DAL: Dynamic Angular Loss for Imbalanced Medical Image Classification

Salman Mohammad SALMAN.MOHAMMAD@UZH.CH
Furkan Kasım FURKAN.KASIM@UZH.CH
Manuel Günther MANUEL.GUENTHER@UZH.CH
Department of Informatics, University of Zurich, Andreasstrasse 15, 8050 Zurich

Editors: Under Review for MIDL 2026

Abstract

Class imbalance remains a major obstacle in medical image classification, where rare but clinically important classes are often overshadowed by majority categories. Despite current strategies for reducing bias, including angular-margin losses such as ArcFace, offer strong discriminative features, their behavior under small amounts of severely imbalanced classes is insufficiently understood. Although previous approaches estimate intrinsic parameters of these losses, their assumptions are based on large amounts of classes, which do not translate into medical cases. We introduce a *dynamic angular loss* (DAL) to generalize these parameters using analytically derived negative-angle statistics, and combine it with a batch-adaptive angular margin and dynamically-weighted cross-entropy. Across three highly imbalanced medical image classification benchmarks, our method consistently achieves superior balanced accuracy scores, and the lowest coefficient of variation of per-class recall. Although this comes with a small performance reduction on majority classes, gains on minority categories are substantial, resulting in a reliable and equitable overall classifier.

Keywords: Image Classification, Class Imbalance, Angular Loss

1. Introduction

Analysis of medical images is a central component of contemporary diagnostic workflows, supporting disease identification, early detection, and follow-up assessment across dermatology, ophthalmology, and other clinical domains. These tasks frequently require the interpretation of subtle and heterogeneous visual patterns, where diagnostic distinctions depend on fine structural and textural cues. Computer-aided diagnosis (CAD) assists medical practitioners to identify known diseases. While some of these diseases appear regularly and many practitioners are able to identify them, the importance of CAD systems comes at highlighting rare cases that practitioners might not be familiar with. Therefore, CAD systems should particularly focus on detecting rare cases, while still recognize common cases reasonably.

Many modern CAD systems are based on deep learning, due to its ability to learn task-specific representations directly from raw clinical images (Esteva et al., 2017; Yu et al., 2018). Convolutional neural networks (CNNs) and related architectures have demonstrated strong performance on large-scale benchmarks, including dermatologist-level skin-cancer classification (Esteva et al., 2017) and high-performing diabetic-retinopathy (DR) detection (Gulshan et al., 2016). However, disease distributions typically reflect population prevalence and clinical datasets are rarely balanced. Therefore, models must operate under pronounced class disparities that influence learned representations (Johnson and Khoshgoftaar, 2019).

Handling such class imbalance is a long-standing challenge in machine learning. It is well-known that predictive models are biased toward majority classes, distort decision boundaries, and – most severely for their practicability – undermine minority-class recognition (Sun et al., 2009; Krawczyk, 2016; Rudd et al., 2016). These issues become more pronounced in medical imaging, where minority categories often represent clinically meaningful but infrequent findings, *e. g.*, melanomas occur notably less often than benign lesions, and exhibit substantial intra-class variability alongside high visual similarity to non-malignant categories (Gessert et al., 2019). This often leads to reduced performance on minority classes, limiting the applicability of deep-learning systems in scenarios requiring reliable detection.

A variety of strategies for handling imbalanced data have been proposed, including class weighting, cost-sensitive learning, over- or undersampling, and data augmentation. Loss function modifications such as focal loss (Lin et al., 2017) and class-balanced loss (Cui et al., 2019) reduce majority class dominance. (Buda et al., 2018) showed that sampling techniques with CNN helped in improving overall performance under various degrees of imbalance.

Angular margin-based softmax formulations provide a mechanism to improve representation discriminability by imposing geometric constraints on the embedding space. Notably, ArcFace (Deng et al., 2022) introduces an angular margin that encourages compact intra-class clusters and stronger inter-class separation. Although these methods achieve strong results in large-scale benchmarks, they are generally developed under relatively balanced class distributions and rely on fixed global margin behavior. In highly imbalanced medical image settings, uniform angular constraints may be insufficient to maintain adequate separation for minority categories, allowing majority classes to dominate the learned embedding geometry.

In this work, we investigate the behavior of angular-margin loss functions under class-imbalanced medical image distributions and address a gap in their theoretical treatment in such settings. Our contributions are threefold. First, we derive analytical upper and lower bounds for the scale parameter s used in angular-margin losses. These bounds provide principled guidance for selecting s when negative-angle distributions diverge from the simplifying assumptions adopted in prior formulations (Zhang et al., 2019). Second, we introduce a hybrid objective that integrates a dynamic angular-margin loss (DAL) with dynamic weighted cross-entropy (DWCE), yielding improved robustness to imbalance when compared with existing imbalance-aware losses. Third, we incorporate the coefficient of variation (CoV) to quantify performance variability across classes. Experiments on two ISIC skin lesion classification benchmarks and the APTOS diabetic-retinopathy dataset demonstrate that the proposed method achieves consistent improvements across multiple tasks and evaluation metrics. Our source code is publicly available.[1]

2. Related Work

Research on class imbalance in medical image classification has motivated a variety of loss functions designed to mitigate majority-class dominance and improve minority-class sensitivity. A central line of work focuses on weighting-based objectives. Focal loss (Lin et al., 2017) suppresses the contribution of well-classified samples and places greater emphasis on harder cases, which are often those belonging to minority categories. More recently, class-wise difficulty-balanced loss (Sinha et al., 2020) proposes adjusting sample contributions

1. Our code will be made public after acceptance.

according to estimated class-wise difficulty, providing an adaptive mechanism that increases gradients for underperforming or underrepresented classes. Complement cross-entropy (Kim et al., 2021) similarly aims to stabilize optimization under skewed distributions by explicitly reducing confidence on incorrect classes rather than solely reinforcing the correct class.

Another class of methods reformulates loss terms using estimates of effective class frequency. Class-balanced loss (Cui et al., 2019) derives weighting coefficients from the effective number of samples, avoiding overly large weights in cases of severe imbalance. AutoBalance (Li et al., 2021) extends this principle by optimizing weighting parameters during training, demonstrating improvements without manual tuning. Label distribution-aware margin (LDAM) loss (Cao et al., 2019) proposes margin adjustments based on class frequencies, enforcing larger margins for minority classes to yield improved generalization under long-tailed distributions. Merge loss (Du et al., 2023) makes use of domain knowledge by aggregating predictions across closely related minority subclasses to stabilize optimization and enhance performance in extreme imbalance settings. Other domain-specific loss formulations have also been introduced in dermatology and ophthalmology. For example, the adaptively weighted balance (AWB) loss (Yue et al., 2023) estimates class-wise weights dynamically during training and has shown gains for multi-center skin-lesion datasets with substantial imbalance. Similar developments have appeared in diabetic-retinopathy analysis, where angular-margin contrastive formulations have been used to improve discriminative separation between DR grades (Zhu et al., 2023). In parallel, several data-centric imbalance strategies have been developed. Balanced MixUp (Galdran et al., 2021) adapts MixUp to produce synthetic samples that preserve the structure of minority classes in medical image settings. Scholz et al. (2024) provide a systematic study demonstrating that imbalance-sensitive loss functions consistently outperform naive cross-entropy across a range of medical classification tasks, reinforcing the need for specialized treatments of clinical class skew.

Beyond weighting and frequency-based corrections, angular margin-based losses provide an orthogonal mechanism by enforcing geometric constraints on the embedding space. Most such methods are developed for the face recognition task (Wang and Deng, 2021), where abundances of classes with large amounts of samples are available for training (Zhu et al., 2021). ArcFace (Deng et al., 2022) adds a fixed additive angular margin to encourage compact intra-class clusters and increased inter-class separation. AdaCos (Zhang et al., 2019) removes the need for manually tuned scale parameters by introducing adaptive scaling of cosine logits. AdaFace (Kim et al., 2022) further incorporates quality-dependent margins that adjust the angular decision boundary based on sample reliability. Such methods are increasingly explored in medical domain. For example, Alirezazadeh and Dornaika (2023) applied them to histopathology-based breast cancer classification, and (Prakash et al., 2025) demonstrate their potential for improving discriminative medical image representations. However, existing angular-margin approaches assume balanced class distributions, leaving a methodological gap that we address in our work.

3. Approach

We extend angular-margin based classification by addressing the limitations of ArcFace under severe class imbalance. We provide a data-driven generalization of the AdaCos

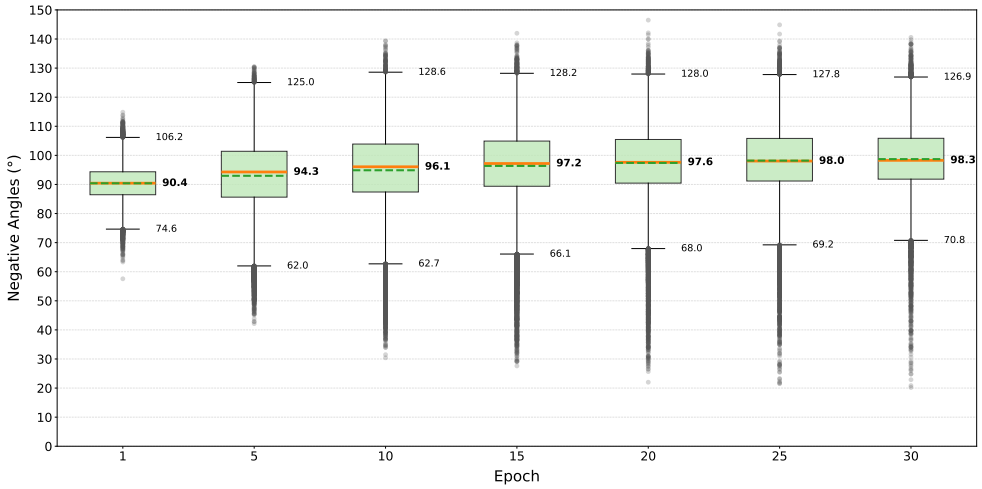


Figure 1: Negative Angles per Epoch. This figure shows box plots of the distribution of ArcFace angles $\theta_{i,j}$ for negative classes on the ISIC 2018 training set during training. Here we show a subset of epochs for $s = 4$, more such plots can be found in the supplemental.

scale parameter s for stabilizing cosine-based softmax, and batch-level mechanisms that compensate for imbalance through dynamic sample weighting or adaptive angular margins.

3.1. Generalized AdaCos Scale Parameter

Let $\mathbf{x}_i$ denote the embedding of sample i with label y_i, and let $\theta_{i,j}$ be the angle between $\mathbf{x}_i$ and classifier weight $\mathbf{W}_j$. The cosine-softmax probability (Deng et al., 2022) is expressed as:

$$P_{i,y_i} = \frac{\exp\big(s\cos(\theta_{i,y_i} + m)\big)}{\exp\big(s\cos(\theta_{i,y_i} + m)\big) + B_i}, \qquad B_i = \sum_{j \neq y_i} \exp\big(s\cos\theta_{i,j}\big), \tag{1}$$

where B_i represents angles $\theta_{i,j}$ of all $C-1$ negative classes $j \neq y_i$. Assuming a margin of $m = 0$, AdaCos determines the optimal scale s by maximizing the curvature of $P_{i,y_i}(\theta)$ at a chosen central angle θ_0, leading to (Zhang et al., 2019):

$$s_0 = \frac{\log B_i}{\cos\theta_0}. \tag{2}$$

The original AdaCos formulation assumes that negative class angles are concentrated at 90°, implying $\forall j \neq y_i : \cos\theta_{i,j} \approx 0$ and yielding $B_i \approx C - 1$ along with the closed-form scale $s_{\text{AdaCos}} = \sqrt{2}\log(C-1)$. While such assumptions are reasonable for face recognition, in our setting with fewer classes and smaller amounts of samples, however, negative angles span a substantially wider interval (see Fig. 1), which violates the assumption $\theta_{i,j} \approx 90°$.

To obtain a more faithful scale estimate, we first assume that all negative angles have a fixed value $\theta_{i,j}$, which we later relax to provide upper and lower bounds for s derived from batch statistics. We rewrite the negative term in (1):

$$B_i \approx (C-1)\exp(s\cos\theta_{i,j}). \tag{3}$$

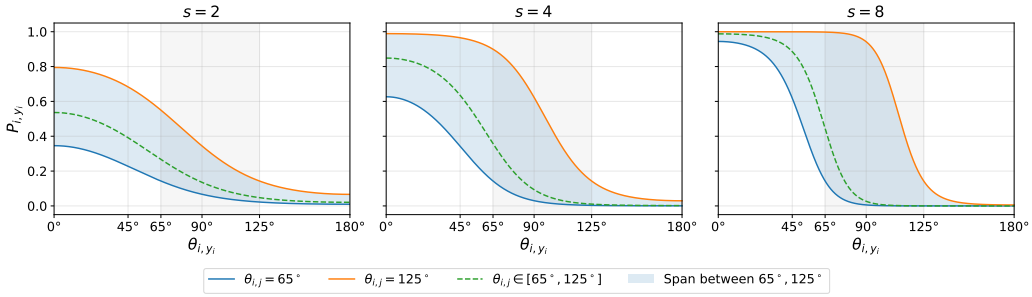


Figure 2: Negative Angle Placement. We show the effect of negative angles $\theta_{i,j}; j \neq y_i$ on the positive-class probability P_{i,y_i} across different scale values s for $C = 7$. The solid lines represent extreme cases, where $\forall j : \theta_{i,j} = 65°$ or $125°$, while the dashed line indicates average cases with linearly scaling $\theta_{i,j}$ between the extremes.

Substituting this approximation into (2) yields the generalized closed-form scale. Specifically, taking the logarithm of (3) gives:

$$\log B_i \approx \log(C-1) + s_0 \cos\theta_{i,j}. \tag{4}$$

Replacing $\log B_i$ in (2) with (4) results in:

$$s_0 = \frac{\log(C-1) + s_0 \cos\theta_{i,j}}{\cos\theta_0}. \tag{5}$$

Multiplying both sides of (5) by $\cos\theta_0$ and collecting terms involving s_0 yields:

$$s_0\big(\cos\theta_0 - \cos\theta_{i,j}\big) = \log(C-1). \tag{6}$$

Solving for s_0 leads to the generalized closed-form scale:

$$\boxed{s_0 = \frac{\log(C-1)}{\cos\theta_0 - \cos\theta_{i,j}}}. \tag{7}$$

Following the original AdaCos formulation (Zhang et al., 2019), we set the reference positive angle to $\theta_0 = 45°$, the geometric midpoint of the feasible positive-angle range $[0, 90°]$. When using $C = 7$ and the empirical negative-angle interval observed in Fig. 1, namely $\theta_{i,j} \in [65°, 125°]$, the above expression produces a data-driven scale range of $s_0 \in [1.40, 6.30]$. Based on this interval, we can select the mid-range value $s = 4$, which provides stable optimization without compromising discriminability, as shown in the appendix.

Fig. 2 illustrates how the positive-class probability P_{i,y_i} evolves as a function of the positive angle θ_{i,y_i} for different choices of the scale parameter s and several representative negative angle configurations. When s is chosen too small, the probability remains significantly below 1 even when the embedding is perfectly aligned with its class center, i.e. $\theta_{i,y_i} \approx 0$. This occurs because $s\cos\theta_{i,y_i}$ in (1) cannot sufficiently dominate the negative logits, even when negative angles $\theta_{i,j}$ are large. Conversely, when s is excessively large, the probability saturates too early: even for relatively large positive angles (far from the class center), the

model assigns very high confidence whenever negative angles are moderately high. Such behavior undesirably compresses the angular decision margin and leads to overconfident predictions and poorly calibrated gradients. Within the empirically observed negative-angle interval $[65°, 125°]$ (see Fig. 1), the choice $s = 4$ provides a favorable balance. As highlighted by the shaded region in Fig. 2, this scale yields well-behaved probability curves: the model achieves high confidence only when θ_{i,y_i} is sufficiently small, while avoiding premature probability saturation for larger angles. This intermediate scaling therefore produces stable optimization dynamics and preserves angular discriminability across classes.

3.2. Batch-Adaptive Angular Margin and Dynamic Weighting

Under class imbalance, ArcFace suffers from two limitations. First, majority classes dominate the optimization because they appear more frequently in the loss and second, a fixed angular margin m imposes identical geometric constraints on all classes, irrespective of their representation in the data. To address both issues simultaneously, we combine a batch-adaptive angular margin (BAAM) with dynamic weighted cross-entropy (DWCE) loss.

Batch-adaptive angular margin. Let a training batch of size N contain N_c samples of class c. We first compute an inverse-frequency score:

$$w_c^{\text{raw}} = \log\left(1 + \frac{1}{N_c}\right), \tag{8}$$

which is larger for minority classes. These scores are then normalized via batch-wise min–max scaling and mapped into a user-specified angular margin range $[m_{\min}, m_{\max}]$:

$$t_c = \frac{w_c^{\text{raw}} - w_{\min}}{w_{\max} - w_{\min}}, \qquad m_c = m_{\min} + (m_{\max} - m_{\min})\, t_c. \tag{9}$$

Each sample inherits the margin of its ground-truth class, m_{y_i}. The ArcFace nominator in (2) is therefore modified as:

$$\exp\big(s\, \cos(\theta_{i,y_i} + m_{y_i})\big), \tag{10}$$

whereas B_i in (1) remains unchanged. Larger m_{y_i} compel minority classes to form more compact discriminative embeddings, mitigating geometric dominance by majority classes.

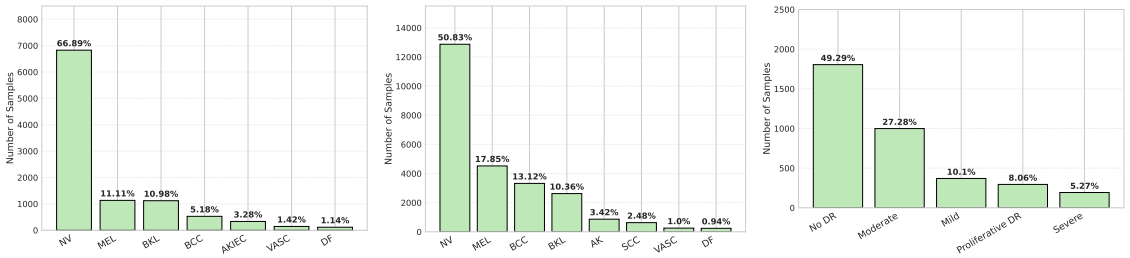
Dynamic weighted cross-entropy. While adaptive margins modify the representation geometry, the optimization process itself remains sensitive to class frequencies. To ensure balanced gradient flow, we incorporate a dynamic weighting factor computed from the batch composition. For each batch, we define the class weight as:

$$\omega_c = \frac{N}{C\, N_c} \tag{11}$$

so that minority classes (with a small N_c) receive proportionally larger loss contributions, while $\omega_c = 0$ for $N_c = 0$. The resulting dynamically weighted cross-entropy (DWCE) loss is:

$$\mathcal{L}_{\text{DWCE}} = -\frac{1}{N} \sum_i \omega_{y_i} \log P_{i,y_i}, \tag{12}$$

where P_{i,y_i} denotes the softmax probability (1) computed using the batch-adaptive margin.



(*a*) ISIC 2018 (*b*) ISIC 2019 (*c*) APTOS

Figure 3: Class Distribution. For each dataset, classes are ordered in decreasing frequency and annotated with their relative proportion (in %) above each bar.

4. Experimental Setup

All experiments are run using a comparable setup, the exact training details can be found in the appendix.

4.1. Datasets

Fig. 3 presents the class distributions for the training sets of the three datasets used in this study. Below, we describe each dataset in detail.

The APTOS 2019 Blindness Detection dataset (Karthik et al., 2019) is designed for diabetic retinopathy (DR) grading and contains 3,662 fundus images, each labeled with one of five severity levels, ranging from no DR to moderate cases and proliferative DR. The images exhibit substantial variations in illumination, contrast, and acquisition conditions due to differences in clinical equipment and screening settings. The distribution is highly imbalanced, with *No DR* and *Moderate* categories accounting for the majority of samples.

The ISIC 2018 dataset (Codella et al., 2019) is built primarily from the HAM10000 collection (Tschandl et al., 2018) and contains 10,208 dermoscopic images originating from multiple clinical centers using diverse dermatoscopy systems, with a native resolution of 600×450 pixels. The dataset covers seven lesion categories, namely dermatofibroma (DF), vascular lesions (VASC), actinic keratosis (AKIEC), basal cell carcinoma (BCC), benign keratosis-like lesions (BKL), melanoma (MEL), and melanocytic nevi (NV). The images contain considerable variability in lesion appearance, lighting, and color. As seen in Fig. 3, the dataset exhibits pronounced class imbalance dominated by the NV category.

The ISIC 2019 dataset expands upon the ISIC 2018 task by adding images from BCN20000 (Combalia et al., 2019), forming a large-scale collection of 25,331 dermoscopic images. ISIC 2019 includes an additional category, squamous cell carcinoma (SCC), leading to eight lesion types in total. Due to its larger scale, multi-source nature, and severe long-tailed distribution, ISIC 2019 represents a more challenging benchmark for evaluating model performance under strong inter-class imbalance.

For ISIC 2018 and ISIC 2019, we use the official challenge-provided test sets (Codella et al., 2018) for final evaluation. The remaining training images are randomly split into 80% training and 20% validation, and the epoch with the highest balanced accuracy on the validation set is selected. For APTOS 2019, no official test set is available, we first split

the released training set into 80% training and 20% test, and then perform a secondary 80%/20% training/validation split on the training portion for model selection. To ensure robust performance estimation, experiments are repeated five times using different random seeds for all three datasets, and we report the average results across the five runs.

4.2. Evaluation Metrics

To robustly assess model performance under the substantial inter-class imbalance present in our datasets (see Fig. 3), we employ complementary metrics. *Accuracy*, *balanced accuracy*, *macro-averaged F1-score* are standard metrics for categorical classification, we include them for completeness and comparability with prior work. The *index of balanced accuracy* (García et al., 2009) IBA_α provides an imbalance-aware alternative to balanced accuracy by incorporating a dominance index α that penalizes disproportionate performance across classes. Following García et al. (2009), we report IBA for $\alpha \in \{0.1, 0.5, 1.0\}$. Superior classifiers achieve higher IBA values, with a maximum of 1.

To characterize the stability of the model across classes, we additionally compute the *coefficient of variation* (CoV) of the per-class recall values (Formby et al., 1999; Chen and Hooker, 2023). Let R_c denote the recall for class $c \in \{1, \ldots, C\}$. The CoV is defined as:

$$\mu_R = \frac{1}{C}\sum_{c=1}^{C} R_c \qquad \sigma_R = \sqrt{\frac{1}{C-1}\sum_{c=1}^{C}(R_c - \mu_R)^2} \qquad \text{CoV} = \frac{\sigma_R}{\mu_R}, \tag{13}$$

Lower CoV values indicate more uniform performance across classes, whereas higher values reflect variability typically associated with bias toward majority classes.

5. Results

Tab. 1 summarizes results on ISIC 2018, ISIC 2019, and APTOS. Our method consistently achieves the **highest balanced accuracy**, outperforming the next-best competitor by more than 3% on each dataset. This improvement is further reinforced by the *index of balanced accuracy (IBA)*, where our approach obtains the best values for all dominance weights ($\alpha \in \{0.1, 0.5, 1.0\}$). These consistent gains demonstrate that our batch-adaptive angular margin and weighting strategy is highly effective at mitigating imbalance and improving performance on rare but clinically important classes.

Although the CB and AWB loss yield the highest macro F1-scores, our method remains competitive and performs only marginally lower. More importantly, our approach achieves the best CoV of per-class recall across all three datasets, indicating the most stable and uniform behavior among classes. Particularly, the classifier avoids collapsing on minority categories and provides more consistent predictive performance across the classes.

Tab. 2 reports the per-class recall (TPR) averaged over five independent runs. A consistent pattern emerges across all datasets: although AWB, PF, CB, or occasionally CE and Focal Loss obtain slightly higher TPRs on the majority and high-frequency classes (*e. g.*, *NV* in ISIC or *No DR* in APTOS), our method achieves competitive performance and typically remains within a narrow margin of these baselines. More importantly and clinically relevant, the most notable gains appear on the minority and severely under-represented

Table 1: Test Set Performance. This table shows test set results across all methods, providing the mean over 5 runs on three datasets. Best results are highlighted in bold.

(a) ISIC 2018

Method	Acc	F1	BAcc	$IBA_{0.1}$	$IBA_{0.5}$	IBA_1	CoV
Cross-Entropy (CE)	0.7779	0.6182	0.6015	0.5451	0.4756	0.3887	0.2612
Focal Loss (Lin et al., 2017)	0.7722	0.6195	0.6095	0.5531	0.4837	0.3969	0.2473
LDAM Loss (Cao et al., 2019)	**0.7864**	0.6603	0.6678	0.6045	0.5379	0.4547	0.2043
CB Loss (Cui et al., 2019)	0.7763	**0.6680**	0.6730	0.6216	0.5524	0.4659	0.1300
CCE Loss (Kim et al., 2021)	0.7738	0.6162	0.6084	0.5536	0.4850	0.3991	0.2339
PF Loss (Du et al., 2021)	0.7357	0.6465	0.6941	0.6435	0.5769	0.4936	**0.0930**
AWB Loss (Yue et al., 2023)	0.7312	0.6368	0.6500	0.5950	0.5250	0.4375	0.1369
DAL (Ours)	0.7443	0.6490	**0.7348**	**0.6832**	**0.6227**	**0.5470**	0.0972

(b) ISIC 2019

Method	Acc	F1	BAcc	$IBA_{0.1}$	$IBA_{0.5}$	IBA_1	CoV
Cross-Entropy (CE)	0.6378	0.4633	0.4562	0.4044	0.3381	0.2551	0.4329
Focal Loss (Lin et al., 2017)	0.6466	0.4792	0.4687	0.4139	0.3470	0.2634	0.4226
LDAM Loss (Cao et al., 2019)	0.6319	0.4759	0.4798	0.4267	0.3642	0.2869	0.3298
CB Loss (Cui et al., 2019)	**0.6472**	0.5105	0.4970	0.4448	0.3752	0.2881	0.3632
CCE Loss (Kim et al., 2021)	0.6467	0.4756	0.4576	0.4053	0.3395	0.2574	0.4415
PF Loss (Du et al., 2021)	0.6308	0.5029	0.5233	0.4721	0.3974	0.3040	0.2701
AWB Loss (Yue et al., 2023)	0.6292	**0.5134**	0.5045	0.4500	0.3805	0.2937	0.3508
DAL (Ours)	0.6114	0.5003	**0.5545**	**0.4993**	**0.4264**	**0.3343**	**0.1872**

(c) APTOS

Method	Acc	F1	BAcc	$IBA_{0.1}$	$IBA_{0.5}$	IBA_1	CoV
Cross-Entropy (CE)	0.8175	0.6163	0.6047	0.5568	0.5139	0.4602	0.4990
Focal Loss (Lin et al., 2017)	**0.8292**	0.6286	0.6111	0.5655	0.5214	0.4664	0.4899
LDAM Loss (Cao et al., 2019)	0.8002	0.6321	0.6287	0.5821	0.5293	0.4697	0.3986
CB Loss (Cui et al., 2019)	0.8028	0.6319	0.6439	0.5889	0.5312	0.4621	0.3092
CCE LOss (Kim et al., 2021)	0.7999	0.5794	0.5748	0.5284	0.4851	0.4309	0.5331
PF Loss (Du et al., 2021)	0.7902	0.6349	0.6410	0.5902	0.5347	0.4653	0.3189
AWB Loss (Yue et al., 2023)	0.8016	**0.6491**	0.6481	0.5993	0.5407	0.4674	0.3047
DAL (Ours)	0.7793	0.6337	**0.6756**	**0.6248**	**0.5659**	**0.4923**	**0.2397**

classes. Across ISIC 2018 (e.g., *AKIEC*, *VASC*, *DF*), ISIC 2019 (e.g., *AK*, *SCC*, *DF*), and APTOS (e.g., *Proliferative DR*, *Severe*), our approach consistently achieves substantially higher TPR values than all other methods. These improvements highlight the effectiveness of our adaptive angular margin and dynamic weighting strategy in addressing strong inter-class imbalance, enabling the model to better recognize rare disease categories without compromising performance on the frequent ones.

An ablation study on the impact of the different parameters of our model on the performance in the ISIC 2018 dataset is presented in the appendix. We show that the dynamic weighting has a high influence on the results, and that the theoretically-derived $s = 4$ parameter yields superior results over larger values.

6. Conclusion

In this paper, we investigated the usability of *dynamic angular-margin losses* (DAL), such as ArcFace (Deng et al., 2022) or AdaCos (Zhang et al., 2019), for improving predictions of minority classes in imbalanced medical image classification tasks. First, we analyzed the scale parameter s of ArcFace and showed that the basic assumption from Zhang et al. (2019) do not translate to medical datasets with a few classes. We derived theoretical upper and

Table 2: Per-Class Recall. This table shows the per-class recall aka. True Positive Rate (TPR) across our three evaluated datasets. Classes are sorted by decreasing number of samples. Bold numbers highlight best values.

(*a*) ISIC 2018

Method	NV	MEL	BKL	BCC	AK	VASC	DF
Cross-Entropy (CE)	**0.9106**	0.5438	0.6497	0.5462	0.6744	0.4988	0.3868
Focal Loss (Lin et al., 2017)	0.9047	0.5380	0.6230	0.5440	0.7162	0.5001	0.4404
LDAM Loss (Cao et al., 2019)	0.9273	0.5480	0.5569	0.6451	0.7209	0.7128	0.5227
CB Loss (Cui et al., 2019)	0.8649	0.6562	0.6775	0.6150	0.6558	0.6285	0.6591
CCE Loss (Kim et al., 2021)	0.9065	0.5166	0.6421	0.5627	0.7132	0.5619	0.3561
PF Loss (Du et al., 2021)	0.7863	0.6561	0.6368	0.6496	**0.7581**	0.6857	0.6863
AWB Loss (Yue et al., 2023)	0.7924	**0.6912**	**0.6820**	**0.6557**	0.5548	0.6176	0.5817
DAL (Ours)	0.7780	**0.6912**	0.6737	0.6344	0.7163	**0.8171**	**0.8091**

(*b*) ISIS 2019

Method	NV	MEL	BCC	BKL	AK	SCC	VASC	DF
Cross-Entropy (CE)	0.8176	0.6093	0.6382	0.4227	0.2470	0.2521	0.3442	0.3187
Focal Loss (Lin et al., 2017)	**0.8186**	0.6273	0.6597	0.4045	0.3011	0.2400	0.3692	0.3055
LDAM Loss (Cao et al., 2019)	0.7798	0.6345	**0.7296**	0.4002	0.2787	0.3623	0.4210	0.4457
CB Loss (Cui et al., 2019)	0.7870	0.6401	0.6910	0.4321	0.2786	0.2824	0.4250	0.4395
CCE Loss (Kim et al., 2021)	0.8326	0.6197	0.6485	0.3995	0.2834	0.2586	0.3141	0.3040
PF Loss (Du et al., 2021)	0.7248	0.6441	0.6518	**0.4945**	0.3401	0.3454	0.4711	0.5142
AWB Loss (Yue et al., 2023)	0.7346	**0.6710**	0.7147	0.4160	0.2695	0.2715	0.4846	0.4901
DAL (Ours)	0.6980	0.6532	0.6232	0.4312	**0.4417**	**0.4125**	**0.5731**	**0.6154**

(*c*) APTOS

Method	No DR	Moderate	Mild	Proliferative DR	Severe
Cross-Entropy (CE)	0.9845	0.8320	0.6108	0.4475	0.1487
Focal Loss (Lin et al., 2017)	**0.9967**	**0.8570**	0.6189	0.4034	0.1795
LDAM Loss (Cao et al., 2019)	0.9862	0.7570	0.5518	0.4742	0.3616
CB Loss (Cui et al., 2019)	0.9718	0.6381	0.7297	0.4698	0.4192
CCE Loss (Kim et al., 2021)	0.9825	0.8050	0.6081	0.3503	0.1282
PF Loss (Du et al., 2021)	0.9761	0.6530	0.7216	0.4746	0.3798
AWB Loss (Yue et al., 2023)	0.9801	0.7040	0.6378	0.4983	0.4205
DAL (Ours)	0.9512	0.6321	**0.7432**	**0.5932**	**0.4723**

lower bounds that work much better in our cases. The proposed loss formulation additionally couples representation-level adaptation (through class-dependent margins) with loss-level balancing (through dynamic weighting). Minority classes are encouraged to produce more compact clusters and simultaneously receive stronger optimization signals, while majority classes are prevented from dominating the shared feature space or the gradient updates. This joint mechanism provides a robust solution to the imbalance problem without altering the underlying ArcFace loss architecture.

We ran experiments on three imbalanced medical image classification tasks and showed that our DAL method improves classification accuracy for underrepresented classes, and provides a more fair model as measured by the coefficient of variation. Furthermore, the results show only little evidence of the common trade-off where gains on minority classes come at the expense of majority-class performance. In general, our method maintains stable recall across the full class spectrum, indicating that the proposed framework yields robust class-wise generalization and effectively handles long-tailed medical image distributions.

In this paper, DAL is applied to medical classification. However, the formulation is generic, and we will apply this to other imbalanced classification tasks in future work. Additionally, we will test combinations of DAL with other types of class balancing techniques.

References

Pendar Alirezazadeh and Fadi Dornaika. Boosted additive angular margin loss for breast cancer diagnosis from histopathological images. *Computers in Biology and Medicine*, 166: 107528, 2023.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Violet (Xinying) Chen and John N Hooker. A guide to formulating fairness in an optimization model. *Annals of Operations Research*, 326(1):581–619, 2023.

Noel C. F. Codella, David A. Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *International Symposium on Biomedical Imaging (ISBI)*, 2018.

Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen W. Dusza, David A. Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael A. 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.

Marc Combalia, Noel C F Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, and Josep Malvehy. BCN20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(10):5962–5979, 2022.

Jie Du, Yanhong Zhou, Peng Liu, Chi-Man Vong, and Tianfu Wang. Parameter-free loss for class-imbalanced deep learning in image classification. *Transactions on Neural Networks and Learning Systems (TNNLS)*, 34(6):3234–3240, 2021.

Zehua Du, Hao Zhang, Zhiqiang Wei, Yuanyuan Zhu, Jiali Xu, Xianqing Huang, and Bo Yin. Merge loss calculation method for highly imbalanced data multiclass classification. *Transactions on Neural Networks and Learning Systems (TNNLS)*, 2023.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 2017.

John P Formby, W James Smith, and Buhong Zheng. The coefficient of variation, stochastic dominance and inequality: a new interpretation. *Economics Letters*, 62(3):319–323, 1999.

Adrian Galdran, Gustavo Carneiro, and Miguel A González Ballester. Balanced-mixup for highly imbalanced medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2021.

Vicente García, Ramón Alberto Mollineda, and José Salvador Sánchez. Index of balanced accuracy: A performance measure for skewed class distributions. In *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*. Springer, 2009.

Nils Gessert, Thilo Sentker, Frederic Madesta, Rüdiger Schmitz, Helge Kniep, Ivo Baltruschat, Rene Werner, and Alexander Schlaefer. Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting. *Transactions on Biomedical Engineering (TBME)*, 67(2):495–503, 2019.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Journal of the American Medical Association (JAMA)*, 316(22):2402–2410, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Karthik, Maggie, and Sohier Dane. Aptos 2019 blindness detection. https://kaggle.com/competitions/aptos2019-blindness-detection, 2019. Kaggle.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yechan Kim, Younkwan Lee, and Moongu Jeon. Imbalanced image classification with complement cross entropy. *Pattern Recognition Letters*, 151:33–40, 2021.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

Mingchen Li, Xuechen Zhang, Christos Thrampoulidis, Jiasi Chen, and Samet Oymak. Autobalance: Optimized loss functions for imbalanced data. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *International Conference on Computer Vision (ICCV)*, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeuRIPS)*, 2019.

Pritesh Prakash, Pankaj Kumar Kandpal, and Ashish Mehta. Advancing medical image classification: Harnessing additive angular margin loss for enhanced discrimination. In *International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies (CE2CT)*. IEEE, 2025.

Ethan M. Rudd, Manuel Günther, and Terrance E. Boult. MOON: A mixed objective optimization network for the recognition of facial attributes. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3), 2015.

Daniel Scholz, Ayhan Can Erdur, Josef A Buchner, Jan C Peeken, Daniel Rueckert, and Benedikt Wiestler. Imbalance-aware loss functions improve medical image classification. *Medical Imaging with Deep Learning (MIDL)*, 2024.

Saptarshi Sinha, Hiroki Ohashi, and Katsuyuki Nakamura. Class-wise difficulty-balanced loss for solving class-imbalance. In *Asian Conference on Computer Vision (ACCV)*, 2020.

Yanmin Sun, Andrew K C Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 23(04):687–719, 2009.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9, 2018.

Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429: 215–244, 2021.

Zhen Yu, Xudong Jiang, Feng Zhou, Jing Qin, Dong Ni, Siping Chen, Baiying Lei, and Tianfu Wang. Melanoma recognition in dermoscopy images via aggregated deep convolutional features. *Transactions on Biomedical Engineering (TBME)*, 66(4):1006–1016, 2018.

Guanghui Yue, Peishan Wei, Tianwei Zhou, Qiuping Jiang, Weiqing Yan, and Tianfu Wang. Toward multicenter skin lesion classification using deep neural network with adaptively weighted balance loss. *Transactions on Medical Imaging (TMI)*, 42(1):119–131, 2023.

Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Dongsheng Zhu, Aiming Ge, Xindi Chen, Qiuyang Wang, Jiangbo Wu, and Shuo Liu. Supervised contrastive learning with angular margin for the detection and grading of diabetic retinopathy. *Diagnostics*, 13(14):2389, 2023.

Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. WebFace260M: A benchmark unveiling the power of million-scale deep face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Table 3: Ablation Study. This table reports test-set performance for ArcFace models under different combinations of scale s, margin m, and dynamic weighted cross-entropy (DWCE). Metrics include overall accuracy (Acc), macro F1-score (F1), balanced accuracy (BAcc), and the imbalance-aware index (IBA) at three settings of α, and the coefficient of variation (CoV). The last column shows the p-values for a paired t-test comparing to our method.

Method	s	m	**DWCE**	**Acc**	**F1**	**BAcc**	**IBA$_{0.1}$**	**IBA$_{0.5}$**	**IBA$_{1}$**	**CoV**	**p-value**
ArcFace	4	0.0	No	0.7731	0.5899	0.5897	0.5558	0.4901	0.4097	0.3115	**4.464e−6**
ArcFace	4	0.1	No	0.7787	0.5847	0.5789	0.5398	0.4763	0.3894	0.3131	**1.576e−6**
ArcFace	4	0.2	No	0.7665	0.5781	0.5708	0.5144	0.4491	0.3675	0.3182	**2.387e−7**
ArcFace	4	0.3	No	0.7778	0.5835	0.5682	0.5121	0.4501	0.3726	0.3240	**1.741e−7**
ArcFace	8	0.0	No	0.7837	0.6385	0.6148	0.5591	0.4887	0.4007	0.3022	**3.482e−7**
ArcFace	8	0.1	No	0.7804	0.6384	0.6134	0.5571	0.4859	0.3969	0.3019	**3.123e−7**
ArcFace	8	0.2	No	**0.7943**	0.6663	0.6380	0.5821	0.5123	0.4251	0.2771	**2.193e−6**
ArcFace	8	0.3	No	0.7798	0.6305	0.6036	0.5477	0.4779	0.3907	0.2947	**1.385e−6**
ArcFace	16	0.1	No	0.7712	0.6116	0.6103	0.5556	0.4854	0.3976	0.2599	**1.441e−6**
ArcFace	16	0.2	No	0.7851	0.6375	0.6101	0.5504	0.4810	0.3943	0.2546	**1.491e−6**
ArcFace	16	0.3	No	0.7824	0.6536	0.6571	0.6034	0.5350	0.4495	0.2114	**1.195e−5**
ArcFace	4	0.0	Yes	0.7553	0.6458	0.7248	0.6767	0.6152	0.5432	0.1068	**4.214e−3**
ArcFace	4	0.1	Yes	0.7665	0.6580	0.7103	0.6594	0.5910	0.5101	0.1138	**1.615e−3**
ArcFace	4	0.2	Yes	0.7595	0.6601	0.7060	0.6494	0.5830	0.5032	0.1223	**1.211e−3**
ArcFace	4	0.3	Yes	0.7684	0.6618	0.6989	0.6411	0.5749	0.4992	0.1049	**5.213e−4**
ArcFace	8	0.0	Yes	0.7524	0.6512	0.7139	0.6619	0.5999	0.5131	0.1198	**2.231e−3**
ArcFace	8	0.1	Yes	0.7559	0.6604	0.7197	0.6698	0.6085	0.5271	**0.0876**	**3.112e−3**
ArcFace	8	0.2	Yes	0.7617	0.6659	0.7143	0.6641	0.6009	0.5203	0.0932	**2.788e−3**
ArcFace	8	0.3	Yes	0.7467	0.6388	0.7010	0.6449	0.5838	0.5010	0.0959	**1.528e−3**
ArcFace	16	0.0	Yes	0.7521	0.6634	0.6910	0.6377	0.5718	0.4894	0.0998	**6.342e−4**
ArcFace	16	0.1	Yes	0.7591	0.6511	0.6890	0.6356	0.5701	0.4927	0.1059	**1.687e−4**
ArcFace	16	0.2	Yes	0.7648	0.6659	0.6979	0.6496	0.5813	0.4988	0.1022	**5.124e−4**
ArcFace	16	0.3	Yes	0.7698	**0.6761**	0.7095	0.6576	0.5889	0.5090	0.0938	**1.442e−3**
AdaCos	dyn	0.0	Yes	0.7163	0.6044	0.7018	0.6502	0.5850	0.5034	0.1348	**1.712e−3**
DAL (Ours)	4	var	Yes	0.7443	0.6490	**0.7348**	**0.6832**	**0.6227**	**0.5470**	0.0930	**-**

Appendix A. Implementation Details

All models are implemented in PyTorch (Paszke et al., 2019) using a ResNet-18 backbone (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015). Following Deng et al. (2022), we replace the original fully connected layer with a lightweight two-layer classification head consisting of a dropout layer, a linear projection from the 512-dimensional backbone output to a 128-dimensional embedding, followed by Batch Normalization which is fed to our classification layer. For baseline models, we use ReLU layer instead of Batch Normalization and a final linear layer producing C outputs. All images are resized to 224×224, and training is performed using stochastic gradient descent (SGD) with a learning rate of 0.001, momentum of 0.9, and weight decay of 0.0001. A batch size of 128 is used, and each model is trained for 30 epochs. To stabilize optimization, the ResNet-18 backbone is frozen for the first five epochs so that only the classification head is updated; thereafter, all layers are fine-tuned jointly. For every dataset, we follow the train/validation/test protocol described previously and select the checkpoint that achieves the highest balanced accuracy on the validation set.

Table 4: Validation Set Performance. This table shows results obtained on the validation sets, providing the mean over 5 runs on two datasets. Best results are highlighted in bold.

(a) ISIC 2018

Method	Acc	F1	BAcc	$IBA_{0.1}$	$IBA_{0.5}$	IBA_{1}	CV
Cross-Entropy (CE)	0.8356	0.6662	0.6531	0.6030	0.5376	0.4559	0.2490
Focal Loss (Lin et al., 2017)	0.8341	0.6739	0.6597	0.6091	0.5454	0.4657	0.2594
LDAM Loss (Cao et al., 2019)	**0.8598**	0.7211	0.7429	0.6976	0.6433	0.5754	0.2008
CB Loss (Cui et al., 2019)	0.8267	**0.7243**	0.7351	0.6838	0.6254	0.5593	0.1664
CCE Loss (Kim et al., 2021)	0.8358	0.6871	0.6770	0.6140	0.5503	0.4705	0.2531
PF Loss (Du et al., 2021)	0.7798	0.6985	0.7649	0.7139	0.6618	0.5958	0.1170
AWB Loss (Yue et al., 2023)	0.7929	0.7169	0.7591	0.7118	0.6589	0.5928	0.1513
DAL (Ours)	0.7743	0.6787	**0.7771**	**0.7321**	**0.6832**	**0.6132**	**0.1434**

(b) ISIC 2019

Method	Acc	F1	BAcc	$IBA_{0.1}$	$IBA_{0.5}$	IBA_{1}	CV
Cross-Entropy (CE)	0.7703	0.6023	0.6013	0.5539	0.4830	0.3945	0.2903
Focal Loss (Lin et al., 2017)	0.7686	0.6067	0.6037	0.5566	0.4838	0.3954	0.2816
LDAM Loss (Cao et al., 2019)	**0.7796**	0.6557	0.6606	0.6149	0.5501	0.4691	0.2557
CB Loss (Cui et al., 2019)	0.7536	0.6670	0.6941	0.6482	0.5823	0.4994	0.1504
CCE Loss (Kim et al., 2021)	0.7671	0.6043	0.5998	0.5524	0.4851	0.4010	0.2977
PF Loss (Du et al., 2021)	0.7357	0.6540	0.7043	0.6584	0.5950	0.5153	0.1607
AWB Loss (Yue et al., 2023)	0.7476	**0.6864**	**0.7117**	**0.6649**	**0.6035**	**0.5267**	0.1644
DAL (Ours)	0.7149	0.6024	0.7068	0.6552	0.5959	0.5167	**0.1255**

Appendix B. Ablation Study of our Method

To better understand the influence of ArcFace design choices under severe class imbalance, Tab. 3 evaluates a wide range of scale values s, margin values m, and the presence or absence of dynamic weighted cross-entropy (DWCE) on ISIC 2018. A clear trend emerges: models trained without DWCE consistently underperform those trained with DWCE, demonstrating that explicit imbalance compensation remains crucial for reliable decision boundaries. Additionally, the results indicate that large values of s or m are not beneficial in this setting, likely due to the difficulty of optimizing angular constraints when class distributions are highly uneven. However, we do observe that for larger values of s, relatively larger m perform better than lower ones. Overall, our selected configuration, which combines a moderate angular parameterization with DWCE, achieves the strongest balanced metrics. These improvements are further confirmed to be statistically significant based on a paired t-test for balanced accuracy, compared against all listed baseline variants.

Appendix C. Validation Set Evaluation

Tab. 4 summarizes the validation set results on the ISIC 2018 and ISIC 2019 datasets, averaged over five independent runs. As opposed to the test set reported in the main paper, the validation set is not split by patient, so different lesions of the same patient may appear in the training and validation sets. This mimics the evaluation setup as performed by Yue et al. (2023). It is interesting to note that, on the validation set, AWB loss and PF loss perform quite competitively compared to our method for the ISIC 2018 dataset, and even surpass its performance on ISIC 2019. However, as shown in Tab. 1, this performance advantage does not translate to the identity-disjoint test sets of ISIC 2018 and ISIC 2019.

Table 5: Per-Class Recall. This table shows the per-class recall aka. True Positive Rate (TPR) on the validation set. Bold numbers highlight best values.

(*a*) ISIC 2018

Method	NV	MEL	BKL	BCC	AK	VASC	DF
Cross-Entropy (CE)	0.9327	0.5925	0.7054	0.6580	0.6157	0.7414	0.3478
Focal Loss (Lin et al., 2017)	**0.9369**	0.5650	0.6786	0.7264	0.5970	0.7672	0.3487
LDAM Loss (Cao et al., 2019)	0.9289	0.6134	0.6930	0.7759	0.5895	0.9123	0.6741
CB Loss (Cui et al., 2019)	0.8920	**0.6740**	0.7046	0.7436	0.5617	0.9142	0.6712
CCE Loss (Kim et al., 2021)	0.9359	0.5733	0.6679	0.7099	0.6113	0.7784	0.3804
PF Loss (Du et al., 2021)	0.7910	0.6718	**0.7369**	0.7899	0.7227	0.9379	0.7048
AWB Loss (Yue et al., 2023)	0.8689	0.6711	0.7321	**0.7956**	0.6468	0.9540	0.6722
DAL (Ours)	0.8086	0.6582	0.6948	0.7834	**0.7612**	**0.9766**	**0.7578**

(*b*) ISIC 2019

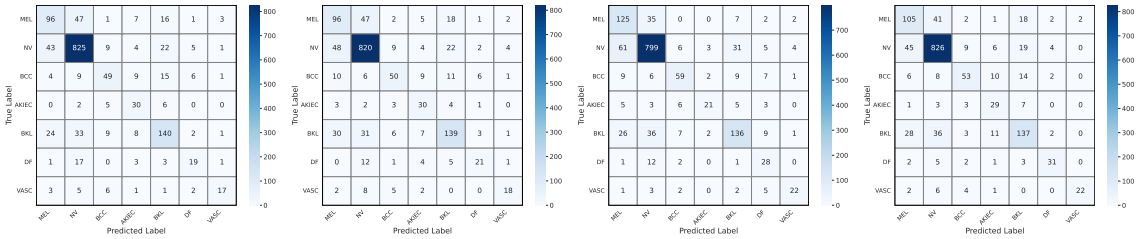
Method	NV	MEL	BCC	BKL	AK	SCC	VASC	DF
Cross-Entropy (CE)	**0.8973**	0.6604	0.7521	0.5962	0.4942	0.4061	0.6536	0.3403
Focal Loss (Lin et al., 2017)	0.8948	0.6377	0.7758	0.5843	0.5395	0.4289	0.6118	0.3504
LDAM Loss (Cao et al., 2019)	0.8821	0.6842	0.8263	0.5910	0.4118	0.4623	0.8284	0.5990
CB Loss (Cui et al., 2019)	0.8366	0.6327	0.7822	0.6133	0.5708	0.5992	0.8284	0.6922
CCE Loss (Kim et al., 2021)	0.8927	0.6388	0.7763	0.6015	0.4704	0.3723	0.7353	0.3221
PF Loss (Du et al., 2021)	0.7866	0.6517	0.7980	0.6147	0.5984	0.5794	0.8984	0.7125
AWB Loss (Yue et al., 2023)	0.7804	**0.6881**	**0.8440**	0.6338	0.5650	0.5992	**0.9216**	0.6615
DAL (Ours)	0.7270	0.6426	0.7365	**0.6367**	**0.6317**	**0.6308**	0.9176	**0.7375**

Tab. 5 presents the per-class recall for both ISIC 2018 and ISIC 2019, averaged over five independent runs, and provides a closer look into the performance of each class on the validation set. A closer inspection reveals that the validation set improvements for the minority classes do not carry over to the test set, as demonstrated in Tab. 2. For example, we observe a significant performance gap in the minority-class TPR on the test sets compared to the validation sets, especially for AWB loss, where the gap is considerably larger. In contrast, PF loss generalizes better from validation to test sets compared to AWB. Although our method underperforms compared to AWB loss on the ISIC 2019 validation results, the test set performance clearly shows that it generalizes more effectively to unseen patients. This suggests that AWB loss may be overfitting to the training set, likely due to data leakage or noise in the dataset, whereas our method learns more discriminative features and achieves stronger generalization. The remaining loss functions primarily benefit the majority classes.

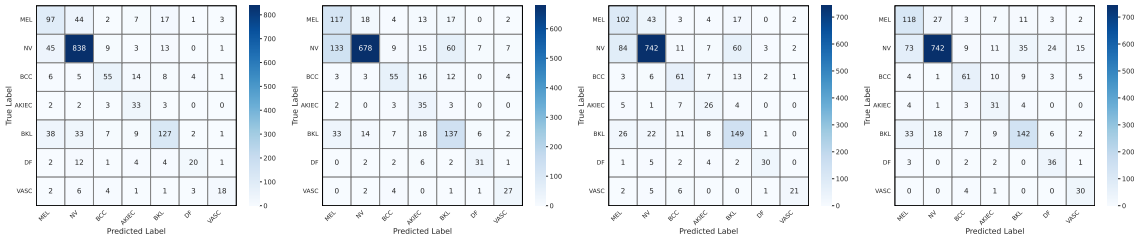
To provide more thorough insight into performance differences, Fig. 4 and Fig. 5 show the confusion matrices for all loss functions using the best-performing model over five independent runs on the test sets of ISIC 2018 and ISIC 2019, respectively. We do not report further evaluation metrics for the APTOS dataset, as it does not include a predefined test set and, therefore, test-set generalization cannot be reliably assessed.

Appendix D. Negative Angle Distributions

In Fig. 1 in the main paper, we have shown how the distribution of negative angles evolve over time, with a fixed parameter of $s = 4$. Using this distribution, we performed a theoretical analysis and to arrive at an optimal value for $s = 4$. We acknowledge that this is a chicken/egg problem, and there is a possibility that other parameters of s yield substantially different angle distributions. Therefore, we performed the same analysis for other s parameters in Fig. 6 to show that the distribution of negative angles $\theta_{i,j}$ is mostly

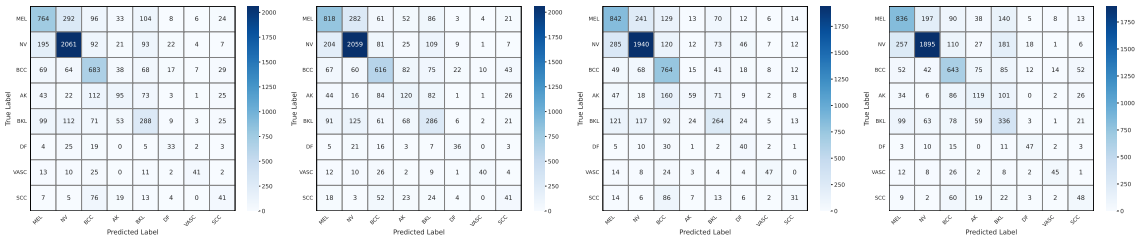


(a) CE (b) Focal (c) LDAM (d) CB

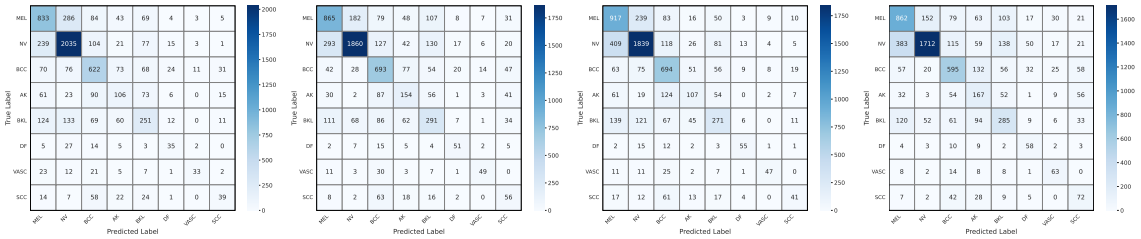


(e) CCE (f) PF (g) AWB (h) Ours

Figure 4: Confusion Matrices for ISIC 2018. We view the confusion matrices of the 7 classes within the ISIC 2018 test set using different loss functions (best of five runs).



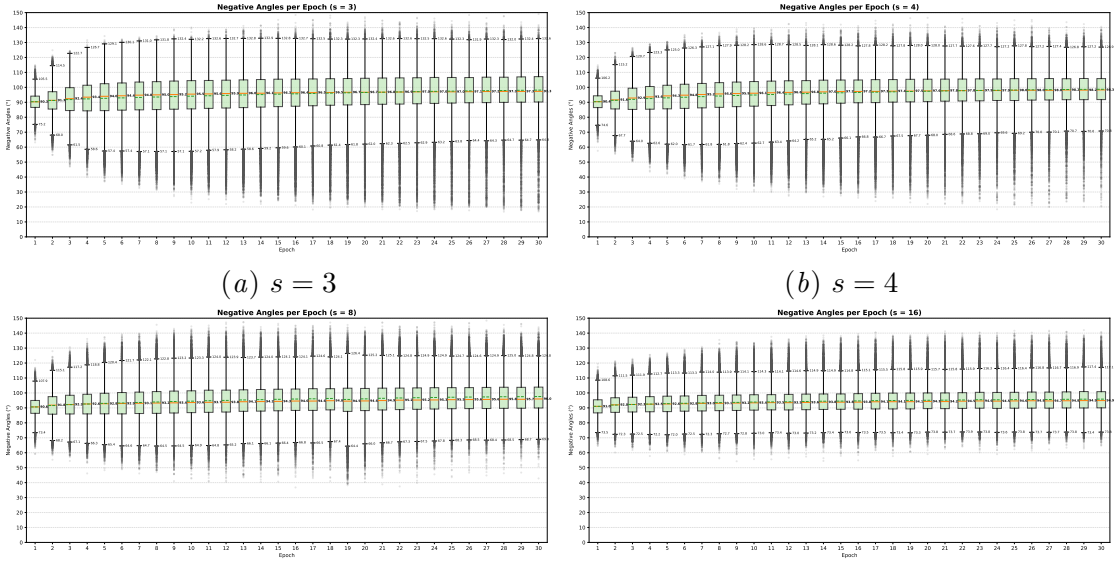
(a) CE (b) Focal (c) LDAM (d) CB



(e) CCE (f) PF (g) AWB (h) Ours

Figure 5: Confusion Matrices for ISIC 2019. We view the confusion matrices of the 8 classes within the ISIC 2019 test set using different loss functions (best of five runs).

independent of the actual s parameter, so even if we would have started our analysis with a larger s value, we would have arrived at similar theoretical bounds and optimum value for s.



(a) $s = 3$ (b) $s = 4$

(c) $s = 8$ (d) $s = 16$

Figure 6: Negative Angles. We show the distribution of negative angles $\theta_{i,j}$ during training on ISIC 2018 dataset with different s parameter values, and including all training epochs.