JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DERMAXAI-NOVA: An Attention-Augmented EfficientNet-B3 Classifier

with Sharpness-Aware Training and Entropy-Based Uncertainty for Skin Lesion Classification

[1]Dr. Suresha D, [2]Eddula Teja Swaroop,[3]Hrishita Nandan Gowda, [4]Jyothi S, [5] Himanshu mankotia.

[1]Professor, [2–5]Final-Year B.E. Students
[1–5]Department of Computer Science and Engineering,
[1–5]Dr. Ambedkar Institute of Technology, Bengaluru, India

Abstract: Dermoscopic skin lesion classification suffers from class imbalance and visual ambiguity and uncertainty. This work assesses DERMAXAI-NOVA, which is an EfficientNet-B3 with CBAM attention and GeM pooling trained with weighted soft-target focal loss, SAM training, MixUp/CutMix data augmentation, and EMA/SWA. On the official ISIC 2018 Task 3 split, the network achieved an accuracy of 83.74 ± 1.11% averaged on 5 random seeds. Balanced accuracy on the same split was 75.69 ± 1.25%, AUC-ROC was 0.9613 ± 0.0039, and melanoma recall was 73.80 ± 3.54%. Removal of the CBAM component resulted in an average change in paired balanced accuracy of −0.05 ± 1.75, and removal of GeM yielded an average change in paired balanced accuracy of −1.86 ± 0.10, but macro AUC improved. Expected calibration error was 0.232 ± 0.014. The network has not been externally and clinically validated, thus, it cannot be used for diagnosis and should be regarded merely as a research prototype.

IndexTerms - class imbalance, EfficientNet, sharpness-aware minimization, skin lesion classification, uncertainty estimation

I. INTRODUCTION

Image classification of skin cancer has been shown to be very good at completing specific tasks [3] but just performing an excellent result at a limited task does not automatically make it ready to perform a full seven-class dermoscopic triage. The data provided by the ISIC 2018 Task 3 is highly imbalanced. The training split suffers from severe class imbalance as well; there are 115 dermatofibroma (DF) images versus 6,705 nevus (NV) images. Thus, many lesion types share similar visual attributes [1], [2]. Hence, evaluation requires high levels of accuracy, and/or class sensitive metrics. We have used many of the known methods described before, such as EfficientNet [4], CBAM Attention [5], GeM Pooling [6], Focal Loss [8], Sharpness-Aware Minimization (SAM) [9], MixUp [10], CutMix [11], and Stochastic Weight Averaging (SWA) [12]. We applied square-root inverse frequency sampling and adaptive class multipliers to our soft-target implementation of Focal Loss. Although, Effective-Number Weighting has also been studied as a part of this study [7].

We do not propose any new architecture or loss. We tested this combination on the official ISIC 2018 task 3 split, with 5 seeds. We report mean values, sample standard deviation, image bootstrapping intervals and calibration statistics. We also describe the two ablation studies we conducted: CBAM + GeM, model weights and TTA and logit post-hoc changes during validation. We explain how the research setup differs from what is deployed. Our contributions are as follows:

1. Five seed evaluation with average ± standard deviation, test-image bootstrap intervals, and calibration metrics;
2. Exploratory paired comparisons of CBAM and GeM ablations, weight variants, TTA, entropy-based uncertainty, and validation-selected logit adjustment;
3. Description of what differs between research evaluation and the deployed setup.

II. DATA AND PREPROCESSING

We use the official ISIC 2018 Task 3 splits: 10,015 images for training, 193 images for validation and 1,512 images for testing [1], [2]. The code looks at the counts and does not split the data again. The classes are AKIEC, BCC, BKL, DF, MEL, NV and VASC. Training counts are 327, 514, 1,099, 115, 1,113, 6,705 and 142, testing counts are 43, 93, 217, 44, 171, 909 and 35. The validation set is small (1 DF, 3 VASC, 8 AKIEC, 15 BCC and 21 MEL images) so per-class estimates become noisy. Images are resized to 260 × 260 and normalized using ImageNet statistics. The training augmentations include synthetic hair and vignetting (each with probability 0.2), horizontal/vertical flip, affine transforms (rotation up to ±180°, translation and scale), color jitter, and random erasing. The weighted sampler chooses probabilities proportional to $1/\sqrt{nc}$ where nc is the number of classes. The validation/test images are only resized and normalized. A heuristic audit identified 5,936 of 11,720 images but the flags did not filter or reweight the data.

III. METHOD

A. Model

We used the EfficientNet-B3 [4] model from torchvision as backbone. It uses ImageNet weights and has a feature map with 1,536 channels. The CBAM block [5] uses channel attention with a shared two-layer MLP with a reduction ratio of 16, and then spatial attention with one 7×7 convolution. Global average pooling is replaced with GeM pooling [6] with a learnable exponent p. That starts at three. 7 logits from a single linear layer and 0.3 dropout. The network therefore results in 11.00 million parameters.

B. Loss

Label smoothing (ε = 0.05) plus MixUp/CutMix keeps the targets soft. This loss, with target distribution q and softmax output p, is

$$L = \sum_k \alpha_k (1 - p_k)^{\gamma} (-q_k \log p_k), \quad \gamma = 2.0 \tag{1}$$

Before computing the loss, the class-multiplier vector α is normalized to mean one. Base multipliers: 2.0 for MEL, 1.5 for BCC and AKIEC, 1.0 for the rest. Each epoch sets a target for those three multipliers: $m_c^{(0)}[1 + 0.5(1 - r_c)]$, r_c is single-view validation recall, and $m_c^{(0)}$ is the base multiplier. The current multiplier moves halfway to that target, capped at three times its base value. Validation holds only 21 MEL, 15 BCC, and 8 AKIEC images, so these updates can be noisy.

C. Training and Weight Selection

Training runs for 30 epochs with all layers unfrozen. AdamW sits inside SAM (ρ = 0.05) [9]. We use a two-epoch linear warm-up to 3×10^{-4}, per-step cosine decay, batch size 32, gradient clipping at 5, and mixed precision. Each training batch has a 0.5 probability of undergoing MixUp or CutMix augmentation. When an augmentation is applied, MixUp (α = 0.4) and CutMix (α = 1.0) are selected with equal probability [10], [11]. EMA (with decay factor 0.998) is updated after each optimization step. SWA accumulates the model weights at the end of each epoch from epoch 22 to 30 [12]. For each seed we analyze raw-best weights (chosen by single-view validation balanced accuracy, with macro AUC as tie-breaker), final EMA and final SWA. Then we select the top candidate based on validation balanced accuracy under 8-view TTA, again breaking ties with macro AUC. We do not use the test set for selection. 4 seeds were selected with raw-best and 1 with EMA-final.

D. Inference and Uncertainty

The eight-view TTA takes an average of logits from four rotations of 90 degrees, each with and without a horizontal flip. We rank test images by entropy of the TTA-averaged softmax and entropy of the mean probabilities of 20 Monte Carlo dropout passes. We apply dropout to the pooled feature vector before the final classifier. Selective-risk behavior is summarized by the area under the risk-coverage curve (AURC). We report expected calibration error (ECE, 15 bins) and negative log-likelihood (NLL) for calibration. We do not set a rejection threshold for the research evaluation.

IV. EXPERIMENTAL SETUP

Seeds 0-3 were trained in Colab, seed 4 and the complete ablations were run in Kaggle. The seedbank notebook has identical code and default settings for each session. The Kaggle run used CUDA on a T4 class GPU. Completed seeds: 5 full-model seeds (0-4), 2 no-CBAM seeds (0-1), 2 no-GeM seeds (0-1). The planned no-SAM were not completed.. We report the mean and sample SD across seeds, along with per-seed 95% confidence intervals from 500 test-image bootstrap resamples. On seeds 0 and 1, ablation differences are paired against the full model. The metrics cover accuracy, balanced accuracy, macro AUC-ROC, average precision, F1, per-class precision and recall, ECE, and NLL.

V. RESULTS

A. Overall and Per-Class Performance

Table I reports the main test metrics over five full model seeds. For each seed we picked the candidate with the highest validation balanced accuracy under eight-view TTA and used macro AUC-ROC to break ties. Four seeds were selected for raw-best and one for EMA-final. The table reports mean sample SD over seeds, not just the best seed. The limits of the per-seed 500-resample test-image bootstrap intervals ranged from 69.83 to 80.77% for balanced accuracy and from 0.9449 to 0.9737 for macro AUC-ROC. Nevi constitute 60.1% of the test set. Thus a classifier that always predicts nevus would obtain an accuracy of 60.1%. In addition to accuracy, we also report balanced accuracy to make sure all classes are considered.

Table I. Primary test metrics across five seeds (mean ± SD; validation-selected weights and eight-view TTA)

Metric	Mean ± SD (n = 5 seeds)
Accuracy (%)	83.74 ± 1.11
Balanced accuracy (%)	75.69 ± 1.25
Macro AUC-ROC	0.9613 ± 0.0039
Macro average precision	0.8342 ± 0.0078
Weighted F1 (%)	83.86 ± 0.91
Macro F1 (%)	76.69 ± 1.20
ECE	0.2319 ± 0.0138
NLL	0.6728 ± 0.0301

Recall was 63.23 ± 2.05% for BKL on average, up to 92.15 ± 1.91% for NV. For MEL, recall hit 73.80 ± 3.54%, precision was 58.21 ± 5.52%, and one-vs-rest AUC-ROC came in at 0.9314 ± 0.0023. BCC recall reached 83.01 ± 2.57%, while AKIEC recall landed at 71.63 ± 4.47%. DF and VASC recall swung more across seeds (SD 6.51 and 10.22 percentage points, respectively), which lines up with their smaller test counts (44 and 35 images).

B. Error Patterns

The mean row-normalized confusion matrix and per-class ROC curves are shown in Fig. 1. NV was identified in approximately 19% of MEL images and MEL was identified in approximately 6% of NV images. Other row-wise confusions: NV predictions for ~20% of DF, ~15% of BKL, ~13% of VASC, and ~7% of AKIEC images; MEL predictions for ~14% of BKL images. These are class-conditional rates, not absolute error counts, and the class frequencies are quite different. The experiment does not show whether these errors are caused by visual similarity or some other model-specific failure mode.

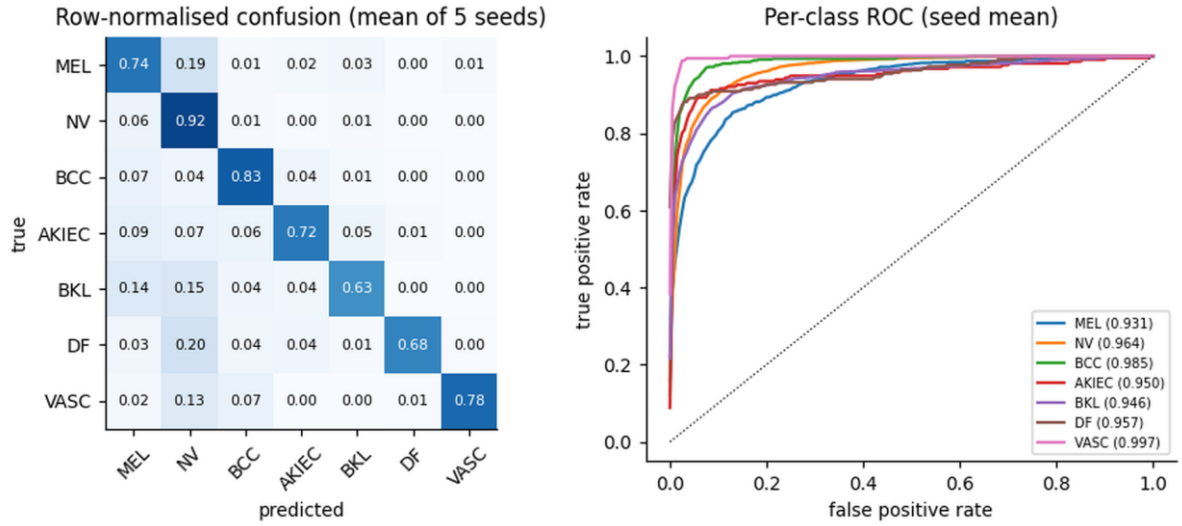

Fig. 1. Test-set row-normalized confusion matrix (left) and per-class ROC curves with AUC-ROC (right), mean over five seeds.

C. Ablations, Weight Variants, and Logit Adjustment

Comparison in Table II is made between the candidate weights, TTA, the selected logit adjustment via validation, and the two component ablations. TTA yielded better balanced accuracy improvement of 0.41 to 0.83 percentage points and macro AUC-ROC improvement of 0.0020 to 0.0035 points for raw-best, EMA and SWA. The EMA with TTA combination yielded the best mean accuracy (84.83%). Balanced accuracy for raw-best and EMA with TTA were approximately the same (75.99% and 75.96%), while for SWA with TTA was worse (79.56%). Each seed was selected based on validation balanced accuracy, not the mean test accuracy among the candidates. The ablations have only two seeds and the Δ columns are paired with full-model seeds 0 and 1 and not the mean five-seed full-model. Without CBAM, the difference in balanced accuracy was −0.05 ± 1.75 points and the difference in macro AUC was +0.0011 ± 0.0095 points. Without GeM, the balanced accuracy difference was −1.86 ± 0.10 points and macro AUC difference was +0.0051 ± 0.0030 points. There are no component benefits demonstrated here since the samples are relatively small. No SAM ablation was performed as planned.

Setting	n	Acc. (%)	Bal. acc. (%)	Macro AUC-ROC
Full model (Table I)	5	83.74 ± 1.11	75.69 ± 1.25	0.9613 ± 0.0039
Raw best, single view	5	83.27 ± 1.30	75.52 ± 1.51	0.9590 ± 0.0033
Raw best, TTA	5	83.85 ± 1.24	75.99 ± 1.33	0.9610 ± 0.0043
EMA, TTA	5	84.83 ± 0.51	75.96 ± 0.78	0.9580 ± 0.0031
SWA, TTA	5	79.56 ± 1.16	75.70 ± 1.32	0.9495 ± 0.0022
Logit adj., val-selected τ	5	83.58 ± 0.98	76.45 ± 1.94	0.9608 ± 0.0046
No CBAM	2	83.60 ± 0.09	76.29 ± 0.34	0.9636 ± 0.0035
No GeM	2	83.57 ± 0.42	74.48 ± 1.32	0.9676 ± 0.0030

D. Uncertainty and Calibration

Entropy-based uncertainty estimation has been studied for skin lesion classification [14], [15]; here, both entropy scores gave only a limited and variable selective-risk signal. The AURC was 12.42 ± 5.39% for TTA entropy and 13.00 ± 5.51% for MC-dropout entropy. Random ordering should have a risk close to the mean test error (about 16.3%), which is a reference level, not a separately trained baseline. The AURC SD (about 5.4 percentage points) was large compared with the accuracy SD (about 1.1), so neither entropy measure showed a consistent advantage across seeds. The ECE was 0.2319 ± 0.0138 and the NLL was 0.6728 ± 0.0301. These measures indicate calibration error, but they do not establish global under- or over-confidence. No post-hoc calibration or rejection threshold was validated; temperature scaling and risk-controlled selective prediction are future work.

VI. Application Configuration

The deployed DERMAXAI FastAPI/React prototype includes rule-based fusion of concern, symptom, and demographic scores extracted from images, as well as Grad-CAM and ABCD-style explanations. The five-seed classifier experiment did not assess these application layer components and no accuracy benefit is attributed to ABCD features. The app runs a previous checkpoint, not the research weights. Configuration varies in checkpoint provenance, inference transforms, and uncertainty scoring:

1. *Checkpoint provenance.* The application defaults to a backend/models/best.pth, but the available materials do not identify the served file by SHA-256 digest or repository commit, so its provenance is unverified.
2. *Inference transforms.* The application uses rotation/flip augmentation rather than the deterministic eight-view TTA used in the research experiment.

3. *Uncertainty threshold.* The configured threshold is 0.8054. It was selected for single-pass entropy according to the description, but it is used on a composite MCUE score that combines Monte Carlo dropout entropy, mutual information, and a Jensen-Shannon term. That composite score doesn't have a calibration yet.

VII. LIMITATIONS

1. That initial evaluation runs five seeds on a single official split. No formal significance tests were performed. We did two ablations, both with two seeds each. We never completed the planned no-SAM ablation.
2. The validation set contained only 193 images, resulting in noisy checkpoint selection and class-multiplier updates, particularly for underrepresented classes.We selected the checkpoints, weight variations, class multipliers per epoch and the τ value. That process was noisy as the validation set had very few examples for the rare classes (1 DF, 3 VASC, 8 AKIEC). We did not make any decisions using the test set.
3. *Calibration.* The ECE of the model shows how well its probabilities are calibrated. The NLL and entropy scores also varied between seeds. Recall for MEL was 73.80 ± 3.54%. No screening operating point or rejection thresholds were validated.
4. *Deployment.* We don't know exactly which checkpoint was used . We don't know the exact hash of the model or repo commit, so we cannot map from the research to the deployed model.
5. We did not test any external data (ISIC 2019/2020, BCN20000). We don't know how generalisable the system is. The system is not clinically validated and should not be used in place of professional diagnosis.

VIII. CONCLUSION

We trained an EfficientNet-B3 with CBAM, GeM, class weighted soft target focal loss, SAM, MixUp/CutMix and EMA/SWA. It achieved an accuracy of 83.74±1.11\%, balanced accuracy of 75.69± 1.25\% and macro AUC-ROC of 0.9613 ±0.0039 on 5 seeds over the official split for the ISIC 2018 Task 3 dataset. The average recall was 73.80 ± 3.54% for MEL, 83.01 ± 2.57% for BCC and 71.63 ± 4.47% for AKIEC. Applying TTA to each candidate weight showed small improvements. But ablation of CBAM or GeM with 2 seeds did not bring any benefits. Logit adjustment did not perform consistently well either. Calibration is definitely a problem: the ECE value is 0.232. Melanoma recall is a problem. This model was trained with a different checkpoint and inference setup than the one deployed in the application. This is a research prototype and has not been validated. We should work on improving the calibration of the model, finish the planned ablations and test on external datasets.

FUNDING

This research received no external funding.

CONFLICTS OF INTEREST

We authors have no conflicts of interest to declare regarding this article.

REFERENCES

[1] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Sci. Data, vol. 5, Art. no. 180161, 2018.

[2] N. C. F. Codella et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)," arXiv:1902.03368, 2019.

[3] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017.

[4] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2019.

[5] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018.

[6] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 7, pp. 1655–1668, 2019.

[7] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017.

[9] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[10] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2018.

[11] S. Yun et al., "CutMix: Regularization strategy to train strong classifiers with localizable features," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019.

[12] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," in Proc. Conf. Uncertainty Artif. Intell. (UAI), 2018.

[13] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. Int. Conf. Mach. Learn. (ICML), 2016.

[14] M. Combalia, F. Hueto, S. Puig, J. Malvehy, and V. Vilaplana, "Uncertainty estimation in deep neural networks for dermoscopic image classification," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2020, pp. 3211–3220, doi: 10.1109/CVPRW50498.2020.00380.

[15] P. Van Molle et al., "Quantifying uncertainty of deep neural networks in skin lesion classification," in H. Greenspan, R. Tanno, and M. Erdt, Eds., Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures, Lecture Notes in Computer Science, vol. 11840. Cham, Switzerland: Springer, 2019, pp. 52–61, doi: 10.1007/978-3-030-32689-0_6.