

An explainable ensemble deep learning model for multiclass skin disease classification using dermoscopic images

Ekramul Hasib¹, Touhidul Islam Masfi¹, S M Abdullah Shafi¹, Abhijit Bhowmik¹, Md. Saef Ullah Miah¹, and S. A. M. Manzur H. Khan¹

Department of Computer Science, American International University-Bangladesh,
Dhaka, Bangladesh
23-51106-1@student.aiub.edu, 22-49559-3@student.aiub.edu, shafi@aiub.edu,
abhijit@aiub.edu, saef@aiub.edu, manzur@aiub.edu

Abstract. Skin disease continues to cause serious concern worldwide, largely because melanoma can turn fatal quickly once diagnosis is delayed. Sorting a lesion correctly among several possible disease categories, and doing so early, has a direct bearing on how efficient treatment can be and on how well care can be tailored to the individual patient. Here we compare five convolutional networks – ResNet18, EfficientNet-B0, DenseNet121, MobileNetV2, and a custom-designed CNN – trained and tested on the HAM10000 dermoscopic image set. Of these, EfficientNet-B0 turned out to be the strongest performer on its own, posting an accuracy of 0.8739, precision of 0.8734, recall of 0.8739, an F1-score of 0.8723, and an AUC of 0.9321. Building on that, we bring EfficientNet-B0 together with MobileNetV2 and ResNet18 inside a weighted soft-voting ensemble, with the goal of cutting down variance and getting a more dependable prediction; doing so lifted performance to 0.8992 accuracy, 0.9089 precision, 0.8992 recall, an 0.8964 F1-score, and a 0.9483 AUC. We also bring in two explainability tools, Grad-CAM++ and LIME, so that the regions of each lesion actually driving the model’s decision can be seen rather than assumed, which helps make the whole pipeline more transparent. It is worth being upfront that these numbers, while encouraging, come from a single public benchmark, so the framework here is best treated as an early research prototype for decision support rather than a diagnostic tool ready for clinical use – testing on independent patient populations that reflect real demographic diversity still has to happen before that claim could be made.

Keywords: Skin Cancer · Deep Learning · EfficientNet-B0 · Ensemble Learning · Explainable Artificial Intelligence (XAI) · dermoscopic Images.

1 Introduction

Few cancers are as common, or carry the same risk of death, as skin disease does worldwide. Melanoma sits at the aggressive end of that spectrum it spreads

fast and accounts for a disproportionate share of the deaths tied to skin cancer overall. Because survival odds drop sharply the longer diagnosis takes, catching the disease early is not a minor detail but the whole game. Yet clinical practice still depends, in large part, on a dermatologist looking at a lesion and making a judgment call, and that judgment can shift depending on the clinician's experience, how ambiguous the lesion happens to look, or simply how stretched the clinic is that day. None of that inspires confidence that diagnosis will be consistent at scale, which is exactly why interest in automated, objective screening tools keeps growing.

Deep learning has quietly rewritten how medical images get analyzed, offering a route to solutions that behave the same way every time rather than varying with who's looking. CNN-based systems, in fact, have already closed much of the gap with board-certified dermatologists on image classification tasks. This [1] took a hybrid route, fusing CNNs with Transformer blocks through what they call a Convolutional Kolmogorov-Arnold Network (CKAN), and their results suggest that pairing the two architecture families captures feature detail that neither one gets to alone. Generative modelling has found a place here too [2] used a diffusion-based classifier, DiffMIC, across both skin- and oral-lesion data and showed it holds up well against standard CNN pipelines. None of this progress would have been possible without the benchmark datasets that keep getting released. This [3] put together BCN20000, a set of 19,424 dermoscopic images pulled from genuinely difficult clinical cases lesions on nails, on mucosal tissue which later became part of the ISIC 2019 Challenge. Another [4] Contributed the SIIM-ISIC 2020 dataset: 33,126 images spanning 2,056 patients across three continents, and notably the first of its kind to attach patient level metadata, letting models be judged in a way closer to how a clinician actually works through a case. In this [5] maintains DermIS, an open image atlas built through a joint effort between the Universities of Erlangen-Nuremberg and Heidelberg, still widely used today for developing and checking new algorithms.

On the more practical side of things, [6] built a lightweight CNN and tested it on HAM10000, making the case that smaller architectures can hold their own against heavier models when the target is a mobile device or an edge deployment. [7], working on architecture search, found NASNet edging out conventional CNNs on precision, recall, and F1-score for the same dataset evidence that automated architecture search isn't just a novelty but a genuinely useful optimization route. This [8] went a step further, wiring DenseNet, MobileNet, and GoogleNet together under an Ant Colony Optimization-guided hyperparameter search, which sped up convergence and improved classification on HAM10000 relative to the individual networks.

Most of the work summarized above zeroes in on tuning one architecture at a time. What we do differently here is build a single pipeline that ties several pieces together benchmarking multiple CNNs side by side, correcting for class imbalance through oversampling and augmentation, fusing models through an ensemble, and layering explainability on top rather than optimizing in isolation. Concretely, this paper puts forward a comparative, interpretable ensemble frame-

work for multiclass skin lesion classification, one that combines oversampling-based imbalance correction, benchmarking across several CNN architectures, a weighted soft-voting ensemble, and explainable AI, all aimed at getting both stronger robustness and clearer interpretability out of dermoscopic image analysis.

The contributions can be summarized as follows:

1. Random oversampling paired with image augmentation is used to correct the class imbalance present in the skin lesion dataset.
2. An ensemble pipeline, validated empirically, is built by combining complementary CNN architectures with explainable AI for robust multiclass classification.
3. Explainable AI is woven into the pipeline so that predictions can be trusted and interpreted rather than taken at face value.

The rest of this paper is organized as follows. The related work is reviewed in Section 2. In Section 3 we describe the dataset and the methodology. Section 4 presents the experimental results and model evaluation. Section 5 discusses the limitations and points out directions for future research. Finally, Section 6 concludes the study.

Table 1. Summary of Skin Disease Classification Studies

Study	Dataset	Method	Key Results	Limitations
[9]	ISIC 2019	Robust deep learning framework for multi-class skin cancer classification	Strong classification performance on imbalanced datasets	Limited dataset diversity; reduced smartphone image performance
[10]	HAM10000	Fuzzy rank-based deep ensemble using Xception, InceptionResNetV2, and MobileNetV2	Improved robustness and accuracy over single CNNs	High computational cost; difficult deployment
[11]	PAD-UFES-20	Smartphone-acquired clinical skin lesion dataset	Public dataset with 2,298 samples and six lesion classes	Limited population diversity; image variability
[12]	PH2	dermoscopic image benchmark with expert annotations	Standard benchmark dataset for evaluation	Small dataset size; limited class diversity
[13]	HAM10000	Large-scale dermoscopic benchmark dataset	Widely used for multiclass lesion classification	Class imbalance; demographic bias

2 Related Works

The literature indicates that ML and AI have been applied extensively in healthcare, assistive computing, secure infrastructures, and explainable analytics. Re-

search on cognitive health begins with AI-based cognitive assessment [22]. Dementia diagnosis has then been explored through comparative ML studies [31]. Interpretable fuzzy systems have further strengthened automated dementia diagnosis [38]. Alzheimer’s disease diagnosis has similarly benefited from transfer-learning ensembles [53]. Other contributions focus on AI- and Internet-of-Healthcare-Things-enabled Alzheimer’s care [29] and pervasive-sensor-based patient management [40]. Inclusive AI for brain-disorder-oriented healthcare adds a wider societal framing [54]. Research on autism includes ML-based psychological assessment [32], IoT-enabled autism screening using belief-rule-based reasoning [58], and fuzzy assistive environments for children [34]. In medical diagnosis, AI has been applied to brain MRI segmentation [45], hybrid brain pathology classification [23], diabetic retinopathy grading [28], breast cancer analysis from thermal imaging [36], and automatic lung screening [26]. Review work also addresses epileptic seizure detection in connected healthcare [52] and ML methods for PCOS detection [55]. Closely related methodological research includes neuronal signal processing [21], neural artefact detection in local field potentials [41], fuzzy ensemble learning for EEG emotion recognition [39], and image-encoding-based deep models for stress classification [50]. COVID-19 studies range from general ML reviews [24] to rough-set prediction of symptomatic cases [37]. Imaging-based work includes one-shot chest X-ray diagnosis [49], superpixel-based infection localisation [51], deep ensemble chest X-ray analysis [57], and hybrid optimisation-based detection [47]. The pandemic context also motivated healthcare robotics [42] and AI-driven stress monitoring [56]. Social-media-based mental health analytics further extend this theme through hybrid deep learning [30] and attention-based explainable depressive text detection [43]. Assistive and ambient intelligence research includes emotion-aware fall monitoring [25], multi-modal early fall detection [46], indoor navigation for neurodegenerative patients [48], and neuro-fuzzy wheelchair control using sEMG [44]. Human-centred analytics also cover explainable engagement analysis [27], sentiment-based student performance modelling [33], and human activity anomaly detection [20].

Work on multiclass skin cancer classification has fanned out in several directions lately new architectures, augmentation tricks, explainability, privacy-conscious training, and models that combine images with other data types. [9] put forward a deep learning framework built specifically to handle multiclass skin cancer classification, and it held up well even on heavily imbalanced data, largely because of improvements made to feature extraction and to the optimization process itself. [10] took a different tack, blending Xception, Inception ResNetV2, and MobileNetV2 through a fuzzy rank-based ensemble to deal with the uncertainty that HAM10000 tends to introduce; their fuzzy aggregation gave a principled way of handling classes that look alike but aren’t.

Datasets, of course, have done a lot of the heavy lifting for this field. PAD-UFES-20 [11] brought in 2,298 smartphone-captured images across six lesion types, each carrying up to 21 clinical metadata fields, specifically built to reflect the kind of diagnosis that happens in healthcare settings without much in the way of resources. PH2 [12] is smaller 200 dermoscopy images of melanocytic

lesions, labelled by dermatologists and accompanied by structural annotations and segmentation maps but it was designed from the outset as a benchmark for evaluating classifiers, and it still serves that role. HAM10000 [13] remains, arguably, the most heavily used benchmark in this space, offering 10,015 dermoscopic images pulled from multiple sources and spread across seven diagnostic categories.

Class imbalance keeps showing up as one of the field’s stubborn problems. [14] Approached it by pairing meta-learning with generative Denoising Diffusion Probabilistic Models (DDPMs) inside an explainable-AI setup, which helped both the balance of classes and the robustness of the resulting classifier; Grad-CAM was used on top to keep the outputs visually interpretable, something that matters if physicians are going to trust the system at all. [15] looked past the image alone, folding in structured clinical free-text alongside the dermoscopic images, and found that removing bias from the text data measurably improved how well the combined model performed. [16] went in a privacy-first direction, training MobileNetV2 across several independent clinical sites through federated learning so that raw patient images never had to leave any single site; the resulting FedAvg-based global model beat centrally trained baselines, which says something about how viable privacy-preserving collaboration can be for real deployment. [17] introduced SkinNet-14, a Compact Convolutional Transformer variant built to work on low-resolution (32×32) dermoscopic images across HAM10000, ISIC, and PAD-UFES, and their results pushed back against the assumption that high-resolution input is a prerequisite for accurate classification – the model held its own even under tight resource constraints. [18] ran a systematic comparison of squeeze-and-excitation, hard-attention, and soft-attention mechanisms bolted onto a pretrained Xception backbone using HAM10000, and found that attention consistently helped, mainly by pulling the model’s focus toward diagnostically relevant lesion regions and away from background clutter. More recently, [19] used transfer learning built around Deep Convolutional Generative Adversarial Networks (DCGANs) to generate additional malignant-lesion samples, boosting how well-represented the minority classes were and, in doing so, showing what generative methods can offer when a dataset is imbalanced to begin with.

3 Methodology

Figure 1 lays out the overall workflow this study follows, moving from dataset preparation and pre-processing through class balancing, model training, ensemble construction, evaluation, and finally an explainable-AI analysis stage.

3.1 Dataset Overview

The dataset underpinning this work is HAM10000 [19], a benchmark that has become something of a default choice for skin lesion classification research. It holds 10,015 dermoscopic images across seven diagnostic categories: melanocytic nevi

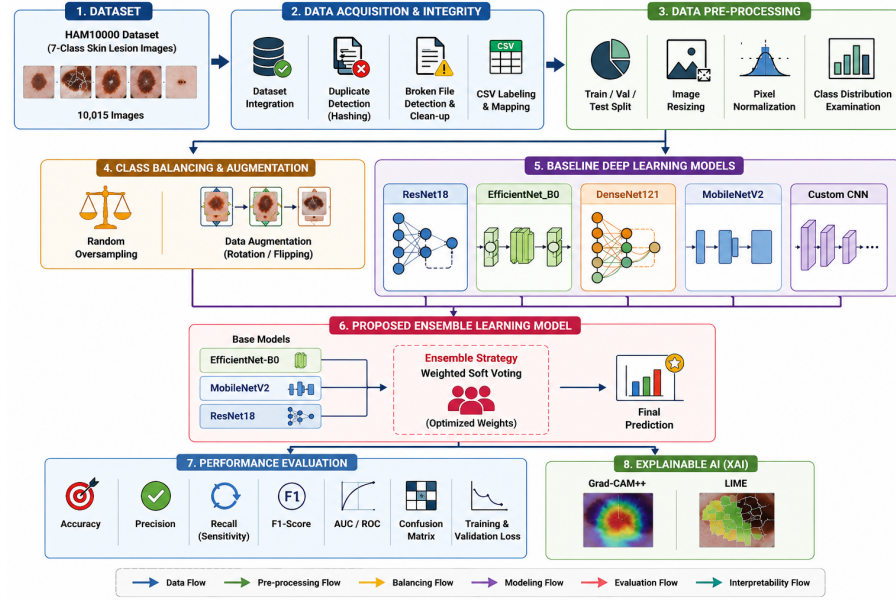


Fig. 1. Proposed deep learning methodology modeling workflow.

(NV) at 6,705 images, melanoma (MEL) at 1,113, benign keratosis-like lesions (BKL) at 1,099, basal cell carcinoma (BCC) at 514, actinic keratoses/Bowen's disease (AKIEC) at 327, vascular lesions (VASC) at 142, and dermatofibroma (DF) at 115. NV dominates the distribution by a wide margin, which is the imbalance problem this study has to work around, but the dataset's scale and diversity are still why it remains a standard reference point.

3.2 Data Pre-processing

Data Collection and Integrity

- **Data Acquisition and Integration:** All image sources were merged into a single repository so the classification pipeline would have one unified dataset to work from.
- **Hashing for Image Duplicate Detection:** Hashing was applied to catch and remove duplicate images redundancy in the training set can quietly bias a model, so this step was treated as essential rather than optional.
- **CSV Labeling & Mapping Wizard:** A master CSV was put together to keep image paths, labels, and lesion categories tied together correctly, reducing the chance of mislabeling anywhere downstream.

Data Splitting and Conventional Preprocessing

- **Data split into training, validation, and test sets.** Images were split 70/15/15 into training, validation, and test sets respectively, giving room for model fitting, hyperparameter tuning, and a final evaluation untouched by either. The split itself was stratified on the diagnostic label (`dx`) using scikit-learn’s `train_test_split` with `random_state = 42`, so class proportions stayed consistent across all three subsets. Duplicates flagged during hashing were removed beforehand, ruling out the possibility of the same duplicate image landing in two different partitions. One caveat worth flagging: HAM10000 contains several images per lesion in some cases (tracked through the `lesion_id` field), and because this split happens at the *image* level rather than the *lesion* level, images belonging to one lesion could technically end up spread across more than one partition. Moving to lesion-level grouping to close that gap entirely is something we’ve left for future work (see Section 5).
- **Uniform Rescaling of Images.** Images were resized to a fixed resolution across the board, mainly so that every architecture being compared would see input of the same shape and size.
- **Normalize and Scale Pixels.** Pixel values were normalized into a common range, which in practice made training more stable, cut down on intensity-driven variance between images, and helped the models converge faster.
- **Review initial class distribution.** Before any training happened, the class distribution was examined directly, mainly to get a concrete sense of just how imbalanced the categories were rather than assuming it from general knowledge of the dataset.

3.3 Class Balancing Using Oversampling and Augmentation

Geometric Data Augmentation (Flip/Rotation/Crop). Training images went through on-the-fly augmentation to widen the variety the model saw: rotations within $\pm 15^\circ$, horizontal flips with probability $p = 0.5$, and random resized crops (scale between 0.9 and 1.0) resized back to 224×224 pixels. This same pipeline touched every training image, including the duplicated minority-class samples that came out of random oversampling – no separate augmentation scheme was used just for those. Augmentation stayed confined to the training set only; validation and test images were left as-is, and the accuracy gains from this step showed up clearly across the multiclass models.

3.4 Reproducibility and Implementation Details

Everything was built in Python using PyTorch and torchvision, with scikit-learn handling the data split and the evaluation metrics. Training ran on a single NVIDIA Tesla T4 GPU through Google Colab, with a fixed seed (`seed = 42`) controlling the stratified split for all five baseline models as well as the ensemble. What wasn’t seeded globally was weight initialization and dataloader shuffling – adding that level of seed control, along with reporting variance across repeated runs, is on the list for future work.

3.5 Baseline Deep Learning Model Selection and Evaluation

Five architectures were trained and benchmarked as baselines: ResNet18, EfficientNet-B0, a custom CNN, DenseNet121, and MobileNetV2. ResNet18 leans on residual connections to sidestep the vanishing-gradient problem; EfficientNet-B0 balances depth, width, and resolution through compound scaling for a more efficient network overall; the custom CNN was built specifically with skin-lesion feature extraction in mind; DenseNet121's dense connectivity encourages feature reuse and keeps information flowing well through the network; and MobileNetV2 trades some capacity for a lightweight, computationally cheap design. All five were run under matched conditions so the comparison would be fair.

3.6 Ensemble Deep Learning Approach for Skin Disease Classification

Rather than settling for the best single model, we combine EfficientNet-B0, MobileNetV2, and ResNet18 into a weighted soft-voting ensemble, chosen because each had already shown strong standalone performance on HAM10000. Each network outputs its own class-probability distribution for a given image, and these get combined into a final prediction as:

$$P_{\text{final}} = w_1 P_{\text{EfficientNet-B0}} + w_2 P_{\text{MobileNetV2}} + w_3 P_{\text{ResNet18}} \quad (1)$$

where w_1 , w_2 , and w_3 are the weights assigned to EfficientNet-B0, MobileNetV2, and ResNet18 respectively, and P is the predicted class-probability distribution from each network. We settled on equal weighting ($w_1 = w_2 = w_3 = 1/3$) deliberately - it's a simple, data-agnostic fusion rule, and it sidesteps the risk of overfitting the fusion strategy to what is, relative to the full dataset, a fairly small validation partition. A more systematic search over (w_1, w_2, w_3) restricted to the validation set, combined with repeated-seed runs to check that the ensemble's gain isn't just a quirk of one particular split, is left for future work (Section 5). No probability calibration was applied before averaging - no temperature scaling, no Platt scaling - meaning the raw softmax outputs were combined directly, on the assumption that the three networks produce reasonably comparable confidence scores to begin with.

3.7 Model Evaluation Parameters

Accuracy, precision, sensitivity, specificity, and F1-score were the metrics used to judge both the baseline models and the proposed ensemble. Unless stated otherwise, precision, recall (sensitivity), and F1-score are reported as support-weighted averages across all seven diagnostic classes, matching the "Weighted Average" row in Table 5. Confusion matrices were used to dig into class-wise performance and to see where misclassifications tended to cluster, while AUC captured how well each model balanced true positives against false positives across different thresholds. Training and validation curves were also watched throughout to catch overfitting early.

3.8 Explainable Artificial Intelligence (XAI) for Model Interpretability

To keep the classifier’s reasoning from being a complete black box, we applied Grad-CAM++ and LIME. Grad-CAM++ in particular was used to highlight exactly which lesion regions the classifier was leaning on when it made each prediction.

4 Results and Findings

This section lays the proposed ensemble side by side with the baseline models, comparing them across accuracy, precision, recall, F1-score, AUC, and what the XAI outputs reveal about interpretability.

4.1 Performance Matrix Analysis (Benchmark Models) Using Weighted Score Method

Table 2 shows DenseNet121 coming out on top among the baselines under the weighted-score approach, hitting 85.58% accuracy and an 85.89% F1-score. EfficientNet-B0 wasn’t far behind at 84.71% accuracy and 85.05% F1-score. MobileNetV2 and the custom CNN both struggled more, and the custom CNN in particular finished at the bottom of the pack (71.26% accuracy, 62.17% F1-score). The gap between these numbers points fairly clearly to class imbalance as the main culprit - it’s the minority-class predictions that seem to be dragging performance down.

Table 2. Performance Comparison of Deep Learning Models Using Weighted Score Method

Model Name	Train Loss	Accuracy	Precision	Recall	F1-Score
ResNet18	0.0028	0.8397	0.8471	0.8397	0.8426
EfficientNet-B0	0.0037	0.8471	0.8566	0.8471	0.8505
MobileNetV2	0.0031	0.8082	0.7731	0.8082	0.7826
DenseNet121	0.0022	0.8558	0.8642	0.8558	0.8589
Custom CNN	0.0645	0.7126	0.5984	0.7126	0.6217

4.2 Impact of Weighted Score Method

Looking at Table 2 again, the weighted-score approach on its own clearly underperforms compared to what oversampling achieves. Class weighting alone wasn’t enough to give minority classes a fair shot at being learned properly, and that showed up directly in weaker predictive capability. Once oversampling was paired with geometric augmentation, though, the picture changed minority classes got meaningfully better representation during training, which fed through into stronger generalization and a more robust model overall.

4.3 Performance Matrix Analysis (Benchmark Models) Oversampling Method

Table 3 lays out results for all five models after 50 epochs of training under identical conditions, tested on the held-out test set. EfficientNet-B0 came out ahead of the pack here too, with 87.39% accuracy, an 87.23% F1-score, and a 93.21% AUC. DenseNet121 and MobileNetV2 trailed it by a similar margin, ResNet18 fell a bit further behind still, and the custom CNN again brought up the rear.

Table 3. Performance Comparison of Deep Learning Models (Using Random Oversampling)

Model Name	Train Loss	Accuracy	Precision	Recall	F1-Score	AUC
ResNet18	0.000762	0.8444	0.8347	0.8444	0.8327	0.9112
EfficientNet-B0	0.000029	0.8739	0.8734	0.8739	0.8723	0.9321
Custom CNN	0.0128	0.7760	0.7623	0.7760	0.7629	0.9145
DenseNet121	0.0019	0.8660	0.8623	0.8660	0.8629	0.9145
MobileNetV2	0.0019	0.8650	0.8600	0.8700	0.8600	0.8901

4.4 Proposed Model

Table 4. Performance of the Proposed Ensemble Model

Model Name	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Ensemble (Proposed)	0.8992	0.9089	0.8992	0.8964	0.9483

Table 4 lays out how the proposed ensemble performed once EfficientNet-B0, MobileNetV2, and ResNet18 were fused through weighted soft voting. The gain over any single model is clear, and the confusion matrix backs that up—performance holds up well across nearly every lesion category, and even the minority classes aren't misclassified all that often. Taken together, the ensemble reached 89.92% accuracy, 90.89% precision, 89.92% recall, an F1-score of 89.64%, and an AUC of 94.83%.

4.5 Proposed Model Confusion Matrix

The confusion matrix in Figure 2 tells much the same story visually—the bulk of predictions sit on the diagonal, which is what you'd want to see across all seven categories. Where things go wrong tends to happen between lesion types that look alike to begin with, MEL, BKL, and NV being the clearest examples.

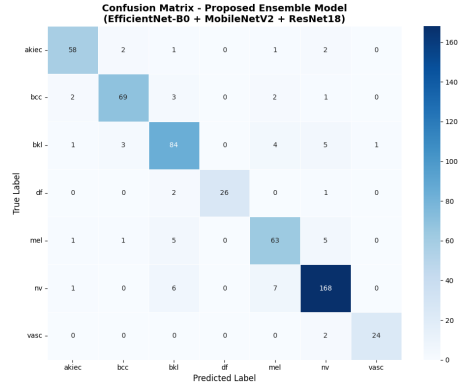


Fig. 2. Confusion matrix of the proposed ensemble model (EfficientNet-B0 + MobileNetV2 + ResNet18)

4.6 Classification Report of Skin Disease

Table 5. Classification Report of the Proposed Ensemble Model

Class	Accuracy	Precision	Recall	F1-Score
AKIEC	0.980	0.920	0.906	0.913
BCC	0.975	0.918	0.897	0.907
BKL	0.944	0.846	0.857	0.851
DF	0.995	0.929	0.897	0.913
MEL	0.953	0.831	0.840	0.835
NV	0.946	0.915	0.924	0.919
VASC	0.995	0.958	0.923	0.940
Weighted Average	0.8992	0.9089	0.8992	0.8964

The class-by-class breakdown in Table 5 is worth reading carefully. The accuracy column is one-vs-rest, $(TP + TN)/N$, pulled from the confusion matrix in Fig. 2, whereas recall is plain class-wise sensitivity, $TP/(TP + FN)$; accuracy sits higher than recall throughout simply because it also credits the model for correctly rejecting instances that belong to other classes. VASC, NV, and AKIEC came out strongest, while MEL and BKL lagged a bit – not surprising, given how visually close these lesion types can be to one another even for a trained eye. The weighted averages, on balance, back up the idea that the ensemble delivers reasonably even performance across the full class set rather than doing well only on the easy majority classes.

4.7 ROC-AUC Curve Analysis

Figure 3 shows the ROC curve for the ensemble, landing at an AUC of 0.9483 – well clear of what a random classifier would manage, and a solid sign that the model separates the classes cleanly.

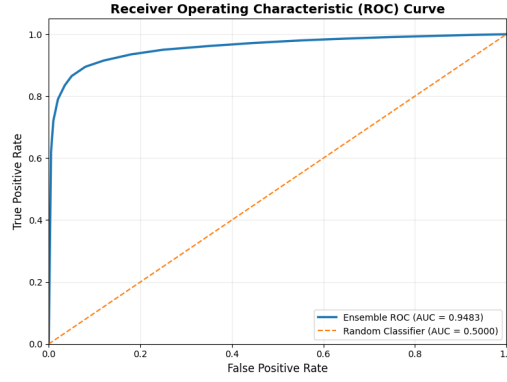


Fig. 3. ROC curve and AUC score of the proposed ensemble model.

4.8 Explainable AI

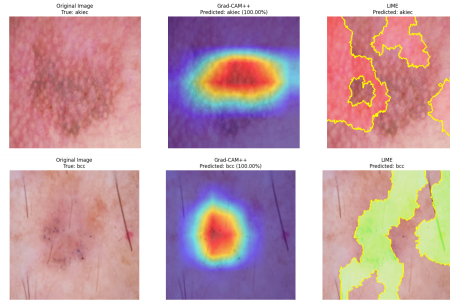


Fig. 4. Grad-CAM++ and LIME explanation for AKIEC and BCC.

Figures 4 through 7 walk through Grad-CAM++ and LIME outputs across all seven classes - AKIEC, BCC, MEL, NV, BKL, DF, VASC - for cases the model got right with high confidence. What stands out is how consistently both methods land inside the lesion boundary rather than drifting into the surrounding skin, which is a reasonably strong signal that the model is actually looking at what a dermatologist would look at, rather than picking up on background artefacts.

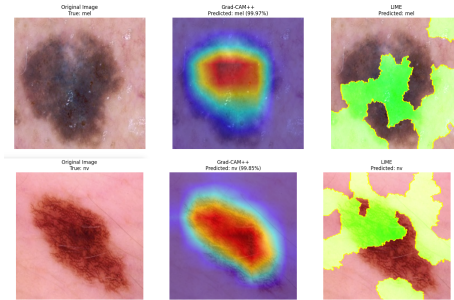


Fig. 5. Grad-CAM++ and LIME explanation for MEL and NV.

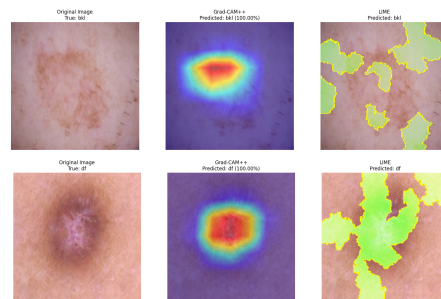


Fig. 6. Grad-CAM++ and LIME explanation for BKL and DF.

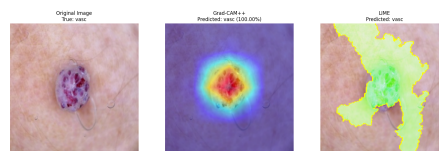


Fig. 7. Grad-CAM++ and LIME explanation for VASC.

5 Limitation and Future Work

The ensemble performs well, but it's not without gaps. For one, everything here rests on HAM10000 alone, and that dataset carries both class imbalance and limited demographic breadth, so how well these results would hold up in a genuinely diverse clinical population is still an open question. The interpretability work, too, stopped at the qualitative level - Grad-CAM++ and LIME give a visual sense of where the model is looking, but nothing here quantifies that with metrics like IoU, pointing-game accuracy, or faithfulness scores. On top of that, the ensemble has no probability calibration built in, the data split happened at the image level rather than the lesion level, and the fusion weights were fixed upfront instead of being tuned empirically. Going forward, the plan is to validate on larger and more demographically varied datasets, improve how well minority classes are handled, bring in explainability metrics that can actually be quantified, add proper seed control across the whole pipeline, and generally push the framework closer to something deployable in a real clinical setting.

6 Conclusion

What this paper sets out to do is build an explainable ensemble for multiclass skin disease classification using dermoscopic images from HAM10000, and on the whole it succeeds at that. Five architectures went into the comparison - ResNet18, EfficientNet-B0, DenseNet121, MobileNetV2, and a custom CNN - and EfficientNet-B0 came out as the strongest individual performer, at 87.39% accuracy and 93.21% AUC. Fusing EfficientNet-B0, MobileNetV2, and ResNet18 into a weighted soft-voting ensemble pushed those numbers higher still: 89.92% accuracy, 90.89% precision, 89.92% recall, an 89.64% F1-score, and a 94.83% AUC, beating every model tested individually. Grad-CAM++ and LIME rounded things out by giving a visual, interpretable window into which lesion regions actually drove each prediction. All told, the ensemble looks like a genuinely promising decision-support tool for computer-aided dermatological diagnosis - though it still needs to prove itself on independent, demographically diverse, multi-centre data before anyone should think about real clinical deployment.

7 Code and Data Availability

This study uses the publicly available HAM10000 dataset [19]. The source code is maintained in a private GitHub repository (https://github.com/EkramulHasib/Skin_Disease/blob/main/Skin_Disease.ipynb) and is not currently publicly accessible; reasonable access requests may be directed to the corresponding author at 23-51106-1@student.aiub.edu, subject to applicable licensing restrictions.

References

1. Agarwal, S., Mahto, A.K.: Skin Cancer Classification: Hybrid CNN-Transformer Models with KAN-Based Fusion. arXiv preprint arXiv:2508.12484 (2025). <https://doi.org/10.48550/arXiv.2508.12484>
2. Krohling, R.A.: Diffusion Models Applied to Skin and Oral Cancer Classification. arXiv preprint arXiv:2504.00026 (2025). <https://doi.org/10.48550/arXiv.2504.00026>
3. Combalia, M., Codella, N.C.F., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., Malvehy, J.: BCN20000: dermoscopic Lesions in the Wild. arXiv preprint arXiv:1908.02288 (2019). <https://doi.org/10.48550/arXiv.1908.02288>
4. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Weber, J., Soyer, H.P.: A Patient-Centric Dataset of Images and Metadata for Identifying Melanomas Using Clinical Context. *Sci. Data* **8**(1), 34 (2021). <https://doi.org/10.1038/s41597-021-00815-z>
5. University of Erlangen–Nuremberg, University of Heidelberg: DermIS – Dermatology Information System. <https://www.dermis.net>, last accessed May 2025
6. Absur, M.N.: Optimizing Deep Learning for Skin Cancer Classification: A Computationally Efficient CNN with Minimal Accuracy Trade-Off. arXiv preprint arXiv:2505.21597 (2025). <https://doi.org/10.48550/arXiv.2505.21597>
7. Gowthami, D., Gowri Shankar, C., Kathiravan, K., Dhanush, K., Dharshni, K.M., Lavanya, C.: Performance Comparison of Deep Learning-Based Classification of Skin Cancer. In: Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART), pp. 1–8 (2025). SciTePress. <https://doi.org/10.5220/0013909400004407>
8. Moustafa, E.: Achieving High-Accuracy Skin Cancer Classification with Deep Learning Optimized by Ant Colony Algorithm. *J. Electr. Syst. Inf. Technol.* **12**(1), 49 (2025). <https://doi.org/10.1186/s43067-025-00243-8>
9. Pacal, I.: A Robust Deep Learning Framework for Multiclass Skin Cancer Classification. *Sci. Rep.* **15**(1), 4938 (2025). <https://doi.org/10.1038/s41598-025-89230-7>
10. Halder, A., Dalal, A., Gharami, S., Wozniak, M., Ijaz, M.F., Singh, P.K.: A Fuzzy Rank-Based Deep Ensemble Methodology for Multi-Class Skin Cancer Classification. *Sci. Rep.* **15**(1), 6268 (2025). <https://doi.org/10.1038/s41598-025-90423-3>
11. Pacheco, A.G.C., Lima, G.R., Salomão, A.S., Krohling, B.A., Biral, I.P., de Angelo, G.G., Alves Jr., F.C.R., Esgario, J.G.M., Simora, A.C., Castro, P.B.C., Rodrigues, F.B., Frasson, P.H.L., Krohling, R.A., Knidel, H., Santos, M.C.S., do Espírito Santo, R.B., Macedo, T.L.S.G., Canuto, T.R.P., de Barros, L.F.S.: PAD-UFES-20: A Skin Lesion Dataset Composed of Patient Data and Clinical Images Collected from Smartphones. *Data Brief* **32**, 106221 (2020). <https://doi.org/10.1016/j.dib.2020.106221>
12. Mendonça, T., Ferreira, P.M., Marques, J.S., Marçal, A.R.S., Rozeira, J.: PH2: A dermoscopic Image Database for Research and Benchmarking. In: Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5437–5440. IEEE, Osaka (2013). <https://doi.org/10.1109/EMBC.2013.6610779>

13. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. *Sci. Data* **5**, 180161 (2018). <https://doi.org/10.1038/sdata.2018.161>
14. Akter, M.S., Shahriar, H., Sneha, S., Cuzzocrea, A.: Multi-Class Skin Cancer Classification Architecture Based on Deep Convolutional Neural Network. In: 2022 IEEE International Conference on Big Data (Big Data), pp. 5404–5413. IEEE (2022).
15. School of Computer Science and Engineering, Yeungnam University: Explainable AI Based Multi-Class Skin Cancer Detection Enhanced by Meta Learning with Generative DDPM Data Augmentation. *Appl. Sci.* **15**(21), 11689 (2025). <https://doi.org/10.3390/app152111689>
16. Subramanian, M., Kalaivani, B., Jeevasree, G., Mathan Kumar, A., Nandhini, P.S.: Skin Cancer Classification and Detection Using Federated Learning. In: Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART), pp. 1–8. SciTePress (2025). <https://doi.org/10.5220/0013586200004407>
17. Al Mahmud, A., Azam, S., Khan, I.U., Montaha, S., Karim, A., Haque, A., Zahid Hasan, Md., Brady, M., Biswas, R., Jonkman, M.: SkinNet-14: A Deep Learning Framework for Accurate Skin Cancer Classification Using Low-Resolution Dermoscopy Images with Optimized Training Time. *Neural Comput. Appl.* **36**(30), 18935–18959 (2024). <https://doi.org/10.1007/s00521-024-10225-y>
18. Chaturvedi, S.S., Gupta, K., Prasad, P.S.: Skin Lesion Analyser: An Efficient Seven-Way Multi-Class Skin Cancer Classification Using MobileNet. In: International Conference on Advanced Machine Learning Technologies and Applications, pp. 165–176. Springer (2020).
19. Liu, X., Yu, Z., Tan, L., Yan, Y., Shi, G.: Enhancing Skin Lesion Diagnosis with Ensemble Learning. In: 2024 4th International Conference on Computer Science and Blockchain (CCSB), pp. 94–98. IEEE (2024). <https://doi.org/10.1109/CCSB63463.2024.10735675>
20. S. W. Yahaya, A. Lotfi, and M. Mahmud, “Towards a data-driven adaptive anomaly detection system for human activity,” *Pattern Recognit. Lett.*, vol. 145, pp. 200–207, 2021.
21. M. Mahmud and S. Vassanelli, “Processing and analysis of multichannel extracellular neuronal signals: State-of-the-art and challenges,” *Front. Neurosci.*, vol. 10, no. JUN, 2016.
22. A. R. Javed, A. Saadia, H. Mughal, T. R. Gadekallu, M. Rizwan, P. K. R. Mad-dikunta, M. Mahmud, M. Liyanage, and A. Hussain, “Artificial Intelligence for Cognitive Health Assessment: State-of-the-Art, Open Challenges and Future Directions,” *Cogn. Comput.*, vol. 15, no. 6, pp. 1767–1812, 2023.
23. B. Deepa, M. Murugappan, M. Sumithra, M. Mahmud, and M. S. Al-Rakhmi, “Pattern Descriptors Orientation and MAP Firefly Algorithm Based Brain Pathology Classification Using Hybridized Machine Learning Algorithm,” *IEEE Access*, vol. 10, pp. 3848–3863, 2022.
24. M. Mahmud and M. S. Kaiser, “Machine learning in fighting pandemics: A COVID-19 case study,” in *COVID-19: Prediction, Decision-Making, and Its Impacts*, Springer, 2020, pp. 77–81.
25. M. J. Al Nahian, T. Ghosh, M. N. Uddin, M. M. Islam, M. Mahmud, and M. S. Kaiser, “Towards Artificial Intelligence Driven Emotion Aware Fall Monitoring Framework Suitable for Elderly People with Neurological Disorder,” *Proc. BI 2020*, vol. 12241 LNAI, pp. 275–286, 2020.

26. H. Mukherjee, P. Sreerama, A. Dhar, S. M. Obaidullah, K. Roy, M. Mahmud, and K. Santosh, "Automatic Lung Health Screening Using Respiratory Sounds," *J. Med. Syst.*, vol. 45, no. 2, 2021.
27. M. A. Rahman, D. J. Brown, N. Shopland, A. Burton, and M. Mahmud, "Explainable Multimodal Machine Learning for Engagement Analysis by Continuous Performance Test," *Proc. HCI 2022*, vol. 13309 LNCS, pp. 386–399, 2022.
28. M. Murugappan, N. Prakash, R. Jeya, A. Mohanarathinam, G. Hemalakshmi, and M. Mahmud, "A novel few-shot classification framework for diabetic retinopathy detection and grading," *Meas.: J. Int. Meas. Confed.*, vol. 200, 2022.
29. S. Jesmin, M. S. Kaiser, and M. Mahmud, "Artificial and Internet of Healthcare Things Based Alzheimer Care During COVID 19," *Proc. BI 2020*, vol. 12241 LNAI, pp. 263–274, 2020.
30. M. H. A. Banna, T. Ghosh, M. J. A. Nahian, M. S. Kaiser, M. Mahmud, K. A. Taher, M. S. Hossain, and K. Andersson, "A Hybrid Deep Learning Model to Predict the Impact of COVID-19 on Mental Health from Social Media Big Data," *IEEE Access*, vol. 11, pp. 77009–77022, 2023.
31. Y. Miah, C. N. E. Prima, S. J. Seema, M. Mahmud, and M. S. Kaiser, "Performance Comparison of Machine Learning Techniques in Identifying Dementia from Open Access Clinical Datasets," *Adv. Intell. Syst. Comput.*, vol. 1188, pp. 79–89, 2021.
32. S. Ahmed, M. F. Hossain, S. B. Nur, M. S. Kaiser, and M. Mahmud, "Toward Machine Learning-Based Psychological Assessment of Autism Spectrum Disorders in School and Community," *Proc. TEHI 2021*, vol. 376, pp. 139–149, 2022.
33. J. Watkins, M. Fabielli, and M. Mahmud, "SENSE: A Student Performance Quantifier using Sentiment Analysis," in *Proc. IJCNN*, 2020, pp. 1–6.
34. A. I. Sumi *et al.*, "fASSERT: A fuzzy assistive system for children with autism using internet of things," *Proc. BI 2018*, vol. 11309 LNAI, pp. 403–412, 2018.
35. M. F. Zohora, M. H. Tania, M. S. Kaiser, and M. Mahmud, "Forecasting the Risk of Type II Diabetes using Reinforcement Learning," in *Proc. ICIEV, icIVPR 2020*, 2020.
36. M. J. Mammoottil, L. J. Kulangara, A. S. Cherian, P. Mohandas, K. Hasikin, and M. Mahmud, "Detection of Breast Cancer from Five-View Thermal Images Using Convolutional Neural Networks," *J. Healthc. Eng.*, vol. 2022, 2022.
37. H. R. Bhapkar, P. N. Mahalle, G. R. Shinde, and M. Mahmud, "Rough sets in COVID-19 to predict symptomatic cases," in *COVID-19: Prediction, Decision-Making, and Its Impacts*, Springer, 2020, pp. 57–68.
38. T. Chen, P. Su, Y. Shen, L. Chen, M. Mahmud, Y. Zhao, and G. Antoniou, "A dominant set-informed interpretable fuzzy system for automated diagnosis of dementia," *Front. Neurosci.*, vol. 16, 2022.
39. T. Dhara, P. K. Singh, and M. Mahmud, "A Fuzzy Ensemble-Based Deep learning Model for EEG-Based Emotion Recognition," *Cogn. Comput.*, vol. 16, no. 3, pp. 1364–1378, 2024.
40. T. M. Niamat Ullah Akhund, M. J. N. Mahi, A. Hasnat Tanvir, M. Mahmud, and M. S. Kaiser, "ADEPTNESS: Alzheimer's disease patient management system using pervasive sensors - early prototype and preliminary results," *Proc. BI 2018*, vol. 11309 LNAI, pp. 413–422, 2018.
41. M. Fabietti, M. Mahmud, A. Lotfi, A. Avarua, D. Gugganmos, R. Xudo, and M. Chiappalone, "Neural Network-based Artifact Detection in Local Field Potentials Recorded from Chronically Implanted Neural Probes," in *Proc. IJCNN 2020*, 2020.

42. M. S. Kaiser, S. Al Mamun, M. Mahmud, and M. H. Tania, "Healthcare robots to combat COVID-19," in *COVID-19: Prediction, Decision-Making, and Its Impacts*, Springer, 2020, pp. 83–97.
43. T. Ghosh, M. H. A. Banna, M. J. A. Nahian, M. N. Uddin, M. S. Kaiser, and M. Mahmud, "An attention-based hybrid architecture with explainability for depressive social media text detection in Bangla," *Expert Syst. Appl.*, vol. 213, 2023.
44. M. S. Kaiser, Z. I. Chowdhury, S. Al Mamun, A. Hussain, and M. Mahmud, "A Neuro-Fuzzy Control System Based on Feature Extraction of Surface Electromyogram Signal for Solar-Powered Wheelchair," *Cogn. Comput.*, vol. 8, no. 5, pp. 946–954, 2016.
45. H. M. Ali, M. S. Kaiser, and M. Mahmud, "Application of Convolutional Neural Network in Segmenting Brain Regions from MRI Data," *Proc. BI 2019*, vol. 11976 LNAI, pp. 136–146, 2019.
46. M. Nahiduzzaman, M. Tasnim, N. T. Newaz, M. S. Kaiser, and M. Mahmud, "Machine Learning Based Early Fall Detection for Elderly People with Neurological Disorder Using Multimodal Data Fusion," *Proc. BI 2020*, vol. 12241 LNAI, pp. 204–214, 2020.
47. A. K. Singh, A. Kumar, M. Mahmud, M. S. Kaiser, and A. Kishore, "COVID-19 Infection Detection from Chest X-Ray Images Using Hybrid Social Group Optimization and Support Vector Classifier," *Cogn. Comput.*, vol. 16, no. 4, pp. 1765–1777, 2024.
48. M. Biswas, A. Rahman, M. S. Kaiser, S. Al Mamun, K. S. Ebne Mizan, M. S. Islam, and M. Mahmud, "Indoor Navigation Support System for Patients with Neurodegenerative Diseases," *Proc. BI 2021*, vol. 12960 LNAI, pp. 411–422, 2021.
49. V. N. M. Aradhya, M. Mahmud, D. Guru, B. Agarwal, and M. S. Kaiser, "One-shot Cluster-Based Approach for the Detection of COVID-19 from Chest X-ray Images," *Cogn. Comput.*, vol. 13, no. 4, pp. 873–881, 2021.
50. S. Ghosh, S. Kim, M. F. Ijaz, P. K. Singh, and M. Mahmud, "Classification of Mental Stress from Wearable Physiological Sensors Using Image-Encoding-Based Deep Neural Network," *Biosensors*, vol. 12, no. 12, 2022.
51. N. Prakash, M. Murugappan, G. Hemalakshmi, M. Jayalakshmi, and M. Mahmud, "Deep transfer learning for COVID-19 detection and infection localization with superpixel based segmentation," *Sustain. Cities Soc.*, vol. 75, 2021.
52. S. Jahan, F. Nowsheen, M. M. Antik, M. S. Rahman, M. S. Kaiser, A. S. M. S. Hosen, and I.-H. Ra, "AI-Based Epileptic Seizure Detection and Prediction in Internet of Healthcare Things: A Systematic Review," *IEEE Access*, vol. 11, pp. 30690–30725, 2023.
53. T. Mahmud, K. Barua, A. Barua, S. Das, N. Basnin, M. S. Hossain, K. Andersson, M. S. Kaiser, and N. Sharmen, "Exploring Deep Transfer Learning Ensemble for Improved Diagnosis and Classification of Alzheimer's Disease," *Proc. BI 2023*, vol. 13974 LNAI, pp. 109–120, 2023.
54. S. Al Mamun, M. S. Kaiser, and M. Mahmud, "An Artificial Intelligence Based Approach Towards Inclusive Healthcare Provisioning in Society 5.0: A Perspective on Brain Disorder," *Proc. BI 2021*, vol. 12960 LNAI, pp. 157–169, 2021.
55. S. Ahmed, M. S. Rahman, I. Jahan, M. S. Kaiser, A. S. M. S. Hosen, D. Ghimire, and S.-H. Kim, "A Review on the Detection Techniques of Polycystic Ovary Syndrome Using Machine Learning," *IEEE Access*, vol. 11, pp. 86522–86543, 2023.
56. S. Jesmin, M. S. Kaiser, and M. Mahmud, "Towards artificial intelligence driven stress monitoring for mental wellbeing tracking during COVID-19," in *Proc. IEEE/WIC/ACM WI-IAT 2020*, 2020, pp. 845–851.

57. A. Paul, A. Basu, M. Mahmud, M. S. Kaiser, and R. Sarkar, “Inverted bell-curve-based ensemble of deep learning models for detection of COVID-19 from chest X-rays,” *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16113–16127, 2023.
58. M. E. Alam, M. S. Kaiser, M. S. Hossain, and K. Andersson, “An IoT-belief rule base smart system to assess autism,” in *Proc. iCEEiCT 2018*, 2018, pp. 672–676.