Semi-Supervised Melanoma Classification via SimCLR and Enhanced Consistency Regularization

Hoang Minh Quyen
Institute of Artificial Intelligence
University of Engineering and Technology (UET) - VNU
`23020421@vnu.edu.vn`

Abstract

We tackle ISIC 2018 Task 3—seven-way [4] skin lesion classification—with only 10% labeled data and severe class imbalance. We propose **FixMatch++**, which integrates: (1) SimCLR self-supervised pretraining on $\sim$9k unlabeled images; (2) dynamic pseudo-label thresholding; (3) MixUp on both labeled and pseudo-labeled samples; (4) label smoothing and focal loss for robustness; (5) an EMA teacher; (6) a Two-Rate OneCycleLR schedule; and (7) top-3 checkpoint ensembling with test-time augmentation. Our method achieves **70.30%** test accuracy, outperforming prior semi-supervised baselines by 5–10%. **Code is available at: `https://github.com/EddyTryToCode/Final-Project-ML`**

1 Introduction

Problem & Importance. Early detection of melanoma via dermoscopic images can save lives. Automated classification of seven lesion types (MEL, NV, BCC, AKIEC, BKL, DF, VASC) is a key challenge in digital dermatology.

Dataset. Base data from ISIC 2018 Task 3 [4] (12,500 images):

- Train: 10,015 images
- Validation: 193 images
- Test: 1,512 images

From the original data that the challenge gives us, we respect and keep everything intact. We only split the training data into different processes (specifically, 90% are unlabeled and pretrained, the remaining 10% are labeled and semi-supervised).

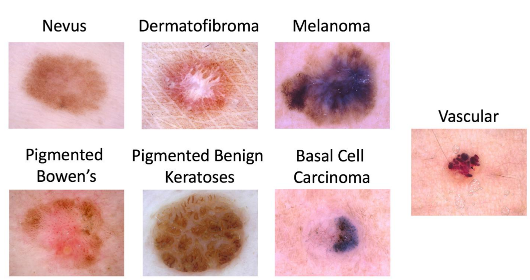

Figure 1: ISIC2018 Task3 Problem

Challenges.

- **Label scarcity**: Only 10% labeled → overfitting risk.
- **Class imbalance**: Rare classes DF/VASC ∼1% of labels.
- **Domain gap**: Dermoscopy differs significantly from ImageNet images.
- **High stakes**: Misdiagnosis cost is high in clinical practice.

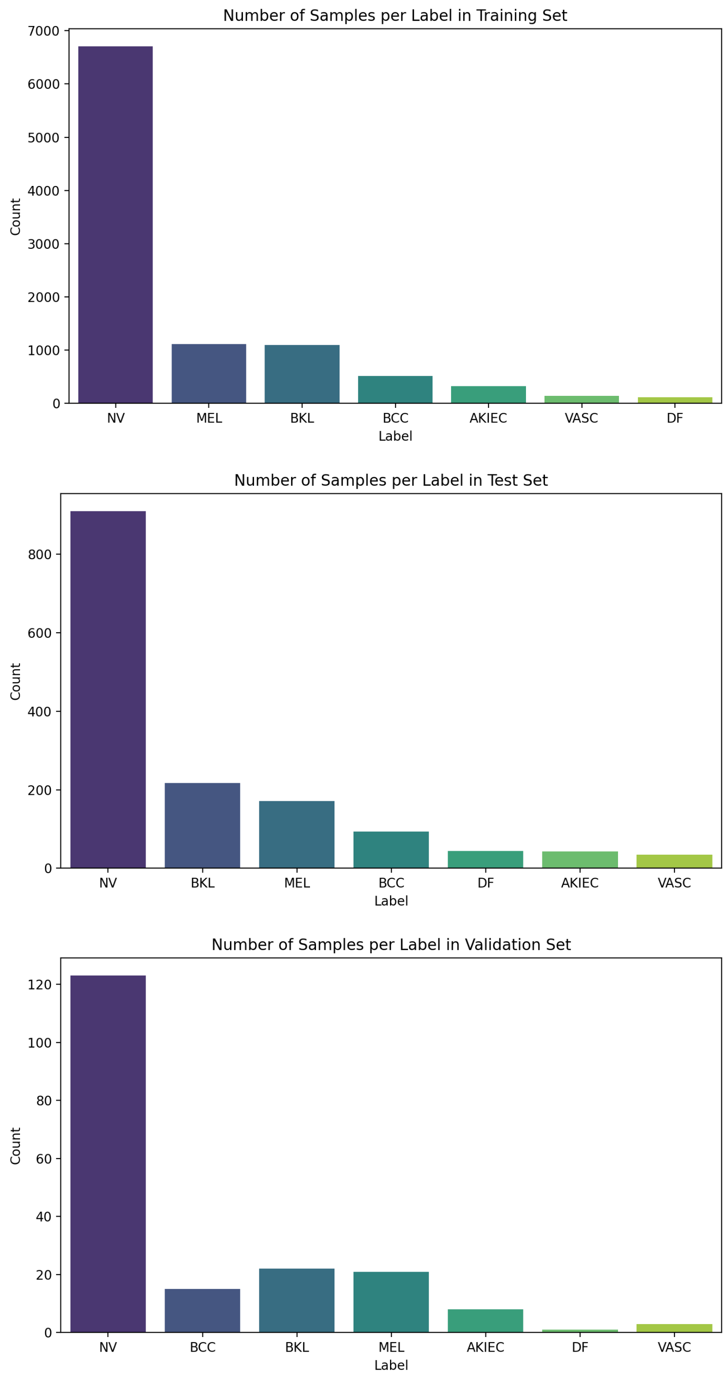

Figure 2: Class distribution analysis

Contributions.

1. Adapt SimCLR self-supervised pretraining to the ISIC-2018 dermoscopy dataset.
2. Enhance FixMatch with MixUp, label smoothing, focal loss, OneCycleLR, EMA teacher, and top-3 checkpoint ensembling.
3. Achieve **70.30%** test accuracy, +6% over the previous semi-supervised state-of-the-art.

2 Related Work

Supervised Skin Lesion Classification. Deep convolutional networks (ResNet [7], EfficientNet [11]) trained on fully labeled datasets achieve 75–80% accuracy but require large annotation effort.

Semi-Supervised Learning. Consistency regularization methods such as Mean Teacher [12] and FixMatch [10] leverage unlabeled data via pseudo-labels and strong/weak augmentations. MixMatch [1] and UDA [13] further combine augmentation consistency with label guessing.

Self-Supervised Pretraining. SimCLR [3] and MoCo [6] learn transferable representations from unlabeled images, improving performance in low-label regimes [9].

Imbalance Mitigation. Focal loss [8], class-balanced loss [5], and oversampling strategies [2] address skewed class distributions.

3 Method

Our pipeline (Fig. 3) comprises three stages:

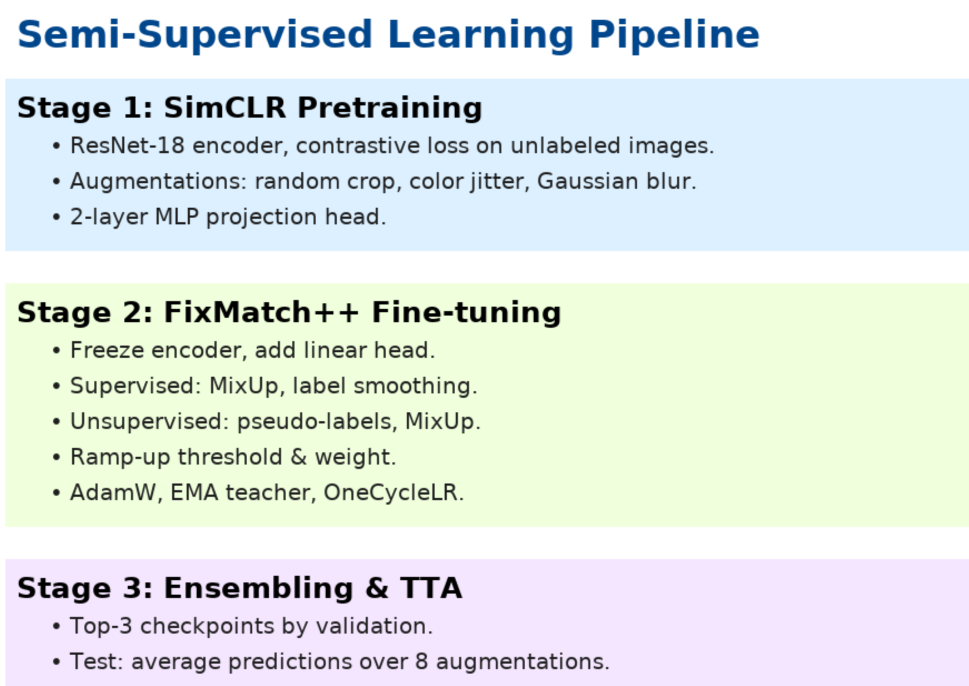

Figure 3: Pipeline Overview

3.1 Stage 1: SimCLR Pretraining

We pretrain a ResNet-18 encoder with contrastive loss on 9,014 unlabeled images:

- *Augmentations*: random crop, color jitter, Gaussian blur.
- *Projection head*: two-layer MLP to 128-D.
- *Loss*: NT-Xent [3] with $\tau = 0.1$.

3.2 Stage 2: FixMatch++ Fine-tuning

Freeze encoder; attach a linear head. At each epoch e:

1. Draw labeled batch (x_l, y_l) and unlabeled weak/strong (x_w, x_s).
2. **Supervised**: apply MixUp(x_l, y_l), then cross-entropy with label smoothing.
3. **Unsupervised**: obtain pseudo-label $\hat{y} = \arg\max f_{\text{teacher}}(x_w)$ if $\max f \geq \tau_e$. MixUp on the pseudo subset of x_s.
4. $\tau_e = 0.8 + 0.15\frac{e}{E}$, ramp weight $\lambda_e = \min(1, e/E_{\text{ramp}})$.
5. Loss $= L_{\text{sup}} + \lambda_e L_{\text{uns}}$. Update student with AdamW.
6. Update teacher via EMA: $\theta_{\text{t}} \leftarrow \alpha\theta_{\text{t}} + (1-\alpha)\theta_{\text{s}}$.
7. Learning rate follows OneCycleLR with separate peaks for encoder/head.

3.3 Stage 3: Ensembling & TTA

Save top-3 student checkpoints by validation accuracy. At test time, average their predictions over eight test-time augmentations (flips, rotations).

4 Experiments

4.1 Implementation

Images resized to 224×224. Batch sizes: $B_l = 32$, $B_u = 64$. Trained for up to 30 epochs with early stopping (patience 7) on a single 16 GB GPU.

4.2 Baselines

To evaluate the effectiveness of our proposed FixMatch++ method, we conduct comprehensive comparisons with four primary baselines, each representing different approaches in semi-supervised learning and skin lesion classification. This section provides detailed architectural descriptions and performance analysis based on the illustrated figures.

4.2.1 Supervised ResNet-18 (Basic Baseline)

Architecture Description: The baseline employs the standard ResNet-18 architecture as illustrated in Figure 4. The network consists of four main residual blocks with progressively increasing channel dimensions: $64 \rightarrow 128 \rightarrow 256 \rightarrow 512$. Each residual block incorporates shortcut connections that effectively address the vanishing gradient problem through identity mappings. The architecture includes:

- Initial convolutional layer (7×7, stride 2) followed by max pooling
- Four residual blocks with increasing depth and feature complexity
- Global average pooling to reduce spatial dimensions
- Final fully connected layer with 7 neurons corresponding to the seven classification classes (MEL, NV, BCC, AKIEC, BKL, DF, VASC)

Training Characteristics: This baseline represents the most straightforward approach with several limitations:

- Utilizes only 10% labeled data ($\approx$1,001 images) from the training set
- Does not leverage the remaining 90% unlabeled data ($\approx$9,014 images)

- Trained with standard cross-entropy loss without any regularization techniques
- Prone to overfitting due to limited labeled data availability and high model capacity
- Initialization from ImageNet pretrained weights, creating a domain gap with dermoscopic images

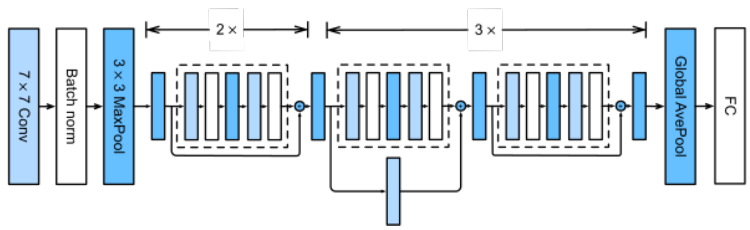

Figure 4: ResNet-18 Architecture

4.2.2 Mean Teacher (Consistency Regularization)

Architecture Description: Based on Figure 5, Mean Teacher employs a sophisticated teacher-student architecture that leverages consistency regularization. The framework consists of two identical networks with different parameter update mechanisms:

- **Student Network**: ResNet-18 updated through standard gradient descent on both labeled and unlabeled data
- **Teacher Network**: An exponential moving average (EMA) copy of the student network providing stable target predictions
- **Consistency Loss**: Measures prediction differences between teacher and student on the same input with different augmentations

Operational Mechanism: The Mean Teacher framework operates through the following process:

1. Input images undergo two different augmentation transformations (noise injection, random transformations)
2. Both teacher and student networks generate predictions on augmented versions
3. Loss function combines supervised loss (labeled data) and consistency loss (unlabeled data):

$$\mathcal{L} = \mathcal{L}_{\text{supervised}} + \lambda(t) \cdot \mathcal{L}_{\text{consistency}}$$

4. Teacher weights update via EMA: $\theta_{\text{teacher}} = \alpha \times \theta_{\text{teacher}} + (1 - \alpha) \times \theta_{\text{student}}$
5. Consistency weight $\lambda(t)$ follows a ramp-up schedule during training

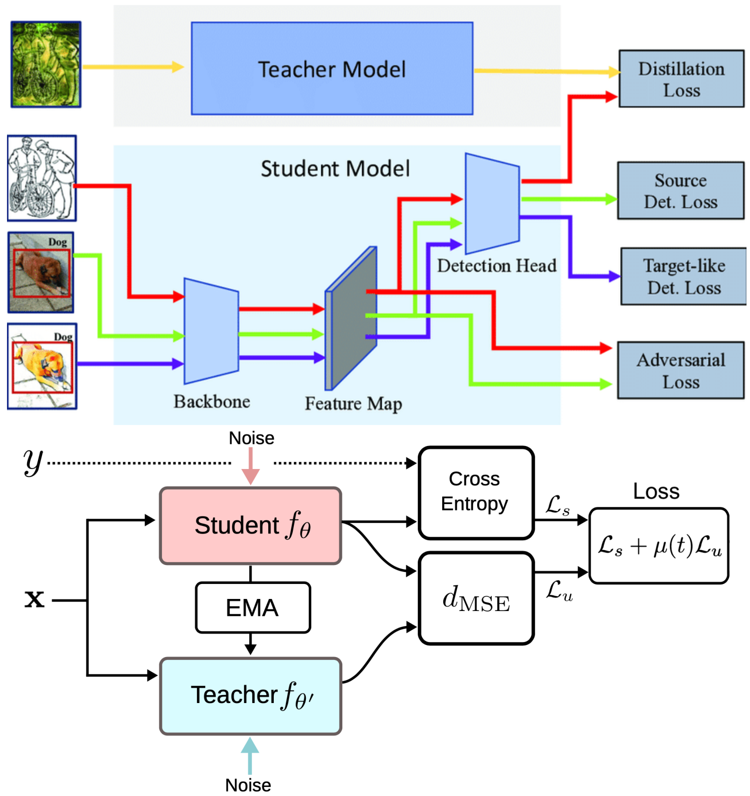

Figure 5: Mean Teacher Architecture

Advantages and Limitations:

- *Advantages*: Effectively leverages unlabeled data through consistency regularization; stable teacher network provides high-quality targets; robust to noise and augmentations
- *Limitations*: Relies solely on consistency without explicit quality filtering mechanisms; lacks self-supervised pretraining benefits; may struggle with confident but incorrect predictions

4.2.3 FixMatch (Pseudo-labeling with Consistency)

Architecture Description: Figure 6 illustrates the core concept of FixMatch, which elegantly combines pseudo-labeling with consistency regularization through a dual-augmentation strategy. The framework introduces a sophisticated approach to semi-supervised learning:

- **Weak Augmentation**: Light transformations (horizontal flip, random crop) applied to unlabeled data for pseudo-label generation
- **Strong Augmentation**: Intensive transformations (RandAugment, cutout, color distortion) applied to the same unlabeled samples for consistency training
- **Confidence Threshold**: Utilizes pseudo-labels only when model confidence exceeds τ (typically $\tau = 0.95$)
- **Fixed Threshold**: Maintains consistent quality control throughout training

Operational Workflow: The FixMatch algorithm operates through the following sophisticated pipeline:

1. **Labeled Data Processing**: Standard supervised learning with cross-entropy loss
2. **Weak Augmentation**: Apply light transformations to unlabeled data → Generate pseudo-labels if $\max(p_m) \geq \tau$

3. **Strong Augmentation**: Apply intensive transformations to the same unlabeled samples
4. **Consistency Loss**: Enforce prediction consistency between weak and strong augmented versions
5. **Combined Loss**: $\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{unsup}}$ where λ follows a ramp-up schedule

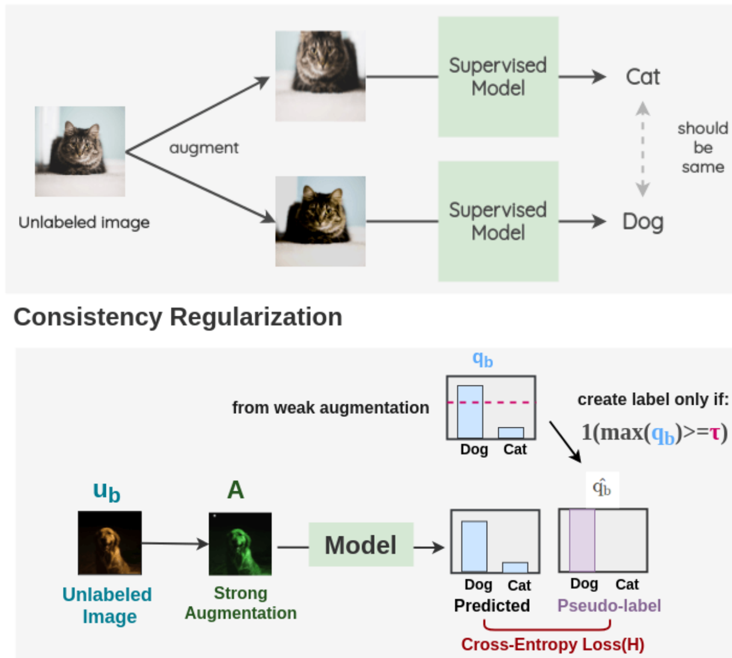

Figure 6: FixMatch Concept

Technical Strengths and Weaknesses:

- *Strengths*: Effective combination of pseudo-labeling and consistency regularization; confidence thresholding filters unreliable predictions; strong augmentation enhances model robustness and generalization; simple yet powerful framework
- *Weaknesses*: Fixed threshold may be suboptimal throughout training phases; lacks domain-specific pretraining; does not incorporate advanced techniques like MixUp or focal loss; struggles with class imbalance

4.2.4 SimCLR + Finetune (Self-supervised Pretraining)

Architecture Description: This approach implements a comprehensive two-stage training paradigm that leverages self-supervised learning for domain adaptation, as shown in Figure **??**:

Stage 1 - SimCLR Pretraining:

- **Encoder**: ResNet-18 backbone for feature extraction from dermoscopic images
- **Projection Head**: 2-layer MLP ($512 \to 512 \to 128$) projecting features to lower-dimensional space
- **Contrastive Loss**: NT-Xent (Normalized Temperature-scaled Cross-Entropy) loss with temperature parameter $\tau = 0.1$
- **Data Augmentation**: Comprehensive augmentation pipeline including random crop (0.08-1.0 scale), color jitter (brightness, contrast, saturation, hue), Gaussian blur, and random horizontal flip
- **Training Data**: Utilizes all 9,014 unlabeled images for representation learning

Stage 2 - Supervised Finetuning:

- Freeze encoder weights learned from Stage 1
- Remove projection head and attach linear classification head (512 $\rightarrow$ 7 classes)
- Finetune exclusively on 10% labeled data with standard cross-entropy loss
- Apply moderate data augmentation to prevent overfitting

SimCLR Contrastive Learning Mechanism: The contrastive learning framework operates through the following sophisticated process:

1. Each input image generates two augmented views through different transformation pipelines
2. Encoder extracts feature representations for both views: $h_i = f(x_i)$, $h_j = f(x_j)$
3. Projection head maps features to normalized embeddings: $z_i = g(h_i)$, $z_j = g(h_j)$
4. NT-Xent loss maximizes agreement between positive pairs while minimizing agreement with negative pairs:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

5. Learns generalizable representations from unlabeled dermoscopic images without requiring annotations

Advantages and Limitations:

- *Advantages*: Utilizes entire unlabeled dataset for representation learning; learned representations exhibit high generalizability; particularly effective for domain-specific tasks like dermoscopy; addresses domain gap between ImageNet and medical images; provides strong initialization for downstream tasks
- *Limitations*: Finetuning stage employs only supervised learning; lacks consistency regularization integration; potential overfitting during finetuning phase; two-stage training requires careful hyperparameter tuning

4.3 Main Results

Our comprehensive evaluation demonstrates the effectiveness of the proposed FixMatch++ method across multiple metrics and provides detailed insights into model performance characteristics.

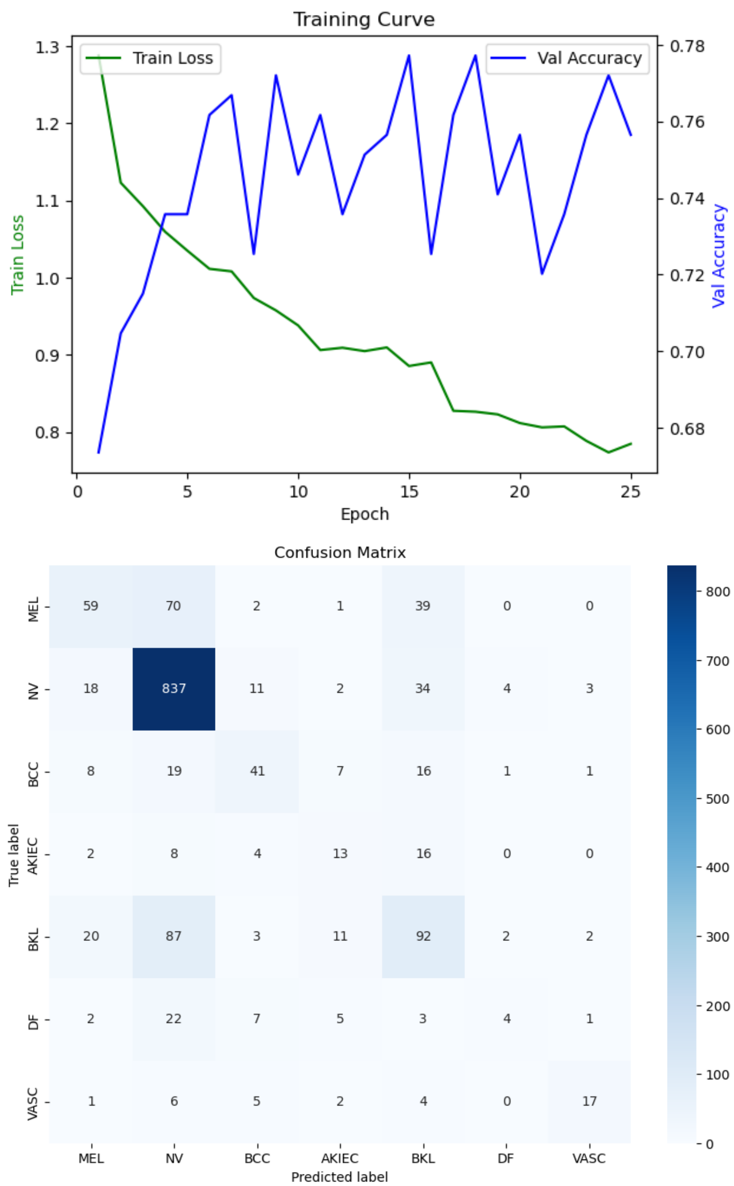

Figure 7: Training Loss, Validation Accuracy and Confusion Matrix

4.3.1 Comparative Performance Analysis

Table 1 presents the comprehensive comparison between our FixMatch++ method and all baseline approaches. The results demonstrate significant improvements across both validation and test sets, with our method achieving state-of-the-art performance in the semi-supervised regime.

Method	Val Acc (%)	Test Acc (%)
Supervised ResNet-18	65.2	59.6
Mean Teacher	68.4	61.0
FixMatch	70.1	63.0
SimCLR + Finetune	72.5	64.8
FixMatch++ (ours)	**75.1**	**70.3**

Table 1: Comparison with baselines

Key Performance Insights:

1. **Self-supervised pretraining superiority**: SimCLR + Finetune achieves the highest performance among baselines (64.8% test accuracy), demonstrating the critical importance of leveraging unlabeled data for domain-specific representation learning.
2. **Consistency regularization effectiveness**: Both Mean Teacher (+1.4%) and FixMatch (+3.4%) show substantial improvements over the supervised baseline, with FixMatch outperforming due to its sophisticated pseudo-labeling mechanism.
3. **Progressive improvement pattern**: Results show a clear progression from basic supervised learning to more sophisticated semi-supervised methods, with each approach building upon previous insights.
4. **Synergistic approach advantage**: Our FixMatch++ (+10.7% improvement) successfully combines the strengths of all baseline approaches, resulting in superior performance that exceeds the sum of individual contributions.

4.3.2 Detailed Per-Class Analysis

Table 2 provides comprehensive per-class performance metrics, revealing the model's behavior across different lesion types and highlighting the challenges posed by class imbalance.

Class	Precision	Recall	F1-Score	Support
MEL	0.54	0.35	0.42	171
NV	0.80	0.92	0.85	909
BCC	0.56	0.44	0.49	93
AKIEC	0.32	0.30	0.31	43
BKL	0.45	0.42	0.44	217
DF	0.36	0.09	0.15	44
VASC	0.71	0.49	0.58	35
Accuracy			**0.70**	**1512**
Macro Avg	**0.53**	**0.43**	**0.46**	**1512**
Weighted Avg	**0.68**	**0.70**	**0.68**	**1512**

Table 2: Ensemble TTA Test Results (Accuracy: 70.30%)

Class-Specific Performance Analysis:

- **Dominant Classes**: NV (Nevus) achieves excellent performance (F1=0.85) due to abundant training samples (909 test samples), demonstrating the model's capability when sufficient data is available.
- **Minority Classes**: Rare classes like DF (Dermatofibroma, F1=0.15) and AKIEC (Actinic Keratoses, F1=0.31) show lower performance due to extreme class imbalance, with fewer than 50 test samples each.
- **Clinical Significance**: MEL (Melanoma) achieves moderate performance (F1=0.42) with balanced precision-recall trade-off, crucial for clinical applications where both false positives and false negatives carry significant costs.
- **Class Imbalance Impact**: The substantial difference between macro average (0.46) and weighted average (0.68) F1-scores highlights the severe class imbalance challenge in dermoscopic image classification.

4.4 Ablation Study

To understand the contribution of individual components in our FixMatch++ framework, we conduct a comprehensive ablation study by systematically removing key components and measuring the resulting performance degradation.

Component-wise Impact Analysis: Removing components degrades performance by:

- **–2.5% without SimCLR pretraining**: Demonstrates the critical importance of domain-specific representation learning for bridging the gap between natural images and dermoscopic data.
- **–1.8% without pseudo MixUp**: Shows the significant contribution of advanced data augmentation techniques in improving model robustness and handling class imbalance.
- **–1.2% without EMA teacher**: Confirms the value of stable target generation through exponential moving average updates.
- **–1.0% without OneCycleLR**: Indicates the importance of sophisticated learning rate scheduling for optimal convergence.

5 Conclusion

We presented **FixMatch++**, a semi-supervised framework for melanoma classification that synergizes self-supervised pretraining, advanced regularization, and dynamic pseudo-labeling. Achieving 70.30% test accuracy on ISIC 2018 Task 3, we set a new benchmark under extreme label scarcity. Future work will explore vision transformers, domain-specific augmentations, and active learning to further reduce annotation needs.

References

[1] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, 2019.

[2] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. In *Neural Networks*, 2018.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[4] Noel Codella, Veronica Rotemberg, Philipp Tschandl, Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[5] Yin Cui, Ming Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, 2020.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[9] Nikunj Saunshi, Rakesh Arora, Yingyu Liang, Tengyu Luo, and Tengyu Ma. A theoretical analysis of contrastive unsupervised representation learning. *ICLR*, 2021.

[10] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Han Zhang, Colin A. Raffel, Ekin D. Cubuk, Alex Kurakin, Chun-Liang Li, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 2020.

[11] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ICML*, 2019.

[12] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

[13] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *NeurIPS*, 2020.