From Cropping Strategies to Multimodal Fusion: A Systematic Evaluation and Ablation Study on ISIC 2019 Skin Lesion Classification

Du Yuxi (2330026036) Hou Shuoran (2330026054) Lu Yunxiao (2330026114)
Group 9

Abstract

Previous studies have shown that GANs can serve as an effective minority class augmentation technique in lung imaging and have a positive effect [5, 9, 11]. Therefore, this project focuses on multi-class skin lesion classification and explores two complementary directions: (1) Analyze whether GAN technology has a similarly positive effect on skin disease classification by adding a small number of GAN-generated samples to the original training set. (2) Combine clinical/demographic metadata (such as age and gender) with image features to examine whether multimodal input can improve the recognition of minority classes. In the end, we concluded that: Combining metadata can enhance the ability to classify skin diseases, with WACC increasing from 0.6050 to 0.6810. In contrast, GAN technology has no positive effect in this aspect, with WACC decreasing from 0.6810 to 0.6416.

1. Introduction

Medical image classification often suffers from severe class imbalance: models tend to favor majority classes, leading to low recall for rare but clinically important lesions and unstable performance across data splits. Improving minority-class recognition and the reliability of evaluation is therefore crucial for practical deployment.

A common direction to mitigate imbalance is to incorporate richer information. Prior work has explored (i) data augmentation, including GAN-based synthesis for minority classes in some medical imaging tasks [5, 9, 11], and (ii) multimodal learning that fuses clinical/demographic metadata with image features. Motivated by these ideas [3, 8], we study skin lesion classification on ISIC 2019 [1] and evaluate whether metadata fusion and GAN-based minority augmentation can improve imbalance-aware performance.

Accordingly, our study focuses on three objectives: (1) compare two multi-view cropping strategies (SS vs. RR) under the same 25-view inference budget; (2) test whether adding clinical/demographic metadata (age, sex, anatomical site) improves minority recognition (SS+Meta); (3) test whether adding 5,000 GAN-synthesized minority-class images to the *training set only* further improves imbalance-aware metrics (SS+Meta+GAN), while keeping validation real-only to avoid evaluation bias.

2. Methodology

2.1. Hypotheses

We focus on two testable questions: (1) whether different image processing methods affect the recognition of minority classes; (2) whether adding metadata (age/gender/anatomical site) to the model can improve the recognition ability of minority classes and overall robustness; (3) after adding 5,000 GAN-generated minority class images to the training set, whether key metrics such as WACC change, and whether these changes reflect an improvement in the recognition of minority classes.

2.2. Dataset

2.2.1. Official ISIC 2019 Images and Labels

We used the official ISIC 2019 dataset. This dataset can be divided into two parts: (1) 25,331 images of skin diseases; (2) a CSV file containing the corresponding image names and 9 categories of data for classification. To enhance discriminative ability, we also integrated clinical/demographic metadata, including age, gender, and anatomical site.

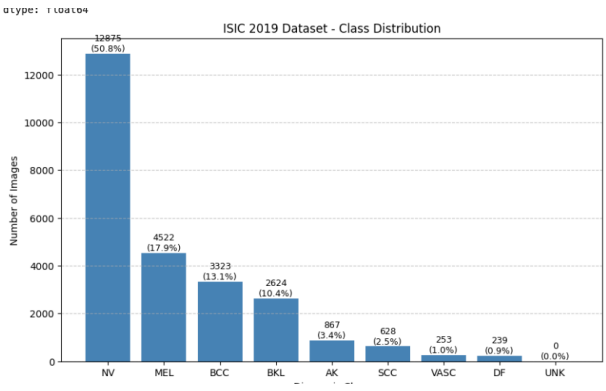

Figure 1. ISIC 2019 dataset class distribution.

2.2.2. GAN-Generated Minority-Class Image Data

We will augment the original training set with 5,000 minority class images generated by GANseg., SCC, VASC and DF). They come with a CSV of real labels, following the same format as the official one (same columns, one-hot encoding), and are integrated into the same dataset as the original (a total of 30,331 samples).

It should be noted that the images we generate are used only for the model's training set and are not included in the test set.

Table 1. Dataset description (all fields).

Feature	Type	Description
image	Categorical	Image identifier (e.g., ISIC_0000000)
age	Numeric	Patient age
sex	Categorical	Patient sex; encoded
site	Categorical	Anatomical site; encoded
AK	Binary (1/0)	Actinic keratosis
BCC	Binary (1/0)	Basal cell carcinoma
BKL	Binary (1/0)	Benign keratosis-like lesions
DF	Binary (1/0)	Dermatofibroma
MEL	Binary (1/0)	Melanoma
NV	Binary (1/0)	Melanocytic nevus
SCC	Binary (1/0)	Squamous cell carcinoma
UNK	Binary (1/0)	Unknown/other
VASC	Binary (1/0)	Vascular lesion

2.3. cGAN-based Data Generation Framework

For the data augmentation of medical images, it is very important to enhance the fidelity and it's relevance to the clinic semantic. The images generated by the standard GAN is not constricted by any clinic semantic, always causing the images to be ambiguous, which means they may lack the pathological details. To further enhance the controllability and the quality of the generated data, the cGAN(Conditional Generative Adversarial Network) is introduced [7]. With the introduction of clinic metadata, the generation manifold is constricted and the training processing is also stabilized.

2.3.1. Input Representation & Conditioning

To bridge the gap between the data structure that neural network accepting and the unstructured clinic data, we transformed the images and the data into a tensor.

Condition vector construction. We constructed a 19 dimensional vector with the diagnosis labels, anatomical site and the patients' ages. Labels and anatomical site are Encoded as One-Hot vectors. The patients' ages are normalized to 0 to 1($Age_{norm} = Age_{current}/Age_{max}$) to prevent the instability in the network.

Image processing. Real images are resized into 256x256 to match the output range of Tanh.

Model integration. In generator, we concatenate the 19-dimensional vector with a 100-dimensional random noise, forming a 119-dimensional vector, giving the input a clear semantics. In the discriminator, due to the difference in the dimensions of 19-dimensional vectors and images, we used spatial extension techniques by replicating the 19-dimensions in spatial dimension and then stacked the 19 feature maps with the 3-channels RGB image, forming an input channel. Thus the discriminator can make judgment about whther the image is true with the clinic background information.

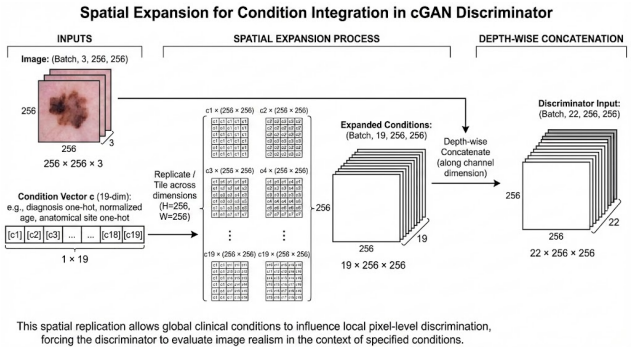

Figure 2. Input of discriminator via spatially concatenated condition.

2.3.2. Generator & Discriminator Architectures

The generator is a 7-layers architectures, aiming at upsampling the vector in to a 256 x 256 RGB images. It has 7 convolution transposed layers, each layers' resolution is doubled than the previous layer, meanwhile the width is decreasing symmetrically. All the layers adapt the ReLU as activation function and batch normalization to stabilize the training process. The output layer adapts the Tanh, mapping the output within the range[-1, 1].

The discriminator is a binary classifier, processing the input of 22 channels tensor. It adapts 7-layers architecture, utilizing a convolution with stride equals to 2 to downsample. The hidden layers use LeakyReLU and batch normalization(except for the first layer). The output layer is a sigmoid function outputting a scaler 0 or 1.

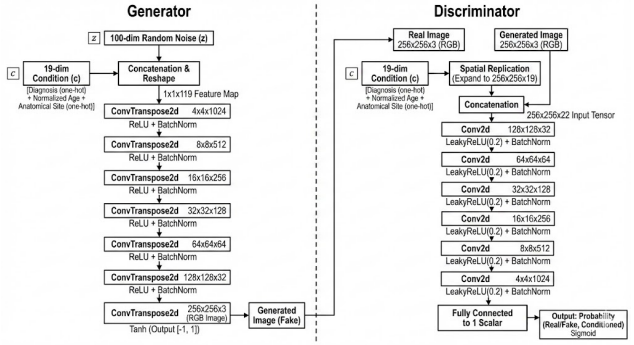

Figure 3. Architecture of cGAN.

2.3.3. Training Strategy

During the training process, we totally trained 50 epochs and 10 checkpoints were kept. In order to find out the model perform the best, we evaluated the log output and visually evaluated the generated images. In the early stages, the generator wins the discriminator, the D(G(z)) is low and after several epochs it rose again.From the stage of stable adversarial interaction between the discriminator and the generator to the end of training, the $Loss_D$, $Loss_G$, $D(x)$ and $D(G(z))$ fluctuated so it is hard to pick a model only depending on the training log. Comprehensive log quantitative analysis and visual qualitative assessment, the images generated in the early stages have noticeable mesh-like artifacts, which originate from the periodic high-frequency noise generated by the generator's upsampling layer, severely impacting visual quality. Ultimately, the model in epoch 49 was chosen to generate the subsequent data due to the artifacts was already shallow, remaining only slight trace. The generated images are cleaner with more natural texture and more contently details than the images generated by previous checkpoints.

2.4. Image Classification Model

This section describes our model and training pipeline. We conduct experiments in a controlled order: (i) train image-only models with two cropping strategies, SS and RR; (ii) compare validation WACC and select the more stable SS setting; (iii) add a metadata branch on top of SS (SS-Meta); and (iv) further include 5,000 GAN-generated images into the training set only (SS-Meta v2) to test whether WACC improves.

2.4.1. Image Preprocessing

To reduce the impact of black borders/irrelevant background, scale variations, and illumination- or device-induced color shifts in dermoscopy images, we perform an offline preprocessing step on the ISIC 2019 official images before training. The pipeline follows "invalid-region cropping $\rightarrow$ geometric standardization $\rightarrow$ color-constancy correction", and the processed outputs are written back to the original directory (overwriting the original images), ensuring a consistent input distribution for SS / RR / SS-Meta / SS-Meta v2.

Black-border Cropping:
We first suppress and crop common black borders and non-informative boundary regions to avoid the model learning background cues. The procedure can be summarized into three steps: (1) a foreground mask is constructed via grayscale conversion, Otsu thresholding, and light Gaussian smoothing; (2) the main region is localized by connected-component analysis, using ellipse fitting to estimate the extent and falling back to a bounding-box approximation when needed; and (3) a conservative reliability check compares inside vs. outside brightness, discarding the crop if the outside-to-inside ratio exceeds **0.3**. We also apply a small inward margin of **0.1** to avoid overly tight crops.

Geometric Standardization via Resizing:
After cropping, we normalize image scale to reduce distribution shifts caused by different original resolutions and acquisition distances. If the image is portrait (`H > W`), we transpose it to maintain a consistent orientation. By default, we preserve the aspect ratio and resize images to a target width of **600** using cubic interpolation; optionally, a fixed resolution of **(450, 600)** can be used in the non-preserving mode.

Illumination Normalization via Color Constancy:
Finally, we apply color-constancy correction to mitigate device/white-balance/lighting-induced color shifts. Saturated pixels are excluded using a small 3×3 dilation mask. The illuminant is estimated using a Minkowski norm with $p = 6$ (without additional smoothing/derivatives in our setting), and RGB channels are rescaled accordingly. The output is clipped to $[0, 255]$ and saved as `uint8`.

2.4.2. Fixed Splits and Metadata Alignment (PKL)

To ensure consistency in 5-fold cross-validation and multimodal training, we generate three key PKL files before training:

(1) `indices_isic2019.pkl`: stores the 5-fold split indices (train/val). All experiments share the same split to ensure fair comparison and reproducibility.

(2) `isic2019_meta.pkl`: organizes per-image metadata (e.g., age, sex, anatomical site) into a directly retrievable feature table. During training, metadata are fetched by image identifier and aligned one-to-one with image samples.

(3) `indices_isic2019_gan_trainonly.pkl`: used only for experiments with GAN samples. In the 5-fold splits, the validation set is forced to contain no GAN samples, while the training set can include all GAN samples, to avoid evaluation bias caused by synthetic data shifting the validation distribution.

2.4.3. CNN Model and Training Configuration

We use EfficientNet-B0 as a unified CNN backbone [10]. The input is a 224×224 RGB image, and the model outputs probabilities over 9 classes. During evaluation, we can optionally exclude the 9th class (*unknown*) from metric computation to analyze its impact on overall performance.

CNN Architectures:

- Training epochs: 15; batch size: 20; optimizer: Adam. [6]
- Initial learning rate (image branch): 1.5×10^{-5}; learning-rate decay: if the validation metric does not improve for 5 consecutive epochs, the learning rate is reduced by a factor of $1/5$; when validation is disabled, decay starts from the 10th evaluation point.

- Class imbalance: class weights are computed from label frequencies in the training set and applied to the cross-entropy loss.
- At the end of each epoch, we evaluate on the validation set and record WACC; the checkpoint with the best validation WACC is saved as the final model for that fold.

Data Loading:

- For each fold, we construct the training or validation datasets based on the fixed split.
- Training data are shuffled and the last incomplete batch is dropped; validation data are not shuffled.

Data Augmentation: During training, we apply strong augmentations to improve generalization, including color jittering, AutoAugment, random flips, rotations ($\pm 180°$), scale jittering (0.8–1.2), affine shear (shear=10), and Cutout (size=16) [2, 4]. The augmentation setting is kept identical across all experiments.

2.4.4. CNN Input Strategy

We adopt different input strategies for training. To avoid comparison bias caused by different numbers of test-time views, we align both strategies to 25 views at inference and use the same fusion rule: the predicted probabilities over the 25 views are averaged.

Same-sized cropping strategy (SS):

- We use a fixed 224×224 window and sample crops on a 5×5 regular grid, yielding 25 views that cover different spatial regions. This strategy emphasizes spatial coverage and reduces sensitivity to lesion location shifts.

Random-resize strategy (RR):

- We crop regions at multiple predefined scales and select several deterministic positions per scale to form 25 views. Each crop is then resized back to 224×224 before being fed into the network. This strategy emphasizes scale robustness and improves stability under lesion size variation and minor localization errors.

2.4.5. SS-Meta: Multimodal Feature Fusion

Based on the SS image-only model, we introduce a lightweight metadata branch to construct SS-Meta. Specifically, we concatenate age (1D), sex (2D one-hot), and anatomical site (8D one-hot) into an 11D vector and apply feature standardization. The metadata branch consists of two fully connected layers (256 units each) with BN/ReLU and dropout of 0.25 to mitigate overfitting. Its output is concatenated with the CNN feature after global average pooling, and the fused representation is fed into a 1024-d fusion FC layer and a final classifier to produce 9-dimensional logits. Note that the implementation outputs 9 logits (including an extra reserved/placeholder class), but evaluation metrics are computed only on the 8 ISIC 2019 diagnosis classes (i.e., excluding c9). To reduce spurious correlations between missing metadata patterns and specific classes, we apply metadata dropout augmentation: each metadata field is randomly set to missing with probability 0.12. We initialize from a trained SS model and fine-tune end-to-end (CNN not frozen), keeping other training settings unchanged for fair comparison; the learning rate is set to 1.5×10^{-5} for the CNN branch and 1.5×10^{-4} for the metadata branch. We train for 15 epochs with a batch size of 20.

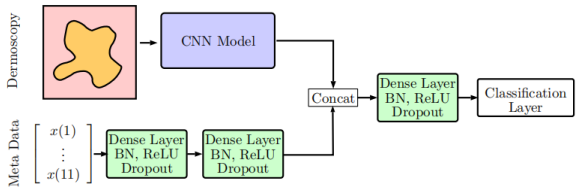

Figure 4. SS-Meta multimodal fusion architecture.

2.4.6. cGAN-Augmented Training

We observed a pronounced class imbalance across the nine categories in ISIC 2019, which may bias the model toward majority classes and affect imbalance-aware metrics such as WACC. To test whether performance improves under a more class-balanced training condition, we add 5,000 cGAN-generated minority-class images to the *training set only* to supplement minority classes (e.g., *SCC* and *VASC*); **these cGAN images are preprocessed using the same pipeline as the ISIC 2019 official images to ensure a consistent input distribution**. We also use a dedicated 5-fold split that strictly keeps the validation set real-only (no GAN samples), thereby avoiding evaluation bias. **For each fold, class weights are recomputed from the effective training set of that fold** (including the injected cGAN samples), matching the actual training distribution in that run.

3. Experimental Results

3.1. Evaluation Protocol & Metrics

To ensure a fair comparison across configurations, we evaluate each fold in a 5-fold cross-validation setting using a unified evaluation script. For fold k, we load the checkpoint that achieves the best validation WACC during training and compute metrics on that fold's validation set only. At inference, we apply the same multi-view (multi-crop) evaluation strategy as in the methodology, fuse the predicted probabilities across views by averaging, and report overall metrics together with per-class metrics and the confusion matrix.

We mainly report WACC, F1, and Mean AUC:

- **WACC (Weighted Accuracy):**

$$\mathrm{WACC} = \sum_{c=1}^{C} w_c \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \quad w_c = \frac{N_c}{\sum_{j=1}^{C} N_j}.$$

- **F1 (Macro-F1):**

$$\mathrm{MacroF1} = \frac{1}{C}\sum_{c=1}^{C}\frac{2\,\mathrm{TP}_c}{2\,\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}.$$

- **Mean AUC:**

$$\mathrm{MeanAUC} = \frac{1}{C}\sum_{c=1}^{C}\mathrm{AUC}_c,$$

where AUC_c is computed one-vs-rest from predicted probabilities for class c.

3.2. Results Table

Table 2 summarizes, for each configuration, the best metrics from the fold that achieves the highest WACC in 5-fold cross-validation (i.e., the best checkpoint of that fold).

Config	WACC	Macro-F1	Mean AUC
SS	0.6050	0.6414	0.9067
RR	0.5959	0.6428	0.9081
SS + Meta	0.6810	0.7007	0.9326
SS + Meta + cGAN	0.6416	0.6585	0.9177

Table 2. Best-fold results across 5-fold cross-validation.

4. Conclusion

Our results lead to three key conclusions: (1) SS achieves a slightly higher WACC than RR (0.6050 vs. 0.5959), suggesting that fixed-grid crops are more robust to lesion mis-centering under a fixed 25-view inference budget; (2) adding metadata yields a clear gain (SS: 0.6050 $\rightarrow$ SS+Meta: 0.6810), supporting that clinical priors (age/sex/site) complement image features; (3) adding 5,000 cGAN-generated minority-class samples reduces WACC (0.6810 $\rightarrow$ 0.6416), indicating that GAN augmentation, as used here, does not improve real-image generalization. Overall, H1 and H2 are supported, while H3 is not.

5. Discussion

Metadata fusion provides a stable improvement (0.6050 $\rightarrow$ 0.6810), likely because age/sex/site offer complementary priors under class imbalance and visually ambiguous classes. In contrast, adding cGAN samples lowers WACC (0.6810 $\rightarrow$ 0.6416), potentially due to: (i) synthetic–real distribution mismatch (color/texture/edge statistics), causing the model to learn synthetic cues; (ii) imperfect clinical semantics or label noise in synthetic images, weakening supervision; and (iii) changing the effective training distribution when injecting many synthetic samples, shifting the decision boundary away from real validation data.

Limitations

- **cGAN quality sensitivity:** artifacts or mismatch can hurt generalization to real images.
- **Metadata generalization risk:** missingness/bias may induce spurious correlations across populations.
- **Unstable summarization:** reporting representative numbers may not reflect fold-to-fold variance.

Planned experiments

1. **Ensemble:** combine models across folds or initializations (or SS/RR) to reduce variance and test stability.
2. **Synthetic quality control:** filter cGAN samples (e.g., discriminator scores or artifact detection) and re-train to verify whether quality drives the drop.

References

[1] Bill Cassidy, Conor Kendrick, Alexandra Brodzicki, Joanna Jaworek-Korjakowska, and Moi Hoon Yap. Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. *Medical Image Analysis*, 75:102305, 2022.

[2] Ekin D. Cubuk, Barret Zoph, Deliang Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] A. Das, V. Agarwal, and N. P. Shetty. Comparative analysis of multimodal architectures for effective skin lesion detection using clinical and image data. *Frontiers in Artificial Intelligence*, 8:1608837, 2025.

[4] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017.

[5] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification, 2018.

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2015.

[7] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

[8] Q. Sun, C. Huang, M. Chen, H. Xu, and Y. Yang. Skin lesion classification using additional patient information. *BioMed Research International*, 2021:6673852, 2021.

[9] S. Sundaram and N. Hulkund. Gan-based data augmentation for chest x-ray classification, 2021.

[10] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019.

[11] Abdul Waheed, Manish Goyal, Deepak Gupta, Ashish Khanna, Fadi Al-Turjman, and Pedro R. Pinheiro. Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. *IEEE Access*, 8:91916–91923, 2020.