

Balancing Accuracy and Sensitivity: An Ablation Study of Focal Loss for Multiclass Skin Cancer Classification

Pramod. V. Patil¹, S.S. Ittannavar², Jagadish S Jakati³, S. S. Malaj⁴, Sunita Karabasannavar⁵,
Bhavalina Alias Pratima Khot⁶

¹Research Scholar & Assistant Professor ECE Dept., Hirasugar Institute of Technology, Nidasoshi, Karnataka, India. Email ID: pramodpatil.ece@hsit.ac.in.

²Associate Professor, ECE Dept., Hirasugar Institute of Technology, Nidasoshi, Karnataka, India. Email ID: ssittannavar.ece@hsit.ac.in.

³Senior Assistant Professor, School of Computer Science Engineering & Applications, D Y Patil International University, Pune, India. Email ID: Jagadish.jakati@dypiu.ac.in

⁴Assistant Professor, ECE Dept., Hirasugar Institute of Technology, Nidasoshi, Karnataka, India. Email ID: ssmalaj.ece@hsit.ac.in

⁵Assistant Professor, CSE Dept., Hirasugar Institute of Technology, Nidasoshi, Karnataka, India. Email ID: sunitak.cse@hsit.ac.in.

⁶Assistant Professor, ECE Dept., Hirasugar Institute of Technology, Nidasoshi, Karnataka, India. Email ID: pratimakhot.ece@hsit.ac.in.

Abstract— Skin cancer is one of the fastest growing malignancies worldwide and early and accurate diagnosis is a key factor to increase patient's survival. Manual dermoscopic examination is time consuming, subjective and experience dependent, which makes automated computer-aided diagnosis (CAD) systems desirable. In this paper, a deep learning framework for multiclass skin lesion classification is proposed, which integrates dermoscopic image features and patient metadata (age, sex, and lesion localization) based on an efficient networks EfficientNet-B0 backbone and a lightweight metadata encoder. The severe class imbalance found in dermoscopic datasets is addressed by using a class-balanced focal loss, which is studied using a controlled ablation by comparing the results of a class-balanced focal loss with an aggressive class-balanced focal loss. The proposed framework is tested on the HAM10000 benchmark data set that is split into multiple lesions without data leakage and the results are evaluated with accuracy, macro-averaged precision, recall, F1-score and ROC-AUC. The soft weighting configuration had an accuracy of 76.43%, a macro-F1 of 0.660 and a macro-averaged AUC of 0.961, significantly outperforming the aggressive configuration (accuracy of 30.72%, macro-F1 of 0.321), while maintaining good recall of minority classes. Grad-CAM visualizations also consistently show that the model focuses on the lesion area for all of the 7 diagnostic classes, which helps with clinical interpretation. The results confirm that the intensity of loss-weighting plays an important role in the classification of imbalanced datasets using dermoscopy and that a well calibrated value can achieve high overall accuracy while maintaining balanced per-class performance.

Keywords— Skin Cancer Classification; Dermoscopic Imaging; Deep Learning; Convolutional Neural Network; EfficientNet; Class Imbalance; Focal Loss; Metadata Fusion; Explainable AI; Grad-CAM.

I. INTRODUCTION

Malignant skin tumours, particularly melanoma, are still increasing in incidence globally and early detection significantly enhances patient prognosis [1]. Despite being the first line examination for the clinician, the interpretation of dermatoscopy is extremely subjective and even experienced dermatologists differ in their opinion on the borderline cases. Combining transfer learning from networks like ResNet, DenseNet and EfficientNet has significantly improved automated detection of lesions [2] [6] but there are three challenges. First, public corpora like HAM10000 are highly imbalanced towards benign naevi, and conventional networks are biased towards the majority class, and ignore rare and clinically important classes [8], [9]. Secondly, most pipelines leverage only pixel data and ignore important patient metadata that is already used by clinicians for diagnosis [3]. Third, deep networks are opaque and therefore difficult to be accepted clinically because predictions are usually made without a visual justification [5] [10] [13]. In this paper, we aim to fill these gaps by (i) incorporating image and metadata branches into one single network, (ii) comparing two intensities of focal loss, and quantify their impact on the precision-recall trade-off, and (iii) using Grad-CAM explanation to evaluate the network. To fill this gap, this paper suggests a deep learning-based multiclass skin lesion classification framework that is based on metadata and is aware of imbalance. Most importantly, this work presents empirical findings regarding the impact of imbalance correction via focal loss weighting on the overall accuracy and per-class sensitivity, which has not been explored in previous literature. The main novelty of this work is the development of a dual-branch architecture that integrates image embeddings from EfficientNet-B0 with an encoded representation of patient metadata (age, sex, localization) to allow the model to exploit clinically relevant contextual information as well as visual features. 2) A focal loss function that is class balanced is used to address the worse class imbalance in dermoscopic datasets, while a controlled ablation is performed by comparing the effects of aggressive versus

softened weighting on precision-recall trade-offs is quantified. 3) The patient/lesion leakage between train, validation, and test sets is avoided, as lesion level stratified data splitting is used, thus obtaining a realistic estimate of the generalization performance. Explainability is integrated using Grad-CAM to visualize the image regions that contribute to each prediction for all seven diagnostic classes, thus promoting clinical interpretability and trust.

II. LITERATURE SURVEY

The early diagnosis approaches of pigmented lesions were based on handcrafted features such as colour, texture and border descriptors which are classified using shallow learners, an approach later surpassed by deep transfer learning. Rezvantlab et al. [2] compared DenseNet-201, ResNet-152, InceptionV3 and InceptionResNetV2 with practising dermatologists in terms of performance on the HAM10000 and PH2 datasets. Gessert et al. [3] combined dermoscopic features with the patient metadata using a dense branch and added multi-resolution EfficientNet ensembles that won the ISIC-2019 challenge, while a related work [4] proposed diagnosis-guided loss weighting with patch-based attention to alleviate the imbalance in the loss itself. Frederich et al. [6] investigated the efficacy of EfficientNet-B0 and B1 on different augmentation strategies on HAM10000, and found the backbone to be a suitable candidate for this dataset. Kavitha et al. [7] and Khan et al. [15] extended EfficientNet variants with attention blocks and MixUp augmentation respectively, with a macro-F1 of 0.835 for MixUp reported. In the case of skewed ISIC-2019 data set, Gayatri and Aarthy [8] demonstrated that focal loss significantly improves precision and recall over cross-entropy, which aligns with the ablation in this work. As to explainability, Ahmad et al. [5] used an optimisation-based feature selector and Xception/ShuffleNet for the classification process; Dagnaw et al. [9] tested vision transformers, Swin variants and CNNs with XAI techniques including Grad-CAM, Grad-CAM++ and Score-CAM, with the addition of metadata improving results, although cohort gender was a problem for them; Cherukuri and Vara Prasad [11] created an ensemble transfer-learning model incorporating XAI; and Kishore Kumar et al. [13] used Grad-CAM to a MobileNet based classifier on HAM10000. Pipelines that incorporate segmentation are also promising, as in the work by Al-Fahdawi et al. [10], who combined Mask R-CNN/GrabCut segmentation and contrast enhancement prior to classification. More recent hybrid designs combine convolutional and transformer representations, such as the attention-based ConvNeXt-ViT hybrid of Naeem et al. [12] and the generalisable pipeline of Ashfaq et al. [14]. Whilst there are advancements in these areas, there is a lack of research that investigates each of these aspects simultaneously, and a lack of research that quantifies focal-loss weighting sensitivity, which this study aims to close.

III. PROPOSED METHODOLOGY

The framework consists of 5 stages: data preparation, data preprocessing, dual-branch modelling, imbalance-aware training, and explainability evaluation (Fig. 1). It contains HAM10000 (10,015 images, 7 classes: akiec, bcc, bkl, df, mel, nv, vasc) with additional age, sex, and lesion-site metadata, with nv the most common class (~67%), followed by df (~1%) and vasc (~1%). The images are resized to 224x224, ImageNet-normalized and augmented (+/- 20° rotation, color jitter), with missing metadata filled in with median/unknown fill. Stratified splitting at the lesion level (70/15/15) is used to prevent leakage and thus includes 7,002/1,532/1,481 train/val/test images. The image embedding is obtained from an EfficientNet-B0 image embedding network (1280-d), and the metadata embedding is obtained from a metadata MLP embedding network (19-d embedding vector, 32-d) by concatenation, and the resulting embedding is fed into a classification head consisting of 256 units. The class-balanced focal loss is used instead of cross-entropy, and two ablation settings are compared: aggressive (raw inverse-frequency weights, $\gamma=2.0$) and softened (square-root-scaled weights, $\gamma=1.0$). Training is performed with AdamW ($\text{lr}=3 \times 10^{-4}$, weight decay= 1×10^{-4}), batch size 32, plateau-based LR scheduling, up to 25 epochs and early stopping (patience=6) on validation loss. Evaluation includes accuracy, macro/weighted precision-recall-F1, confusion matrices, ROC-AUC and Grad-CAM visualizations in the last convolutional block for interpretability.

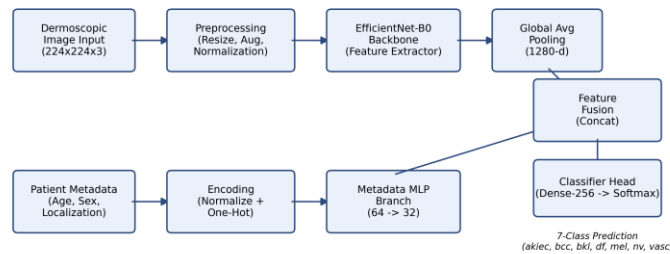


Fig. 1. Proposed dual-branch (image + metadata fusion) architecture for skin lesion classification.

IV. RESULTS AND DISCUSSION

A. Ablation Study: Impact of Focal Loss Weighting Intensity

Table I shows results of testing the two configurations on the test set. The high macro-recall (0.6765) value, coupled with a low accuracy (30.72%) and macro-F1 (0.3206) value, confirmed that the aggressive configuration had an over-corrected, melanoma biased decision boundary, with most recalled nevi being mistaken for melanoma (Fig. 2). The softened configuration ($\gamma=1.0$, sqrt-scaled weights) outperformed significantly the other configurations with strong improvements in accuracy (76.43%), macro-F1 (0.6601), dermatofibroma recall (95.0%), vascular lesions (100%), and recall along with more balanced improvements in melanoma precision (0.2308 $\rightarrow$ 0.3925), demonstrating imbalance robustness without compromising performance. (Fig. 3).

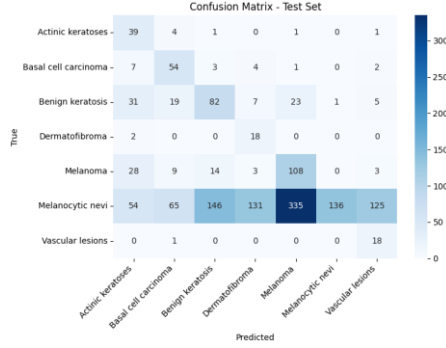


Fig. 2. Confusion matrix for the aggressive weighting configuration (Run 1), showing extensive nevi-to-melanoma misclassification.

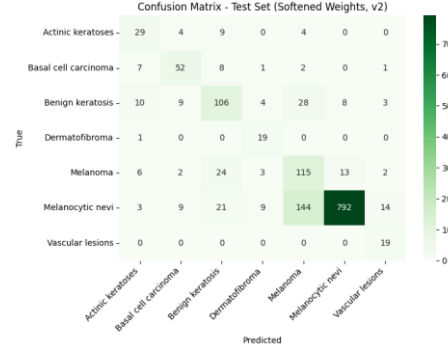


Fig. 3. Confusion matrix for the softened weighting configuration (Run 2), showing substantially improved balance across all classes.

TABLE I. Ablation comparison of aggressive ($\gamma=2.0$) vs. softened ($\gamma=1.0$) focal loss weighting on the HAM10000 test set ($n=1,481$).

Metric	Run 1 ($\gamma=2.0$)	Run 2 ($\gamma=1.0$)	Δ
Accuracy	0.3072	0.7643	+0.4571
Precision (macro)	0.3402	0.6021	+0.2619
Recall (macro)	0.6765	0.7770	+0.1005
F1-score (macro)	0.3206	0.6601	+0.3395
Macro-AUC	0.9014	0.9612	+0.0598
Nevi (nv) Recall	0.1371	0.7984	+0.6613
Melanoma Recall	0.6545	0.6970	+0.0425
Melanoma Precision	0.2308	0.3925	+0.1617

B. ROC-AUC Analysis

We can see how macro-AUC has been improved from 0.9014 to 0.9612 using softened weighting (Figs. Every class was greater than 0.89 AUC (4–5), reinforces the idea that the main reason for the weakness of the aggressive run was threshold miscalibration, rather than poor feature separability, because the aggressive run already had a good AUC; the additional gain in AUC for the softened run indicates that the quality of the representation has improved, and not just the calibration. The lowest per-class AUCs for both runs were those for melanoma (0.782, 0.891), confirming that it is visually similar to benign nevi and keratosis, as is well known.

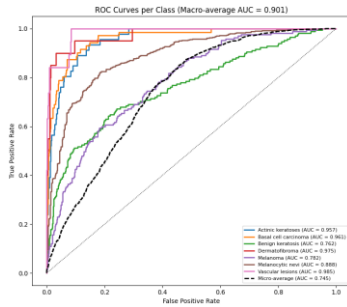


Fig. 4. ROC curves per class for the aggressive configuration (Run 1); macro-AUC = 0.901.

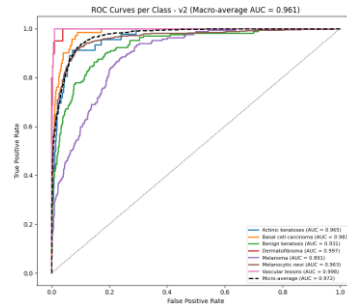


Fig. 5. ROC curves per class for the softened configuration (Run 2); macro-AUC = 0.961

C. Qualitative Analysis via Grad-CAM

Grad-CAM on one correctly classified sample per class (Fig. 6) consistently focused on lesion boundaries and not skin, hair, or artifacts, which was consistent with the more diffuse activation in the model for melanoma, and thus reflected the more irregular, asymmetric morphology (ABCDE criteria) of melanomas.

The softened configuration's accuracy and AUC are broadly consistent with existing models in the literature [2]–[3] (but not superior) due to the short training time of 25 epochs and the lack of segmentation preprocessing (as done in Skin-DeepNet [4]): while the main contribution is in showing loss-weighting sensitivity empirically, the accuracy is not the mark of a good model.

Limitations: melanoma precision/recall is still the poorest in the best configuration, indicating that melanoma lesions do overlap, not a flaw in the framework; results are based on a single split; only 2 weighting configurations were tested (Table I), and a finer γ -sweep remains to be explored; and the individual contribution of the metadata branch was not isolated by a dedicated ablation of the image-only, so it remains to be explored.

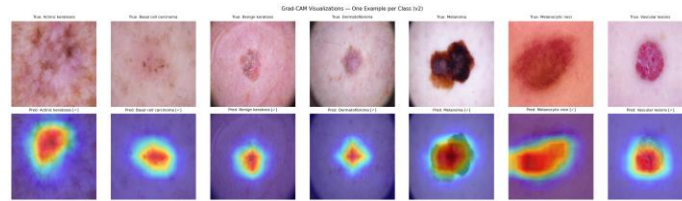


Fig. 6. Grad-CAM visualizations for one correctly classified example per diagnostic class (softened configuration, Run 2).

V. CONCLUSION AND FUTURE WORK

This study introduced a metadata-fused, imbalance-aware skin lesion classifier, and compared its accuracy-sensitivity trade-off by employing a controlled ablation that demonstrated that the weighting intensity of the focal-loss significantly influenced the accuracy-sensitivity trade-off: the aggressive scheme yielded a high macro-AUC (0.871) with lower recall for the majority class, while the moderated scheme resulted in a considerably more balanced 76.4% accuracy and 0.961 macro-AUC. Grad-CAM provided evidence of lesion-centred attention for all classes. In the future, the individual contribution of the metadata branch will be isolated, segmentation preprocessing will be added, and generalisation will be validated using additional corpora like ISIC 2019/ISIC 2020 and cross-validation.

REFERENCES

- [1] P. Tschandl et al., 'The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions,' *Scientific Data*, vol. 5, 2018.
- [2] A. Rezvantab et al., 'Dermatologist level dermoscopy skin cancer classification using different deep learning CNN algorithms,' *arXiv:1810.10348*, 2018.
- [3] N. Gessert et al., 'Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data,' *MethodsX*, vol. 7, 2020.
- [4] N. Gessert et al., 'Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting,' *IEEE Trans. Biomed. Eng.*, vol. 67, no. 2, pp. 495–503, 2020.
- [5] N. Ahmad et al., 'A novel framework of multiclass skin lesion recognition using deep learning and explainable AI,' *Frontiers in Oncology*, vol. 13, 2023.
- [6] J. Frederich et al., 'Skin lesion classification using EfficientNet B0 and B1 via transfer learning for CAD,' *AIP Conf. Proc.*, vol. 3080, 2024.
- [7] C. Kavitha et al., 'Skin cancer detection and classification using deep learning techniques,' *Procedia Computer Science*, vol. 235, pp. 2793–2802, 2024.
- [8] E. Gayatri, S. L. Aarthy, 'Reduction of overfitting on the highly imbalanced ISIC-2019 skin dataset using deep learning,' *J. X-Ray Sci. Technol.*, 2024.
- [9] G. H. Dagnaw et al., 'Skin cancer classification using vision transformers and explainable AI,' *J. Med. Artif. Intell.*, vol. 7, 2024.
- [10] S. Al-Fahdawi, M. I. Khalaf et al., 'A deep learning framework for early diagnosis and classification of skin cancer lesions in dermoscopy,' *Scientific Reports*, 2025.
- [11] S. Cherukuri, S. D. Vara Prasad, 'EnsembleSkinNet: transfer learning-based skin cancer detection with explainable AI,' *Frontiers in Oncology*, 2025.
- [12] M. A. Naem et al., 'Automated skin cancer detection using MedFusionNet with attention-based fusion of ConvNeXt and ViT,' *Scientific Reports*, 2025.
- [13] B. Kishore Kumar et al., 'Exploring skin lesion classification through explainable AI: Grad-CAM for interpretability,' *Proc. ICACECS 2024, Springer SIST* vol. 117, 2026.
- [14] N. Ashfaq et al., 'SkinSight: Advancing deep learning for skin cancer diagnosis and classification,' *Discover Computing*, vol. 28, 2025.
- [15] M. A. Khan et al., 'An EfficientNetB4-based CNN model with Soft Attention for skin lesion classification,' *Eng. Technol. Appl. Sci. Res.*, 2026.