels.

The findings of the proposed framework suggest the exploratory feasibility of automated facial acne grading under data-constrained conditions. To contextualize the present study, Table 6 summarizes representative AI-powered approaches for acne assessment [38], including label distribution learning, object detection, attention-based CNN frameworks, and the proposed transfer learning framework.

Table 6. Comparison of representative AI-powered approaches for acne assessment.

Approach	Representative Study	Prediction Strategy	Main Strengths	Main Limitations
Label distribution learning	Wu et al. [39]	Joint acne image grading and lesion counting	Models acne severity as a distribution rather than a single fixed label, allowing more flexible representation of label ambiguity	Requires reliable label-distribution information and may involve greater annotation complexity
YOLO-based object detection	Azhar et al. [40]	Acne vulgaris type classification and lesion counting	Enables lesion localization and counting, supporting more detailed lesion-level assessment	Requires lesion-level annotation and may be sensitive to image quality and lesion visibility
Skin attention and SFNet	Lin et al. [41]	CNN-based acne severity estimation from facial images	Uses skin attention and deep feature extraction to improve acne severity estimation	Performance may depend on image quality and controlled acquisition conditions
Hybrid attention-guided capsule framework	Rasool et al. [42]	Attention-guided deep learning framework for facial acne severity classification	Enhances feature discrimination and captures spatial relationships among acne-related features	Increased model complexity and limited external validation
Proposed framework	Present study	Global image-level transfer learning with resampling-based class-imbalance assessment and post hoc Grad-CAM visualization	Evaluates model architecture, optimization strategy, validation strategy, and resampling methods under restricted, imbalanced, non-public facial-image conditions	Limited dataset size; image-level classification only; no lesion localization or uncertainty modeling

Unlike lesion-level detection and label-distribution learning approaches, the present study focused on global image-level acne severity classification using transfer learning under restricted, imbalanced, and non-public data conditions. Its main contribution is the systematic evaluation of model architecture, optimization strategy, validation strategy, and resampling-based class-imbalance handling, together with post hoc Grad-CAM visualization for qualitative model interpretability.

Limitations and Future Work

This study has several limitations. First, the dataset included only Fitzpatrick Skin Types III and IV; therefore, the generalizability of the proposed model to other skin types has not yet been established. Future studies should include more diverse populations to evaluate model performance across a broader range of skin tones.

Second, the dataset size was relatively small and imbalanced. Although data resampling was applied to address class imbalance, it did not clearly improve overall model performance, and synthetic images may not fully capture the variability of real-world acne presentations. Third, only frontal facial images were used for model training, which may have contributed to misclassification because side-view information was not included in the model input. In addition, acne severity assessment was based solely on static images without additional clinical context.

In this exploratory study, acne severity labels were assigned by a single board-certified dermatologist to provide expert-based clinical annotation. However, because grading was performed by one expert, inter-rater reliability and discrepancy resolution could not be assessed. Future research should involve multiple independent dermatologists and consensus-based adjudication to improve annotation reliability and clinical reproducibility.

In addition, Grad-CAM visualization showed occasional activation around the lips and adjacent perioral areas, suggesting that model attention was not fully restricted to acne lesions. Although additional lip-region exclusion was explored during preprocessing, it did not lead to clear performance improvement in this dataset. Future studies should further investigate facial-region preprocessing and lesion-focused attention mechanisms to improve interpretability and reduce non-lesion activations. Larger and more diverse datasets, multi-view facial images, lesion-level annotation, and clinical evaluation are also needed to support more robust model validation.

5. Conclusions

This study evaluated a transfer learning-based framework for acne severity classification using restricted, imbalanced, and non-public real-world facial images. The findings suggest the exploratory feasibility of automated facial acne grading under data-constrained conditions, while also showing that the current performance remains insufficient for autonomous clinical decision-making. Systematic evaluation of preprocessing, model architecture, optimizer selection, and validation strategy supported a more structured assessment of model performance. However, resampling-based class-imbalance handling showed limited overall benefit and did not clearly improve classification performance. These results suggest that factors beyond class imbalance, including limited dataset size, subtle visual differences between severity classes, and variability in real-world facial images, may affect model performance. Future studies should validate the framework using larger and more diverse datasets, multi-view facial images, lesion-level annotation, and clinical evaluation before considering clinical implementation.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cosmetics13040166/s1, Figure S1. Representative CycleGAN-generated synthetic acne images at a resolution of 512 × 512 × 3 pixels.

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

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, Belmont report, Council for International Organizations of Medical Sciences: CIOMS, ICH Good Clinical Practice and approved by the Ethics Committee of Mae Fah Luang University (COA No. 52/2025; protocol code EC 22190-18; date of approval: 21 April 2023).

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

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request. These data are not publicly accessible due to restrictions imposed by the Ethics Committee to protect proprietary information of the participants.

Acknowledgments: The authors would like to express their sincere appreciation to Aimusa Arsu, for her valuable professional input in reviewing and evaluating acne treatment products and pharmacological therapies, as well as for providing expert guidance that enhanced the clinical relevance of the treatment recommendations in this study. This study was partially supported by the Biomedical Technology Research Group for Vulnerable Populations, the Center of Excellence in AI and Emerging Technologies, and the Office of Postgraduate Studies, Mae Fah Luang University. APC was supported by Mae Fah Luang University.

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

Abbreviations

The following abbreviations are used in this manuscript:

AAD	American Academy of Dermatology
Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
CNN	Convolutional Neural Network
CycleGAN	Cycle-Consistent Generative Adversarial Network
DST	Dermatological Society of Thailand
EDF	European Dermatological Forum
GANs	Generative Adversarial Networks
Grad-CAM	Gradient-weighted Class Activation Mapping
IGA	Investigator's Global Assessment
MC(AVG)	Macro Average
RMSprop	Root Mean Square Propagation
ROS	Random Over-sampling
SEA	South-East Asia Study Alliance Guidelines
Sgdm	Stochastic Gradient Descent with Momentum
SMOTE	Synthetic Minority Over-sampling Technique
TestAcc	Test accuracy

References

1. Zaenglein, A.L. Acne Vulgaris. *N. Engl. J. Med.* **2018**, *379*, 1343–1352. [CrossRef] [PubMed]
2. Tan, J.K.L.; Bhate, K. A global perspective on the epidemiology of acne. *Br. J. Dermatol.* **2015**, *172*, 3–12. [CrossRef] [PubMed]
3. Hauk, L. Acne vulgaris: Treatment guidelines from the AAD. *Am. Fam. Physician* **2017**, *95*, 740–741. [PubMed]

4. Reynolds, R.V.; Yeung, H.; Cheng, C.E.; Cook-Bolden, F.; Desai, S.R.; Druby, K.M.; Freeman, E.E.; Keri, J.E.; Stein Gold, L.F.; Tan, J.K.L.; et al. Guidelines of Care for the Management of Acne Vulgaris. *J. Am. Acad. Dermatol.* **2024**, *90*, 1006.e1–1006.e30. [CrossRef] [PubMed]
5. Swenson, K.; Graber, E.M.; Thiboutot, D. Over-the-counter acne treatments. In *Cosmetic Dermatology: Products and Procedures*; John Wiley & Sons Ltd.: Hoboken, NJ, USA, 2022; pp. 587–597. [CrossRef]
6. Tassavor, M.; Payette, M.J. Estimated cost efficacy of US Food and Drug Administration–approved treatments for acne. *Dermatol. Ther.* **2019**, *32*, e12765. [CrossRef] [PubMed]
7. Gholizadeh, N.; Rokni, G.R.; Babaei, M. Advantages and Disadvantages of Using AI in Dermatology. *Dermatol. Rev.* **2024**, *5*, e248. [CrossRef]
8. Wang, P.; Fan, E.; Wang, P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognit. Lett.* **2021**, *141*, 61–67. [CrossRef]
9. Han, S.S.; Park, G.H.; Lim, W.; Kim, M.S.; Na, J.I.; Park, I.; Chang, S.E. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network. *PLoS ONE* **2018**, *13*, e0191493. [CrossRef] [PubMed]
10. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
11. Yang, Y.; Guo, L.; Wu, Q.; Zhang, M.; Zeng, R.; Ding, H.; Zheng, H.; Xie, J.; Li, Y.; Ge, Y.; et al. Construction and Evaluation of a Deep Learning Model for Assessing Acne Vulgaris Using Clinical Images. *Dermatol. Ther.* **2021**, *11*, 1239–1248. [CrossRef] [PubMed]
12. Alzahrani, S.; Al-Bander, B.; Al-Nuaimy, W. Attention Mechanism Guided Deep Regression Model for Acne Severity Grading. *Computers* **2022**, *11*, 31. [CrossRef]
13. 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*, 143–153. [CrossRef] [PubMed]
14. 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. [CrossRef] [PubMed]
15. Veni, P.K.; Gupta, A. Revolutionizing Acne Diagnosis with Hybrid Deep Learning Model Integrating CBAM, and Capsule Network. *IEEE Access* **2024**, *12*, 82867–82879. [CrossRef]
16. Huang, N.-C.; Mukundan, A.; Karmakar, R.; Syna, S.; Chang, W.-Y.; Wang, H.-C. Novel Snapshot-Based Hyperspectral Conversion for Dermatological Lesion Detection via YOLO Object Detection Models. *Bioengineering* **2025**, *12*, 714. [CrossRef] [PubMed]
17. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
18. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM.* **2020**, *63*, 139–144. [CrossRef]
19. Frid-Adar, M.; Diamant, I.; Klang, E.; Amitai, M.; Goldberger, J.; Greenspan, H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* **2018**, *321*, 321–331. [CrossRef]
20. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
21. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4793–4813. [CrossRef] [PubMed]
22. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int. J. Comput. Vis.* **2019**, *128*, 336–359. [CrossRef]
23. Kulthanan, K.; Tuchinda, P.; Nitiyarom, R.; Chunharas, A.; Chantaphakul, H.; Aunhachoke, K.; Chularojanamontri, L.; Rajatanavin, N.; Jirapongsananuruk, O.; Vichyanond, P.; et al. Clinical practice guidelines for the diagnosis and management of atopic dermatitis. *Asian Pac. J. Allergy Immunol.* **2021**, *39*, 145–155. [CrossRef] [PubMed]
24. Goh, C.L.; Abad-Casintahan, F.; Aw, D.C.W.; Baba, R.; Chan, L.C.; Hung, N.T.; Kulthanan, K.; Leong, H.N.; Medina-Oblepias, M.S.; Noppakun, N.; et al. South-East Asia study alliance guidelines on the management of acne vulgaris in South-East Asian patients. *J. Dermatol.* **2015**, *42*, 945–953. [CrossRef] [PubMed]
25. Nast, A.; Rosumeck, S.; Erdmann, R.; Alsharif, U.; Dressler, C.; Werner, R.N. Methods report on the development of the European evidence-based (S3) guideline for the treatment of acne—Update 2016. *J. Eur. Acad. Dermatol. Venereol.* **2016**, *30*, e1–e28. [CrossRef] [PubMed]
26. Dinu, J. Transformer-Based Semantic Segmentation Image Model: Face-Parsing. 2021. Available online: https://huggingface.co/jonathandinu/face-parsing (accessed on 12 September 2023).
27. Hosna, A.; Merry, E.; Gyalmo, J.; Alom, Z.; Aung, Z.; Azim, M.A. Transfer learning: A friendly introduction. *J. Big Data* **2022**, *9*, 102. [CrossRef] [PubMed]

28. Davila, A.; Colan, J.; Hasegawa, Y. Comparison of fine-tuning strategies for transfer learning in medical image classification. *Image Vis. Comput.* **2024**, *146*, 105012. [CrossRef]
29. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2223–2232. [CrossRef]
30. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.K.; Wang, Z.; Smolley, S.P. Least Squares Generative Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2794–2802. [CrossRef]
31. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
32. 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. [CrossRef]
33. Gao, Y.; Li, J.; Zhou, Y.; Xiao, F.; Liu, H. Optimization Methods for Large-Scale Machine Learning. In Proceedings of the 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2021, Chengdu, China, 17–19 December 2021; pp. 304–308. [CrossRef]
34. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015. [CrossRef]
35. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the International Joint Conferences on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995; pp. 1137–1145.
36. Yang, C.; Fridgeirsson, E.A.; Kors, J.A.; Reps, J.M.; Rijnbeek, P.R. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *J. Big Data* **2024**, *11*, 7. [CrossRef]
37. Jeong, J.J.; Patel, B.; Banerjee, I. GAN augmentation for multiclass image classification using hemorrhage detection as a case-study. *J. Med. Imaging* **2022**, *9*, 035504. [CrossRef] [PubMed]
38. Tanga, P.; Karampinis, E.; Papakonstantis, M. Al-Powered Approach to Acne. In *Artificial Intelligence Applications in Dermatology: Dermatology ex Machina: Clinical applications*; Springer Nature: Cham, Switzerland, 2026; pp. 211–229. [CrossRef] [PubMed]
39. 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/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10642–10651.
40. Azhar, A.A.; Sutema, I.A.M.P.; Gunawan, P.H. Deep Learning-Based Exploration of YOLOv8 for Acne Vulgaris Type Classification and Lesion Counting. In Proceedings of the 2024 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), Bali, Indonesia, 17–19 December 2024; pp. 1377–1382. [CrossRef]
41. 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 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 9–12 December 2021; pp. 2407–2414. [CrossRef]
42. Rasool, M.U.; Salam, A.; Nazir, A.; Mushtaq, T. Hybrid Attention-Guided Capsule Framework for Automated Facial Acne Severity Classification. *Spectr. Eng. Sci.* **2025**, *3*, 973–1002. [CrossRef]

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.