Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Advancing Dermatological Diagnosis: A Semi-Supervised EfficientNet-B7 Architecture for Skin Lesion Classification Under Annotation Scarcity

Yash Baldania
Charotar University of Science and Technology
Abhishek Patel
abhishekpatel.cse@charusat.ac.in

Charotar University of Science and Technology
Amit Thakkar
Charotar University of Science and Technology

Research Article

Keywords: Medical imaging, Deep learning, Skin lesion classification

Posted Date: June 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10004595/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Advancing Dermatological Diagnosis: A Semi-Supervised EfficientNet-B7 Architecture for Skin Lesion Classification Under Annotation Scarcity

Yash Baldania[1], Abhishek Patel[1*], Amit Thakkar[1]

[1*]Dept. of Computer Science and Engineering, Chandubhai S. Patel Institute of Technology, Charusat, Changa, Gujarat, India.

*Corresponding author(s). E-mail(s): abhishekpatel.cse@charusat.ac.in;
Contributing authors: 23cs003@charusat.edu.in; amitthakkar.it@charusat.ac.in;

Abstract

The skin lesion classification in medical imaging is a difficult task because of the lack of labelled samples, which are enumerated with high disease prevalence variance. The knowledge of labelled samples can be used in conjunction with the discriminative information of unlabeled data, with the aim of addressing the issue. Nonetheless, the creation of an effective semi-supervised learning framework for medical image data classification is hindered by two key challenges. The first one is the estimation of reliable pseudo-labels for the images listed in the unlabeled dataset. The second one is the alleviation of the biases caused by the class imbalance. Accordingly, this paper introduces a semi-supervised framework for medical image data classification. Utilising the HAM10000 dataset, our approach employs a vast EfficientNet-B7 backbone trained on just 2,000 labelled samples, leveraging information enumerated in over 8000 unlabeled images through self-ensembling on Mean Teacher. Encouraging the student model to remain consistent with high-confidence pseudo-labels created by an EMA-based teacher model for strongly augmented inputs. Inspired from FixMatch, the use of focal loss with dynamic class weights and label smoothing is beneficial for supervised learning as it helps with severe class imbalance and mining hard examples. The augmenting technique MixUp further helps provide regularization along with a method called SWA. The convergence occurs within 23 epochs, and achieve an AUC (macro-averaged) of 0.9813 and a macro-averaged F1 score of 0.82. Each of the seven diagnostic classes has a per-class AUC of over 0.95, including the clinically important melanoma (AUC 0.977) and the low-prevalence dermatofibroma (AUC 0.996). This project seeks to design automated diagnostic systems to detect skin cancer in a clinically reliable manner in the early stages.

Keywords: Medical imaging, Deep learning, Skin lesion classification

1 Introduction

The role of Computer Aided Diagnosis is essential for the improvement of efficiency diagnostic and appropriate diagnosis reduction. Hence, it is something that is of great importance and interest to researchers. Over the last few years, many deep learning research works exhibit promising results in the classification of medical images Zhang et al. (2019); Huang et al. (2019); Sun et al. (2021).

But, the performance of models rely heavily on labelled data He et al. (2016). In practice, manually labelling medical images is an expensive option for training deep models. High-quality data which is labelled, is known to be a slow and expert-requiring process by medical professionals et al. (2017). Due to this, good-quality labelled data is scarce for many diseases. However, huge amount of unlabeled data is readily available in hospitals and clinical databases. Semi-supervised learning Rosenberg et al. (2005); Grandvalet and Bengio (2004); et al. (2019, 2020) is useful in such scenarios where the labelled data is limited and dealing is with more unlabeled data.

When labelled data is limited, semi-supervised learning may be able to enhance network performance. At present, most semi-supervised methods that have adopted the consistency enforcing strategy Laine and Aila (2017); Tarvainen and Valpola (2017), which exploits unlabeled data by controlling network predictions consistent under input or network weight perturbations. Based on an input sample, consistency enforcing methods create various perturbed samples of the same input (e.g. by adding Gaussian noise), and then encourage the model predictions to be similar on these samples Laine and Aila (2017). Among the existing methods, former consistency-based methods majorly utilize pre-trained teacher model to create high-quality consistency targets, which is reliable predictions for perturbed versions of the same input. The method of Temporal Ensembling (TE) Laine and Aila (2017), for example, uses the exponential moving average (EMA) of predictions in various epochs as the consistency targets. However, to overcome the limitation of creating consistency targets manually, the Mean Teacher (MT) Tarvainen and Valpola (2017) framework builds an ensembled teacher model that can generate these targets on the fly. This approach has been empirically proven as state-of-the-art on various applications Tarvainen and Valpola (2017); et al. (2019); Perone and Cohen-Adad (2018); et al. (2019).

Thus, we introduced a semi-supervised approach for classifying medical image data by learning from relationships between different samples. We built our model based on established consistency-based methods Tarvainen and Valpola (2017) and developed a complete training system that combines two key techniques: Mean Teacher consistency and self-regularization. We use EfficientNet-B7 as our primary feature extractor to classify skin lesions across multiple categories using the HAM10000 dermatology image data Tschandl et al. (2018); Tan and Le (2019). One major challenge we encountered is the class imbalance. For addressing it, we used three integrated strategies: data augmentation, regularization, and careful filtering of pseudo-labels during training. Our approach uses MixUp augmentation et al. (2018), which creates new training examples by blending pairs of images and their labels. This technique helps minority classes to have more influence on the model's decision-making process, even when we have fewer examples of those classes. We also applied a technique, label smoothing et al. (2016), to prevent the model from becoming overconfident in its predictions. This acts as a safeguard against overfitting to the majority classes. Additionally, we apply various transformations to the images, rotating, flipping, and adjusting brightness and contrast, which, combined with our consistency requirements, help the model learn more diverse features across all disease types. Our proposed work includes a Exponential Moving Average (EMA) teacher model to generate stable predictions on unlabeled data. We adopt a strict confidence threshold ($\tau = 0.95$) before accepting these predictions, which prevents the model from learning from unreliable guesses, especially in rare conditions. Then after, we use cosine annealing Loshchilov and Hutter (2016) for adjusting the learning rate during training and optionally apply Stochastic Weight Averaging (SWA) et al. (2018). These techniques helped the model in finding better solutions for classes which are rare.

The three main contributions done by our methodology are:

Semi-supervised learning has good performance when the data is limited: In this paper, after using supervised learning with a consistency-based teacher modelTarvainen and Valpola (2017), there is a significant improvement in classification accuracy even when the labelled data is limited. For handling class imbalance we used weighted cross-entropy for labelled data. In case of unlabeled data, the consistency Tarvainen and Valpola (2017) is enforced when the teacher model is very confident, which prevents

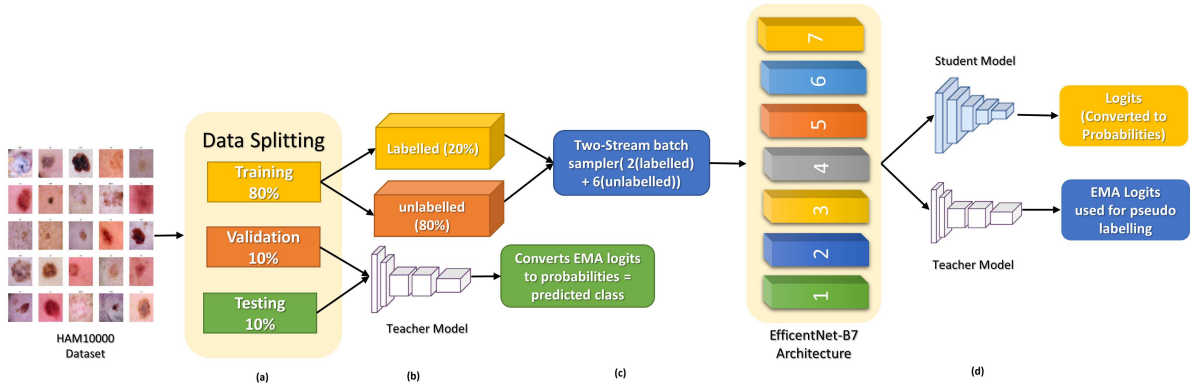


Fig. 1 Semi-supervised learning pipeline architecture for HAM10000 skin lesion classification. The framework utilizes dual networks where (a) data partitioning yields labeled (20%) and unlabeled (80%) subsets, (b) balanced batch sampling feeds student and teacher networks, (c) supervised loss trains the student on labeled samples while the teacher produces pseudo-labels for unlabeled samples via EMA weight averaging, and (d) consistency regularization synchronizes student predictions on strongly-augmented images with teacher predictions on weakly-augmented counterparts.

the model from learning incorrect patterns when the predictions are uncertain.

The framework we built: We have made a framework that brings together several techniques. First of all, for blending images together, we used Mixupet al. (2018) for creating synthetic training samples. Second, we apply strong augmentations (rotations, flips, colour adjustments) et al. (2025) to unlabeled images. Third, we use a confidence-based consistency loss that combines mean squared error with KL divergence and cross-entropy. Then, after, the parameters of the teacher model are progressively updated via exponential moving average (EMA). When these components work together, they create reliable training signals and prevent overfitting, which is crucial when dealing with limited labelled data and unbalanced classes in medical imaging.

Based on our work, what others can do: We have built a pipeline that includes several improvements for increasing feature learning and efficiency: mixed-precision training et al. (2018) for faster computation, cosine annealing Loshchilov and Hutter (2016) with warm restarts for the learning rate, and a high exponential moving average decay rate (0.9999), which keeps the teacher model stable. At the time of validation, teacher model is only used for getting consistent performance during training epochs. This framework give a start to other medical imaging projects which struggle with scarce labelled data.

2 Related Works

2.1 Semi-Supervised Learning (SSL) in Medical Imaging

SSL is increasingly used in the area of medical imaging because of its ability to classify image samples even when the labelled data is limited. One important framework given by Tarvainen and Valpola Tarvainen and Valpola (2017) was Mean Teacher model, where student and teacher models where used; the student model was trained normally, but in the teacher model, weights were calculated as the average of the student's weights over time using EMA. By applying the same concept, the teacher and student models are encouraged to give similar predictions, which helps the network learn useful features even when very few labelled images are available.

Perone and Cohen-Adad Perone and Cohen-Adad (2018) formulated weight-averaged consistency targets for 3D medical imagery, whereas Cui et al. (2019) adapted MT for brain lesion analysis by addressing instabilities in 3D volumetric predictions. Our methodology extends these insights by augmenting teacher reliability through intensified augmentations and adaptive weighting, ensuring the teacher preserves global structural priors despite severe transformations.

Recently, Zhang et al. (2025) developed a Dual Teacher Student system that generates pseudo-labels from two different sources: one on the basis

of temporal averaging and another based on agreement between different model architectures. Using two teachers instead of one, helps reduce a common problem called confirmation bias, where the model keeps reinforcing its own mistakes during self-training. Drawing inspiration from this idea, we encouraged the model to produce consistent outputs for different augmentations of the same image while also stabilising the features the model learns. This dual approach is particularly valuable for skin lesion images, which can look quite different depending on lighting conditions and how zoomed in the camera is, even when it's the same lesion. In the field of histopathology (microscopic tissue analysis), Jiang et al. (2023) showed that iteratively refining pseudo-labels while enforcing consistency can make much better use of unlabeled data. The key was setting strict confidence thresholds to avoid accepting low-quality predictions. From this, we applied a similar philosophy in our work. When we create heavily augmented versions of images with extreme rotations, colour shifts, or crops—the model might make less confident predictions. After the removal of predictions for which the model is not confident, the training process relies totally on reliable pseudo-labels. So that the model learns from better and more accurate samples, ultimately improving the training accuracy in our approach.

2.2 Consistency Regularization Strategies

Consistency regularization is an important concept in modern semi-supervised learning. The idea of consistency regularization is such that the model must give similar predictions even when perturbations are applied for example, rotated, brightened, or blurred). A work by Laine and Aila Laine and Aila (2017) stated Temporal Ensembling, where predictions were collected over time to guide learning process. Later, FixMatch et al. (2020) by Sohn combined the idea with confidence-based pseudo-labels, which states that only confident predictions are used during the training process, which consequently showed that good quality teacher predictions are important when the labelled data is scarce. By following this same approach, the model produces consistent predictions for weakly as well as strongly augmented versions of the same image, ultimately helping the model to learn strong and reliable features from dermoscopic images, which vary in texture, lighting, and lesion shape. However, current studies showed like Fan et al. (2021) stated, that enforcing predictions to be consistent under perturbations is not always ideal; instead, adapting the model to changes can lead to better feature learning. For example, some methodologies focus on preserving relative spatial relationships instead of similar outputs. Whereas, Mai stated RankMatch et al. (2023), which applied ranking-based constraints at the pixel level to improve segmentation quality. From the above ideas, our framework uses augmentation-aware consistency losses, which helped our model to keep lesion structure and spatial relationships intact, even though still producing stable predictions under perturbations. As stated by Wang in PE-MT et al. (2024) showed that using stronger and more varied augmentations improved learning. By adopting a similar approach, we applied geometric and colour transformations which helped in expanding the decision boundary of the model, reducing overfitting and ultimately improving model performance even when the labelled data is scarce.

2.3 Addressing Class Imbalance

Class imbalance is a difficult task in medical imaging. In datasets such as HAM10000, some classes, such as akiec, vasc and df, have very few samples compared to nv. Due to which model focuses more on the majority samples rather than the minority samples. For handling this, we used Focal loss as stated by Lin et al. (2017), which helps the model to focus more on minority samples rather than majority ones. This helps improve the detection of minority classes such as vasc and df, which are often missed when using standard cross-entropy loss. Even some augmentation strategies such as rotation, flipping and shearing, increase the number of training examples, but doesn't address the class imbalance problem fully. But some techniques like MixUp et al. (2018), combine two images to create new training samples, which helps the model generalize better. From this, we used MixUp et al. (2018) and colour jittering during training process. These techniques help the model learn smoother decision boundaries, thereby reducing the tendency of overfitting

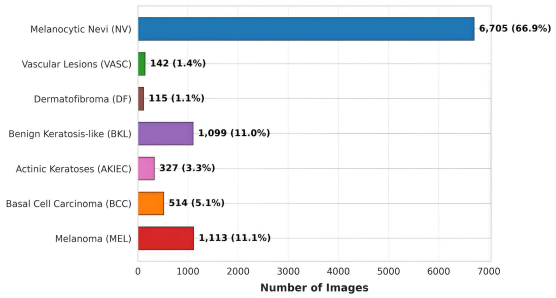


Fig. 2 Class distribution in the HAM10000 dataset showing severe class imbalance. Melanocytic Nevi (NV) comprises 66.9% of the dataset with 6,705 images, while minority classes include Vascular Lesions (VASC) at 1.4% (142 images) and Dermatofibroma (DF) at 1.1% (115 images).

to visual patterns of the majority classes. Additionally, FixMatch-LS et al. (2023) demonstrated that label smoothing mitigates overconfidence in pseudo-labels, especially for scarce classes. Building upon this, our proposed method utilizes an adaptive confidence thresholding mechanism, ensuring that pseudo-labels for minority categories are neither ignored nor excessively trusted during early training phases.

2.4 Optimization and Regularization Techniques

Contemporary optimization strategies play critical roles in stabilizing semi-supervised pipelines. Izmailov et al. (2018) introduced Stochastic Weight Averaging (SWA), which averages model weights across training trajectories to converge to wider optima, yielding superior generalization. We employ SWA in the terminal training stages of our method, enabling robust convergence despite noisy pseudo-label supervision. Parallelized adaptations of SWA and cosine annealing learn ing rate schedules further accelerate convergence while pre serving generalization. In this paper, we used cosine learning rate decay with warm restarts to keep training stable. This was useful when the quality of pseudo-labels changes during training, preventing the model from learning too fast or too quickly. We have also used mixed-precision training allowing us to train larger and deeper models more efficiently without losing accuracy. Mixed-precision also reduces GPU memory usage, allowing to process high-resolution dermoscopic images and also capturing fine lesion details. From all these techniques we effectively handle the problem of class imbalance, improving pseudo-labelling, and also achieving stronger generalization even when labelled data is scarce.

3 METHODOLOGY

3.1 Dataset Description and Preprocessing

The HAM10000 dataset (Human Against Machine with 10000 training images) spans seven diagnostic categories: akiec, bcc, bkl, df, mel, nv, and vasc. (Fig. 2) presents the number and percentage distribution of images across each diagnostic class. A significant class imbalance is observed, with more than two-thirds of the samples belonging to the nv category. The dataset is partitioned into 80% training, 10% validation, and 10% testing sets. The HAM10000 repository Tschandl et al. (2018) contains 10,015 dermatoscopic images across seven diagnostic categories (mel, nv, bcc, akiec, bkl, df, vasc), exhibiting significant class imbalance (e.g., melanocytic nevus $\sim$ 67%, dermatofibroma $\sim$ 1.15%). A subset comprising 2,000 annotated images was designated for supervised learning, while the remaining training data was treated as unannotated for semi-supervised training. All imagery underwent resizing to 480×480 pixels to preserve high-resolution details, with pixel values normalized using ImageNet mean and standard deviation ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$). To counteract class imbalance, class weights were computed from the annotated training set via a balanced-weighting scheme (e.g., inversely proportional to class frequency).

3.2 Model Architecture

EfficientNet-B7 with adversarial training (AdvProp) pre-trained on ImageNet-1K served as the feature extraction backbone, as illustrated in 1. This compound-scaled architecture employs depth coefficient $\alpha = 2.0$, width coefficient $\beta = 1.4$, and resolution coefficient $\gamma = 1.5$, optimizing the relationship between network capacity and computational efficiency. The terminal fully-connected classification layer was modified to output $C = 7$ class logits corresponding to HAM10000 diagnostic categories, initialized with

Xavier uniform initialization. Dropout regularization with rate $p = 0.3$ was applied preceding the classification head to prevent overfitting during fine-tuning.

3.3 Semi-Supervised Learning Framework

A dual-model student-teacher framework et al. (2025) constitutes our implementation:

- Student model parameters denoted as θ_s and teacher parameters as θ_t.
- Teacher updates occur at each iteration via exponential moving average (EMA):

$$\theta_t^{(t)} = \alpha\theta_t^{(t-1)} + (1-\alpha)\theta_s^{(t)}, \tag{1}$$

with EMA decay $\alpha = 0.9999$. For each labeled sample x_i with ground-truth label y_i, we generate two views: a weakly augmented view x_i^w and a strongly augmented view x_i^s. The teacher processes x_i^w while the student processes x_i^s. Supervised loss computation occurs on student output versus ground truth. For each unlabeled sample x_j:

1. Calculate teacher softmax probabilities $p_t(c \mid x_j^w)$.
2. When $\max_c p_t(c \mid x_j^w) \geq \tau$ (with threshold $\tau = 0.95$), establish a pseudo-label

$$\hat{y}_j = \arg\max_c p_t(c \mid x_j^w). \tag{2}$$

3. Otherwise, exclude the sample (omit from unsupervised loss).
4. Calculate student predictions on x_j^s and enforce consistency et al. (2020) between student output and pseudo-label/teacher distribution.

3.4 Loss Functions

Supervised loss (incorporating class balancing and label smoothing): We employ a Focal Loss variant with label smoothing et al. (2024). Let $p_i = p(y_i \mid x_i; \theta_s)$ represent the student's predicted probability for the true class of labeled sample x_i. Let w_{y_i} represent the class weight for class y_i. With smoothing factor $\varepsilon = 0.1$ and $C = 7$ classes, we define the smoothed target q_i as:

$$q_i(c) = \begin{cases} 1-\varepsilon, & \text{if } c = y_i, \\ \frac{\varepsilon}{C-1}, & \text{otherwise.} \end{cases} \tag{3}$$

The supervised loss becomes:

$$L_{\text{supervised}} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i}(1-p_i)^{\gamma}\log p_i \tag{4}$$

with a smoothing penalty (KL divergence) et al. (2023) between student output and q_i. The combined formulation:

$$L_{\text{smooth}} = (1-\alpha_s)L_{\text{focal}} + \alpha_s \text{KL}(p_s \parallel q_i), \tag{5}$$

where $\alpha_s = 0.1$ balances focal loss and label smoothing.

Unsupervised consistency loss: For high-confidence unlabeled samples (set $M = \{j : \max_c p_t(c \mid x_j^w) \geq \tau\}$), the loss becomes et al. (2021):

$$\begin{aligned} L_{\text{cons}} = \frac{1}{|M|}\sum_{j\in M}\Bigg[&\text{CE}(f_s(x_j^s), \\ &\hat{y}_j) + \frac{1}{2}\text{MSE}(p_s(x_j^s), \\ &+ p_t(x_j^w))\frac{1}{2}\text{KL}(p_s(x_j^s) \\ &\parallel p_t(x_j^w))\Bigg], \end{aligned} \tag{6}$$

where f_s represents student logits, p_s and p_t represent student and teacher softmax outputs, CE denotes cross-entropy with the pseudo-label, MSE represents mean-squared error between soft distributions, and KL denotes Kullback-Leibler divergence.

Complete Training Objective: The comprehensive objective combines supervised and unsupervised losses:

$$L_{\text{total}} = L_{\text{smooth}} + \lambda_{\text{cons}}(t) \cdot L_{\text{cons}} \tag{7}$$

where $\lambda_{\text{cons}}(t)$ increases from 0 to $\lambda_{\max}$ over the first T_{ramp} epochs (e.g., linear ramp-up over 80 epochs), then remains constant.

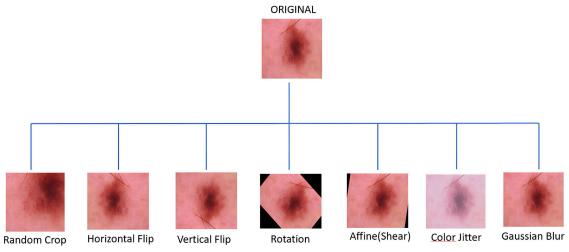


Fig. 3 Data augmentation strategies for skin lesion images, encompassing random crop, horizontal flip, vertical flip, rotation, affine transformation (shear), colour jitter, and Gaussian blur.

3.5 Data Augmentation and MixUp

- Intensive augmentation (for student inputs) et al. (2021, 2025) encompasses random resized crop (scale [0.08, 1.0]), random horizontal and vertical flips, random rotation (0–360°), random affine (shear ±20°, scale [0.8, 1.2]), color jitter (brightness, contrast, saturation ±0.4; hue ±0.1), and Gaussian blur (kernel 5×5, σ from Uniform(0.1, 2.0)).
- Minimal augmentation (for teacher inputs and validation) encompasses resizing to 512×512, centre crop to 480×480, and standard normalization.
- MixUp augmentation et al. (2022) applies with probability $p_{\text{mix}} = 0.5$ on labeled data. For two samples (x_i, y_i) and (x_j, y_j), the mixed sample becomes

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j, \quad \tilde{y} = \lambda y_i + (1-\lambda)y_j, \quad (8)$$

with $\lambda \sim \text{Beta}(\alpha, \alpha)$, $\alpha = 0.4$ (Fig. 3).

3.6 Optimization and Training Protocol

- Batch sampler extracts per iteration: $B_\ell = 64$ labeled samples + $B_u = 64$ unlabeled samples $\rightarrow$ total batch size $B = 128$.
- Optimizer: AdamW with learning rate $\eta_0 = 3 \times 10^{-4}$, weight decay $\lambda_{\text{wd}} = 1 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$.
- Mixed-precision training (autocast) diminishes memory and accelerates training.

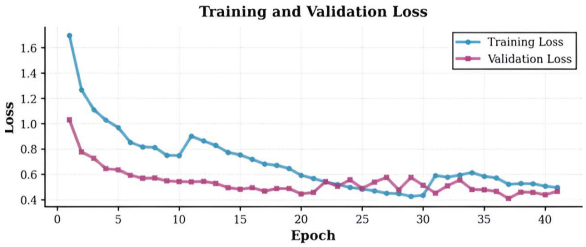


Fig. 4 Train and validation loss trajectories over 41 epochs demonstrating convergence near epoch 25, with terminal training loss approximately 0.5 and validation loss stabilizing at 0.45-0.50.

- Learning rate schedule: Cosine Annealing with Warm Restarts (SGDR) —

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})(1 + \cos(\frac{T_{\text{cur}}}{T_i}\pi)), \quad (9)$$

with $\eta_{\max} = 3 \times 10^{-4}$, $\eta_{\min} = 0$, initial cycle $T_0 = 10$ epochs, and cycle multiplier $T_{\text{mult}} = 2$.

- Following epoch $t_{\text{SWA}} = 120$, we apply Stochastic Weight Averaging et al. (2018): at each epoch $t \geq t_{\text{SWA}}$, update SWA parameters θ_{SWA} by averaging student weights:

$$\theta_{\text{SWA}}^{(t)} = \frac{n_{\text{models}}\theta_{\text{SWA}}^{(t-1)} + \theta_s^{(t)}}{n_{\text{models}} + 1}, \quad (10)$$

where n_{models} increments per snapshot. For final evaluation, we utilize the SWA-averaged model (optionally with teacher-student ensembling) for enhanced robustness and generalization.

4 Results and Discussion

4.1 Training and Validation Loss Behaviour

The model demonstrates stable loss characteristics throughout training (Fig. 4).

Initial Phase (Epochs 1–10): Training Loss (TrL) and Validation Loss (ValL) undergo rapid descent. TrL declines from 1.6969 to 0.7483, and ValL declines from 1.0309 to 0.5431. Validation loss maintains lower values than training loss during this phase, suggesting favourable generalisation without substantial overfitting.

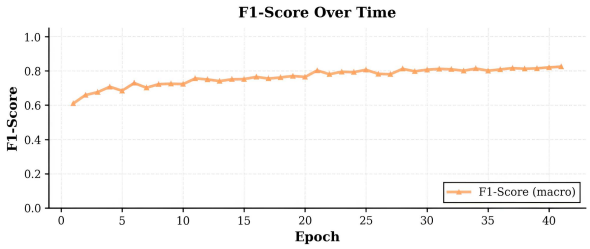


Fig. 5 Macro F1-score trajectory over 41 training epochs, demonstrating steady improvement from 0.62 at epoch 0 to 0.82 at epoch 40, with most gains by epoch 25.

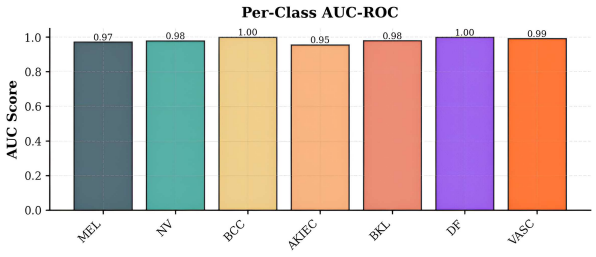


Fig. 6 Per-class AUC-ROC scores demonstrating excellent classification performance across all seven categories. BCC, DF, and VASC attain perfect scores of 1.00, while AKIEC achieves the lowest at 0.95.

Mid-Training Stability (Epochs 11–30): Following the first learning rate reset at Epoch 11, losses stabilize and continue gradual descent. ValL fluctuates but generally remains around 0.45 to 0.55.

Optimal Performance (Epochs 31–41): Substantial improvement characterizes this phase. Minimum Validation Loss of 0.4086 occurs at Epoch 37, indicating optimal generalization. Training loss decreases consistently to 0.4964 by Epoch 41, while validation loss remains low, suggesting a well-trained model without severe overfitting. Terminal ValL reaches 0.4665 at Epoch 41.

4.2 F1 Score Performance

The F1 score (Fig. 5), critical for imbalanced classification, demonstrates strong consistent improvement, reflecting improved precision-recall balance.

Overall F1 Trend: F1 score increases rapidly in early epochs, from 0.6095 (Epoch 1) to above 0.70 by Epoch 4. The model continues to improve steadily, crossing 0.80 around Epoch 21 and then reaching the best performance of 0.83 at epoch 41.

Per-Class F1: Per-class analysis reveals high performance across most classes, with several exceeding 0.80. The highest F1 scores belong to Nevus (NV) and Dermatofibroma (DF), approaching or exceeding 0.85, indicating robust classification across different categories.

4.3 AUC-ROC Analysis

The ROC curve shows how sensitivity and false positive rate change across all thresholds. AUC-ROC quantifies overall discriminative ability (Fig. 6).

Aggregate Performance: Overall, AUC-ROC stability remains high, achieving a maximum of 0.9853 at epoch 37 and stabilising near 0.9816 by the final epoch. This confirms excellent discriminative ability across the dataset.

Per-Class Performance: One-versus-rest ROC curves demonstrate excellent performance across all seven classes. All class-specific AUC values remain very high, with most surpassing 0.96. The model achieves near-perfect discrimination for Dermatofibroma (DF) with an AUC of 0.998, and high scores for Nevus (NV), 0.978 and Basal Cell Carcinoma (BCC), 1.000. Curves cluster near the upper-left corner, indicating high TPR(True Positive rate) at very low FPR(False Positive Rate), confirming high-quality probability outputs.

4.4 PR Curve Analysis

(Fig. 7) presents the Precision–Recall curve, which is especially well suited for evaluating imbalanced multi-class classification tasks, highlighting the relationship between precision and recall.

High Average Precision: The model demonstrates outstanding AP for nearly all classes, with values exceeding 0.95. NV achieves the highest AP at 0.989. Other stable classes include BCC (0.960), DF (0.974), and BKL (0.941).

Table 1 Analysis with 20% labeled data.

Epochs	F1 Score	Training Loss	Validation Loss
1	0.6095	1.6969	1.0309
10	0.7230	0.7483	0.5431
30	0.8074	0.4348	0.5139
40	0.8208	0.5061	0.4376
41	0.8252	0.4964	0.4665

Table 2 Performance Comparison with State-of-the-Art Methods

Method	Precision	Recall	F1-Score	Accuracy
STFL et al. (2024)	80.67%	81.37%	80.93%	81.37%
EfficientNet TL et al. (2022)	88.00%	88.00%	**87.00%**	88.00%
MWNL-CLS et al. (2021)	—	81.90%	—	87.50%
Ours	**91.38%**	**89.90%**	82.52%	**90.75%**

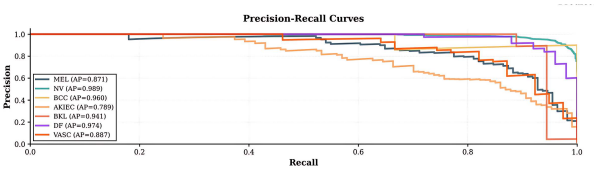


Fig. 7 Precision–recall curves for all skin lesion classes, highlighting strong model performance with high average precision across categories.

Curve Shape: PR curves position high on the plot, meaning the model maintains very high precision even as recall approaches 1.000. This proves critical—most positive predictions remain correct (high precision), and the model identifies a large proportion of true positives (high recall).

Class-Specific Results: Slightly lower AP appears for MEL (0.871) and AKIEC (0.789). While still strong, this suggests the model faces relative challenges balancing Precision and Recall for the least represented or most complex classes. Nevertheless, these high AP values still indicate robust real-world performance for these critical diagnostic categories.

4.5 Training Statistics

Table 1 Epoch-wise analysis of Training loss, Validation loss and F1 score of our method on the HAM10000 dataset.

4.6 Comparative Performance Analysis

The model achieved a maximum area in the (AUC-ROC) curve of 0.9853 on the validation set. This result is superior to several prominent state-of-the-art models applied to the 7-class HAM10000 classification task, which typically report AUC-ROC values in the range of 0.94 to 0.97 for similar multi-class settings. Table 2 compares our method against state-of-the-art approaches on HAM10000.

5 Conclusion

This investigation presented a semi-supervised learning framework integrating Mean Teacher consistency regularization with self-regularized consistency for dermatological lesion classification on the HAM10000 dataset. Leveraging EfficientNet-B7 as the feature extraction backbone, the approach addressed class imbalance through focal loss, label smoothing, and MixUp augmentation while utilizing only 2,000 annotated instances alongside unannotated data. The methodology demonstrated strong performance across diagnostic metrics, including macro-averaged AUC-ROC, precision, recall, F1-score, and specificity, validating the effectiveness of consistency regularization in medical imaging applications with limited annotations. However, overall classification accuracy remains a critical target for improvement

to achieve clinical deployment standards, particularly for challenging minority classes where morphological similarities create diagnostic ambiguity.

Acknowledgments. There is no acknowledgement involved in this work.

Authors contribution. All authors contributed equally to this work.

Funding. No funding is involved in this work.

Data availability. The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests. The authors declare no competing interests.

References

Zhang, J., Xie, Y., Xia, Y., Shen, C.: Attention residual learning for skin lesion classification. IEEE Transactions on Medical Imaging **38**(9), 2092–2103 (2019)

Huang, Q., Li, W., Zhang, B., Li, Q., Tao, R., Lovell, N.H.: Blood cell classification based on hyperspectral imaging with modulated gabor and cnn. IEEE Journal of Biomedical and Health Informatics **24**(1), 160–170 (2019)

Sun, Q., Huang, C., Chen, M., Xu, H., Yang, Y.: Skin lesion classification using additional patient information. BioMed Research International **2021**, 5558442 (2021)

He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of CVPR, pp. 770–778 (2016)

al., G.L.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)

Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-supervised self-training of object detection models. In: WACV, pp. 29–36 (2005)

Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: NIPS (2004)

al., D.B.: Mixmatch: A holistic approach to semi-supervised learning. In: NeurIPS (2019)

al., K.S.: Fixmatch. In: NeurIPS (2020)

Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: ICLR (2017)

Tarvainen, A., Valpola, H.: Mean teachers are better role models. In: NIPS, pp. 1195–1204 (2017)

al., W.C.: Semi-supervised brain lesion segmentation with an adapted mean teacher model. In: IPMI, pp. 554–565 (2019)

Perone, C.S., Cohen-Adad, J.: Deep semi-supervised segmentation with weight-averaged consistency targets. In: DLMIA, pp. 12–19 (2018)

al., L.Y.: Uncertainty-aware self-ensembling model. In: MICCAI, pp. 605–613 (2019)

Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset. Scientific Data **5** (2018)

Tan, M., Le, Q.V.: Efficientnet. arXiv:1905.11946 (2019)

al., H.Z.: mixup. In: ICLR (2018)

al., C.S.: Rethinking the inception architecture. In: CVPR, pp. 2818–2826 (2016)

Loshchilov, I., Hutter, F.: Sgdr. arXiv:1608.03983 (2016)

al., P.I.: Averaging weights leads to wider optima. arXiv:1803.05407 (2018)

al., X.Q.: Mediaug. In: MIUA (2025)

al., P.M.: Mixed precision training. In: ICLR (2018)

al., P.Z.: Dual teacher-student learning. arXiv:2505.11018 (2025)

al., Y.J.: Tumor histopathological ssl. Frontiers in Oncology **12** (2023)

al., Y.F.: Revisiting consistency regularization. arXiv:2112.05825 (2021)

al., H.M.: Rankmatch. In: ICCV, pp. 12876–12886 (2023)

al., W.W.: Pe-mt. In: MICCAI, pp. 379–389 (2024)

al., T.-Y.L.: Focal loss. In: ICCV, pp. 2999–3007 (2017)

al., S.Z.: Fixmatch-ls. Biomedical Signal Processing and Control **84**, 104709 (2023)

al., Q.L.: Semi-supervised medical image classification. IEEE Transactions on Medical Imaging **39**(11), 3429–3440 (2020)

al., A.B.: Class imbalance in skin lesion segmentation. IEEE Access **12**, 109348–109360 (2024)

al., J.S.: Mixup in neuroimaging. NeuroImage: Clinical **36**, 103214 (2022)

al., W.Y.: Self feedback threshold focal learning.

Discover Oncology **15** (2024)
al., K.A.: Skin cancer classification using efficientnet. Neuroscience Informatics **2**(3) (2022)
al., P.Y.: Deep learning on imbalanced datasets. IEEE Transactions on Medical Imaging **40**(10), 2656–2668 (2021)