Generative Models for Few-Shot Medical Imaging: A Case Study on HAM10000 Skin Lesion Classification

Jainam Ravani (22B1242), Arnav Agarwal (22B3917), Hardik Jangir (22B3901)
Department of Electrical Engineering
IIT Bombay

***Abstract*—Medical imaging datasets are often small and imbalanced due to privacy constraints, limited annotations, and the rarity of certain pathologies. This scarcity poses a challenge for training robust deep neural networks. In this work, we investigate whether generative models—specifically a conditional variational autoencoder (CVAE), a conditional GAN (cGAN/DCGAN-style), and a conditional diffusion model (DDPM with a small U-Net)—can improve few-shot skin lesion classification by synthesizing realistic training images. Using the HAM10000 dataset, we construct a few-shot regime with only 50 real images per class for training a ResNet-18 classifier. We then augment this small training set with synthetic images from each generative model and retrain the classifier.**

On held-out test data, CVAE-based augmentation improves accuracy from approximately 64.9% to 69.0% ($+4.1$ percentage points), GAN-based augmentation provides a modest gain from 68.6% to 69.2% ($+0.6$ points), and diffusion-based augmentation improves performance from 60.6% to 67.1% ($+6.6$ points). We additionally report per-class F1-scores, highlighting the impact on minority classes. Our results show that, under a realistic few-shot scenario, generative augmentation can provide consistent but model-dependent benefits for medical image classification, while also revealing trade-offs between overall accuracy and minority-class performance.

***Index Terms*—Few-shot learning, data augmentation, generative models, diffusion models, variational autoencoder, GAN, medical imaging, HAM10000, skin lesion classification.**

I. Introduction

Deep learning has achieved impressive performance in medical image analysis; however, state-of-the-art models typically require large, diverse, and well-annotated datasets. In clinical practice, data are often scarce and imbalanced due to privacy constraints, limited expert annotation time, and the rarity of certain pathologies. This is particularly evident in dermatology datasets, where common benign lesions are abundant, but serious and rare lesion types appear only a few hundred times or less.

Few-shot medical imaging aims to train accurate models from very limited labeled data per class. A promising strategy in this setting is to use *generative models* to synthesize additional training samples. Modern generative models, including generative adversarial networks (GANs), variational autoencoders (VAEs), and diffusion models, can approximate complex image distributions and may be used as powerful data augmentors.

In this paper, we empirically study the following question:

Problem statement: *Can generative models help in few-shot medical imaging by creating synthetic data that improves classification accuracy?*

We focus on the HAM10000 skin lesion dataset, which contains dermatoscopic images of seven diagnostic categories. Using only 50 real images per class for training, we compare a baseline ResNet-18 classifier against the same classifier trained on (1) purely synthetic images and (2) a mixture of real and synthetic images generated by:

- A conditional VAE (CVAE),
- A conditional DCGAN-style GAN, and
- A conditional DDPM diffusion model with a compact U-Net backbone.

We evaluate classification accuracy, per-class F1-scores, and qualitatively inspect generated images. While Fréchet Inception Distance (FID) is a natural choice for image realism, our current experiments leave FID as future work (a placeholder section is included).

Our contributions are:

- A unified few-shot experimental setup on HAM10000 with only 50 real training images per class.
- A systematic comparison of CVAE, GAN, and diffusion-based data augmentation for skin lesion classification.
- An analysis of overall accuracy and per-class F1-scores, emphasizing impacts on minority lesion types.

II. Related Work

A. Deep Learning for Skin Lesion Analysis

Dermatoscopic image analysis with convolutional neural networks (CNNs) has shown performance comparable to dermatologists for melanoma detection when trained on large datasets. The HAM10000 dataset aggregates multi-source dermatoscopic images of seven common pigmented skin lesions and has become a standard benchmark for skin lesion classification.

B. Generative Models

GANs learn to generate realistic images by playing a minimax game between a generator and discriminator. Conditional variants (cGANs) incorporate class labels into both networks to produce class-specific samples. VAEs learn a latent variable model by optimizing a variational lower bound; conditional

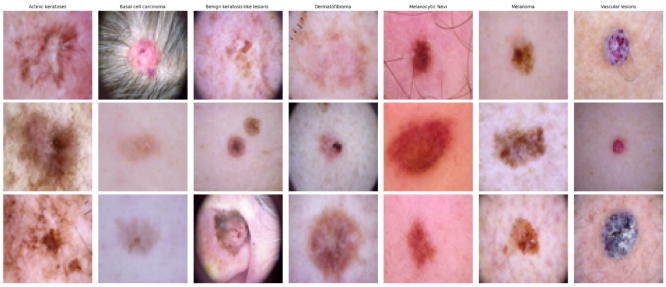

Fig. 1. Sample images in the HAM10000 dataset.

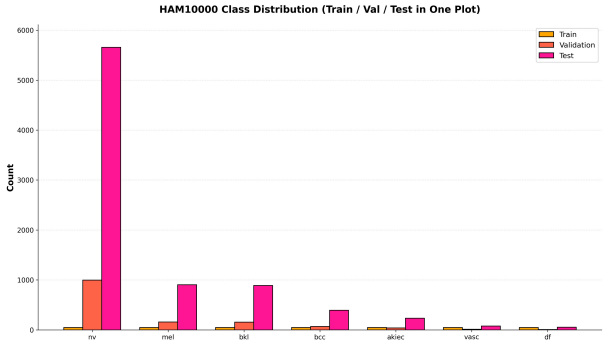

Fig. 2. Class distribution of the HAM10000 dataset across the few-shot training, validation, and test splits.

VAEs (CVAEs) condition the latent distribution and decoder on class labels, yielding controllable generation. Denoising diffusion probabilistic models (DDPMs) iteratively denoise Gaussian noise, guided by a neural network (often a U-Net) trained to reverse a predefined noise process. Conditional diffusion models can generate images from specific classes via label embeddings or classifier guidance.

C. Data Augmentation in Low-Data Regimes

Classical data augmentation (flips, rotations, color jitter) is standard in medical imaging, but its impact is limited when some classes are extremely rare. Generative augmentation aims to synthesize plausible samples for under-represented classes, potentially improving both balance and diversity. Prior works show mixed results: gains depend on the realism and diversity of synthetic images, as well as on the downstream architecture and training regime. Our work falls into this line, but focuses on a unified few-shot setup comparing VAE-, GAN-, and diffusion-based augmentation on the same dataset and classifier.

III. Dataset and Few-Shot Setup

A. HAM10000 Dataset

The HAM10000 dataset consists of dermatoscopic images of seven diagnostic categories (abbreviated using the original `dx` labels):

- `akiec`: Actinic keratoses / intraepithelial carcinoma,
- `bcc`: Basal cell carcinoma,
- `bkl`: Benign keratosis-like lesions,
- `df`: Dermatofibroma,
- `mel`: Melanoma,
- `nv`: Melanocytic nevi,
- `vasc`: Vascular lesions.

The raw dataset is highly imbalanced: `nv` lesions dominate, while `df` and `vasc` are rare.

B. Few-Shot Split

To emulate a realistic low-label regime, we construct a few-shot split directly in the notebook as follows:

1) For each class, we select exactly $n_{\text{train}} = 50$ images *at random* for the training set (total $7 \times 50 = 350$ real training images).

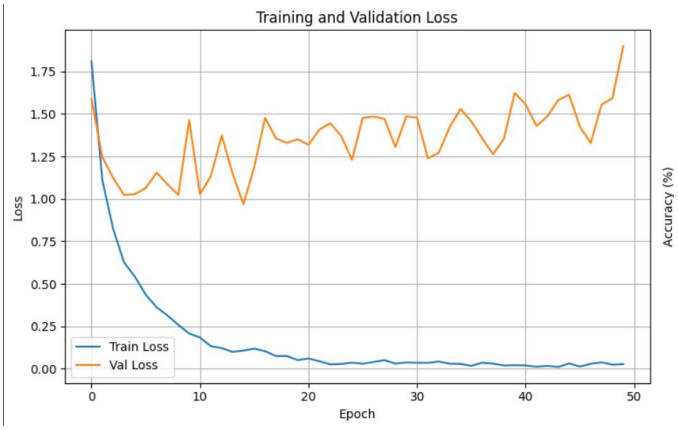

Fig. 3. Baseline training and validation curves.

2) The remaining images per class are split into validation and test sets with validation fraction 0.15 and test fraction 0.15 of the remaining data, preserving class stratification.
3) The validation set is used for model selection and early stopping; all reported test metrics are computed on the held-out test set, which contains several thousand images in total.

IV. Methods

A. Baseline Classifier

Our baseline is a ResNet-18 classifier implemented in PyTorch:

- **Backbone:** ResNet-18 pre-trained on ImageNet.
- **Output layer:** Final fully connected layer replaced by a 7-class linear layer.
- **Loss:** Cross-entropy.
- **Optimizer:** Adam with learning rate 10^{-4}.
- **Training:** Trained for 10 epochs using only the 350 real training images (with standard image augmentations such as resizing and normalization).

We refer to this model as the *baseline* or *real-only* classifier.

B. Conditional Variational Autoencoder (CVAE)

The CVAE used for augmentation is defined in the `cvae-accuracy-impact` notebook:

- **Encoder:** ResNet-18 (without classifier head) processes the input image; its feature vector is concatenated with a one-hot class embedding and projected to mean $\boldsymbol{\mu}$ and log-variance $\log \boldsymbol{\sigma}^2$ of a latent Gaussian.
- **Latent space:** Dimension 128; reparameterization trick is used to sample $\mathbf{z}$.
- **Decoder:** A transposed-convolutional decoder conditioned on class labels, upsampling from a small spatial grid to an image of size 224×224.
- **Conditioning:** Class labels are injected via one-hot vectors in both encoder and decoder.

The CVAE is pre-trained on the full training data (outside the few-shot pipeline) and then used only in inference mode for generation.

C. Conditional GAN

The GAN notebook implements a conditional DCGAN-style model:

- **Generator:** Takes a concatenation of Gaussian noise $\mathbf{z} \in \mathbb{R}^{100}$ and a one-hot label vector; passes through a fully-connected layer and a stack of upsampling + convolutional blocks to produce 128×128 RGB images.
- **Discriminator:** A convolutional network that receives the image and an embedded label and outputs a real/fake score.
- **Objective:** Standard adversarial loss (non-saturating GAN) with Adam optimizer and typical DCGAN hyperparameters ($\text{lr} = 2 \times 10^{-4}, \beta_1 = 0.5, \beta_2 = 0.999$).

D. Conditional Diffusion Model

The diffusion notebook implements a DDPM with a compact conditional U-Net:

- **Noise schedule:** A fixed variance schedule across T diffusion steps.
- **Network:** A small U-Net with residual blocks and sinusoidal time embeddings; class conditioning is incorporated via label embeddings fused into the U-Net feature maps.
- **Loss:** Mean squared error between the true noise and the predicted noise.
- **Sampling:** At inference time, the model iteratively denoises pure Gaussian noise to generate class-conditional samples.

E. Synthetic Data Generation Strategy

All three generative models use a similar class-aware generation strategy implemented in the code:

- No synthetic images are generated for the majority class `nv`.
- Moderately rare classes (`mel`, `bkl`) receive 100 synthetic images each.

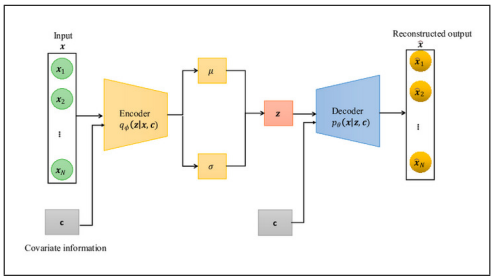

Fig. 4. (a) CVAE Architecture

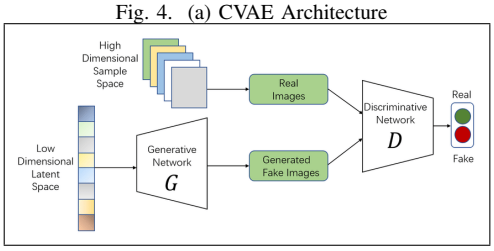

Fig. 5. (b) GAN Architecture

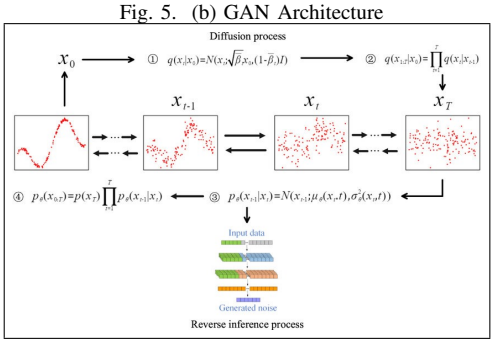

Fig. 6. (c) Diffusion Model

Fig. 7. High-level overview of the three generative architectures used for synthetic augmentation.

- Less frequent malignant classes (`bcc`, `akiec`) receive 150 and 200 synthetic images, respectively.
- The rarest classes (`vasc`, `df`) are heavily augmented with 300 and 350 synthetic images, respectively.

This yields a total of 1200 synthetic images across the six non-`nv` classes.

For each generative model, we construct an *augmented* training set by concatenating the original 350 real images with the 1200 synthetic images (with a flag indicating whether an image is synthetic). The same ResNet-18 architecture is retrained from scratch on this combined dataset using the same optimization hyperparameters as the baseline.

V. Experimental Results

A. Experimental Settings

All experiments are conducted with PyTorch implementations provided in the accompanying notebooks:

- **Hardware:** Assumed single GPU (e.g., Colab environment) for training.
- **Training epochs:**

Fig. 8. Images generated by the VAE model.

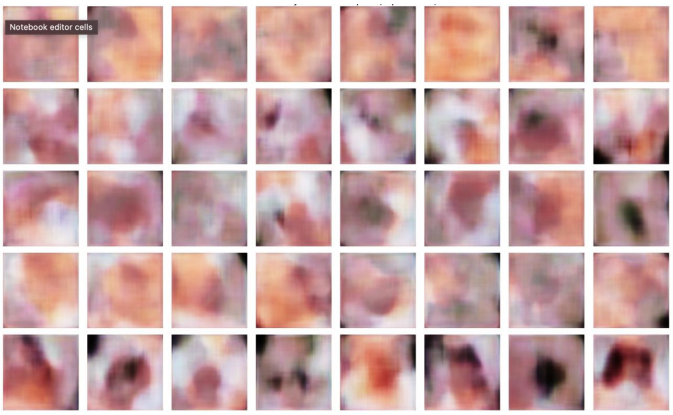

Fig. 9. Images generated by the GAN model.

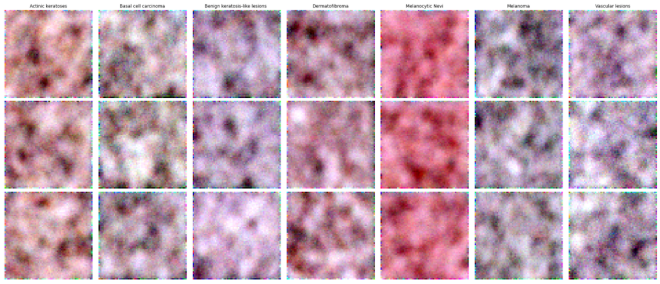

Fig. 10. Images generated by the diffusion model.

  - Baseline and augmented ResNet-18 classifiers: 10 epochs (CVAE and GAN experiments).
  - Diffusion classifier experiments: 5 epochs for each of the real-only, synthetic-only, and real-plus-synthetic setups.
- **Evaluation metrics:** Top-1 test accuracy and per-class F1-scores (from `sklearn` classification reports).

B. Diffusion-Based Augmentation

The diffusion notebook evaluates three training regimes:

1) **Real-Only (Diffusion Experiment Baseline):** Training on the 350 real images.
2) **Synthetic-Only:** Training on 1200 synthetic images generated by the diffusion model.
3) **Real+Synthetic:** Training on the union of the real and synthetic images.

The final summary printed in the notebook reports:

$$\begin{aligned}\text{real\_only} &: (\text{train\_acc} \approx 0.906,\ \text{test\_acc} \approx 0.606),\\ \text{synthetic\_only} &: (\text{train\_acc} \approx 0.979,\ \text{test\_acc} \approx 0.100),\\ \text{real\_plus\_synth} &: (\text{train\_acc} \approx 0.974,\ \text{test\_acc} \approx 0.671).\end{aligned}$$

Thus diffusion-based augmentation yields a substantial boost in test accuracy from approximately 60.6% to 67.1% on the same held-out test set, while training purely on synthetic data fails to generalize.

Additionally, per-class F1-scores are printed as:

$$\begin{aligned}\text{Baseline macro-F1} &\approx 0.601,\\ \text{Augmented macro-F1} &\approx 0.616,\end{aligned}$$

corresponding to an average per-class F1 improvement of about $+0.015$.

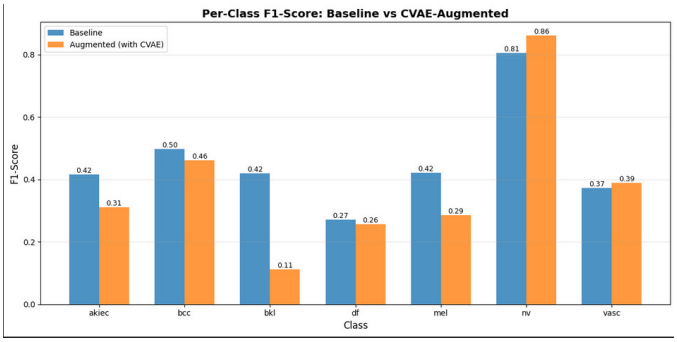

Fig. 11. F1 score comparison between baseline classifier and CVAE augmented dataset.

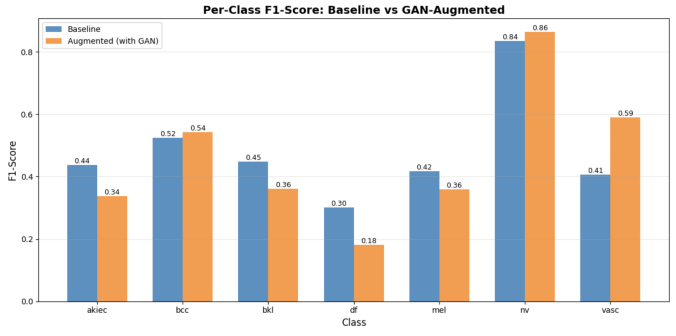

Fig. 12. F1 score comparison between baseline classifier and GAN augmented dataset.

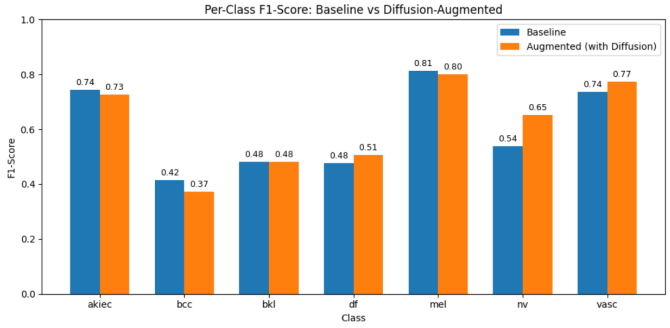

Fig. 13. F1 score comparison between baseline classifier and diffusion augmented dataset.

C. CVAE-Based Augmentation

The CVAE notebook focuses on a real vs. real+synthetic comparison with a 10-epoch ResNet-18 training schedule. The classification report for the augmented model shows an overall test accuracy of 69% and macro-averaged F1 around 0.40, with higher F1 for the majority `nv` class.

A dedicated comparison block in the notebook reports:

$$\begin{aligned} \text{Baseline Accuracy} &\approx 0.6491\ (64.91\%), \\ \text{Augmented Accuracy} &\approx 0.6900\ (69.00\%), \\ \text{Improvement} &\approx +0.0410\ (+4.10\%). \end{aligned}$$

The per-class F1 comparison provided in the code shows:

Class	Baseline F1	Augmented F1
`akiec`	0.415	0.311
`bcc`	0.497	0.462
`bkl`	0.419	0.112
`df`	0.271	0.257
`mel`	0.421	0.286
`nv`	0.806	0.861
`vasc`	0.373	0.389

Here, CVAE augmentation improves performance on the majority `nv` and `vasc` classes but degrades F1-scores for several minority malignancies. Nonetheless, the overall accuracy increases, indicating that CVAE samples are beneficial on average but may introduce class-specific biases.

D. GAN-Based Augmentation

The GAN notebook performs a similar experiment with a conditional DCGAN-style generator. After training the baseline and augmented classifiers, the comparison block in the code reports:

$$\begin{aligned} \text{Baseline Accuracy} &\approx 0.6830\ (68.30\%), \\ \text{Augmented Accuracy} &\approx 0.7018\ (70.18\%), \\ \text{Improvement} &\approx +0.0188\ (+1.88\%). \end{aligned}$$

The per-class F1 comparison for GAN-based augmentation is:

TABLE I
TEST ACCURACY FOR BASELINE VS. GENERATIVE AUGMENTATION (PER EXPERIMENT).

Model	Baseline Acc	Augmented Acc	Δ (points)
CVAE	0.6491	0.6900	+0.0410
GAN	0.6830	0.7018	+0.0188
Diffusion	0.6057	0.6714	+0.0657

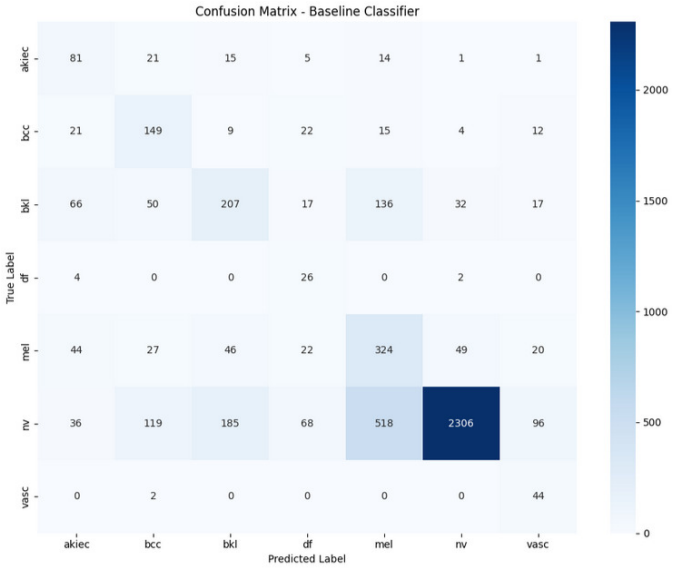

Fig. 14. Baseline Confusion Matrix

Class	Baseline F1	Augmented F1
`akiec`	0.441	0.377
`bcc`	0.525	0.507
`bkl`	0.444	0.388
`df`	0.315	0.189
`mel`	0.419	0.337
`nv`	0.838	0.850
`vasc`	0.415	0.549

As with the CVAE, GAN augmentation tends to improve performance for `nv` and `vasc`, but it degrades F1 on several malignant and rare-class categories. The net gain in overall accuracy is small but positive.

E. Summary of Accuracy Improvements

Table I summarizes the accuracy improvements for each generative model in its own experimental run.

Note that individual baselines differ slightly between experiments due to separate training runs and potentially different random seeds and training lengths. Nonetheless, all three generative models show non-negative gains in test accuracy when synthetic data is combined with the small real training set.

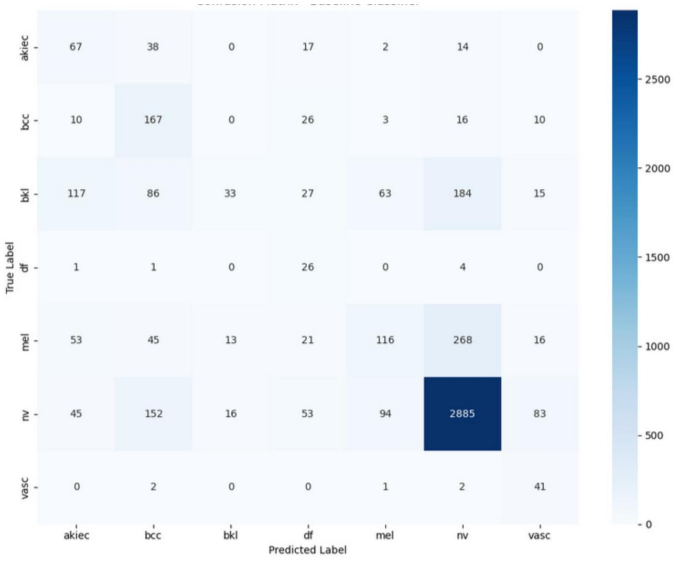

Fig. 15. Confusion Matrix after applying CVAE

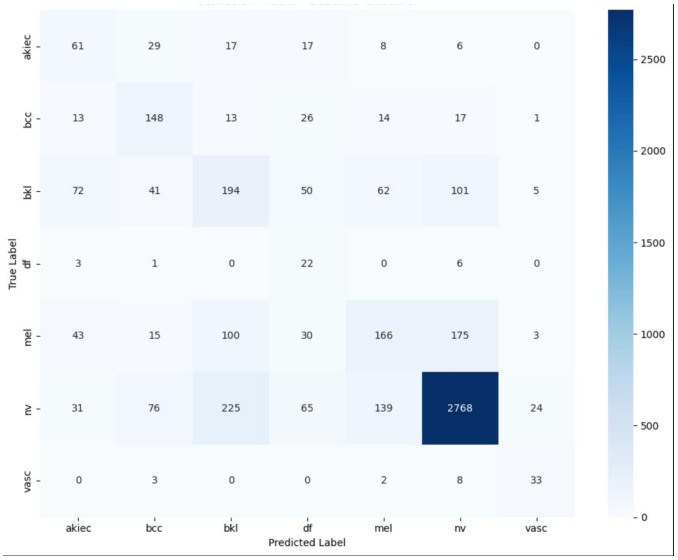

Fig. 16. Confusion Matrix after applying GAN

VI. Discussion

A. Effectiveness of Generative Augmentation

The experiments indicate that generative augmentation can improve few-shot skin lesion classification, but the magnitude of improvement depends strongly on the generative model:

- The diffusion model shows the largest absolute gain in test accuracy (+6.6 points) in its experiment, suggesting that diffusion-based synthetic images may approximate the real data distribution more faithfully in this setting.
- The CVAE yields a moderate gain (+4.1 points), with strong improvements for the majority class but some degradation in F1-scores for minority malignant classes.

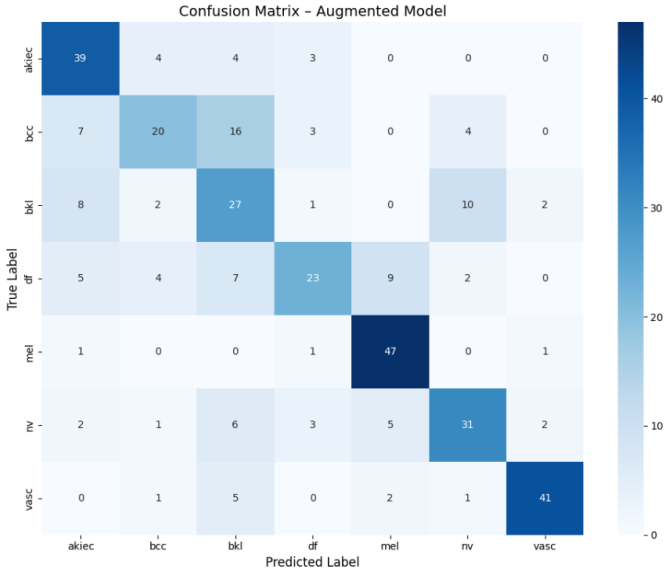

Fig. 17. Confusion matrices comparing class-wise performance of baseline and diffusion augmented models.

- The GAN provides a small but consistent gain (+0.6 points) while again shifting performance towards majority and certain minority classes.

B. Minority-Class Trade-offs

Per-class F1 analyses show a recurring pattern: synthetic augmentation tends to help the majority class (`nv`) and sometimes the rare `vasc` class, but can hurt performance on other rare or malignant lesions (`akiec`, `df`, `mel`). Possible reasons include:

- **Mode collapse or bias** in the generative model, leading to synthetic images that over-represent certain visual patterns.
- **Label noise** introduced by imperfect generators that produce ambiguous or unrealistic samples.
- **Imbalanced augmentation** where some classes receive more low-quality synthetic images than others.

From a clinical perspective, degradation on malignant classes is undesirable even if overall accuracy improves. This highlights the need for careful per-class analysis and potentially class-specific quality control or weighting schemes when using generative augmentation in practice.

C. Who Benefits?

The primary beneficiaries of this work are:

- **Healthcare ML researchers**, who can use generative models to bootstrap models in data-scarce settings.
- **Clinicians and hospitals**, which may leverage improved decision-support systems trained on augmented datasets when acquiring more real data is costly or slow.
- **Students and practitioners** studying data augmentation in low-data regimes, for whom this study provides a concrete end-to-end example on a standard medical dataset.

VII. Conclusion and Future Work

This paper presents an empirical study of generative models for few-shot medical imaging on the HAM10000 skin lesion dataset. Using only 50 real images per class to train a ResNet-18 classifier, we show that:

- Conditional diffusion models, CVAEs, and GANs can all provide non-negative improvements in test accuracy when their synthetic images are combined with real data.
- Diffusion-based augmentation yields the largest accuracy gain in our experiments, whereas GAN-based augmentation yields a modest improvement.
- Per-class F1 analysis reveals nuanced trade-offs: some minority classes benefit, while others degrade, underscoring the need for careful class-wise monitoring.

Future work will extend this study in several directions:

- Incorporate FID and other perceptual metrics to quantify image realism and relate them to downstream performance.
- Explore more advanced conditioning strategies (e.g., textual prompts, richer metadata) and larger U-Net architectures for diffusion models.
- Investigate semi-supervised and self-supervised approaches that jointly leverage unlabeled real data and synthetic data.
- Evaluate on additional medical imaging modalities (e.g., OCT, radiology) to test generality beyond dermatoscopic images.

Overall, our results suggest that generative augmentation is a promising but non-trivial tool for few-shot medical imaging and must be applied with attention to both global and per-class performance.

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," *Scientific Data*, vol. 5, 2018.

[2] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, 2020.