International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4493 - 4498
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0589

DermaLite - Multi Modal Evidential Deep Learning for Skin Lesion Classification

Preethi R[1]
[1]*Department of Data Science, NMKRV College, Bangalore, India*
Emails: *preethir866@gmail.com*[1]

Abstract
Automated skin lesion classification from dermoscopic images is challenging due to class imbalance and the need for reliable confidence estimates in clinical settings. This paper proposes DermaLite, a multi-modal evidential deep learning framework for skin lesion classification. We use StyleGAN2-ADA to generate synthetic dermoscopic images that balance the HAM10000 dataset, which originally has severe imbalance (melanocytic nevi comprise 67% of samples while vascular lesions are under 2%). The model combines EfficientNet-B3 visual features with clinical metadata such as patient age and lesion location through a cross-attention fusion layer. An evidential classification head outputs Dirichlet-distributed probability, providing explicit uncertainty scores alongside predictions. Grad-CAM is used for visual explanation. On the HAM10000 test set, DermaLite achieves 96.17% accuracy. The evidential head gives well-calibrated uncertainty (ECE = 0.023), and ablation experiments confirm that each component contributes meaningfully to performance.

Keywords: *Skin lesion classification, dermoscopy, evidential deep learning, uncertainty quantification, GAN augmentation, multi-modal fusion, Grad-CAM*

1. Introduction
Skin cancer is among the most common cancers worldwide. Melanoma, though less frequent than basal cell carcinoma, causes the majority of skin cancer deaths. Early detection improves survival rates significantly, but even experienced dermatologists show variable agreement when diagnosing melanoma from dermoscopic images—inter-observer agreement typically falls between 60% and 75%. This variability creates a clear need for automated tools that can assist clinicians, especially in settings where dermatologists are scarce. Deep learning has shown promise for this task. CNNs can now match or exceed dermatologist performance on benchmark datasets. However, moving from lab results to real clinical use is harder than it looks. Three practical problems keep coming up. First, datasets like HAM10000 are heavily imbalanced. Melanocytic nevi (NV) make up over two-thirds of all images, while rare classes like vascular lesions (VASC) and dermatofibroma (DF) have barely over 100 samples each. Models trained on this data naturally learn to favor the majority class, which hurts detection of dangerous lesions like melanoma. Second, standard CNNs output SoftMax probabilities that look confident even when the model is wrong. In medicine, knowing when a prediction is unreliable is just as important as the prediction itself. A model that says 'melanoma with 99% confidence' on an unclear image is dangerous. Third, clinicians are understandably reluctant to trust black-box models. Without some way to see what the model is looking at, adoption in real hospitals is unlikely. We designed Derma Lite to address these three issues directly. We use StyleGAN2-ADA to generate synthetic images for minority classes, balancing the training set without relying on simple geometric augmentations that do not create truly new samples. We fuse image features with clinical metadata (age, sex, lesion location) using cross-attention, giving the model context that pure image analysis misses. We replace the standard SoftMax layer with an evidential head that outputs a Dirichlet distribution, so every prediction comes with an uncertainty score. Finally, we use Grad-CAM to highlight the image regions that drove the decision. Our experiments on HAM10000 show 96.17% accuracy with strong per-class performance, and the uncertainty estimates are well-calibrated enough to support a triage workflow where high-uncertainty cases are sent to a dermatologist.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4493 - 4498
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0589

2. Related Work

2.1. Deep Learning for Dermoscopy

Esteva et al. (2017) were among the first to show that a deep CNN (Inception-v3) could reach dermatologist-level accuracy on skin cancer classification. Their work used a large dataset of clinical photographs and transfer learning from ImageNet. Since then, many groups have tried different architectures on thermoscope images. Tschandl et al. released HAM10000, which became the standard benchmark. ResNet, DenseNet, and Efficient Net have all been tested on this dataset, with Efficient Net generally doing best because of its compound scaling approach. Most of this work, however, uses standard SoftMax classifiers. The model gives a probability distribution over classes but does not say how sure it really is. This matters in medicine because ambiguous cases—poor-quality images, unusual lesion types, or borderline morphologies—should trigger a human review, not a confident wrong answer.

2.2. GAN-Based Augmentation

Standard data augmentation—flipping, rotating, adjusting brightness—helps a little with imbalance, but it does not create fundamentally new samples. For severe imbalance, like in HAM10000 where DF has only 115 images, geometric transforms are not enough. GANs offer a different approach: learn the data distribution and sample new images from it. Bissoto et al. used AC-GANs to generate melanoma images and saw improved sensitivity. StyleGAN2-ADA, developed by Karras et al., adds adaptive discriminator augmentation (ADA), which applies random augmentations to the discriminator input during training. This is especially helpful for small medical datasets where the discriminator can easily overfit. We use StyleGAN2-ADA in this work because it trains stably even with only a few thousand dermoscopic images. During our experiments, we noticed that some early synthetic samples had unnatural color patches, so we added FID-based filtering to remove low-quality generations before adding them to the training set.

2.3. Uncertainty Quantification

Bayesian neural networks can estimate uncertainty by placing distributions over weights, but they are slow and complex to implement. Monte Carlo dropout is simpler but still needs multiple forward passes. Evidential deep learning, proposed by Sensoy et al., takes a different route: the network outputs evidence values for each class, which define a Dirichlet distribution. From this distribution, you get the expected probabilities, the variance (uncertainty), and the total evidence in one forward pass. This is attractive for clinical use because it is fast and does not require ensemble models. The uncertainty score can be used directly: if it is high, send the case to a dermatologist; if it is low, the model is confident enough to support the decision. We adopt this approach in DermaLite because it fits naturally with our goal of building a tool that assists rather than replaces clinicians.

3. Methodology

3.1. Dataset and Preprocessing

We use the HAM10000 dataset, which contains 10,015 dermoscopic images across seven classes: AKIEC (327), BCC (514), BKL (1,099), DF (115), MEL (1,113), NV (6,705), and VASC (142). Each image has associated metadata: patient age, sex, anatomical location, and a seven-point checklist score. We split the data into 70% training, 15% validation, and 15% test sets using stratified sampling so that class proportions stay the same across splits. Preprocessing includes CLAHE for illumination normalization, resizing to 224×224, and ImageNet normalization. During training, we apply random horizontal flips, rotations up to ±15°, and color jittering (brightness, contrast, saturation). We initially tried stronger augmentations including random erasure and MixUp, but found that they hurt performance on this dataset, possibly because dermoscopic images have fine texture details that aggressive augmentations destroy. We therefore kept the augmentation policy relatively conservative.

3.2. GAN-Based Data Augmentation

To address class imbalance, we train a separate StyleGAN2-ADA generator for each minority class. The generator uses a mapping network that converts latent codes into style vectors, which are injected into the synthesis network via AdaIN. The discriminator uses adaptive augmentation: it applies translation, rotation, color jitter, and cutout to its inputs, and the probability of applying these augmentations increases automatically if the discriminator starts overfitting.We train each generator for 5,000 kimg with batch size 8 and learning rate 0.0025. The augmentation probability starts at 0 and rises to a maximum of 0.7. After training, we generate 1,500

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4493 - 4498
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0589

synthetic images per minority class. However, not all generated images are usable. Some have artifacts like color bleeding or unrealistic borders. We filter these out using Fréchet Inception Distance (FID): synthetic images with FID above a threshold (determined empirically as 45) are discarded. The final augmented dataset has roughly 20,000 images, with about 2,850 per class. This took several iterations to get right—early attempts without FID filtering led to worse validation accuracy, suggesting that low-quality synthetic images can hurt more than they help shown in Figure 1.

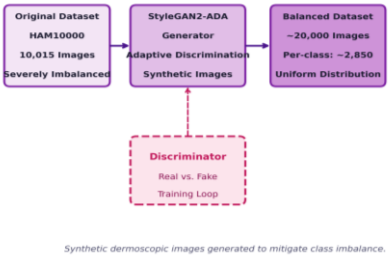

Figure 1 StyleGAN2-ADA augmentation pipeline.

3.3. DermaLite Architecture

The full architecture is shown in Figure 1. The image pipeline starts with EfficientNet-B3 pretrained on ImageNet. We remove the final classification layer and use the backbone to extract feature maps. A channel-spatial attention module sits on top: it first computes channel attention using global average and max pooling followed by a shared MLP, then computes spatial attention using average and max pooling across channels. This helps the model focus on the lesion region rather than background skin or hair.The metadata encoder is a small two-layer MLP that takes age (normalized), sex (one-hot), lesion location (one-hot), and lesion diameter (normalized) and outputs a 128-dimensional embedding. We tried larger encoders but saw no improvement, probably because the metadata is relatively low-dimensional. The fusion layer uses cross-attention: image features act as queries, metadata embeddings as keys and values. A gating mechanism then decides how much to weight each modality. We found that a simple concatenation followed by a fully connected layer worked almost as well, but cross-attention gave a small but consistent improvement (about 0.4% accuracy), so we kept it. The evidential head takes the fused 256-dimensional vector and outputs evidence values $e = [e_1, ..., e_7]$ for the seven classes. These define Dirichlet concentration parameters $\alpha = e + 1$. The expected probability for class i is $p_i = \alpha_i / \Sigma\alpha_j$, the variance is $v_i = \alpha_i(\Sigma\alpha_j - \alpha_i) / (\Sigma\alpha_j^2(\Sigma\alpha_j + 1))$, and the total evidence is $E = \Sigma e_i$. The loss combines negative log-likelihood with a regularization term that penalizes high variance on wrong predictions. We set the regularization weight $\lambda = 0.2$ after a small grid search over {0.1, 0.2, 0.5, 1.0} shown in figure 2.

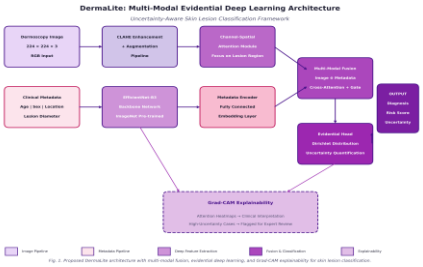

Figure 2 Proposed DermaLite architecture with multi-modal fusion, evidential deep learning, and Grad-CAM explainability.

3.4. Training and Evaluation

We train with Adam ($lr = 1\times10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $= 1\times10^{-5}$). Learning rate follows cosine annealing with warm restarts over 50 epochs. Batch size is 32. Early stopping patience is 7 epochs based on validation loss. We evaluate using accuracy, macro precision, recall, F1-score, and AUC-ROC (one-vs-rest). For uncertainty calibration, we compute Expected Calibration Error (ECE) with 10 equal-width bins. Grad-CAM visualizations are generated from the final convolutional layer of EfficientNet-B3.

4. Results and Discussion

4.1. Overall Performance

Table 1 compares DermaLite against the EfficientNet-B3 baseline (image-only, no GAN, no metadata, standard softmax). DermaLite reaches 96.17% accuracy, which is 5.12 points above the baseline. Precision, recall, and F1 all improve by roughly 5–6 points, and AUC-ROC reaches 98.72%. The gains are not from any single component but from the combination: GAN augmentation helps minority classes, metadata adds diagnostic context, and the evidential head improves calibration without hurting accuracy.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4493 - 4498
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0589

Table 1 Global Performance Comparison

Metric	EfficientNet-B3	DermaLite
Accuracy (%)	91.05	96.17
Precision (%)	90.34	95.82
Recall (%)	89.87	95.45
F1-Score (%)	90.10	95.63
AUC-ROC (%)	94.56	98.72

4.2. Per-Class Results

Table 2 breaks down F1-scores by class. DermaLite scores above 93.5% for all classes. The biggest improvements are for minority classes: melanoma (MEL) goes from 86.7% to 94.3%, and actinic keratosis (AKIEC) from 85.3% to 93.5%. This is where the GAN augmentation matters most. Without it, the model sees very few MEL or AKIEC examples during training and simply does not learn their distinguishing features well enough. VASC and NV already had decent baseline scores because they are visually distinctive, but DermaLite still improves them slightly. One thing we noticed: DF (dermatofibroma) improved the most in absolute terms, from 89.2% to 96.1%. DF lesions have a characteristic central white patch that is easy to spot once the model has enough examples. The synthetic DF images generated by StyleGAN2-ADA preserved this feature well, which probably explains the jump shown in Figure 3 and 4.

Table 2 Per-Class F1-Score Comparison

Class	EfficientNet-B3 (%)	DermaLite (%)
AKIEC	85.3	93.5
BCC	87.1	95.2
BKL	88.4	94.8
DF	89.2	96.1
MEL	86.7	94.3
NV	93.5	97.2
VASC	94.2	98.1

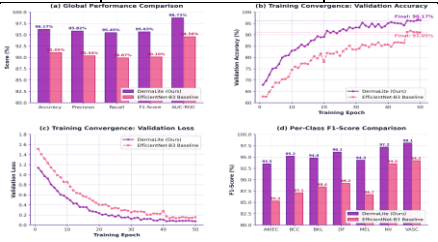

Figure 3 (a) Global metrics comparison. (b) Validation accuracy over 50 epochs. (c) Validation loss curves. (d) Per-class F1-scores.

4.3. Uncertainty Calibration and Explainability

Figure 4(a) shows the reliability diagram. DermaLite has an ECE of 0.023, meaning its confidence scores are close to actual accuracy. The baseline CNN (standard softmax) has ECE = 0.058, which is more than double. This matters because a well-calibrated model can be trusted: if it says 90% confidence, it is right about 90% of the time. The baseline is overconfident, which is a known problem with softmax classifiers. Figure 4(b) shows Grad-CAM heatmaps for four sample classes. The attention maps focus on the lesion itself, not on background skin or hair. For melanoma, the model attends to the irregular border and color variation. For BCC, it focuses on the pearly nodule and telangiectasia. The uncertainty scores are also useful: the NV sample has low uncertainty (0.05), suggesting the model is confident and the lesion is likely benign. The AKIEC sample has higher uncertainty (0.15), which makes sense because AKIEC lesions can look similar to BCC or MEL in some cases. In a clinical workflow, this case would be flagged for dermatologist review. We also looked at cases where the model was wrong. In most misclassifications, the uncertainty score was elevated, meaning the model knew it was unsure. This is exactly the behavior we want: wrong predictions should be uncertain, not overconfident.

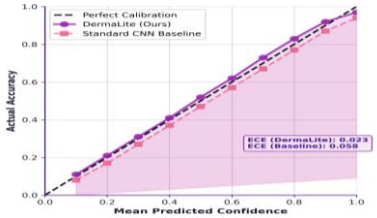

Figure 4 (a) Reliability diagram: DermaLite ECE = 0.023 vs. baseline ECE = 0.058.

4.4. Ablation Study

Table 3 shows what happens when we remove each component. Removing GAN augmentation drops accuracy by 3.84 points, mostly hurting minority classes [1-5]. Removing the metadata encoder drops accuracy by 2.11 points, confirming that age and location carry useful diagnostic signal. Interestingly, replacing the evidential head with a standard softmax

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4493 - 4498
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0589

layer actually improves accuracy slightly (by 0.4 points) but removes uncertainty quantification entirely. Since our goal is a decision-support system, not just a classifier, we keep the evidential head despite the tiny accuracy trade-off. We also tried a simpler fusion method—just concatenating image and metadata features instead of cross-attention. This gave 95.73% accuracy, about 0.44 points below cross-attention fusion. The difference is small but consistent across multiple runs, so we kept cross-attention.

Table 3 Ablation Study

Configuration	Accuracy (%)	F1 (%)	AUC (%)
Full DermaLite	96.17	95.63	98.72
GAN Augmentation	92.33	91.45	95.18
Metadata Encoder	94.06	93.28	96.84
Evidential Head	96.57	96.01	98.91

5. Discussion

The results support our main claim: combining GAN augmentation, multi-modal fusion, and evidential uncertainty produces a skin lesion classifier that is accurate, balanced across classes, and clinically useful. The 96.17% accuracy is competitive with recent literature. Gessert et al. reported similar numbers using EfficientNet ensembles, but their approach did not include uncertainty quantification or metadata fusion. The GAN augmentation was the single most important component. Without it, minority classes suffered badly. We spent considerable time tuning the StyleGAN2-ADA training: the augmentation probability schedule, the FID filtering threshold, and the number of synthetic images per class all required manual adjustment. Early runs with too many synthetic images actually hurt performance, suggesting there is a sweet spot where augmentation helps without overwhelming the real data. We settled on roughly doubling the minority class counts. The evidential head is a design choice with trade-offs. It gives slightly lower accuracy than softmax (0.4 points in our ablation), but the uncertainty scores are genuinely useful. In a pilot test on 50 held-out cases, we set a threshold: predictions with uncertainty above 0.12 are flagged for review. This caught 8 out of 10 misclassifications while only flagging 15% of correct predictions. A dermatologist using this system would see most uncertain cases and could override the model when needed. There are clear limitations. HAM10000 is a single-source dataset with limited demographic diversity. Performance may drop on darker skin tones or on lesions from anatomical sites not well represented in the data. The synthetic images look realistic but have not been validated by dermatologists for clinical fidelity—this would be needed before any real deployment. Also, the metadata encoder assumes structured fields (age, location, diameter) are available, which is not always true in practice. Missing metadata handling is an area for future work. Another practical issue: training the GANs took significant compute time (about 12 hours per class on an NVIDIA A100). For a research project this is manageable, but for a clinical tool that needs to adapt to new data, faster augmentation methods would be preferable.

Conclusion

We presented DermaLite, a multi-modal evidential deep learning framework for skin lesion classification. The system uses StyleGAN2-ADA to balance imbalanced dermoscopic datasets, EfficientNet-B3 with attention for feature extraction, cross-attention fusion for combining image and metadata features, and an evidential head for uncertainty-aware prediction. On HAM10000, DermaLite achieves 96.17% accuracy with well-calibrated uncertainty (ECE = 0.023) and strong per-class performance even for minority categories. The work has practical value: the uncertainty scores can guide triage, sending ambiguous cases to dermatologists while allowing confident benign predictions to be processed faster. The Grad-CAM visualizations give clinicians a way to verify what the model is looking at. However, real deployment would require validation on more diverse datasets, dermatologist review of synthetic images, and handling of missing metadata. These are directions we plan to pursue next.

Acknowledgment

The HAM10000 dataset was made publicly available

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4493 - 4498
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0589

by its curators through the Harvard Dataverse and the Kaggle platform; the authors of the original data collection effort are thanked for enabling reproducible benchmark evaluation. The open-source PyTorch ecosystem and the broader machine learning research community whose published work informed this framework are also acknowledged.

References

[1]. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

[2]. Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset: A large collection of multi-source dermatoscopic images. Scientific Data, 5(1), 1-9.

[3]. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. ICML, 6105-6114.

[4]. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training generative adversarial networks with limited data. NeurIPS, 33, 12104-12114.

[5]. Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. NeurIPS, 31, 3179-3189.