www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(2), April 2025, 158-175. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2025.v16i2.1078

Generative AI and Transformers in Advanced Skin Lesion Classification applied on a mobile device

Erick García-Espinosa, José Sergio Ruiz-Castilla, Farid García-Lamont*
Universidad Autónoma del Estado de México, Centro Universitario UAEM Texcoco, Texcoco, Estado de México, México
E-mails: dev.garcia.espinosa.erick@gmail.com, jsruizc@uaemex.mx, fgarcial@uaemex.mx

Abstract. This research work focuses on the development of a device. Such a device could assist doctors in level 1 and 2 healthcare clinics in Mexico. Because, such clinics lack specialists. The device takes pictures of the patient's skin. The pictures allow to identify diseases and provide a preliminary diagnosis. With the pre-diagnosis it is possible to send the patient to the corresponding specialist. We built a Vision Transformer (VIT) model with a Raspberry Pi 4. The system leverages a dataset augmented by a Generative Adversarial Network (GAN) using Stable Diffusion. The addition of synthetic data significantly improved the performance metrics. Accuracy increased from 90.76% to 92.77%, and the macro average and weighted average F1 scores increased from 0.9076 to 0.9281. Also, improvements were observed in most disease categories. Thus, the model's capacity allows generalization, especially in underrepresented or challenging classes.
Keywords: Generative Adversarial Network, Skin Diseases, Transfer Learning, Transformers

Article Info
Received Jan 26, 2025
Accepted Mar 11, 2025

1 Introduction

In Mexico, the healthcare system is divided into three levels. First-level clinics, mainly in rural areas, provide basic services such as general consultations and vaccinations. However, diagnosing skin diseases like Chickenpox, Measles, Herpes, Melanoma, and Monkeypox is challenging due to the lack of specialists in these facilities. Over 70% of the population seeks care first at these clinics (National Institute of Public Health, 2021), highlighting the need for tools to help general practitioners make better preliminary diagnoses and referrals.

In second-level clinics provide intermediate medical care, and third-level clinics manage complex and specialized cases [1]. First-level clinics, often located in rural areas, are typically staffed by general practitioners who may lack adequate knowledge to diagnose specific conditions, such as skin lesions. These clinics play a crucial role in early disease detection but face significant challenges due to limited resources and specialist availability.

To improve care, mobile clinics have been deployed in rural communities, such as those supported by initiatives such as Rotary International's mobile health care campaigns [2]. These clinics provide essential medical services to underserved populations but typically lack specialists, including dermatologists, onboard. As a result, while these efforts address general healthcare needs, conditions such as skin lesions often remain underdiagnosed or misdiagnosed, highlighting the need for tools that can support accurate preliminary diagnoses in such contexts.

Artificial intelligence (AI) models have demonstrated high precision in the detection of skin lesions, with convolutional neural networks (CNNs) achieving notable success in recent years [3]. More recently, transformer-based models, such as ViT [4], DinoV2 [5], and Swin [6], have surpassed traditional methods, leveraging advanced attention mechanisms to enhance performance. Additionally, generative adversarial networks (GANs) have been employed to generate synthetic images, aiming to augment the dataset by producing images similar to the training set but with small perturbations. This approach has led to improvements in model performance [7].

Despite these advances, there is a gap in implementing AI-based solutions in portable devices suitable for first-level clinics. Most models require high computational resources and are designed for research rather than clinical use in low-resource environments. To address this, we propose a portable device, D.A.N.N, which combines affordable hardware (Raspberry Pi 4) with transformer-based models. This device captures images of skin lesions, provides preliminary diagnoses, and recommends specialist referrals. Different preprocessing and data augmentation techniques were used to improve model accuracy in clinical scenarios.

The main contributions of this work are:

- Portable device development: Integration of accessible hardware and advanced image classification models.
- Synthetic data generation: Use of GANs to enhance model performance.
- Transformers comparison: Evaluation of ViT, DinoV2, and Swin across multiple datasets.
- Explainability: Implementation of LIME for transparent and interpretable predictions.

The article is organized as follows: Section 2 reviews related work and existing methods for skin lesion classification. Section 3 introduces the transformer models used. Section 4 describes the D.A.N.N device's hardware and software. Section 5 discusses the datasets. Section 6 presents experimental results, and Section 7 concludes with findings and future work.

2 Related Work

In the pursuit of improving automated skin disease diagnosis, significant advancements have been made in recent years through the application of deep learning and different algorithms like CNN and Transformers as you can see in Table 1.

- Sumitra et al. (2015) proposed a novel approach for automatic segmentation and classification of skin lesions. By using SVM and k-NN classifiers in combination, they achieved a promising F-measure of 61% on a custom dataset of 726 samples, highlighting the potential of combining classifiers for improved performance.[8]
- AlSuwaidan (2022) explored the efficacy of six CNN architectures for classifying dermatological disorders, with MobileNet outperforming others by achieving an impressive accuracy of 95.7%. This study underscores the effectiveness of MobileNet in handling medical image analysis tasks.[9]
- Tahir et al. (2023) introduced DSCC_Net, a deep learning-based network for skin cancer classification. Tested on public datasets like ISIC 2020, HAM10000, and DermIS, DSCC_Net achieved an AUC of 99.43% and an accuracy of 94.17%, demonstrating its superiority over several baseline models.[10]
- Wei et al. (2023) proposed a hybrid model fusing DenseNet and ConvNeXt to classify skin diseases, which resulted in an accuracy of 95.29% and an F1 score of 89.99% on the HAM10000 dataset. This approach effectively enhanced feature extraction capabilities, crucial for accurate classification [11].
- Cai et al. (2022) developed a multimodal Transformer combining image and clinical metadata for skin disease classification. Their model, tested on the ISIC 2018 dataset, achieved a high accuracy rate, showcasing the benefit of integrating multimodal data in improving diagnostic performance.[12]
- Aldhyani et al. presented a lightweight and efficient model for skin lesion classification, achieving an impressive accuracy of 97.85% on the HAM10000 dataset. This model's design optimizes computational resources while maintaining high accuracy, making it suitable for practical applications.[13]
- Barros et al. constructed a ResNet-152-based model to classify 12 types of skin lesions, achieving an accuracy of 94.50%. By employing data augmentation techniques, they effectively handled the variability within the dataset, enhancing the model's robustness.[14]
- Hammed et al. introduced a Multi-Class Multi-Level (MCML) classification algorithm that achieved a precision of 96.47% on a diverse dataset of 3672 images. This algorithm's multi-level approach allows for detailed and accurate categorization of skin lesions.[15]

Table. 1. Summary of Related Work

Author	Model	Performance
Sumitra et al. (2015)	SVM and k-NN	F-measure: 61%
AlSuwaidan (2022)	MobileNet	95.7%
Tahir et al. (2023)	DSCC_Net	94.17%
Wei et al. (2023)	DenseNet and ConvNeXt	95.29%
Cai et al. (2022)	Multimodal Transformer	93.81%
Aldhyani et al.(2022)	Lightweight Model	97.85%
Barros et al (2019)	ResNet-152	94.50%
Hammed et al (2024)	Multi-Class Multi-Level (MCML)	96.47%

3 Transformers

In this work, we compare three advanced transformer models using various datasets tailored for skin lesion detection. These models were trained on Google Colab Pro, leveraging its computational power and memory capabilities to manage the high resource demands of transformer-based architectures.

The ViT-Base-Patch16-224 model, developed by Google, utilizes a novel approach to image classification by treating images as sequences of patches, enabling it to capture long-range dependencies within the data. This model has demonstrated exceptional performance in large-scale image classification tasks.[4]

Meta's DinoV2 model, which employs self-supervised learning, is designed to learn visual representations from unlabeled data. This approach allows for flexible application across different domains without the need for extensive labeled datasets.[16]

The Swin-Base-Patch4-Window7-224 model by Microsoft introduces a hierarchical vision transformer architecture that processes images at multiple scales. Its innovative shifted window mechanism enhances efficiency and scalability, making it a robust choice for various computer vision tasks.[17]

4 D.A.N.N Device

This section presents the development and implementation details of the D.A.N.N device, where D.A.N.N means Dermatologic Analysis with Neural Networks. The device is designed to operate autonomously, without requiring an internet connection, making it ideal for resource-limited environments such as first-level healthcare clinics. It is lightweight, portable, and powered by a battery, ensuring independent operation. The primary function of the device is to capture high-quality images of skin lesions and provide a preliminary diagnosis using a trained model. The section is organized as follows: Subsection 4.1 describes the hardware components, Subsection 4.2 details the software environment, and Subsection 4.3 explains the deployment process. See Fig.1.

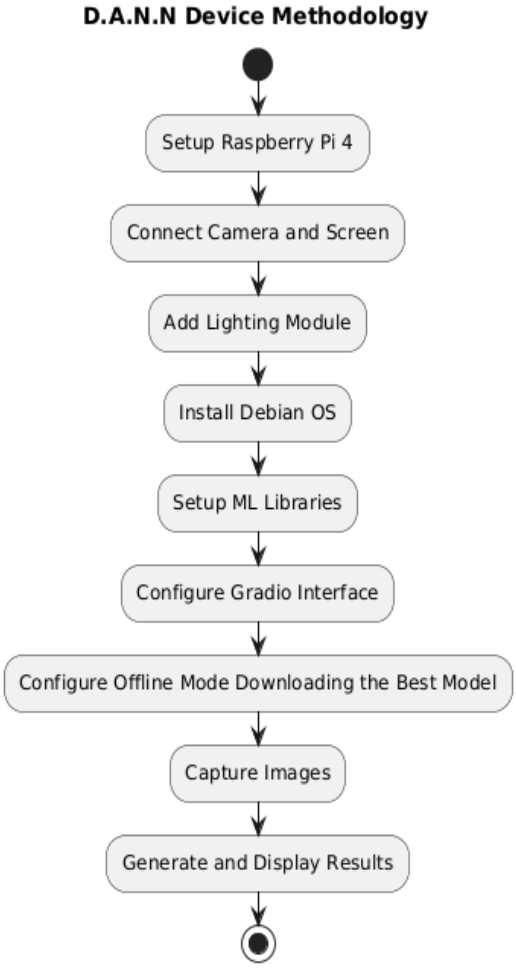

Fig. 1. Methodology of the D.A.N.N Device

4.1 Hardware

For this research, we developed a mobile device called D.A.N.N (Fig. 2), consisting of a Raspberry Pi 4 Model B with 8 GB of RAM. This Raspberry Pi features a Broadcom BCM2711 system on a chip, a quad-core 1.5 GHz ARM Cortex-A72 processor, and a VideoCore VI GPU [18].

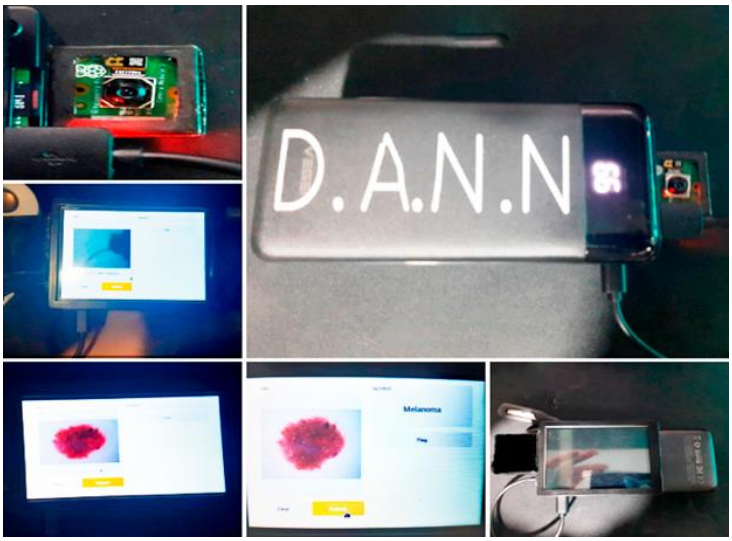

Fig. 2. D.A.N.N Device

In terms of connectivity, the device includes 802.11ac Wi-Fi, Bluetooth 5.0, and Gigabit Ethernet. For video and audio output, it has two micro-HDMI ports that support 4K displays at 60 Hz via HDMI 2.0, a MIPI DSI display port, a MIPI CSI camera port, a 4-pole stereo output, and a composite video port. Additionally, it has 2 USB 3.0 ports and 2 USB 2.0 ports.

The D.A.N.N device is powered by a 20,000 mAh power source that provides a 5V 3A USB-C output and two 5V 3A USB outputs. For user interaction, we use a 3.5-inch touchscreen that operates with a stylus.

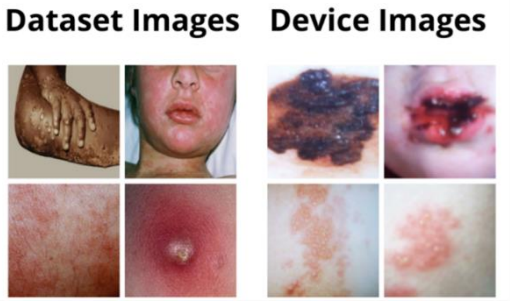

Fig. 3. Comparison of Dataset Images and Images Captured by the D.A.N.N Device

To capture images (see Fig. 3), we incorporated the Raspberry Pi Camera 3 module, which offers a fixed resolution of 11.9 megapixels with a 4608 x 2592-pixel sensor. This module can record in various video modes, such as 2304 x 1296p at 56 fps, 2304 x 1296p at 30 fps in HDR, and 1536 x 864p at 120 fps. Additionally, to ensure the quality of the photographs, we included a small 1.2W USB lamp to provide consistent lighting.

4.2 Software

The device runs on the 64-bit Debian Bookworm Full operating system, which ensures compatibility with the required machine learning frameworks. The software stack includes the Transformers, PyTorch, and TensorFlow libraries, as well as version 2.0.0 of Gradio for creating user-friendly interfaces.

This setup allows seamless interaction with models and enables real-time image capture using the Raspberry Pi Camera 3. The Gradio interface is configured to run on the default browser (Firefox in this case) and is used to display prediction results via a bar graph, offering a visual representation of the model's output.

4.3 Deployment

To make the model accessible for testing, we utilized Hugging Face Spaces [19], which allows interaction with the model from any device connected to the internet. This platform employs Gradio for its interface, and the latest version of Gradio was used due to the absence of hardware limitations in this cloud-based environment.

To set up the Hugging Face Space, it is necessary to create a Hugging Face account. Once the account is created, the configured Space must be uploaded, including all required dependencies, the trained model, and the Gradio interface. This ensures that the model and interface are fully operational and accessible in the cloud The D.A.N.N device itself is optimized for offline operation, requiring internet connectivity only for the initial setup or updates. This design ensures functionality in resource-limited clinical settings. The deployment process includes:

- Initial setup of the Raspberry Pi environment and installation of required libraries.
- Loading the trained model onto the device.
- Configuring the Gradio interface for local operation.
- Testing the device under controlled conditions to verify repeatability of the methodology.

By combining both online access through Hugging Face Spaces and offline functionality via the D.A.N.N device, the system ensures adaptability for a wide range of clinical and testing scenarios.

4.4 Usage

To use the device correctly, we created the following diagram to explain its usage. The strength of this device is that it only needs to be connected to the internet for its initial setup. Afterward, it can operate without an internet connection (see Figure 4)

Fig. 4. Step-by-Step Guide for Using the D.A.N.N Device to Capture Images

5 Dataset

This section provides a detailed overview of the datasets used and created for this research. These datasets include the base dataset, segmented datasets, and datasets augmented with synthetic images.

5.1 Base Dataset

The base dataset is a comprehensive collection of images of skin diseases compiled from multiple sources. These sources include the HAM10000 dataset from 2019, Kaggle, Google Images, Dermnet NZ, Bing Images, Yandex, the Hellenic Dermatological Atlas, and the Dermatological Atlas [20-22]. The dataset, curated to include diverse skin conditions, is publicly

available for download via Kaggle [23]. This dataset serves as the foundation for all subsequent datasets and preprocessing steps.

5.2 Segmented Dataset

The segmented dataset was created by applying the next image segmentation method to the base dataset.

- HSV Color Space Segmentation: Images were converted to the HSV (Hue, Saturation, Value) color space. Regions of interest were identified using specific thresholds: hue (H) between [0, 60], saturation (S) above 90, and value (V) between 10 and 200. This method isolates lesion areas effectively.
- Adaptive Thresholding: After HSV segmentation, images were converted to grayscale, and adaptive thresholding was applied using a block size of 11 and a constant value of 2 to emphasize lesion edges [24].

5.3 Segmented Dataset Using SlimSAM

This dataset was segmented using the "Zigeng/SlimSAM-uniform-50" model [25], applied to the base dataset. Before segmentation, image contrast was enhanced to improve the model's accuracy. SlimSAM performed automatic segmentation, extracting precise lesion regions for further analysis [26].

5.4 Mixed Segmented Dataset

The mixed segmented dataset combines images from both segmentation methods described above. Random selection was used to ensure a balanced representation of segmentation styles, providing diversity for model training and evaluation [26].

5.5 Base + Synthetic Augmentation Dataset

This dataset [27-28]. incorporates synthetic images generated to enhance the base dataset. Using Low-Rank Adaptations (LoRAs), synthetic images were created for the categories Herpes, Measles, Chickenpox, and Monkeypox. Melanoma was excluded due to the abundance of publicly available images for this condition.

- LoRAs Training: The LoRAs were trained using OneTrainer with 60 epochs and 30,000 steps on the base dataset.
- Image Generation: Images were generated using the Fooocus API, guided by a previously trained discriminator model that ensures correct categorization of synthetic images.
- Pre-trained Model: The JuggernautXL V7 model, based on Stable Diffusion XL, served as the pre-trained foundation for LoRAs. The weights and configurations for generating these images are available on our CivitAi profile[29].

Figure 5 illustrates the flow chart for generating synthetic images.

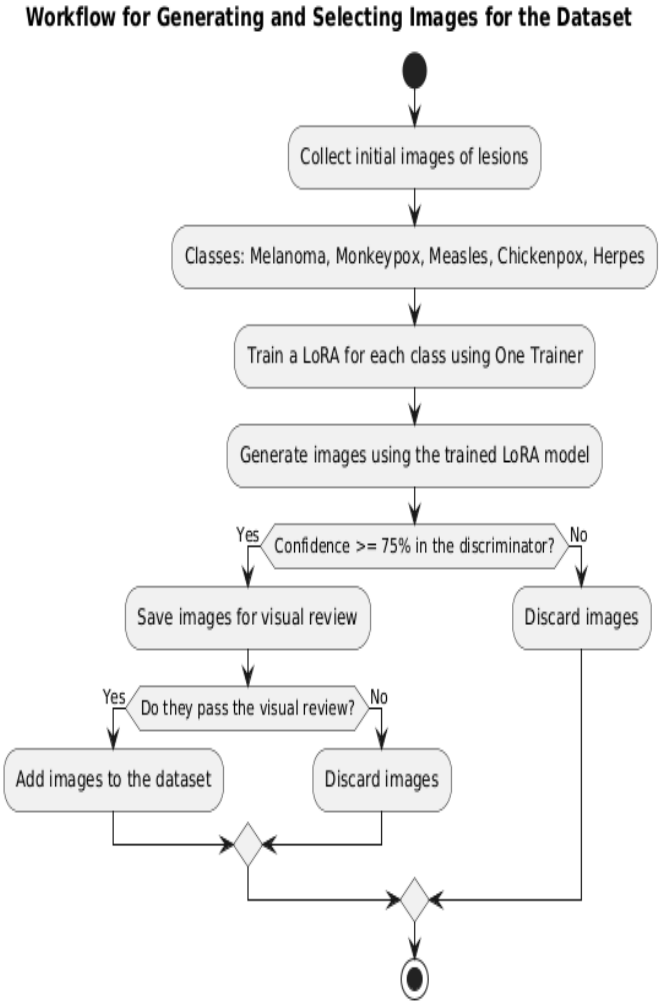

Fig. 5. Workflow for Generating Synthetic Images Using Stable Diffusion GAN

5.6 Base + Synthetic Augmentation for Classes with Fewer Data

This dataset focuses on balancing the dataset by generating only synthetic images specifically for underrepresented categories, namely Chickenpox and Measles. Synthetic images were created only for these categories while retaining original images. This process ensures a balanced categories.

5.7 Dataset Overview

Table 2 provides more details about of the number of images used for training, testing, and validation across all datasets. Importantly, all datasets share the same validation set to maintain consistency in model evaluation. This design ensures that lesion detection is not affected by preprocessing variations, thereby expediting the device's predictions.

Additionally, Figures 6 and 7 present galleries showcasing the datasets' compositions, with Figure 6 highlighting the synthetic images generated.

Table. 2. Datasets Overview

Dataset	Train	Test	Val
Base Dataset	1201	389	100
Segmented Dataset	1131	279	100
Segmented Dataset using SlimSAM	1478	389	100
Mixed Segmented Dataset	1318	319	100
Base + Synthetic Augmentation Dataset	2000	500	100
Base + Synthetic Augmentation less classes	1298	474	100

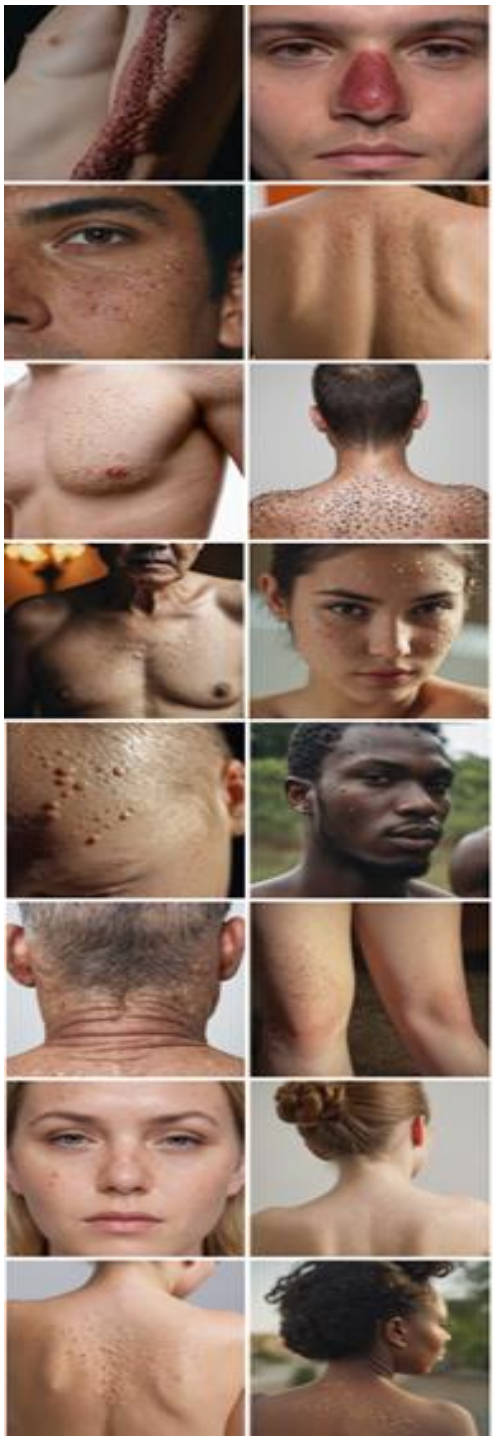

Fig. 6. Gallery of Synthetic Images

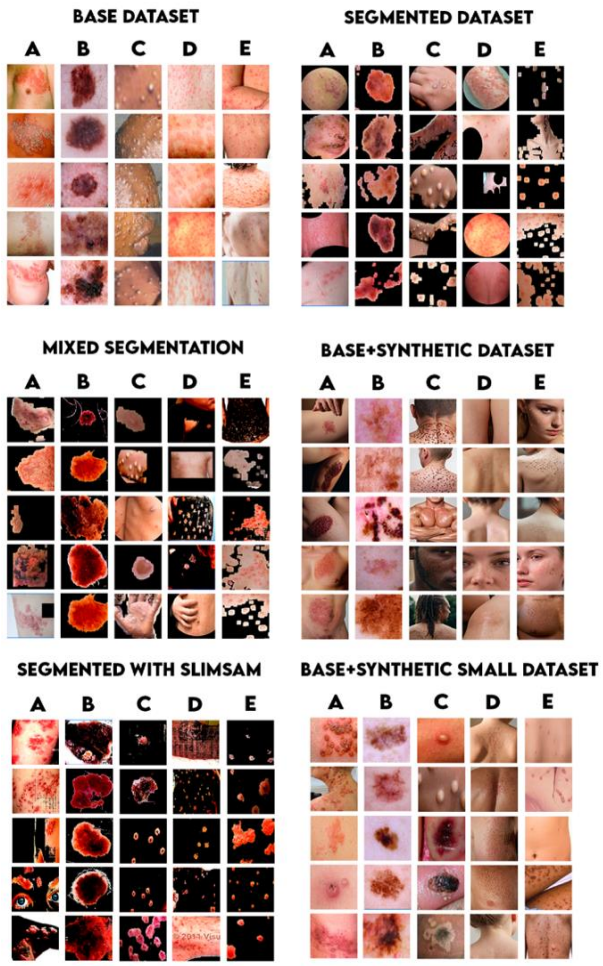

Fig. 7. Gallery of Datasets Used in the Experiment

6 Results

6.1 Model Performance

The performance of the three different transformer models was evaluated across various all datasets to assess their effectiveness in skin lesion detection. The datasets included base, segmented, SlimSAM segmented, mix segmented and Synthetic datasets. Table 3 summarizes the training and validation accuracies, as well as the epoch at which the best performance was achieved for each model and dataset combination.

Table. 3. Transformer Models Performance Across Different Datasets.

Model	Dataset	Train	Validation	Epoch
ViT Transformer	Base	99.6%	90.76%	6
ViT Transformer	Segmented Dataset	99.9%	87.81%	7
ViT Transformer	Segmented Dataset using SlimSAM	99.7%	67.87%	7
ViT Transformer	Mixed Segmented Dataset	99.8%	88.35%	7
ViT Transformer	Base + Synthetic Augmentation Dataset	99.6%	92.77%	6
ViT Transformer	Base + Synthetic Augmentation only for classes with fewer data	99.7%	89.15%	7
DinoV2	Base	96.4%	88.75%	35
DinoV2	Segmented Dataset	96.3%	84.73%	35
DinoV2	Segmented Dataset using SlimSAM	94.0%	76.70%	35
DinoV2	Mixed Segmented Dataset	96.5%	81.52%	35
DinoV2	Base + Synthetic Augmentation Dataset	96.8%	87.50%	35
DinoV2	Base + Synthetic Augmentation only for classes with fewer data	96.1%	89.95%	35
Swin Transformer	Base	96.7%	91.96%	20
Swin Transformer	Segmented Dataset	96.0%	85.15%	20
Swin Transformer	Segmented Dataset using SlimSAM	93.5%	85.04%	20
Swin Transformer	Mixed Segmented Dataset	95.9%	90.76%	20
Swin Transformer	Base + Synthetic Augmentation Dataset	98.2%	89.15%	20
Swin Transformer	Base + Synthetic Augmentation only for classes with fewer data	96.7%	90.36%	20

6.2 Stable Diffusion GAN Performance

In our research, we observed significant improvements when applying synthetic augmentation using Stable Diffusion GAN. This approach allowed us to generate realistic images by leveraging the JuggernautXL model and applying our LoRAs to ensure fidelity in the generated images. One notable advantage of this method is the ability to create a diverse and rich dataset that includes variations in age, ethnicity, skin color, and a wide range of images depicting people with different lesions. Although this process is computationally expensive, it significantly enhances the variety of images available for model training.

However, it is important to note that this approach requires a well-performing pre-trained model to act as a discriminator, ensuring the quality and relevance of the generated images. Despite the computational costs, the benefits of having a richer and more varied dataset outweigh the expenses, making this method a valuable asset in improving the performance of skin lesion detection models.

6.3 Comparative Evaluation of Models

In this section, we present a comparative analysis of the best two models on validation task: the ViT Transformer Base trained on the original dataset (Model 1) and the ViT Transformer Base trained on the dataset augmented with synthetic images (Model 2). The aim is to evaluate the impact of synthetic data augmentation on the models' performance.

6.3.1. *Performance Metrics by Disease*

Table 4 summarizes the precision, recall, F1-scores, sensitivity, specificity, and AUC for each disease on validation set. Model 2 demonstrates notable improvements over Model 1, particularly in recall and sensitivity, which are essential metrics to minimize missed cases in medical diagnosis.

- Monkeypox: Model 2 improved recall from 0.82 to 0.88 and sensitivity from 0.8200 to 0.8800, reducing the likelihood of missed cases while maintaining a high AUC of 0.984322.
- Measles: Recall increased from 0.90 to 0.94, and sensitivity rose from 0.9000 to 0.9400. These improvements indicate better generalization and classification consistency, with AUC increasing from 0.982211 to 0.988744.

Table. 4. Figure caption (9 points font)

Disease	Presicion	Recall	F-1	Sensitivity	Specifity	AUC
Monkeypox (M1)	0.9318	0.8200	0.8723	0.8200	0.9849	0.9832
Measles (M1)	0.8653	0.9000	0.8823	1.0000	0.9648	0.9822
Chickenpox (M1)	0.8301	0.8979	0.8627	0.8979	0.9550	0.9713
Herpes (M1)	0.9387	0.9200	0.9292	0.9200	0.9849	0.9922
Melanoma (M1)	0.9803	1.0000	0.9900	1.0000	0.9949	1.0000
Monkeypox (M2)	0.9565	0.8800	0.9166	0.8800	0.9899	0.9843
Measles (M2)	0.9038	0.9400	0.9215	0.9400	0.9748	0.9887
Chickenpox (M2)	0.8269	0.8775	0.8514	0.8775	0.9550	0.9696
Herpes (M2)	0.9791	0.9400	0.9591	0.9400	0.9949	0.9943
Melanoma (M2)	0.9803	1.0000	0.9900	1.0000	0.9949	1.0000

- Chickenpox: Although precision slightly decreased (from 0.830189 to 0.826923), recall remained steady (0.897959 to 0.877551), and sensitivity slightly dropped (0.897959 to 0.877551). This suggests that augmented data moderately supports this class but may require further refinement.
- Herpes: Model 2 exhibited significant gains in precision (from 0.938776 to0.979167), recall (from 0.920000 to 0.940000), and sensitivity (from 0.920000 to 0.940000). These metrics underline the robustness of Model 2 for this disease.
- Melanoma: Both models achieved exceptional recall (1.00), precision (~0.98), and sensitivity (~1.00), showing no significant differences between the models.

6.3.2. ROC Curves and Confusion Matrices

To evaluate and compare the performance of the two models, ROC curves and confusion matrices were analyzed. The ROC curves illustrate the models' ability to differentiate between classes, providing insights into their overall classification performance and trade-offs between sensitivity and specificity. Confusion matrices offer a detailed view of the classification outcomes, highlighting areas where misclassifications occur and demonstrating improvements in handling challenging cases. These comparative analyses underscore the effectiveness of the enhancements made in Model 2 (See Fig.8).

Table. 4. Figure caption (9 points font)

Disease	Presicion	Recall	F-1	Sensitivity	Specifity	AUC
Monkeypox (M1)	0.9318	0.8200	0.8723	0.8200	0.9849	0.9832
Measles (M1)	0.8653	0.9000	0.8823	1.0000	0.9648	0.9822
Chickenpox (M1)	0.8301	0.8979	0.8627	0.8979	0.9550	0.9713
Herpes (M1)	0.9387	0.9200	0.9292	0.9200	0.9849	0.9922
Melanoma (M1)	0.9803	1.0000	0.9900	1.0000	0.9949	1.0000
Monkeypox (M2)	0.9565	0.8800	0.9166	0.8800	0.9899	0.9843
Measles (M2)	0.9038	0.9400	0.9215	0.9400	0.9748	0.9887
Chickenpox (M2)	0.8269	0.8775	0.8514	0.8775	0.9550	0.9696
Herpes (M2)	0.9791	0.9400	0.9591	0.9400	0.9949	0.9943
Melanoma (M2)	0.9803	1.0000	0.9900	1.0000	0.9949	1.0000

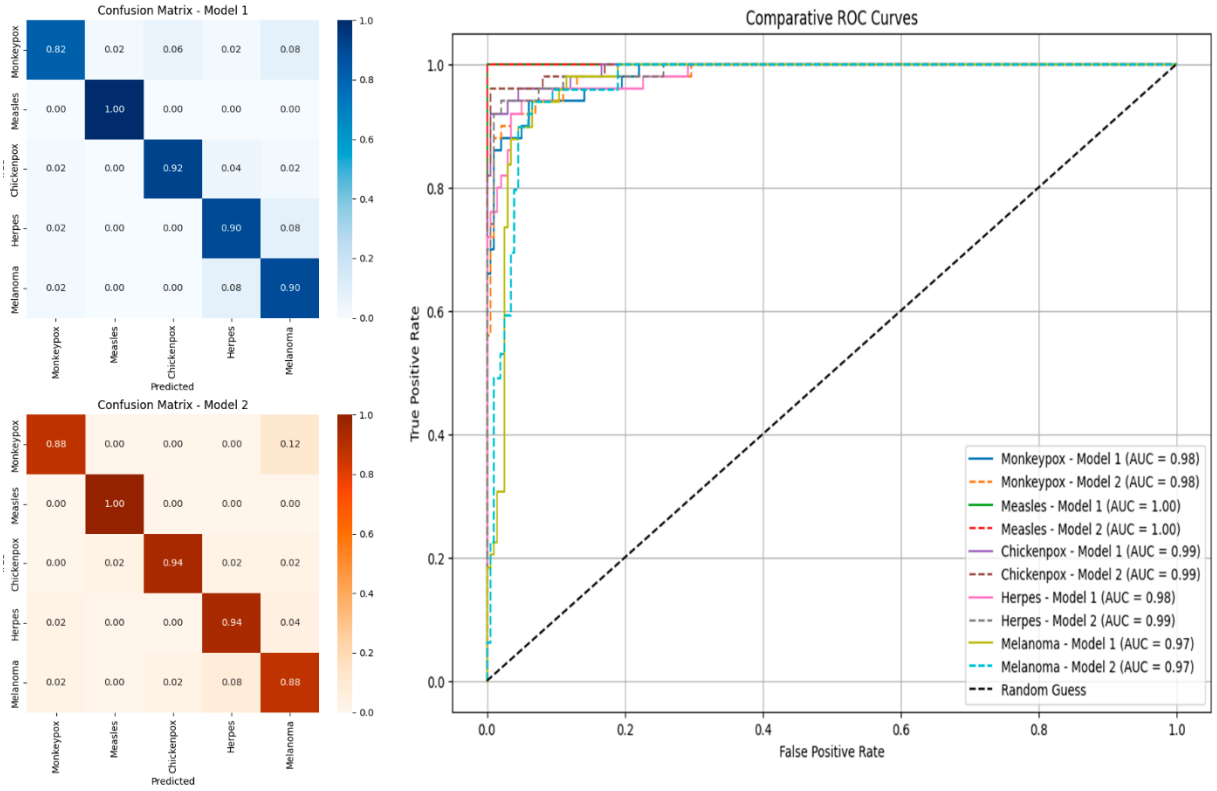

Fig. 8. ROC Curves and Confusion Matrices Comparing the Model Trained on the Base Dataset Against the Model Trained with Synthetic Augmentation

6.3.3. Final Observations on Model Performance

The evaluation of both models highlights the clear advantages of synthetic data augmentation in improving classification performance. Model 2 demonstrated better sensitivity and precision, especially in diseases that were more challenging to classify or underrepresented in the dataset. Misclassification rates were noticeably reduced in key cases, and recall improvements were evident across multiple diseases, showcasing the model's enhanced ability to generalize. Additionally, Model 2 exhibited a faster inference time, completing predictions in 188.55 seconds compared to Model 1's 194.38 seconds. This faster processing underscores Model 2's enhanced computational efficiency alongside its superior diagnostic capabilities, emphasizing the overall value of synthetic data augmentation in advancing medical diagnostic tools.

6.3.4. Final Observations on Transformers Comparison

The comparison of the three transformer models yielded insightful results on their performance in skin lesion detection. All models demonstrated significant promise, particularly when trained on datasets augmented with synthetic images generated via GANs using Stable Diffusion.

- The ViT Transformer achieved the highest validation accuracy of 92.77% when trained on the normal dataset augmented with synthetic images. This indicates that the model can effectively utilize the enhanced dataset to improve its learning and prediction capabilities. However, the model's performance dropped significantly to 67.87% on the SlimSAM segmented dataset, highlighting a potential challenge in handling segmented data with the current segmentation approach.
- DinoV2 also showed robust performance with a best validation accuracy of 89.95% on the synthetic image-augmented dataset, particularly benefiting from the augmentation in underrepresented classes. Its performance on the normal dataset was similarly strong with an accuracy of 88.75%. Additionally, DinoV2 has the advantage of a much faster training time compared to the other two models. Although it did not excel in this particular problem, its excellent performance and quick training make it a great option for solving other problems where training time is a critical factor. However, similar to the ViT

Transformer, DinoV2 struggled with the SlimSAM segmented dataset, achieving a validation accuracy of 76.70%.

- The Swin Transformer displayed consistent and competitive performance across all datasets. It reached its highest validation accuracy of 91.96% on the normal dataset and maintained a high accuracy of 90.36% on the augmented dataset. The slight drop in accuracy to 89.85% with the SlimSAM segmented dataset indicates a relatively better adaptability to segmented data compared to the other models.

6.4 Device Testing

For device testing, the optimal model was chosen, specifically the sixth epoch with base images plus synthetic augmentation. Photos were taken at various distances (Figure 9), concluding that 15 cm is the ideal distance for a perfect composition within the camera's field of view. During the tests, datasets with 10%, 30%, and 50% zoom-outs were used, and a decrease in accuracy and performance was observed as the capture distance increased from 15 cm.

- A 10% zoom-out means taking the photo at approximately 16.5 cm.
- A 30% zoom-out corresponds to taking the photo at approximately 19.5 cm.
- A 50% zoom-out means taking the photo at approximately 22.5 cm.

The results indicate that without zoom-out, an accuracy of 90.76% was achieved. With a 10% zoom-out (16.5 cm), the accuracy was 87.9%. With a 30% zoom-out (19.5 cm), the accuracy was 85.9%, and with a 50% zoom-out (22.5 cm), the accuracy decreased to 72.2%.

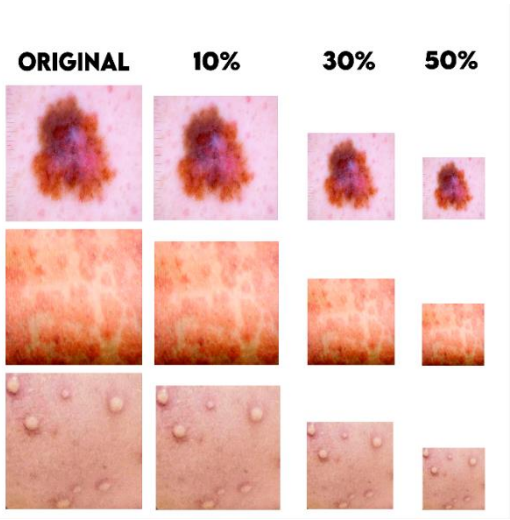

Fig. 9. Images with Progressive Zoom-Out from the Original to 50%

6.5 Explainability

In order to enhance the interpretability of our model and understand the decision-making process behind its predictions, we employed the Local Interpretable Model-agnostic Explanations (LIME) technique. This approach allows us to identify which parts of an image are most influential in the model's decision across all the diseases studied.

Additionally, a gallery of LIME-explained images for the five diseases can be observed in the figure 10. This gallery provides further insight into how the model identifies key features across various skin conditions, ensuring consistency and accuracy in its predictions.

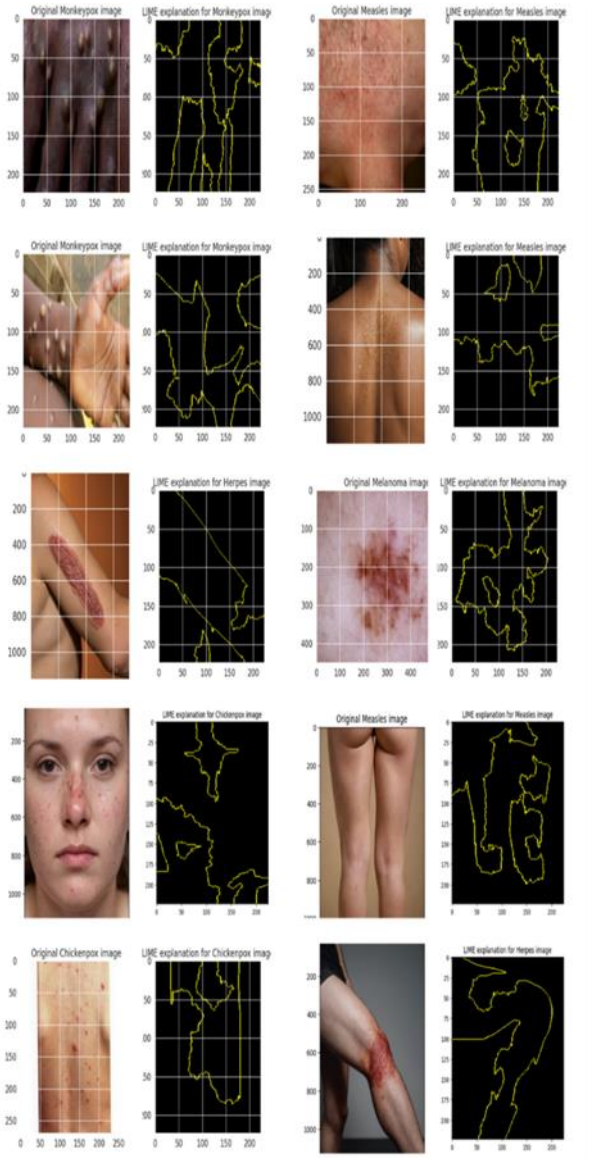

Fig. 10. Visualization of Model Interpretability Using LIME for Disease Detection

This level of explainability is crucial, especially in medical applications, as it provides transparency and trust in the model's predictions. By understanding which parts of an image are driving the model's decisions, healthcare professionals can better interpret the results, ensuring that the model's focus aligns with the clinical understanding of each disease.

4 Conclusions

The application of transformers models for skin lesion detection shows great potential, especially when augmented with synthetic images generated through advanced AI techniques like Stable Diffusion. The augmentation method significantly improves the dataset's diversity and volume, which is crucial for enhancing model performance. This is particularly important given the difficulty in obtaining a large number of images for all types of lesions, as it allows for a more consistent and robust training process, ultimately leading to better validation outcomes.

Our best-performing model, ViT Transformer, achieved a validation accuracy of 92.77% on the dataset augmented with synthetic images. This performance highlights the effectiveness of synthetic augmentation in improving sensitivity, recall, and precision across challenging diseases such as Monkeypox and Measles, while also reducing misclassification rates. Moreover,

the faster inference time of 188.55 seconds compared to the 194.38 seconds of the baseline model underscores the computational efficiency of this approach.

Additionally, the obtained model is relatively lightweight, with a size of only 335 MB, which ensures that the D.A.N.N device can load it quickly. This makes the device more efficient and practical for use in clinics of level 2 or 3 in Mexico. The consistent improvements across diseases and the optimized model size make this tool a promising asset in clinical settings.

To facilitate the testing and application of these models, we have made the best-performing model available on our Hugging Face space[13]. This ensures broader accessibility and provides the opportunity for users to interact with and test the model in a practical setting, potentially contributing to further improvements and applications in clinical environments. This tool can greatly assist healthcare providers, particularly in first-level and second-level clinics, by providing preliminary diagnoses and streamlining the referral process to specialists.

Future Work

Future efforts will focus on testing the D.A.N.N device and the proposed models in real-world medical environments to evaluate their effectiveness and reliability in clinical workflows. Additionally, we aim to explore more robust versions of the transformer architecture, such as the large versions of ViT, to further enhance classification accuracy and scalability. These steps will help refine the system and expand its applicability to a broader range of healthcare settings.

References

1. UNADM (2023). *A tabu search approach to polygonal approximation of digital curves*. *International Journal of Pattern Recognition and Artificial Intelligence*, *14*(2), 243–255.
2. Rotary International. (n.d.). *Rotary mobile clinics bring care to rural Mexico*. Retrieved from https://www.rotary.org/en/rotary-mobile-clinics-rural-mexico. Accessed on December 10, 2024.
3. Khalil, M., Khalil, A., & Ngom, A. (2023). *A Comprehensive Study of Vision Transformers in Image Classification Tasks*. arXiv:2312.01232 [cs.CV]. Retrieved from https://doi.org/10.48550/arXiv.2312.01232. Accessed on December 10, 2024.
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. Google Research, Brain Team. Retrieved from https://arxiv.org/abs/2010.11929. Accessed on December 10, 2024.
5. Kundu, B., Khanal, B., Simon, R., & Linte, C. A. (2024). *Assessing the Performance of the DINOv2 Self-supervised Learning Vision Transformer Model for the Segmentation of the Left Atrium from MRI Images*. arXiv:2411.09598 [eess.IV]. Retrieved from https://doi.org/10.48550/arXiv.2411.09598. Accessed on December 10, 2024.
6. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. Microsoft Research Asia. Retrieved from https://arxiv.org/abs/2103.14030. Accessed on December 10, 2024.
7. Farooq, M. A., Yao, W., Schukat, M., Little, M. A., & Corcoran, P. (2024). *Derm-T2IM: Harnessing Synthetic Skin Lesion Data via Stable Diffusion Models for Enhanced Skin Disease Classification using ViT and CNN*. arXiv:2401.05159 [cs.CV]. Retrieved from https://doi.org/10.48550/arXiv.2401.05159. Accessed on December 10, 2024.
8. Sumitra, Suhil, M., & Guru, D. (2015). *Segmentation and Classification of Skin Lesions for Disease Diagnosis*. In *International Conference on Advanced Computing Technologies and Applications (ICACTA 2015)*.
9. AlSuwaidan, L. (2022). *Deep Learning Based Classification of Dermatological Disorders*. International Conference on Advanced Computing Technologies and Applications.
10. Tahir, M., Naeem, A., Malik, H., Tanveer, J., Naqvi, R. A., & Lee, S.-W. (2023). *DSCC_Net: Multi-Classification Deep Learning Models for Diagnosing of Skin Cancer Using Dermoscopic Images*. *Advanced Intelligent Systems*. https://doi.org/10.1002/aisy.202300211.
11. Wei, M., Wu, Q., Ji, H., Wang, J., Lyu, T., Liu, J., & Zhao, L. (2023). *A Skin Disease Classification Model Based on DenseNet and ConvNeXt Fusion*. *Electronics*, *12*, 438.
12. Cai, G., Zhu, Y., Wu, Y., Jiang, X., Ye, J., & Yang, D. (2022). *A multimodal transformer to fuse images and metadata for skin disease classification*. *European Journal of Dermatology*, *32*(3), 189-197.
13. Aldhyani, T. H. H., Verma, A., Al-Adhaileh, M. H., & Koundal, D. (2022). *Multi-Class Skin Lesion Classification Using a Lightweight Dynamic Kernel Deep-Learning-Based Convolutional Neural Network*. *Diagnostics (Basel)*, *12*(9), 2048. Available: https://n9.cl/theyazn, doi: 10.3390/diagnostics12092048. PMID: 36140447; PMCID: PMC9497471.
14. Barros, D. B., & Silva, N. C. D. (2018). *Skin Lesions Classification Using Convolutional Neural Networks in Clinical Images*. arXiv preprint arXiv:1812.02316. Available: https://n9.cl/barros. Accessed on November 10, 2023.
15. Hameed, N., Shabut, A. M., Ghosh, M. K., & Hossain, M. A. (2020). *Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques*. *Expert Systems with Applications*, *141*, 112961. Available: https://n9.cl/hameed, doi: 10.1016/j.eswa.2019.112961.

16. Meta AI. (2024). *DINOv2: Next Generation Visual Features*. Meta AI. Available: https://dinov2.metademolab.com/. Accessed on August 9, 2024.
17. He, X., Cao, Y., Liu, Z., Bao, J., Zhang, Z., Chen, S., & Yuan, L. (2021). *Masked Autoencoders Are Scalable Vision Learners*. arXiv. Available: https://arxiv.org/abs/2103.14030. Accessed on August 9, 2024.
18. Raspberry Pi Foundation. (2019). *Raspberry Pi 4 Model B*. Raspberry Pi. Available: https://www.raspberrypi.com/products/raspberry-pi-4-model-b/. Accessed on August 9, 2024.
19. DopeErick. (2024). *SkinLesionClassifierSpace*. Hugging Face. Available: https://huggingface.co/spaces/DopeErick/SkinLesionClassifierSpace. Accessed on August 9, 2024.
20. ISIC Archive. (2023, November 8). *ISIC Challenge Datasets*. Available: https://n9.cl/isic-challenge. Accessed on November 10, 2023.
21. Usatine, R. P., & Madden, B. D. (2023). *Interactive Dermatology Atlas*. Dermatlas.net. Available: https://n9.cl/dermatlas. Accessed on November 10, 2023.
22. Hellenic Dermatological Atlas. (2023). *Home | Hellenic Dermatological Atlas*. Available: https://n9.cl/hellenic. Accessed on November 10, 2023.
23. DEVDOPE. (2024, May 1). *Skin Disease Lightweight Dataset*. Kaggle. Available: https://www.kaggle.com/datasets/devdope/skin-disease-lightweight-dataset. Accessed on August 9, 2024.
24. DEVDOPE. (2024). *Skin Disease Lightweight Dataset - Segmented*. Kaggle. Available: https://www.kaggle.com/datasets/devdope/skin-disease-lightweight-dataset-segmented. Accessed on August 9, 2024.
25. Chen, Z., Fang, G., Ma, X., & Wang, X. (2023). *SlimSAM: 0.1% Data Makes Segment Anything Slim*. arXiv. Available: https://doi.org/10.48550/arXiv.2312.05284. Accessed on August 9, 2024.
26. DEVDOPE. (2024). *Skin Disease Variations Dataset*. Kaggle. Available: https://www.kaggle.com/datasets/devdope/skin-disease-lightweight-dataset-segmented. Accessed on August 9, 2024.
27. DEVDOPE. (2024). *Synthetic Skin Disease Dataset/Real and Synthetic*. Kaggle. Available: https://www.kaggle.com/datasets/devdope/skin-disease-lightweight-dataset-segmented. Accessed on August 9, 2024.
28. DEVDOPE. (2024). *Synthetic Skin Disease Dataset/Only Synthetic*. Kaggle. Available: https://www.kaggle.com/datasets/devdope/skin-disease-lightweight-dataset-segmented. Accessed on August 9, 2024.
29. DevDope. (2024). *DevDope Profile*. Civitai. Available: https://civitai.com/user/DevDope. Accessed on August 9, 2024.