Dual-Penalty Conformal-Aware Loss for Reliable Skin Lesion Classification Under Class Imbalance

Anum Fatima[1[0009−0008−8455−4970]], Moi Hoon Yap[1[0000−0001−7681−4287]], Neil D. Reeves[2[0000−0001−9213−4580]], Michael Odling-Smee[3], David C. Bond[3], and Connah Kendrick[1[0000−0002−3633−6598]]

[1] School of Computing and Mathematics, Manchester Metropolitan University, UK
[2] Medical School, Faculty of Health and Medicine, Lancaster University, UK
[3] Aire Innovate Limited, UK

Abstract. Reliable uncertainty estimation is key for clinical adoption of deep learning, particularly in settings with severe class imbalance. Conformal prediction is a post-training technique that produces a set of plausible labels when a model is uncertain, enabling safer decision support. However, it treats the trained model as fixed, and models trained with standard loss functions, such as cross-entropy and focal loss, are more confident on common than rare classes. As a result, post-hoc conformal prediction yields a high coverage gap, leaving minority classes undercovered across three variants (standard, class-conditional, and clusterwise) and a range of calibration sizes. To address this, we propose a conformal-aware loss guided by two class-specific, threshold-guided penalties that directly shape per-class confidence during training: a coverage penalty to maintain sufficient true-class confidence for each class, and an efficiency penalty to discourage overly large prediction sets. On the ISIC 2019 challenge dataset, our proposed method reduces the coverage gap across nearly all settings, with the largest gains on rare classes, improving empirical minority class coverage towards 90% target while keeping prediction sets clinically compact. The same observation holds on Diabetic Foot Ulcer Challenge 2021 dataset and Diabetic Retinopathy Detection dataset, suggesting that training time confidence shaping can improve reliability of conformal prediction. Code is available at `https://github.com/anumfatima427/cp-under-imbalance`

Keywords: Conformal prediction · Uncertainty quantification · Coverage gap.

1 Introduction

Classification using deep learning has shown promising performance across different domains, including skin lesion analysis [9], diabetic retinopathy (DR) grading from retinal fundus images [14], and diabetic foot ulcer assessment [24]. However, these tasks share a common challenge: the most clinically critical classes are severely underrepresented, making reliable uncertainty estimation essential where predictions directly impact decisions.

Conformal prediction (CP) [22] is a distribution-free framework that wraps any pre-trained classifier to produce prediction sets with finite-sample coverage guarantees, providing multiple possible classifications rather than a single output. A classifier can flag genuine ambiguity, e.g., *melanoma*, *nevus* rather than committing to one label. Given a user-specified error rate, split CP guarantees that the prediction set contains the true label for at least (e.g, 90%) of future test samples, assuming that calibration and test data are exchangeable [2].

Since CP treats the model as fixed, it cannot change the way the model distributes its error across classes. This motivates training the model itself to produce similar non-conformity score distributions, where class confidences can be similar across all classes, rather than attempting to correct them afterwards. Several works have explored conformal-aware (CA) training strategies. Stutz et al. [19] proposed conformal training, which simulates CP on mini-batches during training, but applies a global penalty that does not differentiate between classes. Einbinder et al. [7] train with conformal-inspired hold-out data to improve conditional coverage, but apply it globally on balanced benchmarks. Gibbs et al. [10] introduced adaptive conformal inference, which adapts the miscoverage rate over time for sequential data; however, it targets longitudinal settings that adapt based on a single value rather than per-class calibration. Fisch et al. [8] proposed CP Sets with limited false positives, and Wang et al. [23] extended this with conformal loss controlling prediction, a training time technique that reduces false positives in prediction sets. More recently, class-adaptive conformal training [15] learns per-class penalty weights using Augmented Lagrangian optimization, but requires a complex constrained training procedure with additional hyperparameters.

Across these methods, the penalty is applied either globally or through constrained optimization. We observe that the root cause of the coverage failure is not the CP algorithm itself but the underlying model, in which training on uneven classes leads to lower confidence in those classes. Classifiers trained with standard losses produce vastly different nonconformity score distributions across classes, making it difficult for any single threshold to serve all classes well. Our contributions are:

- We propose a CA loss that adds two class-specific, threshold-guided penalties to cross-entropy (CE) to encourage sufficient true-class confidence while discouraging overly large prediction sets.
- We provide an empirical evaluation of standard, class-conditional, and clusterwise CP on three imbalanced medical imaging datasets, showing that standard CP consistently undercovers minority classes.

2 Methodology

2.1 Dataset and Preprocessing

ISIC2019 challenge dataset [13,20,4,5] contains 33,569 dermoscopic images across eight diagnostic categories, including rare malignant and pre-malignant classes,

such as squamous cell carcinoma (SCC), and actinic keratosis (AK) and benign but infrequent vascular lesions (Vasc) and dermatofibroma (DF) classes, the standard protocol split the dataset into 25,331 train images and 8,238 test images. To assess generalizability beyond dermatology, we use two additional imbalanced medical datasets:

- Diabetic Foot Ulcer Challenge 2021 (DFUC2021) [24,3]: 11,689 diabetic foot ulcer images across four wound pathology classes, including None/Control, Infection, Ischaemia, and Both, split into 5,955 train and 5,734 test images.
- Diabetic Retinopathy Detection (DDR) [14]: 12,522 retinal fundus images with five annotated DR severity grades including No DR, Mild, Moderate, Severe, and Proliferative, split into 10,017 train and 2,505 test images.

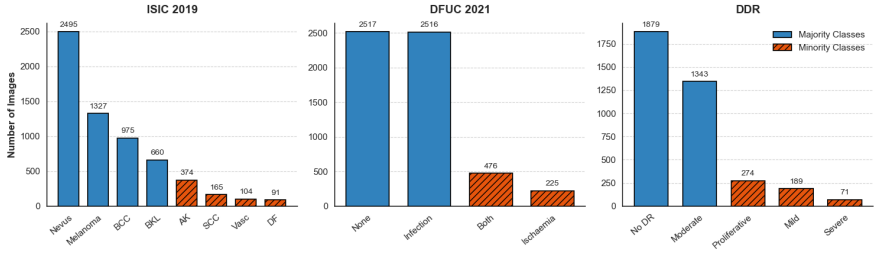

Fig. 1. Class distribution across ISIC2019, DFUC2021, and DDR test sets. Critical minority classes (orange) are severely underrepresented compared to majority classes (blue).

Figure 1 illustrates class imbalance across three datasets, i.e., less than 4% for DF, Vasc, and SCC in ISIC2019, Severe DR comprises only 1.9% of DDR, and Ischaemia comprises 3.9% of DFUC2021. We split the dataset into train and validation using stratified random sampling. To prevent models from collapsing to majority class predictions, we employed weighted random sampling during training. Sampling weights were assigned inversely proportional to empirical class frequencies. All images were standardized to 224×224 pixels and augmented using random horizontal and vertical flips, alongside color jittering.

2.2 Model Architecture and Training Setup

We use ResNet-50 [12] architecture with ImageNet pre-trained weights. The final fully connected layer is replaced with a K-way classifier ($K = 8$ for ISIC2019, $K = 5$ for DDR (excluding ungradable images) and $K = 4$ for DFUC2021). Modern networks are often overconfident on rare cases [11], especially under domain shift [17,16]. We test whether our penalties regularize these biased priors under imbalance.

We compare our proposed CA loss against standard CE and Focal loss. All three configurations share an identical training pipeline. Models are optimized using Adam (initial learning rate of 10^{-4}), cosine annealing schedule, batch size of 32 for 50 epochs with a patience of 10.

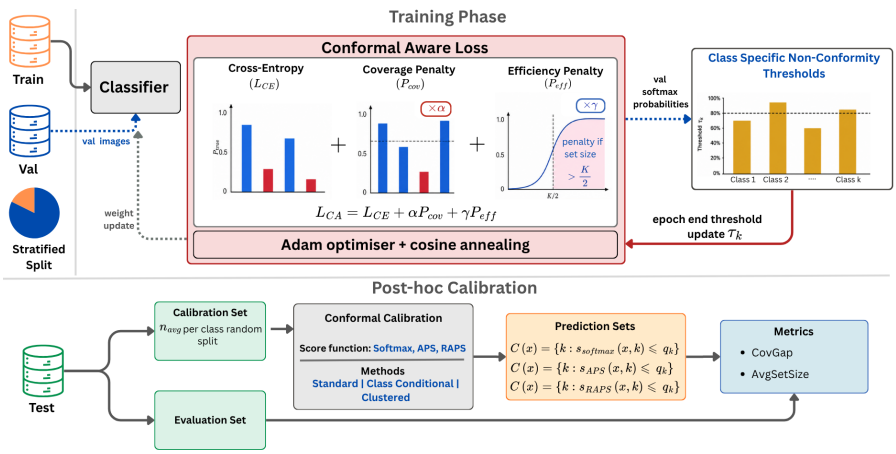

Fig. 2. Overview of the proposed conformal-aware (CA) framework. **(Top)** During training, CA loss (L_{CA}) uses class-specific nonconformity thresholds (τ_k)computed from the validation set at the end of each epoch. The training objective combines standard cross-entropy loss (L_{CE}) with a coverage penalty (P_{cov}) and an efficiency penalty (P_{eff}). **(Bottom)** After training, standard split conformal calibration is applied on a held-out subset for evaluation.

2.3 Conformal-Aware (CA) Training

Standard loss functions such as CE are blind to the set-based nature of conformal prediction (CP). Consequently, applying post-hoc CP [1] to a trained model often yields overly large prediction sets and poor minority-class coverage. Our conformal-aware (CA) loss (Fig. 2) addresses this during training through a dual penalty driven by class-specific non-conformity thresholds: (i) a coverage penalty (P_{cov}) that maintains sufficient true class confidence and (ii) an efficiency penalty (P_{eff}) that suppresses overly large prediction sets.

Threshold computation. At the end of each epoch, for every sample (x, y) in validation set $\mathcal{D}_{\text{val}}$, we compute the non-conformity score $1 - \hat{\pi}_y(x)$, where $\hat{\pi}_y(x)$ is the softmax probability assigned to true class. Scores are grouped by class label, and the threshold for class k is set as:

$$\tau_k = Q_{1-\epsilon}\left(1 - \hat{\pi}_y(x) : (x, y) \in \mathcal{D}_{\text{val}}, y = k\right), \tag{1}$$

where $Q_{1-\epsilon}$ denotes the $1-\epsilon$ quantile with $\epsilon = 0.1$ (90th percentile of class-wise non-conformity scores). The thresholds are initialized to 0.5, which means that the model gives 50% confidence to the true class and is updated only after all training batches in an epoch have been processed.

Coverage Penalty. To encourage sufficient confidence in the true class, we penalize samples whose true class probability falls below the class-specific floor $(1 - \tau_{y_i})$. The hinge term allows the system to penalise scores too uncertain for

the true class. For a mini-batch of size N

$$P_{cov} = \frac{1}{N} \sum_{i=1}^{N} \max(0, (1 - \hat{\pi}_{y_i}(x_i)) - \tau_{y_i}), \tag{2}$$

where $\hat{\pi}_{y_i}(x_i)$ is the softmax probability assigned to the true class y_i and τ_{y_i} is the corresponding class-specific threshold.

Efficiency Penalty. Although P_{cov} encourages confidence in the true class, it does not prevent the model from predicting high probabilities to several classes at once, which leads to large and uninformative prediction sets. To discourage this behavior, we introduce P_{eff} based on a differentiable approximation of the prediction set size. For a sample x_i, the set size is defined as:

$$|C(x_i)| = \sum_{k=1}^{K} \sigma \left(\frac{\hat{\pi}_k(x_i) - (1 - \tau_k)}{T} \right), \tag{3}$$

where $\hat{\pi}_k(x_i)$ is the predicted softmax probability for class k, τ_k is the class-specific threshold, σ is the sigmoid function, and T controls the sharpness of approximation. Each sigmoid term acts as a soft indicator of whether class k would be included under the threshold rule $\hat{\pi}_k(x_i) \geq 1 - \tau_k$, and summing over classes yields a smooth estimate of set size. We then penalize only overfull sets:

$$P_{eff} = \frac{1}{N} \sum_{i=1}^{N} \max \left(0, |C(x_i)| - \frac{K}{2} \right), \tag{4}$$

where N is the size of the mini-batch and K is the total number of classes. This hinge term is 0 when the estimated set size is at $K/2$. Thus, it increases with the number of available classes, thereby discouraging overly broad prediction sets. CA loss (L_{CA}) sums the standard CE loss (L_{CE}) with P_{cov} and P_{eff}:

$$L_{CA} = L_{CE} + \alpha P_{cov} + \gamma P_{eff}, \tag{5}$$

where hyperparameters α and γ controls the strength of P_{cov} and P_{eff} respectively. We set $\alpha = 1.5$ and $\gamma = 0.15$ and approximate the set-size indicator (Eq. 3) with a sigmoid sharpness $1/T$, $T = 0.02$.

2.4 Conformal Prediction Setup

Probability Computation. After training, we apply the model to the test set to obtain softmax probability predictions $\hat{\pi}_y(x)$ for each class y. These softmax predictions are pre-computed once and used for all subsequent CP experiments.
Calibration. For each experiment, we randomly sample a calibration subset of size $n_{avg} \times K$ from predefined test split, where $n_{avg} \in \{50, 100, 150\}$ controls the overall calibration budget relative to the number of classes K. The remaining test samples are used for evaluation. This sampling is not class-balanced, since the subset is drawn from naturally imbalanced test distribution. The number

of calibration examples per class varies across classes and random seeds, with minority classes often contributing fewer samples. We use $\epsilon = 0.1$ to target 90% marginal coverage.

Non-Conformity Score Calculation. For each calibration sample (x_i, y_i) where $i = 1, \ldots, n_{\text{avg}} \times K$, we compute a conformal score from pre-computed softmax predictions. We evaluate three score functions: softmax, Adaptive Prediction Sets (APS), and Regularized Adaptive Prediction Sets (RAPS). For softmax baseline, the non-conformity score for class k is defined as $s_{softmax}(x_i, k) = 1 - \hat{\pi}_k(x_i)$ where $\hat{\pi}_k(x_i)$ is the predicted softmax probability for class k. APS [18] uses cumulative predicted softmax probabilities up to the rank of class k, typically yielding smaller prediction sets. RAPS extends APS with an additional rank-based regularization term to further discourage unnecessarily large prediction sets.

Threshold Calibration. Using the conformal scores computed on calibration set, we determine a threshold to define prediction sets $\mathcal{C}(x_j)$ for each evaluation sample x_j with true label y_j. The calibration procedure differs across:

- *Standard (STD)*[2] computes a single global threshold. However, this is averaged over all classes, potentially masking class-wise disparities.
- *Class-conditional (CC)* [21] uses a separate threshold per class, ensuring per-class coverage, but for rare classes with few calibration samples, the threshold becomes trivially large, causing those classes to appear in every prediction set.
- *Clusterwise (CW)* [6] groups classes into clusters with a shared per-cluster threshold, pooling related classes for more stable rare-class thresholds.

Prediction Set Formation. For a new test image X_{test} with unknown label, we construct the prediction set by including all classes whose conformal score falls below the calibrated threshold.

2.5 Evaluation Metrics

We evaluate all methods using two metrics [6], where n_{test} is the number of evaluation samples and y_i is the true label for sample x_i. **Class-conditional coverage** computes this separately per class. **CovGap** averages the absolute deviation of the class coverage from the target level $(1 - \epsilon)$ as:

$$CovGap = 100 \times \frac{1}{|K|} \sum_{k=1}^{K} \left| \text{Coverage}_k - (1 - \epsilon) \right|, \tag{6}$$

where Coverage_k is the marginal coverage for class k, K is the total number of classes, and 0% indicates perfect class-conditional coverage and higher values indicate disparity. **Average set size** measures the sharpness of the prediction set:

$$AvgSize = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |\mathcal{C}(x_i)|, \tag{7}$$

Table 1. Comparison of CP performance using RAPS across three medical datasets. We evaluate STD, CC, and CW calibration across varying calibration sizes (n_{avg}). Best overall results for each dataset and ties are in **bold**, and best results within each calibration method are *italicized*.

			$n_{avg} = 50$		$n_{avg} = 100$		$n_{avg} = 150$	
Dataset	Method	Loss	CovGap↓	AvgSize↓	CovGap↓	AvgSize↓	CovGap↓	AvgSize↓
ISIC 2019	STD	CE	20.3 (0.6)	2.1 (0.1)	19.1 (0.8)	2.2 (0.1)	19.2 (0.5)	2.2 (0.0)
		Focal	17.2 (0.7)	**1.9 (0.1)**	16.6 (0.7)	**2.0 (0.1)**	16.9 (0.6)	**2.0 (0.0)**
		CA (Ours)	*15.3 (0.7)*	2.1 (0.1)	*15.5 (0.6)*	**2.0 (0.1)**	*15.7 (0.2)*	**2.0 (0.0)**
	CC	CE	5.8 (0.4)	4.8 (0.3)	5.7 (1.1)	4.1 (0.3)	3.9 (0.4)	4.1 (0.2)
		Focal	5.8 (0.4)	4.7 (0.3)	4.2 (0.3)	*3.7 (0.3)*	3.6 (0.5)	*3.5 (0.2)*
		CA (Ours)	**5.0 (0.5)**	*4.7 (0.1)*	**4.2 (0.2)**	3.9 (0.3)	**3.2 (0.2)**	3.7 (0.1)
	CW	CE	15.0 (0.4)	2.6 (0.1)	9.3 (0.7)	2.8 (0.1)	8.4 (1.2)	3.1 (0.3)
		Focal	*13.8 (0.6)*	2.4 (0.2)	8.4 (0.7)	*2.7 (0.2)*	6.1 (1.4)	*2.9 (0.2)*
		CA (Ours)	*13.0 (0.4)*	*2.4 (0.1)*	*7.8 (1.1)*	3.2 (0.2)	*5.2 (0.9)*	3.5 (0.2)
DFUC 2021	STD	CE	8.2 (0.7)	1.9 (0.1)	8.5 (0.4)	**1.9 (0.0)**	8.7 (0.4)	**1.9 (0.0)**
		Focal	9.6 (0.5)	2.0 (0.0)	10.2 (0.6)	2.0 (0.0)	10.6 (0.6)	2.0 (0.0)
		CA (Ours)	*5.2 (0.4)*	**1.9 (0.0)**	*5.3 (0.2)*	**1.9 (0.0)**	*5.5 (0.2)*	**1.9 (0.0)**
	CC	CE	4.8 (0.3)	3.1 (0.1)	3.3 (0.4)	2.4 (0.1)	2.9 (0.5)	2.2 (0.1)
		Focal	5.1 (0.4)	3.0 (0.1)	3.8 (0.3)	2.3 (0.0)	3.2 (0.4)	2.2 (0.1)
		CA (Ours)	**4.5 (0.6)**	*2.9 (0.1)*	**2.7 (0.3)**	*2.2 (0.1)*	**2.1 (0.2)**	**2.1 (0.0)**
	CW	CE	8.0 (0.9)	1.9 (0.1)	6.9 (0.7)	**1.9 (0.0)**	7.0 (0.9)	1.9 (0.1)
		Focal	7.5 (0.9)	2.1 (0.0)	8.3 (0.9)	**1.9 (0.0)**	8.2 (1.3)	2.0 (0.0)
		CA (Ours)	*5.1 (0.8)*	**1.9 (0.0)**	*3.8 (0.5)*	2.0 (0.0)	*4.9 (0.7)*	**1.9 (0.0)**
DDR	STD	CE	17.9 (0.9)	**1.4 (0.0)**	17.6 (0.4)	**1.4 (0.0)**	17.1 (0.9)	**1.4 (0.0)**
		Focal	12.9 (1.2)	1.6 (0.1)	13.6 (1.4)	1.5 (0.1)	13.7 (1.5)	1.5 (0.0)
		CA (Ours)	*8.7 (0.5)*	**1.4 (0.0)**	*8.5 (0.5)*	**1.4 (0.0)**	*8.8 (0.5)*	**1.4 (0.0)**
	CC	CE	5.2 (0.3)	4.1 (0.3)	*3.9 (0.7)*	2.8 (0.2)	3.6 (0.6)	2.6 (0.1)
		Focal	5.5 (0.4)	3.0 (0.1)	4.8 (0.4)	2.9 (0.1)	4.2 (0.6)	2.4 (0.1)
		CA (Ours)	**4.7 (0.4)**	*2.9 (0.1)*	**3.9 (0.1)**	*2.5 (0.1)*	**3.1 (0.4)**	*2.2 (0.1)*
	CW	CE	16.4 (2.0)	*1.6 (0.1)*	14.0 (1.6)	1.7 (0.1)	11.9 (2.1)	*1.7 (0.1)*
		Focal	11.0 (1.8)	1.8 (0.2)	10.6 (1.4)	1.9 (0.1)	9.1 (0.9)	1.8 (0.0)
		CA (Ours)	*9.2 (1.7)*	*1.6 (0.1)*	*7.6 (0.5)*	*1.6 (0.1)*	*6.5 (0.7)*	1.7 (0.2)

where smaller sets indicate more precise predictions, while larger sets reflect greater uncertainty.

3 Results and Discussion

CA loss preserves base classification performance, achieving Macro-F1 of 0.85 on ISIC2019, 0.57 on DFUC2021, and 0.62 on DDR. Table 1 shows CA loss improves coverage-efficiency balance on imbalanced datasets. On ISIC2019, STD calibration achieves 90% marginal coverage but conceals severe minority class undercoverage, CE baseline yields a *CovGap* of 19.2% even at the largest calibration size ($n_{avg} = 150$). CC calibration removes most of this disparity (*CovGap* of 3.9%) but at the cost of an inflated prediction set (4.1). Our CA loss improves *CovGap* under all three calibration strategies on ISIC2019, reducing STD from 19.2 to 15.7, CC from 3.9 to 3.2, and CW from 8.4 to 5.2 at $n_{avg} = 150$. The same trend holds on DFUC2021 and DDR.

Figure 3 (top) shows that our proposed CA loss achieves the lowest *CovGap* across all calibration sizes. The clearest clinical benefit of CA is its protection of rare, high risk classes (Figure 3 (bottom)). On ISIC2019, the CE baseline falls below the 90% target for several minority classes, dropping to 66% for 'Vasc'. CA recovers coverage for (AK, SCC, DF, and Vasc). The same recovery holds for (Ischaemia, Both) in DFUC2021, and (Severe DR, Mild) in DDR dataset. By keeping rare conditions covered, CA reduces the class-conditional coverage gap for skin lesion analysis, with the same trend on other modalities.

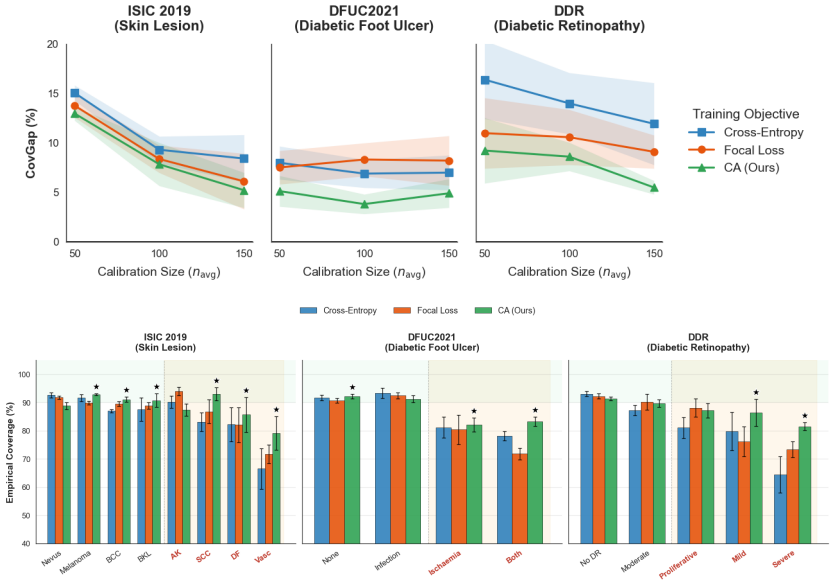

Fig. 3. (Top) Impact of calibration size (n_{avg}) on coverage gap (CovGap). Shaded bands denote standard error. **(Bottom)** Per class empirical coverage. CE and Focal loss undercover critical minority classes (right, red labels). CA loss (green) recovers coverage toward the 90% target. ⋆ marks classes where CA improves coverage over baseline loss functions. Error bars denote standard error.

4 Conclusion

We studied conformal prediction under severe class imbalance in classification tasks and showed that marginal coverage can conceal undercoverage of rare classes. Our CA loss improves class-wise reliability and discourages overly broad prediction sets during training, improving the empirical coverage of critical minority pathologies on ISIC2019 and on other imbalanced datasets (DFUC2021 and DDR). Future work will extend this with multi-label classification and segmentation tasks.

Disclosure of Interests. The authors have no competing interests relevant to the content of this article.

References

1. Angelopoulos, A., Bates now, S., Malik, J., Jordan, M.I.: Uncertainty sets for image classifiers using conformal prediction. arXiv preprint arXiv:2009.14193 (2020)
2. Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning **16**(4), 494–591 (03 2023). https://doi.org/10.1561/2200000101, https://doi.org/10.1561/2200000101
3. Cassidy, B., Reeves, N.D., Pappachan, J.M., Gillespie, D., O'Shea, C., Rajbhandari, S., Maiya, A.G., Frank, E., Boulton, A.J., Armstrong, D.G., et al.: The dfuc 2020 dataset: analysis towards diabetic foot ulcer detection. touchREVIEWS in Endocrinology **17**(1), 5 (2021)
4. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: 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: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
5. Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)
6. Ding, T., Angelopoulos, A., Bates, S., Jordan, M., Tibshirani, R.J.: Class-conditional conformal prediction with many classes. Advances in neural information processing systems **36**, 64555–64576 (2023)
7. Einbinder, B.S., Romano, Y., Sesia, M., Zhou, Y.: Training uncertainty-aware classifiers with conformalized deep learning. In: Advances in Neural Information Processing Systems. vol. 35, pp. 22380–22395 (2022)
8. Fisch, A., Schuster, T., Jaakkola, T., Barzilay, R.: Conformal prediction sets with limited false positives. In: International Conference on Machine Learning. pp. 6514–6532. PMLR (2022)
9. Gessert, N., Nielsen, M., Shaikh, M., Werner, R., Schlaefer, A.: Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. MethodsX **7**, 100864 (2020)
10. Gibbs, I., Candes, E.: Adaptive conformal inference under distribution shift. Advances in Neural Information Processing Systems **34**, 1660–1672 (2021)
11. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. p. 1321–1330. ICML'17, JMLR.org (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. International Skin Imaging Collaboration: Isic 2019 challenge. https://challenge.isic-archive.com/landing/2019/ (2019), accessed: 2026-06-18
14. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences **501**, 511–522 (2019)
15. Marani, B.E., Silva-Rodriguez, J., Ayed, I.B., Vakalopoulou, M., Christodoulidis, S., Dolz, J.: Class adaptive conformal training. arXiv preprint arXiv:2601.09522 (2026)
16. Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., Lucic, M.: Revisiting the calibration of modern neural networks. Advances in neural information processing systems **34**, 15682–15694 (2021)

17. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Curran Associates Inc., Red Hook, NY, USA (2019)
18. Romano, Y., Sesia, M., Candes, E.: Classification with valid and adaptive coverage. Advances in neural information processing systems **33**, 3581–3591 (2020)
19. Stutz, D., Cemgil, A.T., Doucet, A., et al.: Learning optimal conformal classifiers. arXiv preprint arXiv:2110.09192 (2021)
20. 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 (2018)
21. Vovk, V.: Conditional validity of inductive conformal predictors. In: Asian conference on machine learning. pp. 475–490. PMLR (2012)
22. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world. Springer (2005)
23. Wang, D., Wang, P., Ji, Z., Yang, X., Li, H.: Conformal loss-controlling prediction. IEEE Transactions on Neural Networks and Learning Systems **36**(2), 2973–2983 (2024)
24. Yap, M.H., Cassidy, B., Pappachan, J.M., O'Shea, C., Gillespie, D., Reeves, N.D.: Analysis towards classification of infection and ischaemia of diabetic foot ulcers. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). pp. 1–4. IEEE (2021)