Synthetic Data Generation for Long-Tail Medical Image Classification: A Case Study in Skin Lesions

Jiaxiang Jiang
Intel
jiaxiang.jiang@intel.com

Mahesh Subedar *
Google
msubedar@gmail.com

Omesh Tickoo
Intel
omesh.tickoo@intel.com

Abstract

Long-tailed class distributions are pervasive in multi-class medical datasets and pose significant challenges for deep learning models which typically underperform on tail classes with limited samples. This limitation is particularly problematic in medical applications, where rare classes often correspond to severe or high-risk diseases and therefore require high diagnostic accuracy. Existing solutions—including specialized architectures, rebalanced loss functions, and handcrafted data augmentation—offer only marginal improvements and struggle to scale due to their limited and largely deterministic variability. To address these challenges, we introduce a diffusion-model-driven synthetic data augmentation pipeline tailored for medical long-tailed classification. Our approach features a novel inpainting diffusion model combined with an Out-of-Distribution (OOD) post-selection mechanism to ensure diverse, realistic, and clinically meaningful synthetic samples. Evaluated on the ISIC2019 skin lesion classification dataset, one of the largest and most imbalanced medical imaging benchmarks, our method yields substantial improvements in overall performance, with particularly pronounced gains on tail classes with more than 28% *improvement on the class with the fewest samples. These results demonstrate the effectiveness of diffusion-based augmentation in mitigating long-tail imbalance and enhancing medical classification robustness.*

1. Introduction

Despite the impressive progress brought by modern deep neural network (DNN) architectures [1–4], image classification performance in conventional benchmarks is approaching saturation. However, these gains rely heavily on large, balanced datasets, a condition rarely met in real-world medical imaging. Medical images like dermoscopy inherently exhibit long-tailed label distributions, where a few dominant "head" classes contain most samples, while numerous clinically important "tail" classes are severely under-represented. Models trained on such skewed data often achieve competitive overall accuracy yet fail to generalize to tail classes, which is particularly concerning because these classes may correspond to rare but high-risk diseases. Consequently, improving recognition performance under long-tailed distributions has become a critical challenge, and long-tailed medical image classification (LTMIC) has emerged as an important research direction within the computer vision community.

Existing solutions for LTMIC can be grouped into three categories [5, 6]: loss function design, model improvement, and class re-balancing. Cross-entropy loss is the most widely used loss function in deep learning model training. However, in LTMIC setting, this loss ignores the class imbalance in number of samples for each class. Therefore, variations of training loss functions are designed to capture the intrinsic imbalance in the training set [7–12]. Instead of adjusting training loss function directly, [13] proposes to adjust the prediction logits of a neural network model, which uses an calibration function to adjust the logits. The calibration function is learnt by matching the calibrated prediction to some pre-defined more balanced class distribution.

Model improvement methods design specific model architectures to solve LTMIC. Most LTMIC model can be divided into two stages, feature representation leaning and classifier. Therefore, some methods aim to design better learning representation model [14, 15] and another methods are mainly to design suitable classifier [16–20]. [14] learns discriminative features by using the overall distance among all samples pairs within one training batch and [15] enlarges the inter-class distance and minimize the largest distance between intra-class samples in the feature space. DataFuse[20] shows introducing external data can result in a better feature representation. [18] proposes to design the classifier with two branches, one is a conventional learning branch and another is a re-balancing branch, to target the imbalance of the training set. [17] explores dis-

*Work done while at Intel.

tillation technique to design the long-tail classifier. [18] defines a loss sharpness in addition to loss value itself as the objective function to mitigate the imbalance problem. MRE[19], MME[21], MME[22] ensemble individual classifiers in a multi-expert framework. Empirically, these ensembling methods achieve the state-of-the-art performance mainly by reducing the model uncertainties to obtain robust predictions. However, all model improvement methods and special loss function design methods share a common a commonality: they acknowledge the data imbalance and focus on the training of models [23]. They both require meticulous design and thus are too sensitive to hyperparameters and have limited generalization capability. For example, one carefully designed loss function works well on one model but fails on another model. Also, different ensembles of models can have quite different results even on the same dataset.

Another group of methods are class-rebalancing methods which aim to re-balance the negative influence of the imbalance of the training set by adjusting data distribution. One of the most studied methods in this group are resampling based methods [24–26] which have been explored to re-balance classes by adjusting the number of samples per class in each sample batch for model training. Instead of simply resampling the training set, recent work [27] develops a class specific data augmentation method to adjust the data set distribution. [27] shows that adding more samples to adjust the distribution can increase average classification performance as well as class-wise performance. However, the augmentation method proposed in [27] is too simple and it does not capture distribution of each class .

Recent development of image generation techniques [28–30] show great potential of generating high quality natural images. Those methods utilize diffusion models [28, 29] to produce visually appealing images both with or without conditions. There is a large amount of following work [31–33] in pushing the limits of generating better photorealistic images with high FID and PSNR scores. Despite its success in generating visually appealing images for digital design, applications of diffusion models in limited data scenarios especially domain specific scenarios remain under-explored.

[34] and [23] explore how to make diffusion models to generate images for the long-tailed dataset. [34] proposes a class-balancing diffusion model (CBDM) to generate diverse and good quality images for each class in the long-tailed datasets. CBDM[34] demonstrate its diverse generation capability especially for tail classes. Based on CBDM [34], [23] show that not all samples are equally important in training diffusion models for long-tailed datasets. [23] show that approximately-in-distribution (AID) samples play more important role than in-distribution (ID) and out-of-distribution (OOD) samples. However, both [34] and [23] show great performance on natural image long tail datasets where the distances between different class distributions are relatively large than medical long tail datasets such as dermatoscope imagery.

We propose a synthetic data generation approach to solve the Long-Tail Medical Image Classification (LTMIC) problem, as illustrated in Fig 1. First, we extract the area of interest from dermatoscopy images using the method described in [35]. Subsequently, we generate a set of synthetic images using our fine-tuned inpainting diffusion model. Specific details regarding the diffusion model architecture and fine-tuning techniques are described in section 2.

A medical Out-of-Distribution (OOD) method [36] is utilized to filter out OOD samples for each class. Following this filtering process, we combine the synthetic samples with the original dataset to train a classifier. We demonstrate that the inclusion of synthetic samples significantly improves classification results, particularly for the tail classes. Since this paper introduces a novel synthetic data generation approach, we utilize only a simple CNN-based classifier [19] in our experiments without further data augmentation or model ensembles.

In summary, the main contributions of this work are as follows:

1. We propose a novel synthetic data generation pipeline incorporating an inpainting diffusion model and OOD filtering to address the LTMIC problem;
2. We introduce a novel inpainting diffusion model architecture specifically designed for medical image generation;
3. We present a novel OOD-based sampling strategy to fine-tune diffusion models on medical long-tail datasets.

2. Approach

2.1. Diffusion Model Background

In the standard diffusion model [28], during training, the forward process is characterized by the following equation with a batch of input images $\mathbf{x}_0$.

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t\mathbf{I})) \tag{1}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$ is calculated through pre-defined variance scheduler $\{\beta_t \in (0,1)\}_{t=1}^{T}$. Next, a diffusion model parameterized by parameter θ to get the posterior probability distribution $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, t)$. Instead of modeling this posterior probablity directly, [30] uses a UNet[37] with parameter θ_d to model ϵ_t which is the noise added to the original image to get $\mathbf{x}_t$. [38] shows that the diffusion model parameter θ_d can be finetuned with few images with the following objective function to get high quality images:

$$L = \mathbb{E}_{t\sim[1,T]}[||\epsilon_t - \epsilon_{\theta_d}(\sqrt{\tilde{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\tilde{\alpha}_t}\epsilon_t, t, y_{emb})||^2] \tag{2}$$

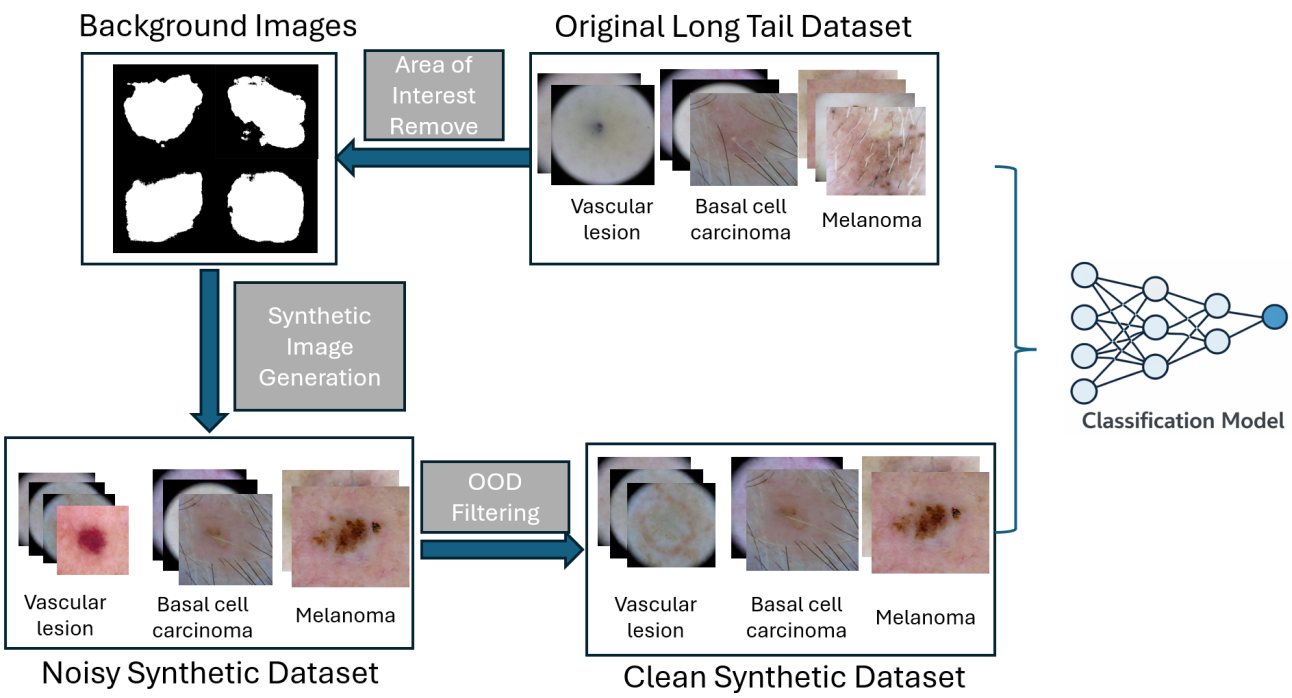

Figure 1. The overal pipeline of our synthetic data generation approach. First, a segmentation method is used to get background images of the original long tail dataset. Next, a finetuned inpaint diffusion model is used to generate noisy synthetic images which can contain OOD image samples. Then an OOD filtering is applied to noisy synthetic dataset to get clean synthetic dataset. Last, the clean synthetic dataset is combined with the original long-tail dataset to train a final classification model.

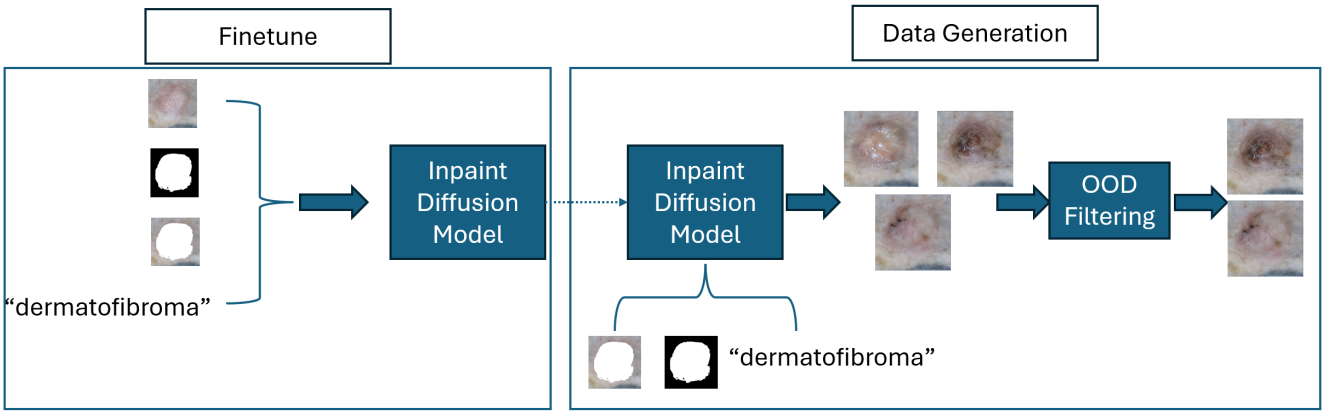

Figure 2. Inpaint diffusion model finetuning and data generation. First, we finetune an inpainting diffusion model using the original long tailed training dataset, their associated binary area of interest masks from a segmentation model, background of original dermoscopy images, and their class labels. Next, the finetuned diffusion model is used to generate synthetic samples. The finetuned model takes input the background images of the original training set images and their associated class labels. Finally, an OOD method is used to filter out OOD samples to make the synthetic dataset cleaner to the classification model.

where ϵ_{θ_d} is the noise estimated by the UNet with parameter θ_d and y_{emb} is the embedding of the image class label y. During reverse process, noise is estimated by the finetuned UNet at each time step t, and new sample is generated by subtracting estimated ϵ_t iteratively from pure noise sample $\mathbf{x}_T \sim \mathcal{N}(0, I)$.

2.2. Method Overview

The overal data generation pipeline is illustrated in Fig 1. Mathematically, for general image classification dataset with N number of data samples and a set of classes C, $\mathcal{D}_{\mathcal{LT}} = \{(x_i, y_i)\}_{i=1}^{N}$ where x_i represents i_{th} image and $y_i \in C$ represents i_{th} image label. In the long-tail setting,

$|c_1| \geq |c_2| \geq \cdots \geq |c_M|$ for each $c_j \in C$, M represents total number of different classes of the dataset, $|c_j|$ represents number of samples in c_j, and M represents total number of classes in C. In this paper, we create a synthetic $\mathcal{D}_{syn}$ which has the same set of classes C as $\mathcal{D}_{\mathcal{LT}}$, For each class c_j^{syn} in $\mathcal{D}_{syn}$, $|c_j^{syn}| = max\{0, |c_1| - |c_j|\}$. Because of stochasticity in diffusion model-based data generation approach, it is not guaranteed that all samples in $\mathcal{D}_{syn}$ are good samples to train the downstream classification model. Thus $\mathcal{D}_{syn}$ is a noisy synthetic dataset and an OOD method is used to filter out noisy samples to get clean synthetic dataset $\mathcal{D}_{syn}^{clean}$. For j_{th} class c_j^{clean} in $\mathcal{D}_{syn}^{clean}$, $|c_j^{clean}| = \gamma \times |c_j^{syn}|$ where $\gamma \in [0, 1]$ is a hyperparameter and it represents the percentage of clean samples in each class in $\mathcal{D}_{syn}$. Last, the augmented dataset $\mathcal{D}_{syn}^{clean} \cup \mathcal{D}_{\mathcal{LT}}$ is used to train the classification model from [19].

Since the main innovations of this paper are focusing on data generation part of the pipeline so this section will mainly describe how to acquire inpaint diffusion model and use it to generate synthetic samples to train the classification model. As illustrated in fig 2, an inpaint diffusion model is finetuned with the original dermoscopy image x_i, binary area of interest mask b_i, background of original dermoscopy image d_i, and their class label y_i. The finetune process makes the inpaint diffusion model learn the distribution of masked out region conditioned on d_i and y_i. After finetuning, the inpaint is used to generate images following the standard reverse process as described in [23, 30, 34].

2.3. Proposed Inpaint Diffusion Model

2.3.1. Inpaint Diffusion Model Architecture

Our inpaint diffusion model is illustrated in Fig 3. During finetuning, the model takes input of dermoscopy image $\mathbf{x}$, binary segmentation image of area of interest $\mathbf{b}$, dermoscopy background image $\mathbf{d}$, and class label $\mathbf{y}$. One of the main challenges in multi-class domain specific image generation for image classification is to generate images distinguishable enough between different classes while also within the real distribution of label $\mathbf{y}$. The general text embedding models are not suitable for highly specialized description of skin diseases. Hence, we train a simple fully connected class embedding network as shown in Fig 3. The class embedding network is parameterized by learnable parameters θ_c. Both original image $\mathbf{x}$ and dermoscopy background image $\mathbf{d}$ are encoded into latent space by a pre-trained Variational Auto Encoder (VAE) and noise ϵ_t is added to get noisy latents $\mathbf{z}_{xt}$ and $\mathbf{z}_{dt}$ respectively at diffusion time step t. The binary mask $\mathbf{b}$ is rescaled as $\mathbf{b}^*$ to match the shape of noisy latents $\mathbf{z}_{xt}$ and $\mathbf{z}_{dt}$. $\mathbf{z}_{xt}$, $\mathbf{z}_{dt}$, and $\mathbf{b}^*$ are concatenated into noisy latent $\mathbf{z} = [\mathbf{z}_{xt}, \mathbf{z}_{dt}, \mathbf{b}^*]$ as the input to the diffusion UNet model parameterized by θ_d. Therefore, the total learnable parameters include $\theta = \theta_d \cup \theta_c$. Thus the funetune objective function becomes:

$$L = \mathbb{E}_{t \sim [1,T]}[||\epsilon_t - \epsilon_\theta(\mathbf{z}, t, y_{emb})||^2] \tag{3}$$

2.3.2. Finetuning with sample selection

In order to make the diffusion model learn the distribution of each class equally well, we apply an oversampling strategy to make sure the number of samples of each class is the same. The idea is to assign a weight w_i^j for $i_t h$ sample in class c_j. Then we repeat samples in class $c_j, j = 2, 3, \ldots, M$ based on w_i^j to make sure the number of finetuning samples for each class is equal to $|c_1|$. Next, we will describe how we compute sample weights w_i^j.

As [23] suggests that different samples can have different importance in training the generative model and approximately-in-distribution (AID) samples, which locate slightly far from the center of the distribution play the most important role. Therefore, we utilize the self-supervised medical anomaly detection method [36] to train for each class $c_j, j = 2, 3, \ldots, M$ using images $\mathbf{x}$ from $\mathcal{D}_{\mathcal{LT}}$. After training, for each class c_j, we compute anomaly score $\mathcal{A}_i^j$ for each image $\mathbf{x}_i^j$ in this class. Then weight w_i^j is computed as

$$w_i^j = \frac{1}{e^{|\bar{\mathcal{A}}^j - \mathcal{A}_i^j|}} \tag{4}$$

where $\bar{A}^j$ is the median value of $\bar{A}_i^j, i = 1, 2, \ldots |c_j|$. $|\bar{\mathcal{A}}_i^j - \mathcal{A}_i^j|$ is the distance measure of how far sample i is from AID samples. The samples that are closer in distance are given higher sampling probability during the finetuning step.

3. Experimental Results

ISIC2019 skin lesion dataset is used to test our method. The dataset consists 25331 images of 8 classes: Melanoma (MEL), Melanocytic Nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis (AKIEC), Benign Keratosis (BKL), Dermatofibroma (DF), Vascular Lesion (VASC), and Squamous Cell Carcinoma (SCC). The number of images for each class vary significantly: MEL images is 4,522; NV is 12,875; BCC is 3,323; AKIEC is 867; BKL is 2,624; DF is 239; VASC is 253; and SCC is 628. Fig.4 shows example images from two tail classes DF and VASC, and synthetic images from the corresponding input image.

The 5 fold cross validation is used for the experiments. We split the dataset into 5 equal folds and each class is split proportionally into 5 equal folds. For each round, we use 4 folds to finetune our generative model, generate synthetic samples, and is then combined with the clean synthetic dataset to train a classification model. The remaining 1 fold is used to get the test results from the trained classification model. We iterative this process five times to get the average performance of the classification models.

To evaluate our approach across all eight classes, we utilize the four metrics defined in the ISIC 2019 challenge:

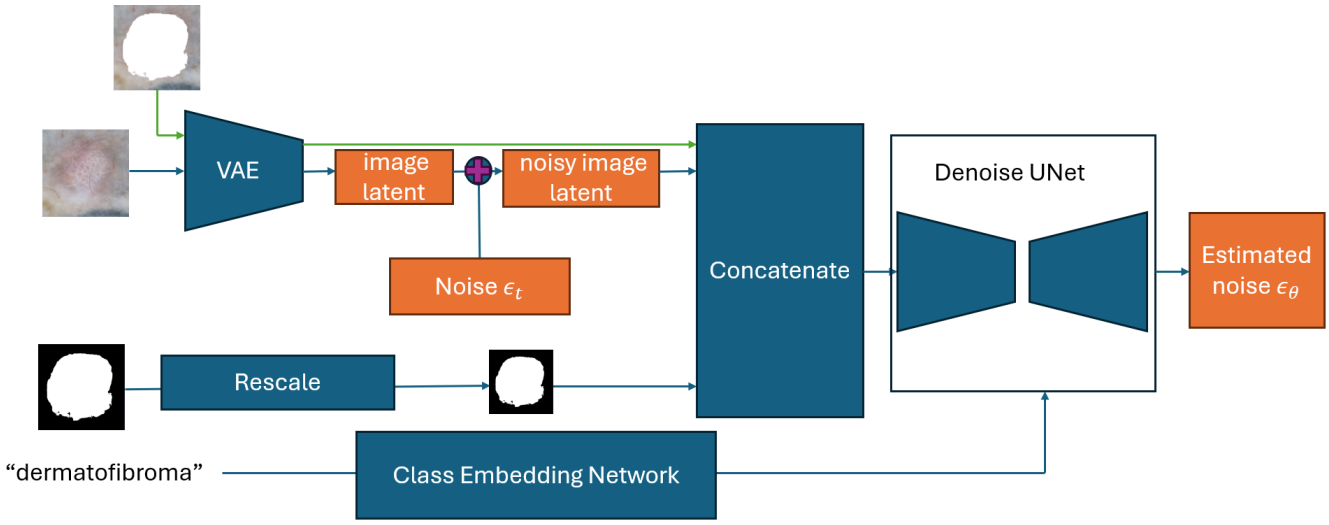

Figure 3. Inpaint diffusion model architecture.

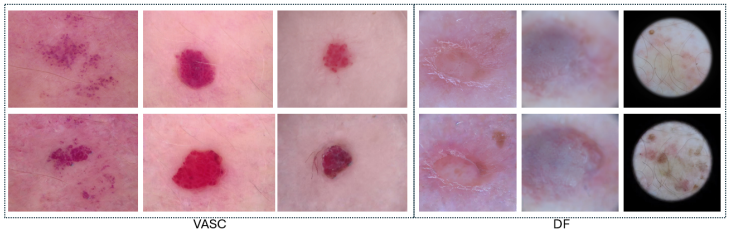

Figure 4. Example images (top row) from the dataset and the corresponding synthetic images (bottom row) from our inpaint diffusion model. Images are selected from two classes with the least number of samples: VASC and DF.

Balanced Multiclass Accuracy (BMA), sensitivity, specificity, and F1 score. In this multiclass context, these metrics are derived from the classification confusion matrix. Specifically, BMA is calculated as the mean recall across all classes (i.e., the average of the diagonal elements divided by the total positive instances per class). All reported metrics represent the average results across 5-fold cross-validation.

Table 1 compares our approach against current SOTA methods. For a fair comparison, we utilize the single classification model from [19] without multi-resolution ensembling (MRE) (MRE Without Ensembling). Our synthetic data generation approach surpasses all multi-model ensembling methods [19, 21, 22] across all four metrics by a significant margin. Notably, our method outperforms [20] by approximately 4% in BMA without requiring any external data. Among diffusion-based synthetic data approaches, our specific diffusion architecture and sampling strategy yield superior performance compared to [34, 38]. Furthermore, the results highlight that OOD filtering consistently improves performance across all evaluation metrics.

We further analyze how synthetic data influences classification performance across individual classes using class-wise accuracy metrics, as shown in Table 2. We can clearly observe that classes with fewer original samples (tail classes) get great benefit from the synthetic data. For example, *DF*—the tail class with the fewest samples—shows an accuracy improvement of over 28%, significantly exceeding the overall average performance gain.

Another key finding is that OOD filtering tends to improve head-class performance more than tail-class performance. This occurs because head classes already possess sufficient clean training data; consequently, synthetic samples without filtering introduce disproportionate noise for these classes. In contrast, tail classes benefit so substantially from the increased sample volume that they can tolerate higher levels of noise within the synthetic data than head classes can.

As mentioned above, OOD filtering plays a role in balancing head and tail classes classification accuracy. Thus, for the ablation study, we investigate how the hyperparameter γ, the percentage of clean samples that we keep in each class affects the final classification performance. As Fig.5 shows, specificity almost does not change with γ. For other three evaluation metrics, as γ increase from 0 to 1 (0 means use original training dataset only and 1 means no OOD filtering for synthetic dataset), the performance first

Table 1. Benchmarking results for our methods against SOTA methods on ISIC2019 dataset. BMA, F1 score, sensitivity, and specificity are used to capture the performance of various methods.

Methods	BMA	F1 score	sensitivity	specificity
MRE[19]	0.757	0.726	0.693	0.957
CNNE[21]	0.753	0.700	0.692	0.944
DataFuse[20]	0.762	0.674	0.690	0.951
MME[22]	0.637	0.596	0.624	0.933
CBDM[34]	0.739	0.706	0.739	0.957
Dreambooth[38]	0.740	0.719	0.740	0.957
MRE Without Ensembling	0.677	0.700	0.677	0.956
Our Approach Without Filtering	0.800	0.760	0.800	0.963
Our Approach	**0.802**	**0.780**	**0.802**	**0.965**

Table 2. Impact of Synthetic Data Augmentation on Per-Class Classification Accuracy.

Skin Lesion Classes	MRE Without Ensembling	Our approach without filtering	Our approach
MEL	0.694	**0.759**	0.758
NV	0.753	0.762	**0.793**
BCC	0.792	0.864	**0.882**
AKIEC	0.604	0.709	**0.724**
BKL	0.527	0.655	**0.668**
DF	0.611	**0.897**	0.880
VASC	0.876	**0.969**	0.965
SCC	0.561	**0.772**	0.745

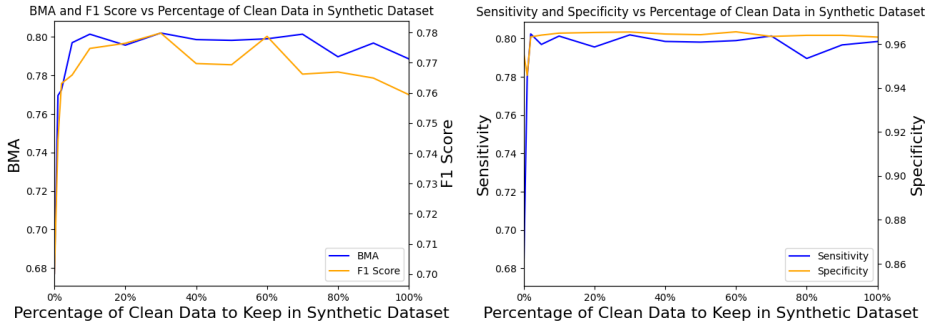

Figure 5. Ablation study on how γ, the percentage of clean samples affect the final classification performance. Specificity is the least sensitive metric to the hyperparameter. For other three evaluation metrics, the performance significantly improved with even a little amount of synthetic data. Then as γ increases, the performance stays flat or decrease.

get improved sharply and then those metrics stay flat or decrease. Based on Fig 5, the best γ value should be between $0.2 \sim 0.6$.

4. Conclusion

In this paper, we proposed a novel synthetic data generation approach to address the challenges of LTMIC, using skin lesion classification as a primary case study. Our method integrates a novel diffusion model with out-of-distribution (OOD) filtering to ensure the generation of robust, in-distribution synthetic samples. Experimental results demonstrate that fine-tuning downstream models with our synthetic data consistently outperforms current SOTA methods.

While our approach is model-agnostic and can be integrated with any architecture, treating data generation and classification as independent processes may be suboptimal. Therefore, our future work will focus on the joint optimization of these two components to further enhance LTMIC performance.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 1

[2] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[3] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[4] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022. 1

[5] L. Ju, S. Yan, Y. Zhou, Y. Nan, X. Xing, P. Duan, and Z. Ge, "Monica: Benchmarking on long-tailed medical image classification," *arXiv preprint arXiv:2410.02010*, 2024. 1

[6] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 9, pp. 10795–10816, 2023. 1

[7] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in neural information processing systems*, vol. 32, 2019. 1

[8] Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, and B. Chang, "Disentangling label distribution for long-tailed visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6626–6636, 2021.

[9] G. R. Kini, O. Paraskevas, S. Oymak, and C. Thrampoulidis, "Label-imbalanced and group-sensitive classification under overparameterization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18970–18983, 2021.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[11] J. Ren, C. Yu, X. Ma, H. Zhao, S. Yi, *et al.*, "Balanced meta-softmax for long-tailed visual recognition," *Advances in neural information processing systems*, vol. 33, pp. 4175–4186, 2020.

[12] J. Tan, C. Wang, B. Li, Q. Li, W. Ouyang, C. Yin, and J. Yan, "Equalization loss for long-tailed object recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11662–11671, 2020. 1

[13] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun, "Distribution alignment: A unified framework for long-tail visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2361–2370, 2021. 1

[14] B. Kang, Y. Li, S. Xie, Z. Yuan, and J. Feng, "Exploring balanced feature spaces for representation learning," in *International conference on learning representations*, 2020. 1

[15] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *Proceedings of the IEEE international conference on computer vision*, pp. 5409–5418, 2017. 1

[16] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020. 1

[17] T. Li, L. Wang, and G. Wu, "Self supervision to distillation for long-tailed visual recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 630–639, 2021. 1

[18] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9719–9728, 2020. 1, 2

[19] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution efficientnets with meta data," *MethodsX*, vol. 7, p. 100864, 2020. 2, 4, 5, 6

[20] P. Georgiadis, E. V. Gkouvrikos, E. Vrochidou, T. Kalampokas, and G. A. Papakostas, "Building better deep learning models through dataset fusion: A case study in skin cancer classification with hyper-datasets," *Diagnostics*, vol. 15, no. 3, p. 352, 2025. 1, 5, 6

[21] S. Zhou, Y. Zhuang, and R. Meng, "Multi-category skin lesion diagnosis using dermoscopy images and deep cnn ensembles," *línea], ISIC Chellange*, 2019. 2, 5, 6

[22] W.-X. Tsai, Y.-C. Li, and C. H. Lin, "Skin lesion classification based on multi-model ensemble with generated levels-of-detail images," *Biomedical Signal Processing and Control*, vol. 85, p. 105068, 2023. 2, 5, 6

[23] J. Shao, K. Zhu, H. Zhang, and J. Wu, "Diffult: Diffusion for long-tail recognition without external knowledge," *Advances in Neural Information Processing Systems*, vol. 37, pp. 123007–123031, 2024. 2, 4

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002. 2

[25] Z. Zhang and T. Pfister, "Learning fast sample reweighting without reward data," in *Proceedings of the*

IEEE/CVF international conference on computer vision, pp. 725–734, 2021.

[26] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," *arXiv preprint arXiv:1910.09217*, 2019. 2

[27] S. Ahn, J. Ko, and S.-Y. Yun, "Cuda: Curriculum of data augmentation for long-tailed recognition," *arXiv preprint arXiv:2302.05499*, 2023. 2

[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020. 2

[29] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020. 2

[30] R. Robin, B. Andreas, L. Dominik, E. Patrick, and O. Björn, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 2, 4

[31] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023. 2

[32] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36479–36494, 2022.

[33] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023. 2

[34] Y. Qin, H. Zheng, J. Yao, M. Zhou, and Y. Zhang, "Class-balancing diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18434–18443, 2023. 2, 4, 5, 6

[35] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024. 2

[36] Y. Cai, W. Zhang, H. Chen, and K.-T. Cheng, "Medianomaly: A comparative study of anomaly detection in medical images," *Medical Image Analysis*, p. 103500, 2025. 2, 4

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015. 2

[38] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023. 2, 5, 6