INSTITUIÇÃO EXECUTORA

Centro de Informática UFPE

FADE UFPE

MCTI FUTURO

COORDENADORA

Softex

APOIO

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO

GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

ACORDO DE PARCERIA Nº 05/23 FADE/UFPE/SOFTEX - RESIDENCIAL IC13 (CONVÊNIO Nº 02/2023 UFPE) 23076.125530/2022-28

Lux.AI

Training Lesion classification on Synthetic images using Stable diffusion and transformers

1. INTRODUCTION

With recent advancements in AI, various learning-based models are increasingly being applied to medical imaging [1]. By evaluating classification approaches [2] and pre-processing techniques [3], we explored how these methods can be utilized in clinical dermatology, specifically focusing on deep supervised architectures. Recently, large models like transformers have been proposed for image classification tasks [4]. Although originally designed for natural language processing (NLP) [5], these architectures have been adapted for image-based tasks [6], leveraging their self-attention mechanisms to focus on critical local features within an image.

In dermatology, the local attributes of lesions often hold essential information for accurate diagnosis. To address this, we aim to apply and compare transformer-based architectures within the clinical dermatology context [6,7,8,9,10,11]. Specifically, we evaluate these attention-based networks against previously trained classification architectures [2,12,13,14,15].

Additionally, we investigate the use of AI-generated images to augment the training of these models [16]. Since transformers require a significant amount of data due to their large number of parameters, the limited size of dermatology datasets poses a challenge. To mitigate this, we supplement training with images generated using Stable Diffusion and compare the performance of different training strategies.

2. METHODOLOGY

Previously, we assessed and validated the use of CNN architectures in dermatology [2], testing models on public dermoscopic and clinical datasets. By experimenting with various architectures such as ConvNext [13], ResNet [14], DenseNet [15], and EfficientNet [12], we gained insights into how different models perform under varying challenges in clinical dermatology. Additionally, by evaluating preprocessing techniques [3], we identified strategies to address challenges in diverse testing scenarios. To build on these analyses, we now aim to evaluate how the latest transformer-based architectures can tackle the complex task of identifying skin lesions. Specifically, we trained with the transformer architectures: ViT [6], SWin Transformer [7], EVA-02 [8], EdgeNeXT [9], Deit III [10], and Beit v2 [11].

We utilized various representative transformer and convolutional architectures to establish a robust foundation for comparison. Furthermore, we conducted experiments using synthetically generated malignant images created with Stable Diffusion [16]. By gradually increasing the ratio of real to synthetic malignant images in the dataset, we gained valuable insights into the impact of synthetic data on model performance.

In our experiments, we utilized the recently released ISIC24 clinical dataset [17], which comprises 401,059 samples. The most significant challenge posed by ISIC24 is its extreme class imbalance: only 393 images are positive samples (malignant lesions), while 400,666 are negative (benign lesions). This represents an almost 1:1000 ratio, making it particularly difficult to develop a reliable model with such limited positive data. However, overcoming this challenge is crucial as it reflects the type of disparities commonly encountered in real-world scenarios, including the multiclass imbalances in our own dataset.

To address the extreme class imbalance in the ISIC24 dataset, we implemented the following strategies:

1. **Balanced Batches:** For each training iteration, we ensured that batches were balanced by including a proportional representation of positive and negative samples. Given the limited number of 393 positive images, many were repeated across iterations and epochs. This effectively acts as a simple oversampling method for the positive class.
2. **Negative Class Downsampling:** We reduced the number of negative samples in the training set to mitigate the class imbalance and allow the model to focus more on learning from the underrepresented positive class.
3. **Synthetic Image Generation:** To augment the training data, we generated approximately 6,000 synthetic malignant images using a Stable Diffusion model fine-tuned on ISIC24's malignant image subset. These synthetic images were gradually introduced into the training set during experiments to assess their impact on model performance. For more details on the fine-tuning process, refer to our sprint review or the wandb fine-tuning logs.

Baseline and augmentations

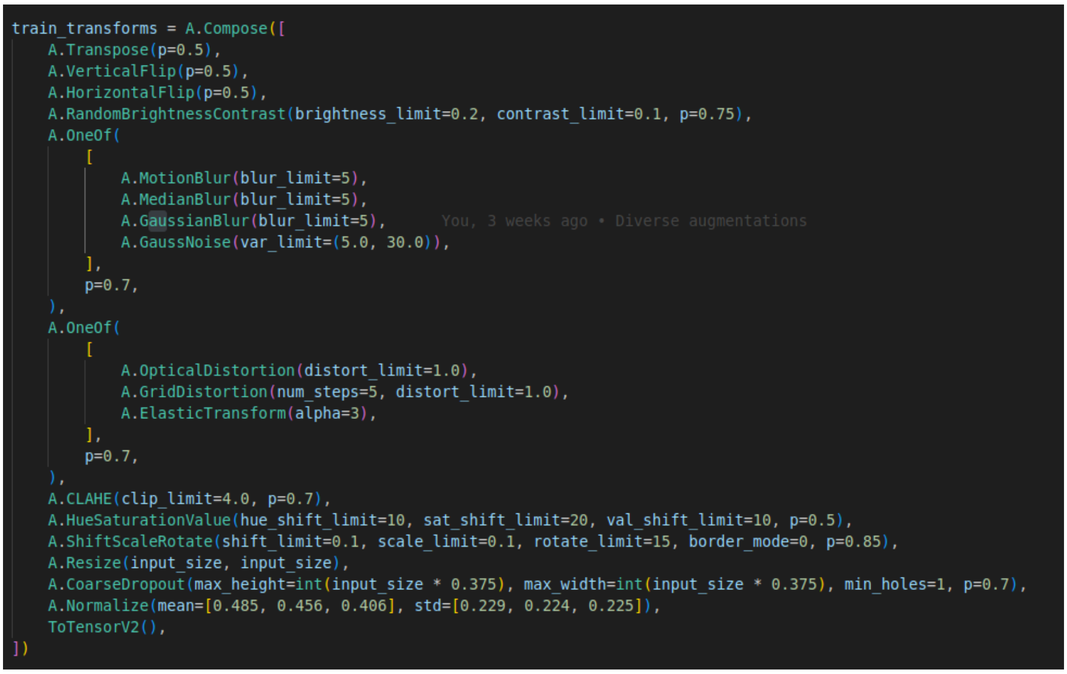
As baseline we applied a simple pipeline on ISIC24 as it is, using a shuffled data loader.

- Augmentation: Simple Random Horizontal Flip
- Single Positive images (for training): ~294
- Single Negative images (for training): ~300500

New augmentation pipeline on ISIC24 using a balanced data loader, based on previous research [2,3].

Each batch received the same amount of negative/positive images, essentially repeating the positive ones (i.e a 1024 batch has 512 negative and 512 negatives; 512 - 393 = 119 positive images were repeated) (Figure 1)

- Single Positive images (for training): ~294
- Single Negative images (for training): ~300500



```
train_transforms = A.Compose([
    A.Transpose(p=0.5),
    A.VerticalFlip(p=0.5),
    A.HorizontalFlip(p=0.5),
    A.RandomBrightnessContrast(brightness_limit=0.2, contrast_limit=0.1, p=0.75),
    A.OneOf(
        [
            A.MotionBlur(blur_limit=5),
            A.MedianBlur(blur_limit=5),
            A.GaussianBlur(blur_limit=5),      You, 3 weeks ago • Diverse augmentations
            A.GaussNoise(var_limit=(5.0, 30.0)),
        ],
        p=0.7,
    ),
    A.OneOf(
        [
            A.OpticalDistortion(distort_limit=1.0),
            A.GridDistortion(num_steps=5, distort_limit=1.0),
            A.ElasticTransform(alpha=3),
        ],
        p=0.7,
    ),
    A.CLAHE(clip_limit=4.0, p=0.7),
    A.HueSaturationValue(hue_shift_limit=10, sat_shift_limit=20, val_shift_limit=10, p=0.5),
    A.ShiftScaleRotate(shift_limit=0.1, scale_limit=0.1, rotate_limit=15, border_mode=0, p=0.85),
    A.Resize(input_size, input_size),
    A.CoarseDropout(max_height=int(input_size * 0.375), max_width=int(input_size * 0.375), min_holes=1, p=0.7),
    A.Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]),
    ToTensorV2(),
])
```

Figure 1 - Augmentations applied on ISIC24

In addition to data augmentation, we evaluated various batch compositions to address the class imbalance in the ISIC24 dataset. Table 1 provides an overview of the data distribution for each training set configuration. **Balanced + Augmentation** maintains the original class proportion by applying strong augmentations to negative samples, effectively balancing the dataset. **Balanced Downsampled** achieves a similar balance by reducing the number of negative samples while applying augmentation. **Super-Downsampled** adjusts class proportions of the original dataset and applies augmentation equally to both negative and positive samples to improve representation. Additionally, we tested various proportions of synthetic and real images, using Stable

Diffusion-generated images to enhance the representation of negative samples in the ISIC24 dataset (Table 2).

Table 1 - Training distributions on ISIC24

	Default	Balanced + aug	Balanced downsampled	super-downsampled
Positive (malignant)	294	294	296	303
Negative (benign)	300500	300500	1803	600

Table 2 - Training distributions using synthetic images from stable diffusion

	synthetic 1:1	Synthetic 1:3	Synthetic 1:15	Synthetic only
Positive (malignant)	303	189	43	0
Positive synthetic	303	567	645	5971
Negative synthetic	1800	800	680	6000

3. EXPERIMENTS AND DISCUSSIONS

Generated images

As mentioned earlier, we used Stable Diffusion-generated images to enhance the dataset's representation of malignant samples. Figures 2 and 3 illustrate examples of AI-generated images simulating skin lesions. For the input prompt, we instructed the model to generate images of "Malignant skin lesions."

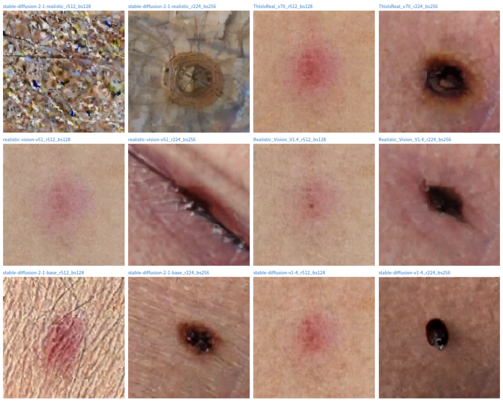

stable-diffusion-2-1-realistic_r512_bs128
stable-diffusion-2-1-realistic_r224_bs256
ThisIsReal_v70_r512_bs128
ThisIsReal_v70_r224_bs256
realistic-vision-v51_r512_bs128
realistic-vision-v51_r224_bs256
Realistic_Vision_V1.4_r512_bs128
Realistic_Vision_V1.4_r224_bs256
stable-diffusion-2-1-base_r512_bs128
stable-diffusion-2-1-base_r224_bs256
stable-diffusion-v1-4_r512_bs128
stable-diffusion-v1-4_r224_bs256

Figure 3 - AI generated images of “Malignant Skin lesions”

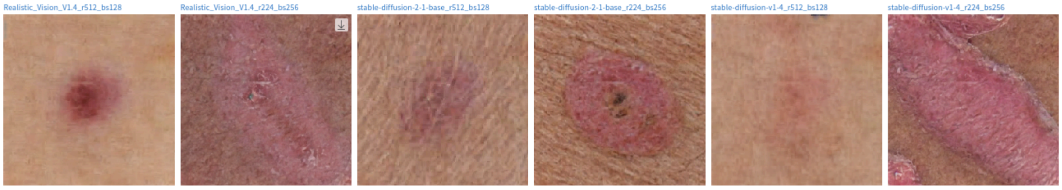

Realistic_Vision_V1.4_r512_bs128
Realistic_Vision_V1.4_r224_bs256
stable-diffusion-2-1-base_r512_bs128
stable-diffusion-2-1-base_r224_bs256
stable-diffusion-v1-4_r512_bs128
stable-diffusion-v1-4_r224_bs256

Figure 4 - Resulting images after running for 50 epochs

Scores Transformers vs CNNs

Convolutional and transformer-based architectures delivered very similar results. Swin Transformer models occasionally exhibited fluctuations in metrics, suggesting that the optimizer might have been trapped in a local minimum. Similar behavior was observed in other architectures, likely due to the warmup phase. The EdgeNeXT consistently outperformed most other models, including convolutional ones, in the majority of experiments. Convolutional architectures demonstrated faster convergence across epochs compared to transformer-based models. This is likely due to the "data-hungry" nature of transformer backbones, which require more data and have higher parameter counts and complexity. The overall accuracy on the test and validation set is illustrated in Figure 4.

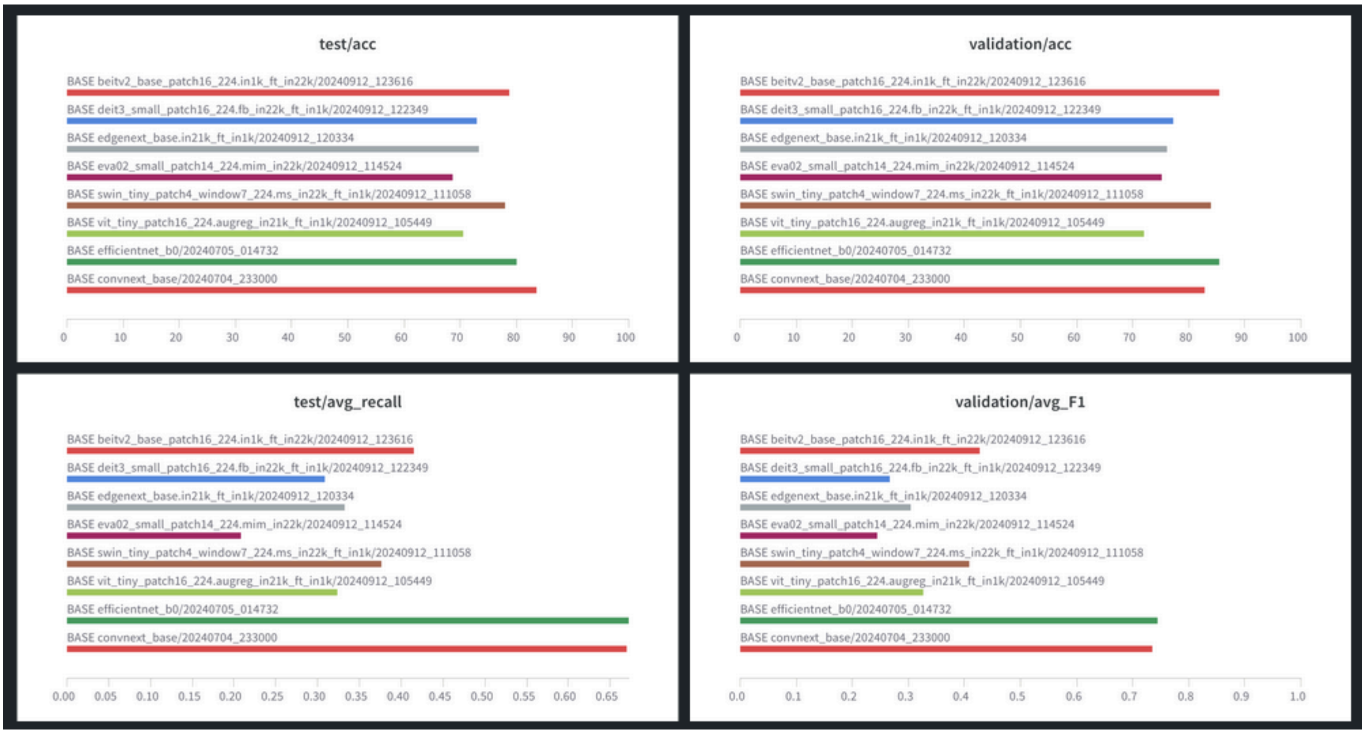


Figure 4 - Scores of CNNs and transformers models executed on ISIC24 test set

Figure 5 illustrates an example of how transformer models focus on local attributes. The heatmaps of the EfficientNet and ViT networks, based on activations generated from an input image, highlight the regions influencing the model's prediction. Transformers tend to prioritize local features, concentrating activations on the lesion area. In contrast, CNNs extract more general features by considering the global context, allowing surrounding areas to influence the model's decision. Future research could explore integrating the local focus of transformers with the broader contextual understanding of CNNs to combine the strengths of both approaches in analyzing image attributes.

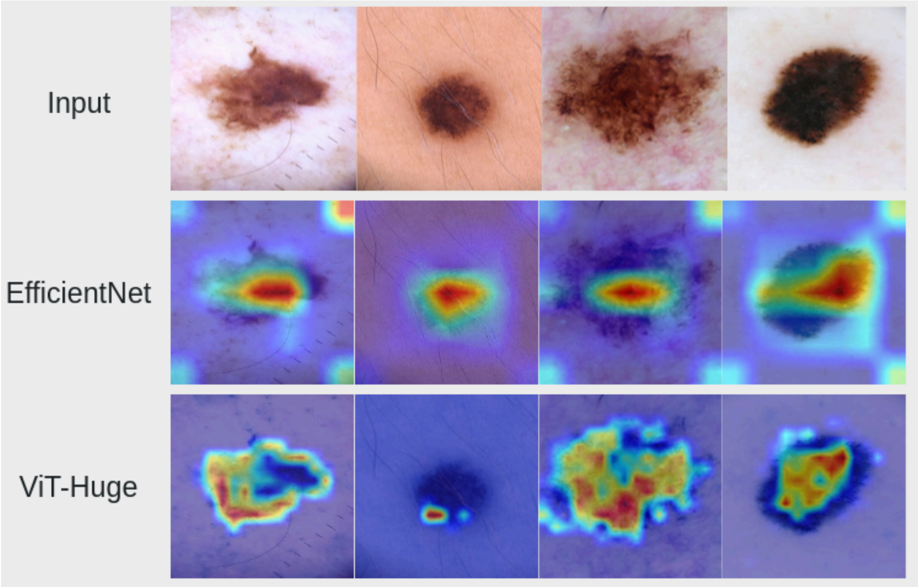


Figure 5 - Activation maps of EfficientNet (middle row) and ViT (bottom row). The red areas indicate the most relevant regions that contribute to the model's decision for the predicted class.

Initializing on synthetic images

Experiments revealed a clear trend: as the proportion of synthetic images increased, validation metrics generally declined**.** This behavior is illustrated in Figure 6. Models trained with synthetic data achieved higher **Recall**, which is particularly valuable in health-related applications where minimizing false negatives is critical. A 1:1 ratio of synthetic to real images yielded promising results, achieving comparable performance to models trained without synthetic data while improving metrics like **Recall** and **F1-score**. Models trained with synthetic data tended to converge faster but exhibited poorer validation metrics, potentially indicating overfitting. This may suggest that the generated images lacked sufficient diversity.

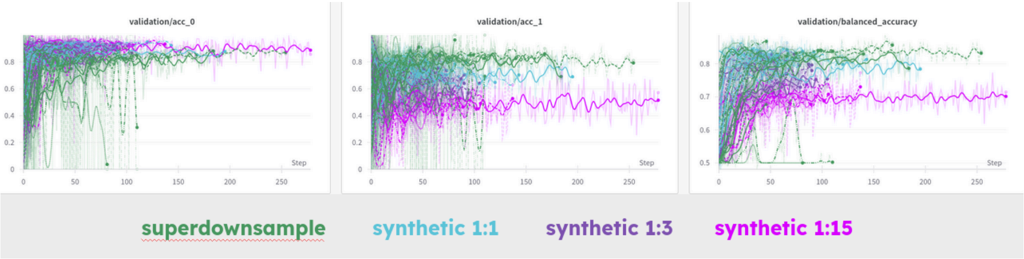


Figure 6 - Network scores varying synthetic distribution

Public set ISIC24

As an additional experiment, we tested our variations of transformers and CNNs on the ISIC24 Challenge test set (Table 3). We evaluated the models on both the public and test sets, comparing the performance of different architectures and varying proportions of synthetic samples. Scores were computed using the challenge's specific metric [17]: only predictions with an AUC > 0.8 contribute to the final score, which ranges from 0.0 to 0.2, with higher values indicating better performance. These results allowed for a comparison with competing methods and confirmed trends observed in our previous experiments, reinforcing our conclusions.

Table 3 - Results on ISIC24 test set

Modelo	Variação de treinamento	Private Score (pAUC 0.8)	Public Score (pAUC 0.8)
EdgeNeXt Base	**superdownsampled**	**0.127**	**0.149**
EdgeNeXt Base	real/synthetic 1:1	0.048	0.049
EdgeNeXt Base	**real/synthetic 1:3**	**0.123**	**0.135**
EdgeNeXt Base	real/synthetic 1:15	0.044	0.040
EdgeNeXt Base	synthetic only	0.026	0.025
EfficientNet B0	**superdownsampled**	**0.110**	**0.128**
EfficientNet B0	real/synthetic 1:1	0.085	0.085
EfficientNet B0	**real/synthetic 1:3**	**0.092**	**0.112**
EfficientNet B0	real/synthetic 1:15	0.093	0.095
EfficientNet B3	synthetic only	0.060	0.060
Ensemble of highlighted models (bold)		0.130	0.147

Preliminary Conclusions

These findings highlight the need for further research into different architectures for classification and synthetic image generation in dermatology. Our approach used synthetic images for training and real images for evaluation, as this was the only viable method available at the time, apart from expert dermatological assessment. Enhancing the diversity and quality of synthetic images is crucial to maximizing their potential in medical AI applications.

However, we can further explore and evaluate alternative training strategies for transformers by leveraging advancements in synthetic data generation techniques. For instance, distinguishing between different classes of malignant and benign lesions could help reduce intra-class feature variability. Additionally, optimizing proprietary diffusion models tailored specifically for dermatological images could yield better results, as Stable Diffusion is primarily optimized for generating general-purpose images rather than medical-specific data.

Full metrics report are available at:

https://wandb.ai/tic13/transformers/reports/Transformers-vs-Convolutional-performance-on-real-and-synthetic-ISIC24--VmlldzoxMDAyMDY0OA?accessToken=dudlz99gq13ls4b33kkmtly0nr79wydsl8165e8d848va7tnijcrmfw4k106ou9f

4. REFERENCES

[1] Pinto-Coelho, Luís. "How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications." *Bioengineering* 10.12 (2023): 1435.

[2] Skin lesion classification on dermoscopy and clinical images using deep learning. LuxAI. Technical Report, 2024. https://luxai.cin.ufpe.br

[3] Data Pre-processing and Augmentation on Dermoscopy Images for Skin Lesion Classification. LuxAI. Technical Report, 2024. https://luxai.cin.ufpe.br

[4] Khan, Salman, et al. "Transformers in vision: A survey." *ACM computing surveys (CSUR)* 54.10s (2022): 1-41.

[5] Vaswani, Ashish, et al. "Attention Is All You Need.(Nips), 2017." *arXiv preprint arXiv:1706.03762* 10 (2017): S0140525X16001837.

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

[7] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[8] Fang, Yuxin, et al. "Eva-02: A visual representation for neon genesis." *Image and Vision Computing* 149 (2024): 105171.

[9] Zheng, Jiahao, et al. "Lightweight Vision Transformer with Spatial and Channel Enhanced Self-Attention." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[10] Touvron, Hugo, Matthieu Cord, and Hervé Jégou. "Deit iii: Revenge of the vit." *European conference on computer vision*. Cham: Springer Nature Switzerland, 2022.

[11] Peng, Zhiliang, et al. "Beit v2: Masked image modeling with vector-quantized visual tokenizers." *arXiv preprint arXiv:2208.06366* (2022).

[12] Koonce, Brett, and Brett Koonce. "EfficientNet." *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization* (2021): 109-123.

[13] Liu, Zhuang, et al. "A convnet for the 2020s." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[14] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[15] Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[16] Farooq, Muhammad Ali, et al. "Derm-t2im: Harnessing synthetic skin lesion data via stable diffusion models for enhanced skin disease classification using vit and cnn." *arXiv preprint arXiv:2401.05159* (2024).

[17] Nicholas Kurtansky, Veronica Rotemberg, Maura Gillis, Kivanc Kose, Walter Reade, and Ashley Chow. ISIC 2024 - Skin Cancer Detection with 3D-TBP. https://kaggle.com/competitions/isic-2024-challenge, 2024. Kaggle.