LS+: Informed Label Smoothing for Improving Calibration in Medical Image Classification

Abhishek Singh Sambyal(✉)[1], Usma Niyaz[1], Saksham Shrivastava[1], Narayanan C. Krishnan[2], and Deepti R. Bathula[1]

[1] Department of Computer Science & Engineering, Indian Institute of Technology Ropar, Punjab, India
{abhishek.19csz0001, usma.20csz0015, 2022aim1012, bathula}@iitrpr.ac.in
[2] Department of Data Science, Indian Institute of Technology Palakkad, Kerala, India
ckn@iitpkd.ac.in

Abstract. Deep Neural Networks (DNNs) exhibit exceptional performance in various tasks; however, their susceptibility to miscalibration poses challenges in healthcare applications, impacting reliability and trustworthiness. Label smoothing, which prefers soft targets based on uniform distribution over labels, is a widely used strategy to improve model calibration. We propose an improved strategy, Label Smoothing Plus (LS+), which uses class-specific prior that is estimated from validation set to account for current model calibration level. We evaluate the effectiveness of our approach by comparing it with state-of-the-art methods on three benchmark medical imaging datasets, using two different architectures and several performance and calibration metrics for the classification task. Experimental results show notable reduction in calibration error metrics with nominal improvement in performance compared to other approaches, suggesting that our proposed method provides more reliable prediction probabilities. Code is available at https://github.com/abhisheksambyal/lsplus.

Keywords: Calibration · Label smoothing plus · Medical imaging · Deep neural networks · Reliability.

1 Introduction

Deep neural networks (DNNs) have demonstrated outstanding performance across various medical image tasks, including classification, segmentation, and detection [11]. However, modern DNNs are prone to miscalibration, compromising the reliability and trustworthiness of their predictions – critical factors in healthcare applications [10]. Therefore, addressing the issue of miscalibration and enhancing model calibration is of utmost importance.

Various approaches including data augmentation [24], ensemble [12], label smoothing [16, 23], focal loss [15], entropy-based regularization and feedback calibration during training [13, 21], have been proposed to mitigate DNN miscalibration. While some of these approaches involve varying the inputs to the

DNN [24], others focus on changing the true label distribution [17, 29]. Studies have demonstrated the effectiveness of smoothing true labels during training for improving calibration [16]. Probabilities from DNNs serve as confidence indicators for predictions; High probabilities signify stronger belief in a predicted class, crucial in fields like medical diagnosis. However, interpreting the DNN results is incomplete without taking into account the model calibration [9, 22]. Calibration ensures that assigned probabilities accurately reflect the true likelihood of events. Without proper calibration, interpretations based solely on probabilities may be misleading or unreliable.

Contribution. Miscalibration [6] is defined as the disparity between the true confidence (accuracy) and the predicted confidence (output probability). Achieving perfect calibration entails bringing the predicted confidence score close to accuracy. To address this, we propose Label Smoothing Plus (LS+) a novel and simple extension to label smoothing that substitutes the hard labels with informed smoothened versions computed from the validation set. The contributions of the paper are outlined as follows:

1. We introduce a simple yet effective approach to enhance model calibration by altering the true label distribution with a surrogate distribution computed from the class-wise accuracy on the validation set.
2. Our proposed method improves calibration with better or on par-performance when compared to other popular approaches on three medical imaging datasets.
3. Using retention curves and density plots of correct and incorrect predictions, we observed that our method provides reliable and interpretable scores for model reject/second opinion, which is essential for safety-critical applications.

2 Related Work

Post-hoc calibration — Use a hold-out data set (calibration/validation set) to calibrate the confidence scores of a neural network. Several well-studied calibration methods include Platt scaling [20], isotonic regression [28], and temperature scaling (TS) [6]. Weight scaling [5] is an alternative version of TS for medical imaging tasks that explicitly optimizes the ECE measure to improve calibration. Additionally, class-distribution-aware vectors [8] for TS and label smoothing are used to address class-wise overconfidence. Meta-calibration [3] proposes differentiable ECE-driven calibration to obtain well-calibrated and highly accurate models.

Train-time calibration — An alternative approach that directly generates calibrated DNN models. Explicit confidence penalty (ECP) [19] leverages the entropy of the predicted distribution to regularize the loss function. Both Label smoothing (LS) [16, 23] and Focal loss (FL) [15] implicitly regulate the network output probabilities, encouraging their distribution to closely resemble the uniform distribution. Furthermore, auxiliary loss functions in conjunction with negative log-likelihood (NLL) are used to improve calibration. The difference

between Confidence and Accuracy (DCA) [14] serves as an auxiliary loss, penalizing the model when the cross-entropy loss is reduced but the accuracy remains unchanged. Multi-class Difference in Confidence and Accuracy [7] broadens the scope of DCA by considering the calibration of every class, not solely the top-predicted class.
Our current work proposes a more informed strategy to enhance model calibration; the alignment of predicted probabilities (confidence) with accuracy is achieved by incorporating class specific priors derived from a separate validation set to account for current calibration level of the model.

3 Methodology

3.1 Preliminaries

Consider a multi-class classification problem comprising of K classes. Let $\hat{\mathbf{p}} = [\hat{p}_1, \ldots, \hat{p}_K]$ and $\mathbf{y} = [y_1, \ldots, y_K]$ be the predicted class distribution (confidence scores) of a deep neural network (DNN) and the ground truth one hot label encoding for an instance x respectively.
Calibration — A well-calibrated classifier generates confidence scores that align with the actual frequency of correct predictions. Formally, we can define calibration for a perfectly calibrated model for all classes as, $\mathbb{P}(y = y^* | \hat{\mathbf{p}}[y] = \hat{p}) = \hat{p}$, where, $y \in \text{argmax}_k\ y_k$, $y^* \in \{1, \cdots, K\}$, $\hat{\mathbf{p}}[y]$ is the confidence that sample x belongs to class y. [7]
Hard Labelling (HL) — DNN is conventionally trained using only the cross entropy (CE) loss defined as $CE(\mathbf{y}, \hat{\mathbf{p}}) = -\sum_k y_k \log \hat{p}_k$, which reduces to $\log \hat{p}_k$ if x is labeled k. Minimizing CE loss is equivalent to maximizing the log-likelihood of the correct label. Often, the optimization is continued until $\hat{p}_k$ is very close to y_k. As a result the DNN may suffer from over-fitting causing over confident predictions, leading to poor generalization and miscalibration.
Label Smoothing (LS) — An approach to mitigate miscalibration is to replace the one-hot encoded (*hard*) label vector with a smoothened (*soft*) label vector $\mathbf{y}' = (1 - \alpha)\mathbf{y} + \alpha\mathbf{u}$, where $\mathbf{u}$ is a fixed distribution (typically uniform). Thus, label smoothing strategy involves minimizing $\mathcal{L}_{LS}$ defined as

$$\mathcal{L}_{LS} = H(\mathbf{y}', \hat{\mathbf{p}}) = -\sum_{k=1}^{K} y'_k \log \hat{p}_k = (1 - \alpha)\ CE(\mathbf{y}, \hat{\mathbf{p}}) + \alpha\ CE(\mathbf{u}, \hat{\mathbf{p}}) \tag{1}$$

As the $CE(\mathbf{u}, \hat{\mathbf{p}})$ term penalizes the deviation between prediction ($\hat{\mathbf{p}}$) and prior ($\mathbf{u}$) distributions, it can be expressed using Kullback-Leibler (KL) divergence: $CE(\mathbf{u}, \hat{\mathbf{p}}) = D_{KL}(\mathbf{u}, \hat{\mathbf{p}}) + H(\mathbf{u})$. As $H(\mathbf{u})$, the entropy of $\mathbf{u}$, is a constant, the label smoothing cost function simplifies to [23]:

$$\mathcal{L}_{LS} = (1 - \alpha)\ CE(\mathbf{y}, \hat{\mathbf{p}}) + \alpha\ D_{KL}(\mathbf{u}, \hat{\mathbf{p}}) \tag{2}$$

Algorithm 1 Pseudocode of LS+

Input: A training dataset $\mathcal{D}_T = \{(x_i, y_i)\}_{i=1,\cdots,N}$, a validation dataset $\mathcal{D}_\mathcal{V}$, number of classes K, number of training epochs T, pre-trained model $\mathcal{M}$
Class-wise accuracy vector: $\mathcal{V}^{acc} = \mathcal{M}(\mathcal{D}_\mathcal{V})$, where $\mathcal{V}^{acc} \in \mathbb{R}^K$ and each component of the vector corresponds to the accuracy associated with the class $k \in K$
Compute new, class-specific label distribution set $\{\mathbf{v}^1, \ldots, \mathbf{v}^K\}$ *using Eqn (3)*
Minimize $\mathcal{L}_{LS+}$ over training data using the new distribution computed from $\mathcal{D}_\mathcal{V}$
for $t = 0$ **to** $T - 1$ **do**
    For each training instance i that belongs to class k, choose the corresponding informed prior $\mathbf{v}^k$
    $\mathcal{L}_{LS+} = (1 - \alpha) \cdot CE(\mathbf{y}, \hat{\mathbf{p}}) + \alpha \cdot D_{KL}(\mathbf{v}^k, \hat{\mathbf{p}})$
end for

3.2 Label Smoothing Plus (LS+)

There are two drawbacks with vanilla label smoothing. Firstly, the approach does not take into account the DNN's current calibration level. As a result, forcible application of label smoothing to an already well-calibrated DNN may worsen its calibration. Secondly, the uniform prior does not take into account class-wise calibration levels (poorly and well-calibrated classes are treated alike). We propose Label Smoothing Plus (LS+) that addresses these two drawbacks in one go. LS+ replaces the uniform prior $\mathbf{u}$, with an informed class specific prior $\mathbf{v}^k = [v_1^k, \ldots, v_K^k]$ for $k = \{1, \ldots, K\}$, that is estimated on a separate validation set. In particular, the element v_j^k in the informed prior $\mathbf{v}^k$ for class k is defined as

$$v_j^k = \begin{cases} \mathcal{V}_k^{acc} & \text{if } j == k \\ (1 - \mathcal{V}_k^{acc}) \cdot \frac{1}{K-1} & \text{otherwise} \end{cases} \tag{3}$$

where, $\mathcal{V}_k^{acc}$ is validation set accuracy for class k using the pretrained (without label smoothing) model $\mathcal{M}$. *Example: For a pre-trained, three-class classification model with 60% validation accuracy for a specific class creates a label vector* $[0.6, 0.2, 0.2]$*, which coerces the model to generate class prediction probabilities to match the validation accuracy.*

Learning the priors on the validation set ensures unbiased estimates and takes into account the current model calibration status. Furthermore, the smoothening of the prior is also dependent on the class accuracy. Priors of classes that are already accurately predicted by the model are smoothened to lesser extent than those of classes that are not accurately predicted. During training, the informed prior $\mathbf{v}^k$ corresponding to the ground truth class label for the instance x is used in place for a fixed uniform prior $\mathbf{u}$. In theory, $\mathbf{v}^k$ may be computed periodically after every few training iterations. However, we compute it only once before LS+ is applied. The complete pseudo-code for LS+ is presented in Algorithm 1.

4 Experiments and Results

Datasets — We evaluate LS+ using three benchmark datasets curated for medical image classification: (i) *Chaoyang - Histopathology dataset* [30] consists of colon slides with a patch size of 512×512. It is a multiclass ($K = 4$) dataset that is divided into training and testing sets consisting of 4021 and 2139 images respectively. Furthermore, we partitioned the training set into train (90%) and validation (10%). (ii) A *Minimalist Histopathology Image Analysis (MHIST) dataset* [26] comprises of 3,152 histopathology images of colorectal polyps. It is a binary class ($K = 2$) dataset with images of size 224×224. The training and test sets consist of 2175 and 977 samples, respectively. Here, we partitioned the training dataset into train (80%) and validation (20%). (iii) *International Skin Imaging Collaboration (ISIC - 2018)* [4, 25] is a multi-class dataset ($K = 7$; highly imbalanced) of dermoscopic images of skin with a size of 600×450. It consists of separate train/validation/test sets with 10015/193/1512 samples, respectively. *The performance on the separate test set in all the three datasets facilitates an unbiased evaluation of LS+ and other approaches.*

Network Architectures and Implementation Details — We used two popular image classification architectures: ResNet-34 and ResNet-50, implemented using Tensorflow 2.4. These models are ImageNet pretrained and were specifically chosen for their effectiveness on small biomedical datasets [2, 27]. During training, all images are resized to 224×224 dimension. We used Adam optimizer with a learning rate set to $1e{-}3$, batch size of 8, and standard data augmentation techniques [18]. For training LS+, we used $\alpha = 0.5$.

Baseline Methods — We compare LS+ with the following models: (a) Conventional classification using cross-entropy loss with one-hot encoded labels (Hard Labels), (b) cross-entropy loss with label smoothing (LS) [23], (c) focal loss ($\gamma = 3$) (FL) [15] that provides implicit regularization and two auxiliary loss methods - (d) difference between confidence and accuracy (DCA) [14], and (e) multi-class difference in confidence and accuracy (MDCA) [7].

Evaluation Metrics — We use several metrics to evaluate the models. Performance of the models is measured using accuracy (ACC), area under receiver operating characteristic (AUROC), precision, recall, F1-score. Similarly, a comprehensive comparison of calibration is achieved using expectation calibration error (ECE), adaptive calibration error (ACE), static calibration error (SCE), cross-entropy error (CE) and brier loss (Brier) [22].

4.1 Results

Calibration performance comparison with SOTA — Table 1 provides a quantitative comparison of our method with SOTA approaches on Chaoyang, MHIST and ISIC-2018 datasets, respectively. These results demonstrate that validation accuracy-based label smoothing provides significant and consistent reduction across all calibration error metrics. Remarkably, this improvement in calibration was achieved without compromising performance. In fact, marginal

Table 1: **Quantitative Results.** Performance and Calibration results on the test set of three benchmark datasets. The reported values are the average of 3 runs and given as percentages (%) with SD (σ) as subscript. ↑: Higher is better, ↓: Lower is better. Architectures: R34 *(ResNet-34)*, R50 *(ResNet-50)*; Datasets: D1 *(Chaoyang)*, D2 *(MHIST)* and D3 *(ISIC)*.

D1	Method	ACC ↑	AUROC ↑	Precision ↑	Recall ↑	F1 ↑	ECE ↓	ACE ↓	SCE ↓	CE ↓	Brier ↓
R34	HL	$81.50_{1.2}$	$\mathbf{94.08}_{0.5}$	$75.91_{1.1}$	$74.58_{1.3}$	$75.10_{1.3}$	$11.33_{2.5}$	$11.21_{2.7}$	$06.26_{1.2}$	$74.21_{15.2}$	$29.74_{2.5}$
	LS [23]	$81.91_{0.4}$	$93.84_{0.6}$	$76.94_{0.9}$	$74.59_{0.8}$	$75.51_{0.7}$	$03.67_{0.9}$	$03.80_{0.8}$	$03.70_{0.2}$	$50.65_{2.8}$	$26.43_{1.1}$
	FL [15]	$81.89_{1.1}$	$94.07_{0.5}$	$76.58_{2.2}$	$\mathbf{75.68}_{1.1}$	$75.68_{1.6}$	$08.34_{6.2}$	$08.34_{6.1}$	$05.78_{3.3}$	$52.57_{4.2}$	$28.16_{2.3}$
	DCA [14]	$81.91_{0.7}$	$93.86_{0.3}$	$76.63_{0.6}$	$73.41_{1.3}$	$74.57_{1.1}$	$09.27_{1.4}$	$09.06_{1.6}$	$04.94_{0.8}$	$60.34_{3.8}$	$27.75_{1.3}$
	MDCA [7]	$81.52_{1.5}$	$93.13_{1.2}$	$76.55_{1.5}$	$74.91_{1.6}$	$75.45_{1.4}$	$10.72_{2.7}$	$10.58_{2.9}$	$05.99_{1.2}$	$81.92_{22.9}$	$29.42_{2.4}$
	Ours	$\mathbf{82.28}_{0.7}$	$94.02_{0.2}$	$\mathbf{77.30}_{1.3}$	$75.36_{1.4}$	$\mathbf{75.99}_{0.7}$	$\mathbf{02.81}_{0.7}$	$\mathbf{03.13}_{1.1}$	$\mathbf{03.49}_{0.4}$	$\mathbf{49.76}_{2.2}$	$\mathbf{25.66}_{0.9}$
R50	HL	$80.79_{0.5}$	$93.20_{0.5}$	$75.51_{0.4}$	$73.98_{0.3}$	$74.56_{0.2}$	$09.26_{4.5}$	$09.16_{4.6}$	$05.71_{1.6}$	$72.81_{22.7}$	$29.93_{2.3}$
	LS [23]	$80.62_{1.5}$	$93.04_{0.6}$	$75.41_{1.4}$	$73.92_{1.4}$	$74.48_{1.4}$	$03.51_{0.4}$	$04.27_{0.6}$	$\mathbf{03.66}_{0.3}$	$53.77_{2.2}$	$27.91_{1.4}$
	FL [15]	$80.52_{3.0}$	$93.47_{1.0}$	$76.30_{2.0}$	$72.60_{2.6}$	$73.73_{2.6}$	$04.16_{0.6}$	$04.26_{0.9}$	$04.06_{1.6}$	$53.98_{8.1}$	$27.70_{3.5}$
	DCA [14]	$79.82_{0.4}$	$92.75_{0.3}$	$74.82_{1.0}$	$72.32_{1.7}$	$73.18_{1.1}$	$13.89_{0.4}$	$13.88_{0.4}$	$07.43_{0.2}$	$92.42_{3.1}$	$33.30_{0.4}$
	MDCA [7]	$79.88_{2.0}$	$92.44_{0.6}$	$75.22_{2.5}$	$71.03_{3.2}$	$72.26_{3.0}$	$11.85_{4.3}$	$11.70_{4.5}$	$06.76_{2.0}$	$86.82_{27.9}$	$32.88_{4.2}$
	Ours	$\mathbf{81.44}_{1.7}$	$\mathbf{93.56}_{0.6}$	$\mathbf{76.39}_{2.0}$	$\mathbf{74.76}_{1.2}$	$\mathbf{75.20}_{1.9}$	$\mathbf{03.33}_{0.5}$	$\mathbf{03.45}_{0.7}$	$04.26_{0.8}$	$\mathbf{53.06}_{5.2}$	$\mathbf{27.17}_{2.7}$
D2	Method	ACC ↑	AUROC ↑	Precision ↑	Recall ↑	F1 ↑	ECE ↓	ACE ↓	SCE ↓	CE ↓	Brier ↓
R34	HL	$77.14_{3.6}$	$84.32_{2.7}$	$78.27_{5.8}$	$73.63_{1.8}$	$74.14_{2.4}$	$17.40_{3.6}$	$17.24_{3.7}$	$17.98_{3.7}$	$101.40_{24.6}$	$38.97_{6.8}$
	LS [23]	$78.68_{1.5}$	$87.16_{1.0}$	$78.45_{3.4}$	$76.31_{2.2}$	$76.51_{1.4}$	$06.50_{1.6}$	$06.78_{1.4}$	$08.13_{2.2}$	$45.92_{2.4}$	$29.71_{1.2}$
	FL [15]	$80.32_{1.0}$	$87.10_{1.4}$	$80.13_{1.1}$	$76.63_{1.2}$	$77.70_{1.2}$	$12.01_{2.0}$	$12.11_{2.3}$	$11.76_{2.5}$	$48.25_{0.6}$	$31.55_{1.2}$
	DCA [14]	$77.83_{1.1}$	$85.83_{1.0}$	$77.10_{0.8}$	$73.94_{1.7}$	$74.86_{1.6}$	$08.51_{1.5}$	$08.46_{1.9}$	$09.01_{2.1}$	$51.02_{5.2}$	$31.45_{1.8}$
	MDCA [7]	$80.25_{1.7}$	$87.45_{1.4}$	$79.38_{1.9}$	$77.44_{2.2}$	$78.12_{2.0}$	$12.10_{2.4}$	$11.91_{2.2}$	$12.28_{2.5}$	$63.28_{11.6}$	$31.36_{0.8}$
	Ours	$\mathbf{81.48}_{0.9}$	$\mathbf{87.69}_{0.7}$	$\mathbf{80.77}_{1.3}$	$\mathbf{78.70}_{0.6}$	$\mathbf{79.47}_{0.8}$	$\mathbf{05.68}_{0.8}$	$\mathbf{06.42}_{0.8}$	$\mathbf{06.75}_{1.0}$	$\mathbf{44.78}_{0.8}$	$\mathbf{28.43}_{0.7}$
R50	HL	$77.21_{4.5}$	$83.63_{4.5}$	$75.71_{4.7}$	$74.22_{6.1}$	$74.65_{5.8}$	$12.17_{3.0}$	$11.89_{3.2}$	$12.33_{2.6}$	$61.24_{11.1}$	$34.38_{5.8}$
	LS [23]	$80.73_{1.4}$	$86.84_{1.5}$	$80.43_{2.6}$	$77.95_{0.5}$	$78.62_{0.7}$	$04.57_{1.9}$	$05.33_{1.6}$	$06.17_{1.4}$	$45.38_{3.5}$	$28.60_{2.2}$
	FL [15]	$77.25_{1.2}$	$84.01_{1.0}$	$77.60_{2.3}$	$72.81_{2.8}$	$73.65_{2.7}$	$10.61_{3.6}$	$10.74_{3.5}$	$11.93_{2.5}$	$51.62_{1.5}$	$34.07_{1.0}$
	DCA [14]	$79.40_{3.0}$	$85.50_{3.1}$	$78.97_{3.2}$	$75.61_{3.7}$	$76.61_{3.7}$	$07.99_{1.1}$	$08.08_{1.1}$	$09.14_{1.0}$	$60.43_{11.5}$	$30.94_{4.1}$
	MDCA [7]	$77.62_{2.4}$	$84.22_{2.8}$	$76.22_{2.7}$	$74.97_{2.3}$	$75.44_{2.4}$	$09.91_{3.0}$	$09.82_{3.3}$	$10.05_{3.0}$	$60.54_{10.5}$	$33.08_{4.0}$
	Ours	$\mathbf{81.45}_{0.9}$	$\mathbf{88.29}_{0.4}$	$\mathbf{80.57}_{1.2}$	$\mathbf{79.04}_{1.2}$	$\mathbf{79.60}_{1.0}$	$\mathbf{04.03}_{0.6}$	$\mathbf{04.27}_{0.7}$	$\mathbf{05.80}_{1.2}$	$\mathbf{42.40}_{0.9}$	$\mathbf{26.87}_{0.5}$
D3	Method	ACC ↑	AUROC ↑	Precision ↑	Recall ↑	F1 ↑	ECE ↓	ACE ↓	SCE ↓	CE ↓	Brier ↓
R34	HL	$74.25_{0.3}$	$\mathbf{92.37}_{0.8}$	$63.76_{2.9}$	$52.91_{0.9}$	$55.85_{1.1}$	$15.38_{4.7}$	$15.30_{4.7}$	$04.78_{1.2}$	$112.51_{30.2}$	$40.77_{3.1}$
	LS [23]	$73.19_{1.1}$	$88.61_{3.6}$	$63.69_{3.6}$	$49.94_{2.9}$	$52.69_{4.3}$	$08.84_{3.0}$	$09.52_{2.9}$	$03.37_{0.5}$	$85.98_{8.4}$	$39.45_{1.8}$
	FL [15]	$74.01_{1.8}$	$90.69_{1.4}$	$64.95_{4.4}$	$52.02_{3.9}$	$55.90_{3.9}$	$04.19_{2.1}$	$04.44_{2.4}$	$03.01_{0.7}$	$77.03_{5.4}$	$37.06_{2.3}$
	DCA [14]	$\mathbf{74.37}_{1.5}$	$91.13_{1.3}$	$\mathbf{65.64}_{3.8}$	$53.27_{5.0}$	$\mathbf{57.38}_{3.8}$	$12.42_{3.3}$	$12.25_{3.4}$	$04.08_{0.8}$	$90.67_{13.8}$	$38.70_{4.0}$
	MDCA [7]	$72.62_{1.5}$	$90.15_{2.9}$	$61.78_{4.4}$	$\mathbf{53.41}_{5.2}$	$55.74_{3.2}$	$13.44_{5.5}$	$13.45_{5.4}$	$04.56_{1.2}$	$100.40_{20.4}$	$41.70_{4.7}$
	Ours	$74.03_{0.8}$	$90.02_{0.4}$	$61.28_{5.9}$	$48.36_{2.4}$	$50.23_{1.6}$	$\mathbf{03.72}_{0.5}$	$\mathbf{03.36}_{0.6}$	$\mathbf{02.04}_{0.2}$	$\mathbf{76.85}_{0.7}$	$\mathbf{36.66}_{0.2}$
R50	HL	$72.84_{0.3}$	$\mathbf{89.14}_{0.8}$	$61.75_{1.6}$	$49.38_{1.7}$	$52.92_{2.1}$	$15.41_{3.1}$	$15.36_{3.0}$	$04.91_{0.9}$	$137.89_{18.8}$	$43.10_{2.2}$
	LS [23]	$73.28_{1.8}$	$88.24_{2.4}$	$60.08_{2.5}$	$49.56_{6.0}$	$51.41_{5.3}$	$\mathbf{05.31}_{0.2}$	$\mathbf{06.45}_{0.9}$	$02.80_{0.5}$	$85.35_{1.1}$	$38.44_{1.1}$
	FL [15]	$71.96_{1.8}$	$88.73_{2.2}$	$59.11_{1.5}$	$\mathbf{51.07}_{3.6}$	$53.48_{2.9}$	$06.54_{4.1}$	$06.74_{3.8}$	$03.28_{0.7}$	$105.18_{19.7}$	$41.11_{3.5}$
	DCA [14]	$72.67_{0.8}$	$88.77_{3.0}$	$61.35_{3.7}$	$48.97_{2.8}$	$52.12_{2.1}$	$14.02_{6.7}$	$13.81_{6.5}$	$04.38_{1.8}$	$112.82_{36.0}$	$42.49_{5.1}$
	MDCA [7]	$73.68_{1.0}$	$89.13_{2.1}$	$61.09_{3.6}$	$50.56_{2.8}$	$53.95_{2.9}$	$17.65_{6.2}$	$17.62_{6.2}$	$05.44_{1.5}$	$143.59_{43.3}$	$43.32_{4.8}$
	Ours	$\mathbf{73.77}_{0.9}$	$89.01_{1.5}$	$\mathbf{63.12}_{1.4}$	$51.04_{1.4}$	$\mathbf{55.03}_{1.5}$	$06.96_{0.6}$	$06.95_{1.0}$	$\mathbf{02.63}_{0.2}$	$\mathbf{83.61}_{3.6}$	$\mathbf{37.38}_{1.1}$

improvement can be observed in majority of the performance metrics across different architectures and datasets. Even for the highly imbalanced ISIC dataset, our model provides notable enhancement across all calibration metrics with minimal effect on performance, further solidifying the effectiveness of our approach.

Uncertainty-based Retention Curves — To assess the reliability of the models, we plot the accuracy of a model as a function of its retention rate. As the fraction of predictions retained is increased, ground truth labels are replaced

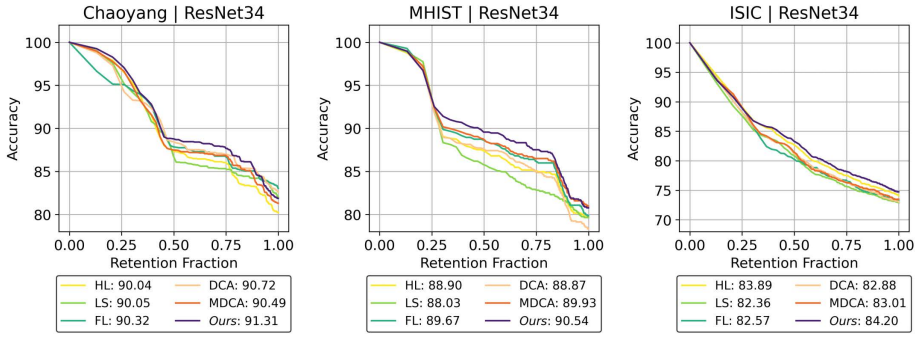

Fig. 1: **Retention Curves.** Accuracy as a function of retention fraction along with the area under the retention curve (R-AUC) values using ResNet-34 for all three datasets. HL - *Hard Labels*, LS - *Label Smoothing*, FL - *Focal Loss*, DCA - *Difference between Confidence and Accuracy* and MDCA - *Multi-class Difference in Confidence and Accuracy.*

with predicted labels in decreasing order of prediction scores, providing a comprehensive view of error distribution across the dataset. *For a zero retention fraction, we opt for the predicted label vector ($\boldsymbol{\Omega}$) to be the same as the ground truth ($\boldsymbol{G}$), resulting in 100% accuracy. As we increase the retention fraction, we replace the label vector $\boldsymbol{\Omega}$ with the fraction of the original predicted labels from samples having the highest predicted probability. We continue the substitution process until the entire label vector is replaced with the predicted labels.* The area under this accuracy-retention curve (R-AUC) serves as a metric for evaluating the quality of uncertainty estimates (predicted confidence scores) [1], with a higher value indicating models with better predictions. Figure 1 exhibits superior reliability of our proposed validation accuracy based label smoothing model, making it more suitable for medical image analysis. Additional plots for ResNet-50 are shown in the supplementary material.

Clinical Significance of Predicted Confidence Scores — To gain deeper insights into model calibration, we distinguish between the confidence scores assigned to correct and incorrect classified samples in Figure 2. Ideally, the confidence scores of the correctly predicted samples should be close to 1 (indicating high certainty), while incorrect classified samples should move away (reflecting uncertainty). The density plots associated with the majority of SOTA approaches (except *FL*) exhibit left-skewed distributions for correct predictions (green), signifying high confidence levels of these models. Undesirably, these models also express high confidence in their incorrect predictions (red). *FL* exhibits contrasting behaviour with right-skewed distributions for both correct and incorrect predictions indicating overall low confidence levels. Our proposed approach strikes the right balance by assigning relatively high scores for correctly classified samples while adeptly conveying uncertainty associated with incorrectly classified samples with low scores. This nuanced approach positions our model as a reliable

and trustworthy solution that increases the likelihood of expert medical intervention when the model lacks confidence. Additional plots for different datasets and architectures are shown in the supplementary material.

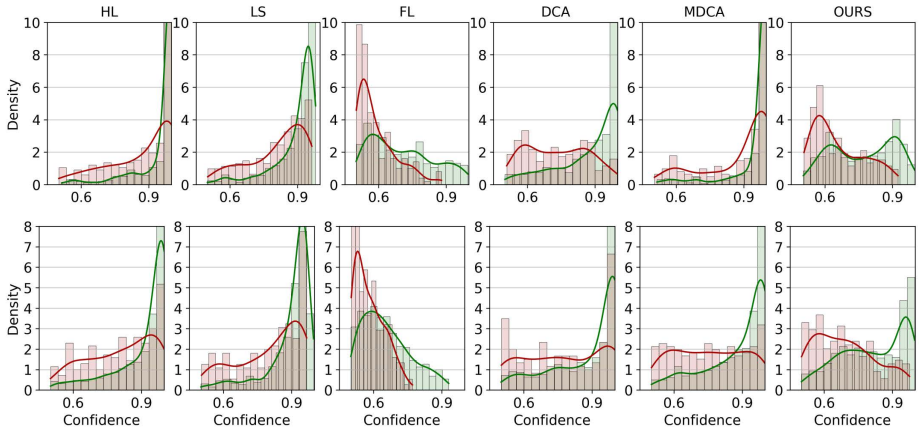

Fig. 2: Comparison of density plots for correct (green) and incorrect (red) classification confidences for ResNet-34 (top) and ResNet-50 (bottom) using MHIST dataset. The area under the histogram integrates to 1. We have clipped the y-axis in all the plots to better visualize the trends.

5 Conclusion

We propose an informed label smoothing strategy (LS+) that addresses the shortcomings of the traditional version by taking into consideration the model's current calibration status as well as class-wise calibration levels. This is achieved by replacing the uniform prior with an informed class-specific prior estimated from the class accuracy on a separate validation set. Experimental results from three benchmark medical image classification tasks show that LS+ provides significant improvement in calibration. Consistent improvement across multiple performance and calibration metrics using two different architectures as well as higher R-AUC values along with density plots exhibit reliability and clinical readiness of LS+. Our present study assumes that both the validation and test sets stem from the same distribution. In medical imaging, heterogeneity of population, scanners and acquisition protocols presents a shift in distribution. Hence, our future efforts will be directed towards adapting LS+ to excel in out-of-distribution (OOD) scenarios.

Acknowledgments. The support and the resources provided by PARAM Sanganak under the National Supercomputing Mission, Government of India at the Indian Institute of Technology Kanpur are gratefully acknowledged.

Disclosure of Interests. The authors have no competing interests in the paper as required by the publisher.

References

1. Andrey, M., Neil, B., Yarin, G., Mark, G., Alexander, G., German, C., et al.: Shifts: A dataset of real distributional shift across multiple large-scale tasks. In: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2021) vii
2. Azizi, S., et al.: Big self-supervised models advance medical image classification. In: IEEE International Conference on Computer Vision (ICCV) (2021) v
3. Bohdal, O., Yang, Y., Hospedales, T.: Meta-calibration: Learning of model calibration using differentiable expected calibration error. Transactions on Machine Learning Research (TMLR) (2023) ii
4. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic) (2019) v
5. Frenkel, L., Goldberger, J.: Calibration of medical imaging classification systems with weight scaling. In: Medical Image Computing and Computer Assisted Intervention (MICCAI) (2022) ii
6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning (ICML) (2017) ii
7. Hebbalaguppe, R., Prakash, J., Madan, N., Arora, C.: A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) iii, v, vi
8. Islam, M., Seenivasana, L., Ren, H., Glocker, B.: Class-distribution-aware calibration for long-tailed visual recognition. In: International Conference on Machine Learning (ICML), Uncertainty and Robustness in Deep Learning Workshop (2021) ii
9. Jungo, A., Reyes, M.: Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI) (2019) ii
10. Kompa, B., Snoek, J., Beam, A.L.: Second opinion needed: communicating uncertainty in medical machine learning. npj Digital Medicine (2021) i
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2012) i
12. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems (NeurIPS) (2017) i
13. Larrazabal, A.J., Martínez, C., Dolz, J., Ferrante, E.: Maximum entropy on erroneous predictions: Improving model calibration for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI) (2023) i
14. Liang, G., Zhang, Y., Wang, X., Jacobs, N.: Improved trainable calibration method for neural networks. In: British Machine Vision Conference (BMVC) (2020) ii, v, vi

15. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) i, ii, v, vi
16. Müller, R., Kornblith, S., Hinton, G.: When does label smoothing help? In: Advances in Neural Information Processing Systems (NeurIPS) (2019) i, ii
17. Murugesan, B., Liu, B., Galdran, A., Ayed, I.B., Dolz, J.: Calibrating segmentation networks with margin-based label smoothing. Medical Image Analysis (2023) i
18. Niyaz, U., Sambyal, A.S., Bathula, D.R.: Leveraging different learning styles for improved knowledge distillation in biomedical imaging. Computers in Biology and Medicine (2024) v
19. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., Hinton, G.E.: Regularizing neural networks by penalizing confident output distributions. In: International Conference on Learning Representations (ICLR), Workshop Track Proceedings (2017) ii
20. Platt, J.: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Advances in Large Margin Classifiers (2000) ii
21. Qin, Y., Wang, X., Lakshminarayanan, B., Chi, E.H., Beutel, A.: What are effective labels for augmented data? improving calibration and robustness with autolabel. In: IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) (2023) i
22. Sambyal, A.S., Niyaz, U., K., N.C., Bathula, D.R.: Understanding calibration of deep neural networks for medical image classification. Computer Methods and Programs in Biomedicine (2023) ii, v
23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016) i, ii, iii, v, vi
24. Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2019) i
25. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1), 180161 (Aug 2018) v
26. Wei, J., Suriawinata, A., Ren, B., Liu, X., Lisovsky, M., Vaickus, L., et al.: A petri dish for histopathology image analysis. In: International Conference on Artificial Intelligence in Medicine (AIME) (2021) v
27. Wen, Y., Chen, L., Deng, Y., Zhou, C.: Rethinking pre-training on medical imaging. Journal of Visual Communication and Image Representation (2021) v
28. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery (2002) ii
29. Zhang, C., Jiang, P., Hou, Q., Wei, Y., Han, Q., Li, Z., Cheng, M.: Delving deep into label smoothing. IEEE Transactions on Image Processing (2021) i
30. Zhu, C., Chen, W., Peng, T., Wang, Y., Jin, M.: Hard sample aware noise robust learning for histopathology image classification. IEEE Transactions on Medical Imaging (TMI) (2022) iv