Taming the Tail: Leveraging Asymmetric Loss and Padé Approximation to Overcome Medical Image Long-Tailed Class Imbalance

Pankhi Kashyap[1]
pankhikashyap.research@gmail.com
Pavni Tandon[1]
19d070044@iitb.ac.in
Sunny Gupta[1]
sunnygupta@iitb.ac.in
Abhishek Tiwari[1]
abhishektiwari.bm@gmail.com
Ritwik Kulkarni[2,3]
ritwik@oraiclebio.com
Kshitij Sharad Jadhav[1]
kshitij.jadhav@iitb.ac.in

[1] Indian Institute of Technology, Bombay Mumbai, India
[2] University of Helsinki
[3] Oraicle Biosciences LTD

Abstract

Long-tailed problems in healthcare emerge from data imbalance due to variability in the prevalence and representation of different medical conditions, warranting the requirement of precise and dependable classification methods. Traditional loss functions such as cross-entropy and binary cross-entropy are often inadequate due to their inability to address the imbalances between the classes with high representation and the classes with low representation found in medical image datasets. We introduce a novel polynomial loss function based on Padé approximation, designed specifically to overcome the challenges associated with long-tailed classification. This approach incorporates asymmetric sampling techniques to better classify under-represented classes. We conducted extensive evaluations on three publicly available medical datasets and a proprietary medical dataset. Our implementation of the proposed loss function is open-sourced in the public repository: https://github.com/ipankhi/ALPA.

1 Introduction

Medical image classification is a crucial component in the development of effective diagnostic as well as prognostic tools [28]. The utility of these tools often relies on the ability to manage and interpret large volumes of medical imaging data. However, a pervasive challenge encountered in these datasets is the prevalence of a long-tailed distribution—a scenario where the majority of data samples belong to a few dominant classes, while the remaining

classes have significantly fewer samples [23]. This imbalance poses significant challenges in training accurate classifiers, as conventional machine learning algorithms often struggle to learn from classes with limited samples [16]. The existence of long tails in medical image datasets can be attributed to several factors, such as the rarity of certain medical conditions or diseases leading to a limited number of samples available for those classes [27]. As a result, these classes have few positive examples, making them challenging to detect and classify accurately. Furthermore, data collection in medical imaging is often biased towards common and easily accessible conditions, resulting in an uneven representation of different classes [9], [12]. The challenges posed by long-tailed class distributions in medical image classification have thus prompted researchers to explore various solutions.

In their survey "Deep Long-Tailed Learning," Zhang *et al* grouped existing solutions into three main categories: class re-balancing, information augmentation, and module improvement [38]. These were further classified into nine sub-categories; Re-sampling methods, such as oversampling and undersampling, involve altering the class distribution in the training set [6], [21]. Class-sensitive learning methods, like re-weighting [19], [8] ,[34] and re-margining [5] aim to re-balance training loss values for different classes promoting equitable learning, while logit adjustment techniques [25] aim to re-calibrate the output probabilities of the classifier to account for the imbalanced class distribution. Transfer learning aims to enhance model training on a target domain by transferring knowledge from a source domain [10] and data augmentation techniques diversify datasets by either applying transformations directly to existing data or by utilizing generative AI methods, such as Generative Adversarial Networks (GANs) and Diffusion models, to create new samples [30], [33], [32]. Representation learning methods aim to learn more discriminative feature representations that can better separate different classes [15], [37], while classifier design involves optimizing the architecture and parameters of the classifier to improve its performance on long-tailed datasets by transferring geometric structures from head classes to tail classes [20]. Decoupled training techniques decouple the training of the classifier into two stages: a representation learning stage and a classifier learning stage [14]. Finally, ensemble learning methods combine multiple classifiers, each trained on different subsets of the data or with different techniques, to improve classification performance [40], [18].

Loss functions play a crucial role in guiding model training. Class-sensitive loss functions are designed to mitigate the adverse effects of class imbalance by adjusting the contribution of each class to the overall loss calculation. These loss functions aim to ensure that the model does not disproportionately prioritize majority classes over minority ones during training. By doing so, they help alleviate the challenges associated with skewed class distributions and improve the model's ability to generalize across all classes. Focal loss, introduced by [19], is a classic strategy to mitigate long-tailedness in classification tasks by dynamically adjusting the weighting of different examples during training to focus more on hard-to-classify samples. Similarly, class-balanced loss [8] assigns weights to different classes inversely proportional to their frequencies. Asymmetric loss [2] and asymmetric polynomial loss [11] are variants of loss functions designed to penalize misclassifications of minority classes more heavily than majority classes.

2 Our contribution

Polynomial expansions allow for the modeling of higher-order interactions between variables that linear models typically miss, thus providing a more nuanced and detailed depiction of

data behaviors. Additionally, this method can be particularly useful in healthcare image analysis domains where capturing non-linear patterns is essential for predicting outcomes with high accuracy. By incorporating polynomial terms, models can approximate a wider range of functions, thereby adapting more effectively to the underlying complexities of the dataset [17].

The Padé approximation [35] is a mathematical technique that approximates a function through a ratio of two polynomials rather than relying solely on polynomial expansions. In earlier works, learnable activation functions based on the Padé approximation have shown promising performance [26], [3]. This method is particularly effective in modeling functions with singularities and provides a more accurate approximation over certain intervals. By applying the Padé approximation to the BCE loss function, we aim to achieve a more precise representation of the loss landscape, enabling our model to adjust more effectively to the true distribution of training data. Asymmetric focusing addresses the imbalance between the positive and negative classes by applying different weights to the loss contributions of each class. This technique is crucial in long-tail scenarios, where the minority class requires greater emphasis to ensure sufficient model sensitivity towards less frequent conditions.

In our research,

- We introduce a novel approach to address the challenge of long-tailed medical image classification by proposing a Padé expansion-based polynomial loss function.
- Furthermore, by implementing an asymmetric focus, this loss function demonstrates enhanced classification performance for under-represented classes compared with other loss function-driven techniques in long-tailed problems.
- We rigorously tested the efficacy of our method (**A**symmetric **L**oss with **P**adé **A**pproximation [**ALPA**]) across three publicly available medical image datasets in addition to a proprietary medical image dataset.

3 Related Work

The development of loss functions tailored for imbalanced datasets has been a focal point of research. The standard cross entropy loss is a commonly used loss function for classification tasks, defined as:

$$\begin{cases} L_{CE}^{+} = -\sum_{i=1}^{K} y_i \log(\hat{y}_i), \\ L_{CE}^{-} = -\sum_{i=1}^{K} (1-y_i) \log(1-\hat{y}_i), \end{cases} \tag{1}$$

where K is the number of classes, and y_i and $\hat{y}_i$ represent the ground-truth and estimated probabilities for class i respectively. However, when dealing with imbalanced datasets, the cross entropy loss (Equation 1) treats all class samples equally and does not consider the imbalanced distribution. Thus, it tends to prioritize majority classes, leading to suboptimal performance on minority classes. Lin *et al* [19] proposed Focal Loss (Equation 2) as a modification, which dynamically adjusts the loss weights based on the predicted probabilities. This enables Focal Loss to down-weigh the loss assigned to well-classified examples and focus more on difficult-to-classify instances. It is formulated as follows:

$$\begin{cases} L_{Focal}^{+} = \alpha_{+}(1-\hat{y})^{\gamma} \log(\hat{y}) \\ L_{Focal}^{-} = \alpha_{-}\hat{y}^{\gamma} \log(1-\hat{y}) \end{cases} \tag{2}$$

where α_+ and α_- are the balancing factors for positive and negative losses, respectively, and γ is the focusing parameter. Notably, setting $\gamma = 0$ yields the binary cross-entropy loss. However, Focal Loss uses the same focusing parameter γ for both positive and negative losses. This can lead to suboptimal performance, especially in scenarios where the tail classes require different treatment compared to the head classes.

The Asymmetric Loss [2] introduces an asymmetric weighting scheme to alleviate the weaknesses of the Focal Loss. Equation 3 assigns different focusing parameters for positive and negative losses, allowing for separate optimization of the training of positive and negative samples. It is defined as:

$$\begin{cases} L^+_{ASL} = (1-\hat{y})^{\gamma_+} \log(\hat{y}) \\ L^-_{ASL} = \hat{y}^{\gamma_-} \log(1-\hat{y}) \end{cases} \tag{3}$$

where γ_+ and γ_- are the focusing parameters for positive and negative losses respectively.

The Class-Balanced (CB) Loss [8] is another technique aimed at mitigating the challenges posed by class imbalance in training datasets and is formulated as follows:

$$L_{CB} = -\frac{1}{K} \sum_{k=1}^{K} \frac{1-\beta^{\gamma}}{1-\beta} \cdot y_k^{\gamma} \cdot \log(\hat{y}_k) \tag{4}$$

where γ is the focusing parameter and β is a hyperparameter controlling the balance between the effective number of samples for each class and the average effective number of samples. Unlike traditional loss functions, CB loss (Equation 4) introduces a mechanism to dynamically adjust the weights of different classes during the training process. This adjustment is based on the effective number of samples for each class, thereby ensuring that minority classes receive higher weights compared to majority classes. In [13] Jamal *et al* shows that class-balanced loss can underperform due to the domain gap between head and tail classes. Similarly, the Label-Distribution-Aware Margin (LDAM) Loss [5] is a loss function designed to enhance the discriminative power of deep neural networks by explicitly maximizing the margins between different classes. Unlike traditional loss functions like cross-entropy, LDAM loss focuses on optimizing the margins between classes in the feature space, thereby promoting better class separation and improved generalization performance. However, negative eigenvalues can persist in the LDAM loss landscape for tail classes due to insufficient data representation, leading to directions of negative curvature [31], making it inefficient for achieving effective generalization on tail classes.

4 METHOD

4.1 Padé approximants for BCE loss

The BCE loss can be decomposed into C-independent binary classification subproblems:

$$L_{\text{BCE}} = \frac{1}{C} \sum_{i=0}^{C} \left(y_i L^+ + (1-y_i) L^- \right), \quad y_i \in \{1,0\} \tag{5}$$

where $L^+ = -\log(\hat{y}_i)$ is for the positive class, and $L^- = -\log(1-\hat{y}_i)$ is for the negative class. Here, $\hat{y}_i$ is the prediction probability after the sigmoid function. We first define L_{BCE} in Padé approximant form. For positive classes where $y_i = 1$, we set the polynomial expansion point

to be 1; for negative classes where $y_i = 0$, we set the expansion point to 0. Thus, Padé approximants for the positive and negative classes for a single sample are:

$$\begin{aligned} L^{+}_{\text{Padé}} &= \frac{a_0 + \sum_{m=1}^{M} a_m \hat{y}^m}{1 + \sum_{n=1}^{N} b_n \hat{y}^n}, \\ L^{-}_{\text{Padé}} &= \frac{c_0 + \sum_{m=1}^{M} c_m (1-\hat{y})^m}{1 + \sum_{n=1}^{N} d_n (1-\hat{y})^n} \end{aligned} \tag{6}$$

where $\hat{y}$ represents the prediction probability of a single sample, while M and N represent the orders of the numerator and denominator polynomials, respectively, and a_0, a_m, b_n, c_0, c_m, and d_n are coefficients of Padé approximants.

4.2 Derivation of the coefficients

The conventional Padé approximation of order m/n tends to reproduce the Taylor expansion of order [35] $m+n$, and the coefficients are found by setting

$$\frac{P(x)}{Q(x)} = A(x) \tag{7}$$

where $P(x)$ is a numerator polynomial of order m, $Q(x)$ is the denominator polynomial of order n of Padé approximant , and $A(x)$ is the Taylor expansion of order $m+n$.
In terms of the Taylor Series Expansion, L^{+} and L^{-} are:

$$\begin{cases} L^{+}_{\text{Taylor}} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{(\hat{y}_i - 1)^k}{k} \\ L^{-}_{\text{Taylor}} = -\sum_{k=1}^{\infty} \frac{\hat{y}_i{}^k}{k} \end{cases} \tag{8}$$

In line with previous research that highlights the effectiveness of the first-degree polynomial [17], we adopt the first-order Padé approximation for our loss function. This approach sets both the numerator's and the denominator's orders to one, and for deriving the coefficients, we equate them with the respective Taylor series expansion of the second order.
The first order of 6 would be:

$$\begin{cases} L^{+}_{\text{Padé}} \approx \frac{a_0 + a_1 \hat{y}_i}{1 + b_1 \hat{y}_i} \\ L^{-}_{\text{Padé}} \approx \frac{c_0 + c_1 (1-\hat{y}_i)}{1 + d_1 (1-\hat{y}_i)} \end{cases} \tag{9}$$

Expanding 8 up to second order:

$$\begin{cases} L^{+}_{\text{Taylor}} \approx (\hat{y}_i - 1) - \frac{1}{2}(\hat{y}_i - 1)^2 \\ L^{-}_{\text{Taylor}} \approx -\hat{y}_i - \frac{1}{2}\hat{y}_i{}^2 \end{cases} \tag{10}$$

By equating 9 and 10, we obtain the values of the coefficients $a_0 = -1.5$, $a_1 = 1.5$, and $b_1 = 0$, and the coefficients $c_0 = -1$, $c_1 = 1$, and $d_1 = 0$.

4.3 Addition of asymmetric focusing mechanism and balancing factors

Allowing separate optimization of the positive and negative samples, we add balancing factors and asymmetric focusing mechanism from 3. Our proposed asymmetric loss based on

Padé approximation becomes,

$$L_{\text{ALPA}} = \sum_{i=1}^{N} \left[\alpha y_i (1-\hat{y}_i)^{\gamma_{\text{pos}}} L^{+}_{\text{Padé}} + \beta (1-y_i) \hat{y}_i^{\gamma_{\text{neg}}} L^{-}_{\text{Padé}} \right] \cdot W_i \tag{11}$$

where N is the number of labels, $\hat{y}_i$ is the predicted probability and y_i is the binary target label for the i-th sample, α and β are balancing parameters, γ_{pos} and γ_{neg} are focusing parameters, $L^{+}_{\text{Padé}}$ and $L^{-}_{\text{Padé}}$ are the Padé Approximation forms for positive and negative predictions, respectively. W_i is the weight for the i-th sample, calculated as $(1-pt_i)^{\gamma}$, where pt_i is the predicted probability adjusted for the target label such that $pt_i = y_i\hat{y}_i + (1-y_i)(1-\hat{y}_i)$, and γ is the summation of focusing parameters, with γ_{pos} applied for positive targets and γ_{neg} for negative targets.

We studied the effects of hyperparameters α, β, γ_{pos} and γ_{neg} on the loss function and evaluated the loss function using the best-performing combination of values on the datasets used in this study.

4.4 Gradient Analysis

Gradients play a pivotal role in the training process, guiding the adjustments of network weights with respect to the input logit z. In this section, following the work of [2], we provide a comprehensive analysis of the loss gradients of ALPA compared to established loss functions such as Cross Entropy, Focal Loss, and Asymmetric Loss.
For ALPA, we have $L^{-}_{\text{ALPA}} = (\hat{y}_i)^{\gamma_{\text{neg}}+1}$, thus, the negative gradient equation for the ALPA function is given by:

$$\frac{dL^{-}_{\text{ALPA}}}{dz} = \frac{dL^{-}_{\text{ALPA}}}{d\hat{y}_i} \cdot \frac{d\hat{y}_i}{dz} = (\hat{y}_i)^{\gamma_{\text{neg}}+1} \cdot (1-\hat{y}_i) \cdot (\gamma_{\text{neg}}+1)$$

where $\hat{y}_i = \frac{1}{1+e^{-z}}$ represents the predicted probability for the input logit z and γ_{neg} is the focusing parameter for negative targets.

The results of the gradient analysis are shown in Figure 1.

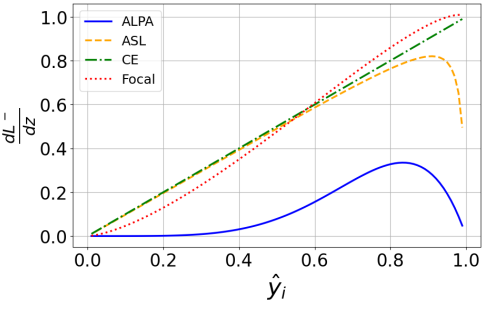

Figure 1: Comparison of loss gradients for ALPA ($\gamma_{\text{neg}} = 4$), ASL ($m = 0.01$, $\gamma_{\text{neg}} = 0.01$), CE ($m = 0$, $\gamma_{\text{neg}} = 0$), and Focal Loss ($\gamma = 0.5$)

We observe that the gradient for ALPA increases moderately as the probability $\hat{y}_i$ approaches 1. This suggests that our proposed loss function provides a consistent learning

Table 1: Details of long-tailed medical datasets.

Dataset	Classes	Samples	Imbalance Ratio
APTOS2019	5	3,662	10
DermaMNIST	7	10,015	58
BoneMarrow	17	147,904	621.79
Oraiclebio	52	3643	164

signal across the probability spectrum. It neither penalizes very harshly for misclassifications (when $\hat{y}_i$ is low) nor relaxes too much when the classification is correct (when $\hat{y}_i$) is high). Thus, ALPA appears to be a good choice for consistent learning across all probabilities. By focusing on harder examples and not over-penalizing the correctly classified ones, it achieves better generalization compared to other losses.

5 Experimental Setup

5.1 Datasets

The APTOS 2019 BD dataset [4] includes data from individuals diagnosed with varying levels of Diabetic Retinopathy (DR), categorized into five classes: No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR. The DermMNIST dataset [36] comprises 450x600 pixel images of various skin diseases classified into seven categories: Melanoma, Melanocytic Nevus, Basal Cell Carcinoma, Actinic Keratosis, Benign Keratosis, Dermatofibroma, and Vascular Lesion. The BoneMarrow dataset [24] contains expertly annotated cells from bone marrow smears of 945 patients, classified into 17 types including Basophil (BAS), Blast (BLA), Erythroblast (EBO), and more. The Oraiclebio dataset, which remains proprietary, includes 3,643 images of oral regions featuring 52 classes of precancerous and cancerous lesions.

Details of these datasets are summarized in Table 1, where the imbalance ratio, defined as N_{max}/N_{min} (with N representing the sample count per class), illustrates the significance of the long-tailed distribution. For experimentation, each dataset was split 80-20 into training and testing sets, and a 5-fold cross-validation strategy was used during training to enhance model reliability.

5.2 Implementation

We use ConvNeXT-B [22] as the backbone for the proposed loss. We resize the input images as 256 x 256 and exploit the data augmentation schemes following the previous work [1, 7]. We train our networks using the Adam optimizer with 0.9 momentum and 0.001 weight decay. The batch size is 128, and the initial learning rate is set to $1e{-}4$. Our networks are trained on PyTorch version 2.2.1 with RTX A6000 GPUs. We use accuracy, balanced accuracy and F1-score as evaluation metrics for this study.

6 Results

In this section, we present experimental results validating the effectiveness of the proposed **ALPA** function. We first analyze the impact of hyperparameters on the loss function and

then compare **ALPA** with state-of-the-art loss functions like Asymmetric Loss, Focal Loss and Cross Entropy.

6.1 Effect of the hyperparameters

To evaluate the effect of hyperparameters, we experimented as follows:

- **Loss v1**: Hyperparameters were randomly set as $\alpha = 1$, $\beta = 1$, $\gamma_{pos} = 0$, and $\gamma_{neg} = 4$. This is indicated as Loss v1 in Table 2.
- **Loss v2** ($L_{\mathbf{ALPA}}$): Using random search, hyperparameters were optimized within the ranges α and β (0.5 to 2), and γ_{pos} and γ_{neg} (0 to 5). Final values were $\alpha = 0.875$, $\beta = 1.625$, $\gamma_{pos} = 0$, and $\gamma_{neg} = 4$. This is indicated as Loss v2 in Table 2.
- **Loss v3**: Incorporating Hill Loss [39] following [29], we added $\lambda - \hat{y}_i$ to L^- ($\lambda = 1.5$), optimizing via random search to $\alpha = 1.25$, $\beta = 2$, $\gamma_{pos} = 3$, and $\gamma_{neg} = 2$. This is indicated as Loss v3 in Table 2.

Results on the APTOS2019 dataset for these settings are shown in Table 2. We focused on detecting crucial cases like Proliferative DR and examined the performance of underrepresented classes to proceed with Loss v2. From here on, Loss v2 is referred to as L_{ALPA}.

Table 2: Performance metrics for different versions of Loss.

Classes	Number of training samples	Loss v1		Loss v2		Loss v3	
		Acc	F1-score	Acc	F1-score	Acc	F1-score
No DR	1454	98.86	0.98	98.01	0.98	98.86	0.98
Mild	786	64.71	0.62	58.82	0.59	26.47	0.37
Moderate	302	88.26	0.76	86.38	0.79	92.96	0.79
Severe	230	2.78	0.5	33.33	0.44	36.11	0.39
Proliferative DR	157	23.08	0.36	47.69	0.58	30.77	0.45

6.2 Comparison with existing methods

We compare our proposed loss function with state-of-the-art methods such as ASL, Focal Loss, LDAM, and CE on the datasets listed in Section 5.1. Results on the publicly available datasets are presented in Tables 3, 4, and 5, while the results for LDAM loss functions can be found in the supplementary materials. **ALPA** consistently excels in classes with fewer samples while maintaining competitive accuracy in classes with higher representation. In terms of balanced accuracy, **ALPA** surpasses all other loss functions across the three public datasets.

Table 3: Comparison of different loss functions on the APTOS2019 dataset.

Classes	Number of training samples	ALPA		ASL		CE		FOCAL	
		Acc	f1-score	Acc	f1-score	Acc	f1-score	Acc	f1-score
No DR	1454	98.01	0.98	98.58	0.98	98.86	0.97	**99.43**	0.94
Mild	786	**58.82**	0.59	41.18	0.50	16.18	0.27	10.29	0.16
Moderate	302	86.38	0.79	90.61	0.77	**97.18**	0.76	88.73	0.75
Severe	230	**33.33**	0.44	11.11	0.17	8.33	0.14	8.33	0.15
Proliferative DR	157	**47.69**	0.58	43.08	0.57	16.92	0.28	30.77	0.41
Balanced Accuracy			**0.65**		0.57		0.47		0.48

Table 4: Comparison of different loss functions on the DermaMNIST dataset.

Classes	Number of training samples	ALPA		ASL		CE		FOCAL	
		Acc	f1-score	Acc	f1-score	Acc	f1-score	Acc	f1-score
akiec	256	4.23	0.08	**16.9**	0.29	0.00	0.00	1.41	0.03
bcc	406	19.44	0.31	**26.85**	0.37	5.56	0.10	**26.85**	0.38
bkl	882	**35.94**	0.45	18.43	0.28	17.97	0.26	1.84	0.04
df	88	**33.33**	0.39	22.22	0.35	3.70	0.07	7.41	0.10
mel	885	**10.09**	0.18	8.33	0.15	0.44	0.01	1.32	0.03
nv	5375	98.50	0.86	96.62	0.84	**99.70**	0.82	98.27	0.82
vasc	120	**72.73**	0.35	54.55	0.19	36.36	0.37	81.82	0.42
Balanced Accuracy			**0.39**		0.35		0.23		0.31

Table 5: Comparison of different loss functions on the Bone Marrow dataset.

Classes	Number of training samples	ALPA		ASL		CE		FOCAL	
		Acc	f1-score	Acc	f1-score	Acc	f1-score	Acc	f1-score
BAS	348	54.84	0.68	**55.91**	0.68	46.24	0.61	51.61	0.66
BLA	9569	85.94	0.87	87.06	0.88	**89.93**	0.88	87.52	0.88
EBO	21883	95.94	0.96	96.23	0.96	96.21	0.96	**96.34**	0.96
EOS	4719	**97.34**	0.97	97.25	0.96	**97.34**	0.97	96.99	0.97
FGC	41	**83.33**	0.77	66.67	0.73	**83.33**	0.77	50.00	0.60
HAC	339	58.57	0.71	67.14	0.77	**72.86**	0.81	71.43	0.80
KSC	38	**100.00**	1.00	75.00	0.75	75.00	0.86	**100.00**	0.89
LYI	54	36.36	0.40	**36.98**	0.44	27.77	0.38	9.09	0.14
LYT	20911	94.13	0.94	**94.92**	0.94	94.75	0.94	94.62	0.94
MMZ	2479	39.58	0.46	36.98	0.45	56.60	0.52	**58.33**	0.56
MON	3230	74.81	0.78	**79.38**	0.79	76.91	0.77	79.01	0.78
MYB	5238	68.76	0.70	70.20	0.70	61.26	0.68	**73.09**	0.73
NGB	7967	67.82	0.69	69.42	0.71	66.12	0.71	**72.06**	0.74
NGS	23628	**94.24**	0.91	92.86	0.92	93.32	0.92	93.65	0.93
PEB	2196	75.55	0.75	77.76	0.78	73.71	0.76	**78.68**	0.78
PLM	6137	**93.90**	0.93	91.49	0.93	92.16	0.93	91.69	0.93
PMO	9546	85.58	0.86	**88.40**	0.86	88.19	0.86	84.03	0.85
Balanced Accuracy			**0.77**		0.75		0.76		0.76

In the Oraiclebio dataset, **ALPA** delivers the highest accuracy in 23 classes, achieving a balanced accuracy of 51.06% and performing similarly to ASL, which attains a balanced accuracy of 52%, demonstrating its robustness across varying data support levels. Meanwhile, CE averages 47.92% accuracy, and Focal Loss performs significantly worse, with an average balanced accuracy of just 22%. Thus, **ALPA** stands out for its ability to handle diverse and imbalanced datasets effectively.

7 Conclusion

In this study, we present a Padé approximation-based loss function with asymmetric focusing, tailored for multi-class classification tasks with long-tailed distributions. Our proposed loss function demonstrates competitive and superior performance on long-tailed datasets when benchmarked against previous state-of-the-art approaches. We believe that our findings can serve as a valuable resource for future research, offering a foundation for further development and integration into new studies.

8 Future work

The learning process of the model is intrinsically tied to data representation. Modifying the loss function alone, however, may have limited potential for performance improvement. To enhance class-wise accuracy, integrating loss functions with data augmentation strategies and data generation pipelines presents a promising approach. Data augmentation artificially expands the training dataset by creating modified versions of existing images, while data generation pipelines synthesize entirely new samples. These methods can help balance class representation and bolster model robustness. A more thorough exploration of these techniques in future work could offer substantial benefits in addressing class imbalance.

Acknowledgements

This work was supported through the funded project by Oraicle Biosciences LTD (OraiBio) to Indian Institute of Technology, Bombay.

References

[1] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3478–3488, 2021.

[2] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification, 2021.

[3] Francis Benistant. Innovative activation functions: Harnessing the power of padé approximants in artificial neural networks : Application to autoencoder on the mnist dataset. 02 2023.

[4] Jyostna Devi Bodapati, Veeranjaneyulu Naralasetti, Shaik Nagur Shareef, Saqib Hakak, Muhammad Bilal, Praveen Kumar Reddy Maddikunta, and Ohyun Jo. Blended multimodal deep convnet features for diabetic retinopathy severity prediction. *Electronics*, 9(6):914, 2020.

[5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss, 2019.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL http://dx.doi.org/10.1613/jair.953.

[7] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019.

[8] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. URL https://arxiv.org/abs/1901.05555.

[9] Kamana Dahal and Mohd Hasan Ali. A hybrid gan-based dl approach for the automatic detection of shockable rhythms in aed for solving imbalanced data problems. *Electronics*, 12(1):13, 2022.

[10] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty, 2019.

[11] Yusheng Huang, Jiexing Qi, Xinbing Wang, and Zhouhan Lin. Asymmetric polynomial loss for multi-label classification, 2023.

[12] Saeed Iqbal, Adnan N. Qureshi, Jianqiang Li, Imran Arshad Choudhry, and Tariq Mahmood. Dynamic learning for imbalanced data in learning chest x-ray and ct images. *Heliyon*, 9, 2023. URL https://api.semanticscholar.org/CorpusID:259040001.

[13] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7610–7619, 2020.

[14] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition, 2020.

[15] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021.

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

[17] Zhaoqi Leng, Mingxing Tan, Chenxi Liu, Ekin Dogus Cubuk, Xiaojie Shi, Shuyang Cheng, and Dragomir Anguelov. Polyloss: A polynomial expansion perspective of classification loss functions. *arXiv preprint arXiv:2204.12511*, 2022.

[18] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax, 2020.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[20] Bo Liu, Haoxiang Li, Hao Kang, Gang Hua, and Nuno Vasconcelos. Gistnet: A geometric structure transfer network for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8209–8218, October 2021.

[21] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009. doi: 10.1109/TSMCB.2008.2007853.

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[23] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world, 2019.

[24] C Matek, S Krappe, C Münzenmayer, T Haferlach, and C Marr. An expert-annotated dataset of bone marrow cytology in hematologic malignancies. *The Cancer Imaging Archive*, 2021.

[25] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment, 2021.

[26] Alejandro Molina, Patrick Schramowski, and Kristian Kersting. Padé activation units: End-to-end learning of flexible activation functions in deep networks, 2020.

[27] A Obukhov. Proceedings of the computational methods in systems and software, 2021.

[28] Andreas S. Panayides, Amir Amini, Nenad D. Filipovic, Ashish Sharma, Sotirios A. Tsaftaris, Alistair Young, David Foran, Nhan Do, Spyretta Golemati, Tahsin Kurc, Kun Huang, Konstantina S. Nikita, Ben P. Veasey, Michalis Zervakis, Joel H. Saltz, and Constantinos S. Pattichis. Ai in medical imaging informatics: Current challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, 24(7):1837–1857, 2020. doi: 10.1109/JBHI.2020.2991043.

[29] Wongi Park, Inhyuk Park, Sungeun Kim, and Jongbin Ryu. Robust asymmetric loss for multi-label long-tailed learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2711–2720, 2023.

[30] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning, 2017.

[31] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, et al. Escaping saddle points for effective generalization on class-imbalanced data. *Advances in Neural Information Processing Systems*, 35:22791–22805, 2022.

[32] Zakaria Rguibi, Abdelmajid Hajami, Dya Zitouni, Amine Elqaraoui, Reda Zourane, and Zayd Bouajaj. Improving medical imaging with medical variation diffusion model: An analysis and evaluation. *Journal of Imaging*, 9(9):171, 2023.

[33] Youssef Skandarani, Pierre-Marc Jodoin, and Alain Lalande. Gans for medical image synthesis: An empirical study. *Journal of Imaging*, 9(3), 2023. ISSN 2313-433X. doi: 10.3390/jimaging9030069. URL https://www.mdpi.com/2313-433X/9/3/69.

[34] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition, 2020.

[35] Eric W. Weisstein. Padé approximant. https://mathworld.wolfram.com/PadeApproximant.html, n.d. From MathWorld–A Wolfram Web Resource.

[36] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[37] Zhixiong Yang, Junwen Pan, Yanzhan Yang, Xiaozhou Shi, Hong-Yu Zhou, Zhicheng Zhang, and Cheng Bian. Proco: Prototype-aware contrastive learning for long-tailed medical image classification, 2022.

[38] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey, 2023.

[39] Youcai Zhang, Yuhao Cheng, Xinyu Huang, Fei Wen, Rui Feng, Yaqian Li, and Yandong Guo. Simple and robust loss design for multi-label learning with missing labels. *arXiv preprint arXiv:2112.07368*, 2021.

[40] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition, 2020.