Advancing Skin Cancer Diagnosis: Leveraging Vision Transformers and Self-Supervised Learning for Robust Melanoma Classification

Arrun Sivasubramanian*†, Sai Lohitaksh Reddy Devireddy*, Ishani Arya†
*Department of Applied Mathematics and Statistics, Johns Hopkins University
†Department of Computer Science, Johns Hopkins University

***Abstract*—Skin diseases affect over a third of the global population, yet their impact is often underestimated. Automating the classification of these diseases with just visual cues is essential for supporting timely and accurate diagnoses and reducing diagnosis time with the assistance of dermatologists. However, a generalized model fine-tuned to accurately classify skin diseases usually fail on malignant disease detection such as melanoma and raise several false alarms for benign rashes. Automated melanoma classification using dermoscopic images has gained significant traction with the advent of deep learning models for feature extraction. This research explores the use of Vision Transformers (ViT) and DinoV2, focusing on their ability to classify lesions from the ISIC 2020 dataset. By integrating metadata such as age, gender, and anatomical site with advanced augmentation techniques, we address challenges such as class imbalance and noise in imaging data. Our results demonstrate the effectiveness of metadata integration and transformer-based models in improving sensitivity and overall classification performance, paving the way for more accurate and scalable diagnostic tools.**

***Index Terms*—Skin Disease Classification, Vision Transformers, Swin Transformers, DinoV2, GradCAM, SHAP.**

I. Introduction

THE skin protects the body from pollutants, heat, and UV radiation [1]. Skin cancer is one of the most prevalent types of cancer globally, with over 5 million cases diagnosed annually in the United States alone, indicating a higher prevalence than previously thought [2]. These skin disorders become progressively dangerous as time passes. Dermatologists think it can be treated if the harm is detected in time, but things may get complicated when they depend solely on manual methods to identify disorders. The main reason for this is because there are several forms of illnesses. Furthermore, physical diagnosis with just visual cues may be complicated since many skin illnesses have similar visual features, further complicating diagnosis and medical treatment. [3].

The severity and symptoms of these skin issues vary greatly, with some skin diseases being hereditary while outside influences cause others. Over 3000 acute and chronic skin disorders affecting persons of various ages and genders have been recorded [4]. They might be temporary or permanent and can be unpleasant or lethal in a few cases, like melanoma. Though they can be treated with medication, lotions, ointments, or lifestyle modifications [5], they can significantly burden patients through decreased quality of life, confidence, and higher costs.

In recent years, unsupervised feature extraction from pictures has relied heavily on deep learning (DL) approaches, particularly convolutional neural networks (CNNs) [6]. Academics have produced several CNN designs to increase the performance in areas with high availability and diverse annotated data [7]. They have also been crucial in medical image-based classification and analysis [8], [9]. In the considerable data age, high-performance GPUs have enabled mapping a large dataset on a network for enhanced CNN implementation [10]. These aspects have helped minimize human error and variability in medical diagnosis, leading to better patient safety and satisfaction, as well as diagnostic efficiency and accuracy. Despite advancements, several challenges persist:

- **Class Imbalance:** Datasets like ISIC 2020 exhibit significant skewness, with benign lesions vastly outnumbering malignant ones.
- **Noisy Data:** Dermoscopic images often contain artifacts such as hair and uneven lighting, which hinder model performance.
- **Limited Use of Metadata:** Patient-specific metadata, such as age, gender, and anatomical site, is often underutilized despite its diagnostic relevance.

II. Motivation of this study

In a previous study carried out in the lab [12], we discovered that transformers are arguably better in the skin disease classification tasks, as it performed better on a 31-class skin disease classification task on a dataset that combines Atlas Skin disease dataset and ISIC 2018 dataset, which was first used by Abdul Rafay et. al [11]. However, the model raised several false positives while doing predictions for the 'nevus' class, as it predicted them to be 'actinic keratosis'. Also, when transfer learning of the weights was done and was used for prediction on the humans against Machine challenge (the HAM10000 dataset), the model yielded several nevus samples as melanoma, thus classifying a benign rash as malignant. This leads to several concerns and the requirement to focus on building a robust melanoma classifier that can be used for real-time applications.

This study aims to address these challenges by applying Vision Transformers (ViT) and DinoV2 to the ISIC 2020 dataset that solely contains samples that belong to melanoma and benign rashes. ViT, with its self-attention mechanism

after extracting the patch and position embeddings, excels in capturing global image features. In contrast, DinoV2 - a way to train with by leveraging the self-supervised learning (SSL) to extract robust feature representations can yield accurate results. Also, the availability of metadata helps us integrate them to enhance model sensitivity and generalizability. The key contributions of this research include:

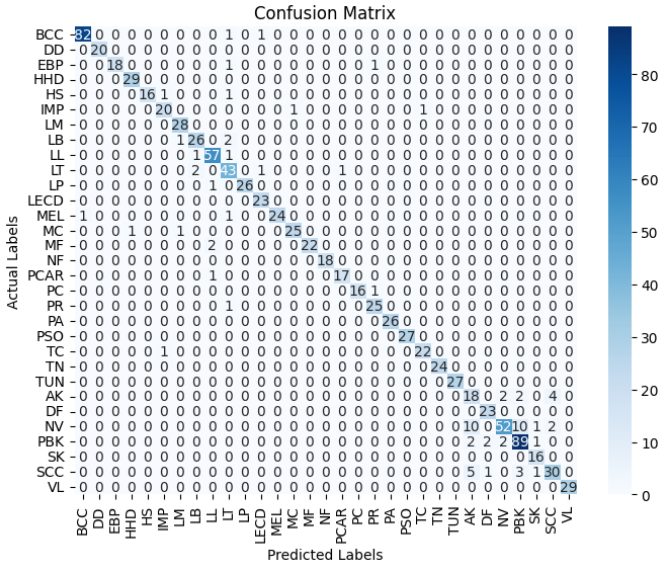


Fig. 1: Confusion matrix for the trained ViT-Base model on unaugmented data.

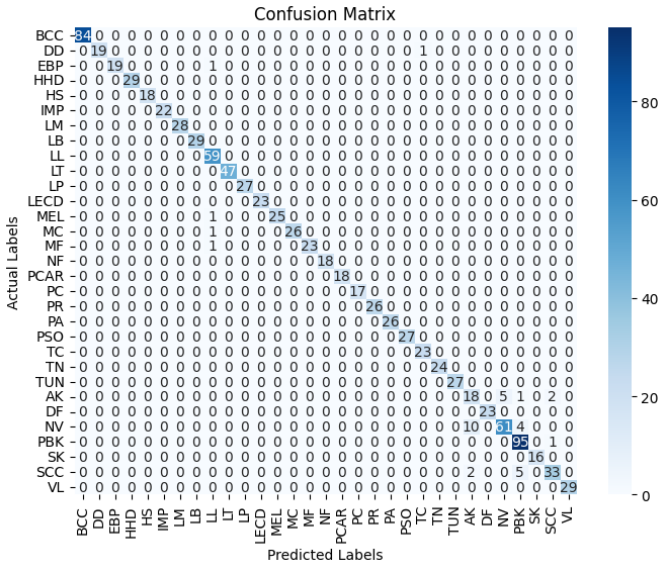


Fig. 2: Confusion matrix for the trained DinoV2-Base model on unaugmented data.

The major contributions of this work are:

- Developing a framework for integrating patient metadata into image-based models for improved classification performance.

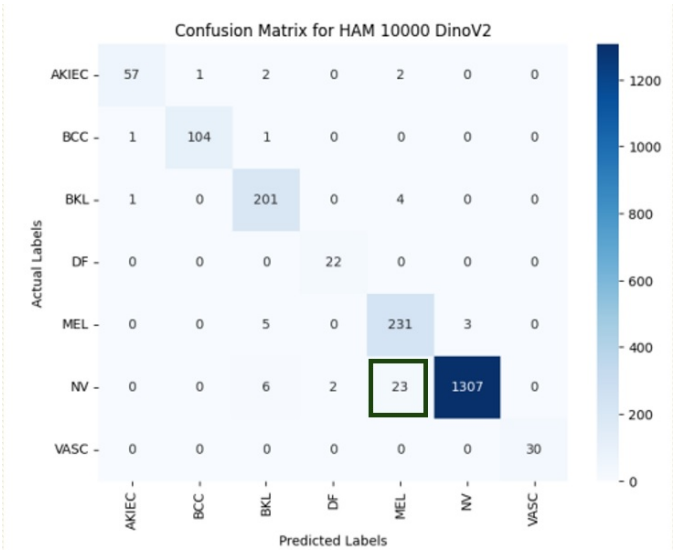


Fig. 3: Confusion matrix of HAM10000 with 'Nevus' misclassified as 'Melanoma'

- Implementing novel augmentation strategies and weighted loss functions to address class imbalance and dataset variability.
- Evaluating the performance of ViT and DinoV2 against state-of-the-art benchmarks with and without metadata for melanoma classificaiton.

The manuscript is structured as follows: Section III contains the related works done in the literature and the relevant gaps discovered and addressed. Section IV outlines the suggested technique, data curation, and experimental setup, whereas Section **??** discusses the outcomes and models for the actual dataset, the explanation for the outputs for selected samples using XAI frameworks, and the outputs of the best-performing transformer architectures for the smaller datasets to test robustness. Section VII concludes the work by summarizing it and describes the advantages that medical professionals could leverage.

III. Related Works

The epidermis protects internal organs, which can become scarred or injured as a result of infections or other causes such as increased pollution and poor nutrition. People frequently disregard warning signs of a skin problem, and the majority of current skin disease detection and treatment approaches rely on clinician-performed biopsies. Due to the difficulty of diagnosing SDCs in clinical settings, the prevalence of skin problems has increased, necessitating timely and precise identification [16]. With the emergence of large-scale datasets such as ISIC 2018, [13] HAM 10000 [14] and Dermnet [18], various works in literature apply deep learning models that can capture accurate features for feature classification using convolution and transformers.

Transformers have shown to be fairly adept at processing complex visual input. Their greater performance over CNNs in a variety of visual tasks has driven this revolution. They have developed into a powerful replacement, processing image

Fig. 4: Sample images of each of the 31 classes (with abbreviations) of the SDC dataset [11].

patches via self-attentional processes. As evidenced by the literature, there has been much research on refining transformer topologies due to their efficiency in tasks such as picture classification, particularly skin disorders.

The models have been trained on smaller benchmark datasets to perform SDC, even though transformer performances on benchmark datasets have improved and some works have used XAI to demonstrate their effectiveness. These datasets solely focus on diseases that are prevalent and do not include all places where diseases occur in the human body or the various geographic areas where these diseases occur. This could result in a rare disease being diagnosed as a well-known illness that shares the same outward signs. Accurately classifying a greater number of diseases with more samples per class using a single transformer model is crucial given the increasing number of skin disease cases that correspond to a variety of infection categories. Thus, this study aims to address the challenges associated with skin disease classification by leveraging state-of-the-art Vision Transformers (ViT) and DinoV2 models. These transformer architectures are applied to the ISIC 2020 dataset, focusing on melanoma and benign rashes. ViT, with its self-attention mechanism, excels in capturing global image features through patch and position embeddings, while DinoV2, employing self-supervised learning (SSL), extracts robust feature representations to enhance classification accuracy. Additionally, the availability of patient metadata is utilized to improve the model's sensitivity and generalizability. This approach ensures a robust methodology that improves prediction accuracy, enables better diagnosis and prognosis, and is extensible to datasets featuring prominent dermatological problems.

Along with experiments with state-of-the-art models, Grad-CAM and SHAP—two XAI frameworks that help dermatologists, physicians, and medical specialists comprehend and visualize the areas of the image prioritized by each transformer to automate diagnosis—are used to expose the black-box nature of the trained models. In addition to giving dermatologists additional information, such as regions of incidence that might be missed because of human error, this would aid in their more accurate diagnosis of the ailment. Additionally, heatmaps can include details on severity, as well as the quantity and rate of spread, following cross-validation with patient clinical data.

IV. Methodology

A. Dataset Description

Abdul Rafay and Waqar Hussain [11] initially curated the dataset by combining the majority classes (categories with more than 80 samples) of the Atlas Dermatology and ISIC 2018 datasets, containing 3,399 and 561 images, respectively, to obtain a total of 4,910 samples. The dataset was split into an 80:20 train-test split. In our study, the train data was further split into a 90:10 split, resulting in an overall train-validation-test split of 72:8:20.

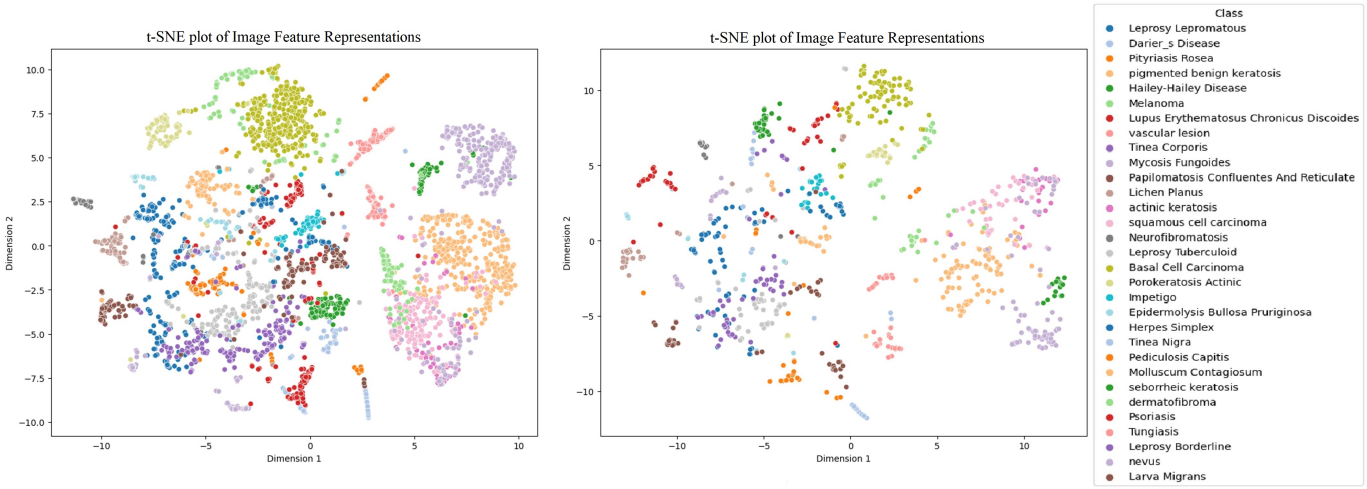


Fig. 5: t-SNE plot of the train (left) and test (right) data.

According to Atlas Dermatology, there were 561 distinct skin problems, some of which had insufficient data to build and train a deep model. There were still just nine or ten samples available for a number of groups. Consequently, a cutoff point was determined to manually curate the dataset, gathering information from classes that had a minimum of 80 cases. Following the filtration process, the dataset contained 3,399 samples in 24 classifications. ISIC 2018, the second source, recognized nine different kinds of skin conditions. Two of these nine categories, meanwhile, were already present in the Atlas Dermatology dataset. The two classes were removed from the nine prior to the merger based on the results of the screening.

TABLE I: Sample distribution of the main dataset.

	Train	Validation	Test	Total
Raw Data	3,524	392	994	4,910
Augmented Data	35,240	3,920	994	49,100

In addition to this dataset, two smaller benchmark datasets with pictures of common skin conditions have also been taken into consideration for evaluating the robustness various transformer designs for the SDC task. The HAM10000 dataset is a vast collection of multi-source dermatoscopic images of common pigmented skin lesions that offers useful resources for classification and research. It includes image samples covering important diagnostic categories such as actinic keratoses and other pigmented lesions. 10,015 photos from 7 classes are included. Dermnet, another dataset, is a library of photos used to identify and categorize different skin conditions. For research and diagnostic reasons, a varied group of dermatologists maintains a 19,500-image, 23-class dataset that includes photos of various skin disorders. Table II contains the number of samples present in the additional datasets that are benchmarked in this work.

Once experiments on the dataset with transformers was performed, and the focus purely shifted to binary class melanoma diagnosis, the ISIC 2020 was considered. This was done as the dataset is a comprehensive collection of 33,126 dermoscopic images of benign and malignant skin lesions from over 2,000 patients, curated to support machine learning challenges in melanoma classification. This dataset includes both melanoma and comparative benign lesion images from the same patients, enabling advanced studies in lesion classification and risk stratification. It is designed to aid researchers in developing models for early melanoma detection, providing unique identifiers for patients to facilitate cross-referencing and data organization. The dataset has been widely used in benchmarks and competitions, showcasing its significance in dermatological research and deep learning applications in medical imaging

TABLE II: Sample distribution of the additional datasets.

Dataset	Train	Validation	Test	Total
HAM10000	7,211	801	2,003	10,015
Dermnet	13,950	1,550	4,000	19,500
ISIC 2020	26,501	3,313	3,312	33,126

B. Exploratory Data Analysis (EDA) for melanoma

EDA on the ISIC 2020 dataset revealed critical patterns and biases in the dataset:

- **Age Distribution:** Patients range from 0 to 90 years, with a higher prevalence of malignant lesions observed in older age groups.
- **Gender Distribution:** Male and female patients are approximately balanced, though certain lesion characteristics vary by gender.
- **Anatomical Site:** Lesions are predominantly located on the torso and lower extremities, with significant variability in malignancy rates across sites.

Visualization techniques, including histograms, pie charts, and heatmaps, were used to understand feature correlations. For instance, age and anatomical site were found to be weakly correlated, suggesting that their independent integration into the model could provide complementary information. Thus, missing samples were filled and the metadata was homogenized before training.

Fig. 6: Visualization of missing data across key features in the ISIC 2020 dataset, highlighting incomplete entries for features like *sex*, *age*, and *anatomy*

C. Preprocessing Pipeline

To prepare the data for training, the following steps were performed:

- **Image Resizing:** Images were resized to 224×224 pixels to match model input requirements.
- **Normalization:** Pixel values were scaled to the range [0,1].
- **Hair Removal:** Morphological operations followed by inpainting were used to remove hair artifacts. This involved applying a blackhat filter with a kernel size of 17×17, followed by inpainting to fill the removed regions:

$$\text{Hair-Free Image} = \text{Inpaint}(I_{thresholded})$$

- **Weighted Loss:** A class-weighted cross-entropy loss function was employed to address class imbalance:

$$\text{Loss} = -\sum_{i=1}^{N} w_i y_i \log(\hat{y}_i)$$

where $w_i = \frac{N}{n_i}$ is the weight for class i, y_i is the true label, and $\hat{y}_i$ is the predicted probability.

D. Transformer Networks used

Transformers have paved the way for the integration of text and image data in multimodal applications by outperforming traditional CNNs in computer vision tasks such as object detection and image classification. Research focuses on improving their architecture, scaling them to larger datasets, and investigating their potential for addressing a range of visual comprehension challenges, including crucial biomedical applications, as they continue to have an impact on the computer vision environment. To our knowledge, no prior research has been conducted using transformers like DinoV2 on a dermatology job, which is what makes the proposed study unique. Furthermore, using the largest SDC dataset, this dataset enabled us to thoroughly examine SDC alongside other well-known transformers.

1) Vision Transformers: Image classification has shown a great deal of interest in ViTs [15] due to their remarkable performance and scalability. By splitting an image into non-overlapping patches and linearly embedding them into a series of tokens, which are then processed by transformer layers after combining with the corresponding position embeddings of the tokens, ViTs are able to capture both local and global dependencies in a single attention mechanism, in contrast to typical CNNs that are excellent at capturing local features through hierarchical convolution layers. The Vision Transformers architecture as initially suggested is explained in Figure 8. Equations 1 and 2 provide the equation of the output calculated by the multi-head self-attention block on the embeddings.

It has been shown that the design can adapt to images of varying sizes without requiring significant changes. High transfer learning capabilities of ViT models pre-trained on large-scale datasets enable them to be fine-tuned on smaller datasets for specific image classification applications. However, they could be computationally expensive and need a lot of training data to function properly. Because ViTs rely on dense attention layers, they typically have higher memory and computational costs despite their advantages. Additionally, their effectiveness frequently depends on the availability of large-scale pre-training datasets to reduce overfitting, particularly when working with smaller or more specialized datasets.

$$\text{MHSA}_{Q,K,V} = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O \tag{1}$$

$$\text{head}_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \tag{2}$$

2) DinoV2: The self-DIstillation with NO labels (DINO) [17] is a sophisticated self-supervised learning approach for training models that improves computer vision by reliably detecting specific objects inside pictures and video frames. Many academics and organizations have concentrated their efforts on self-supervision learning (SSL) models in recent years. They generate labels using a semi-automatic method that entails watching a labelled dataset and estimating part of the data from that batch based on the characteristics. Some SSL systems circumvent these issues by employing DINO, which employs SSL and knowledge distillation methods. It enables extraordinary features to develop, such as robust object component recognition and robust semantic and low-level picture understanding. Figure 9 explains how the choice of curating such a dataset is made.

DINOv2 addresses the issue of training larger models with more data by enhancing stability through regularization approaches inspired by the similarity search and classification literature and incorporating efficient PyTorch 2 and xFormers techniques. The teacher-student model for training is shown in Figure 10. It leads to quicker, more memory-efficient

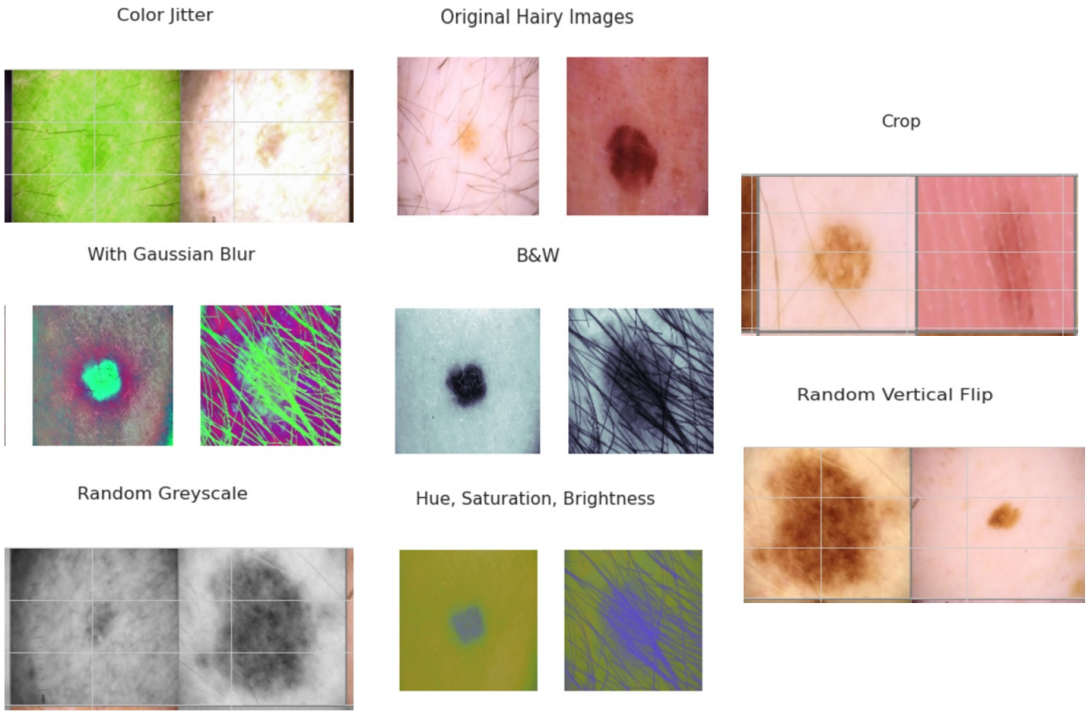


Fig. 7: Augmentation techniques employed for upsampling the train data.

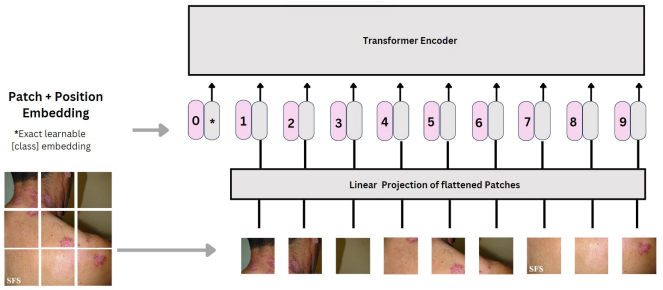


Fig. 8: Architecture diagram of Vision Transformers [15] for SDC.

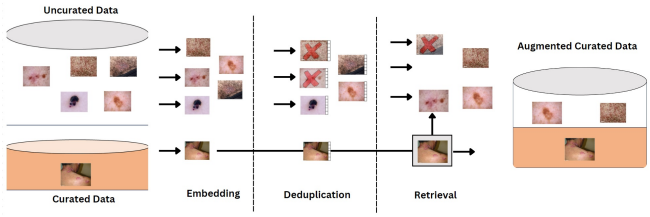


Fig. 9: Data curation for the Semi-supervised learning mechanism of DinoV2.

training with the potential for data, model size, and hardware scaling. In addition to the approaches, the researchers also applied parameters such as the iBOT Masked Image Modeling (MIM) loss term, the curriculum learning strategy to train the models in a meaningful order from low to high-resolution images, softmax normalization, KoLeo regularizers (which improve the nearest-neighbour search task), and the L2-norm for normalizing the embeddings are some of the strategies DINOv2 adopted to improve their results.

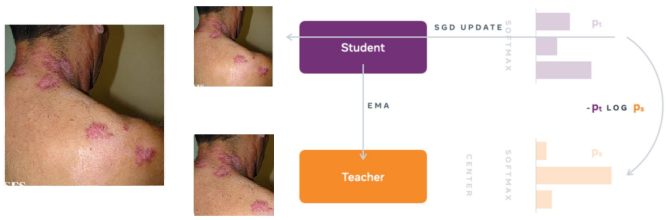


Fig. 10: Teacher student model training approach of DinoV2.

E. XAI for explainability

An important area of AI research and development is explainable artificial intelligence, or XAI, which aims to make AI systems more transparent and interpretable so that humans can understand how they make decisions. The black box problem, which commonly plagues complex machine learning models like deep neural networks, is resolved by XAI. By providing insights into the reasoning behind the predictions or conclusions made by AI systems, XAI fosters a sense of accountability and trust. It also assists users in identifying and reducing biases, errors, and unexpected behaviors in AI applications. From feature attribution to visualization, XAI uses a variety of techniques and methods to help professionals and non-experts alike better understand and use AI systems.

While SHAP provides a more versatile method that can be used with a variety of machine learning models and is especially effective at identifying feature significance, Grad-CAM is especially useful for showing deep neural network

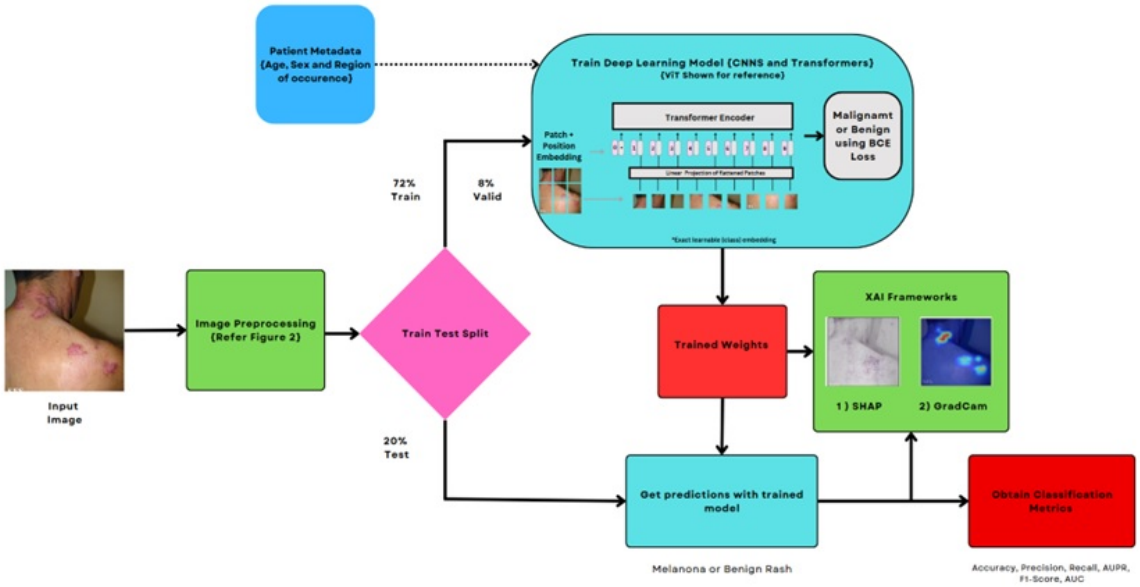


Fig. 11: Methodology of the proposed pipeline of the study.

judgments in image-related tasks. Since the comprehensive approaches provide different but complimentary insights, both tactics advance the broader topic of XAI by enhancing the openness and interoperability of AI systems.

F. Experimental Setup

The experiments were carried out on a system with 16GB RAM and 8GB vRAM GPU(NVIDIA RTX 3060. The images were fed in batches of 8, and trained with the weighted binary cross entropy loss for the melanoma classification task and the categorical cross entropy loss for the 31-class classification task. The equations of the loss functions and the metrics are provided below:

$$\text{Loss}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \quad (3)$$

$$\text{Loss}_{\text{CCE}} = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(p_{ij}) \quad (4)$$

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (5)$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (6)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (7)$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

The models are trained using the methodology in Figure 11

V. Results and Discussion

The results of the work are shown in Table III. Initial experiments were conducted without metadata integration to evaluate the standalone performance of ViT and DinoV2. Integrating metadata significantly improved performance metrics for both models. ViT with metadata integration surpassed existing benchmarks, achieving higher accuracy and F1-scores than CNN-based models reported in the literature. These results underscore the potential of transformer-based architectures for melanoma detection.

TABLE III: Performance Comparison of Models

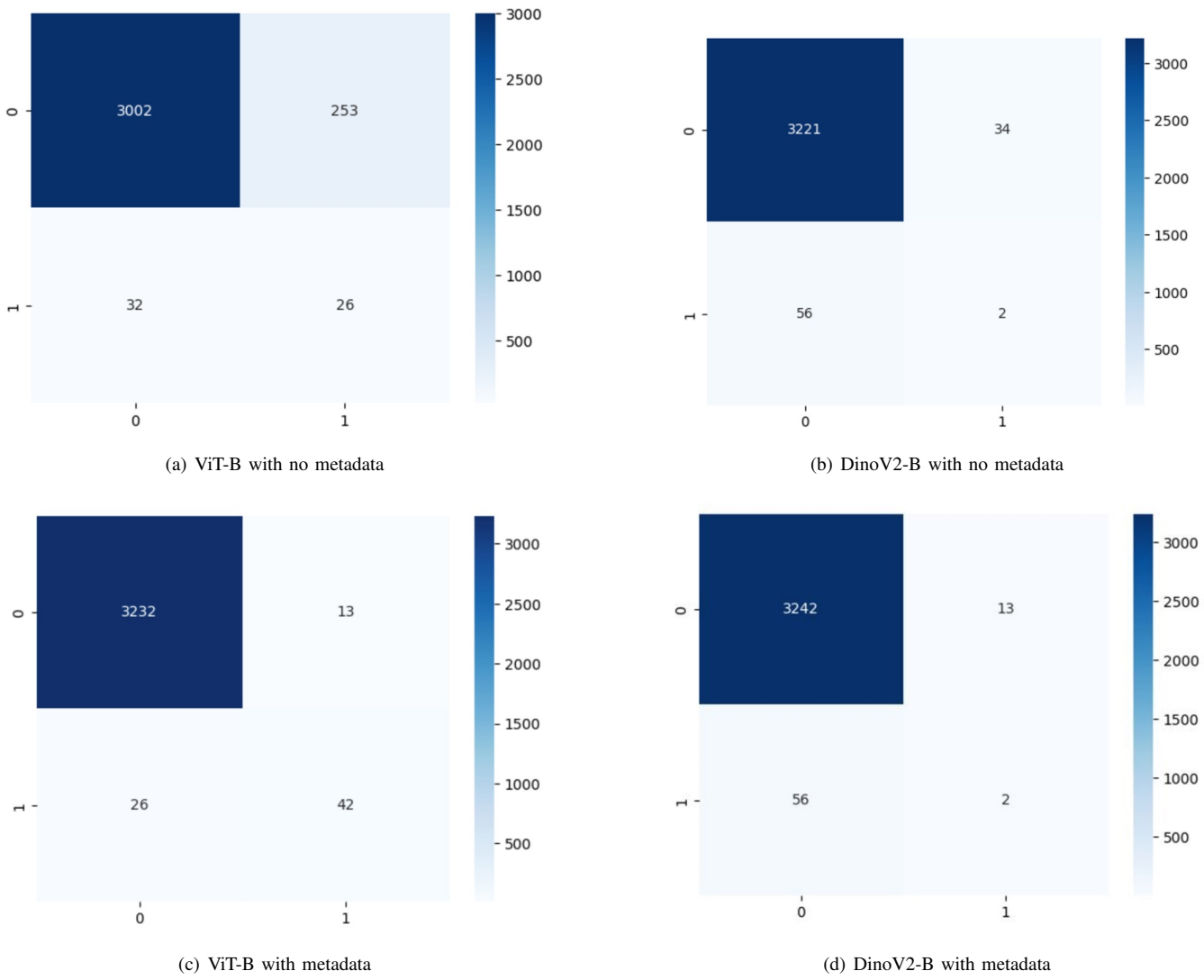
Model	Accuracy	Precision	Recall	F1-Score
ViT (Baseline)	0.9140	0.5413	0.6853	0.5545
DinoV2 (Baseline)	0.9728	0.5192	0.5120	0.5144
ViT with Metadata	0.9903	0.8798	0.8048	0.8383
DinoV2 with Metadata	0.9792	0.5582	0.5152	0.5221

To evaluate the impact of individual components, ablation studies were conducted by selectively removing metadata, augmentation, and weighting techniques. The results demonstrated that metadata integration contributed the most significant performance boost, followed by augmentation.

VI. Discussion

The results demonstrate the effectiveness of transformer-based architectures for melanoma classification, particularly when augmented with metadata and advanced preprocessing techniques. The key findings are as follows:

- **ViT vs. DinoV2:** Vision Transformer consistently outperformed DinoV2 in all metrics, highlighting the importance of pretrained transformers for medical imaging tasks.
- **Metadata Integration:** Adding age, gender, and anatomical site information significantly improved recall and F1-scores, particularly for the minority malignant class.



(a) ViT-B with no metadata

(b) DinoV2-B with no metadata

(c) ViT-B with metadata

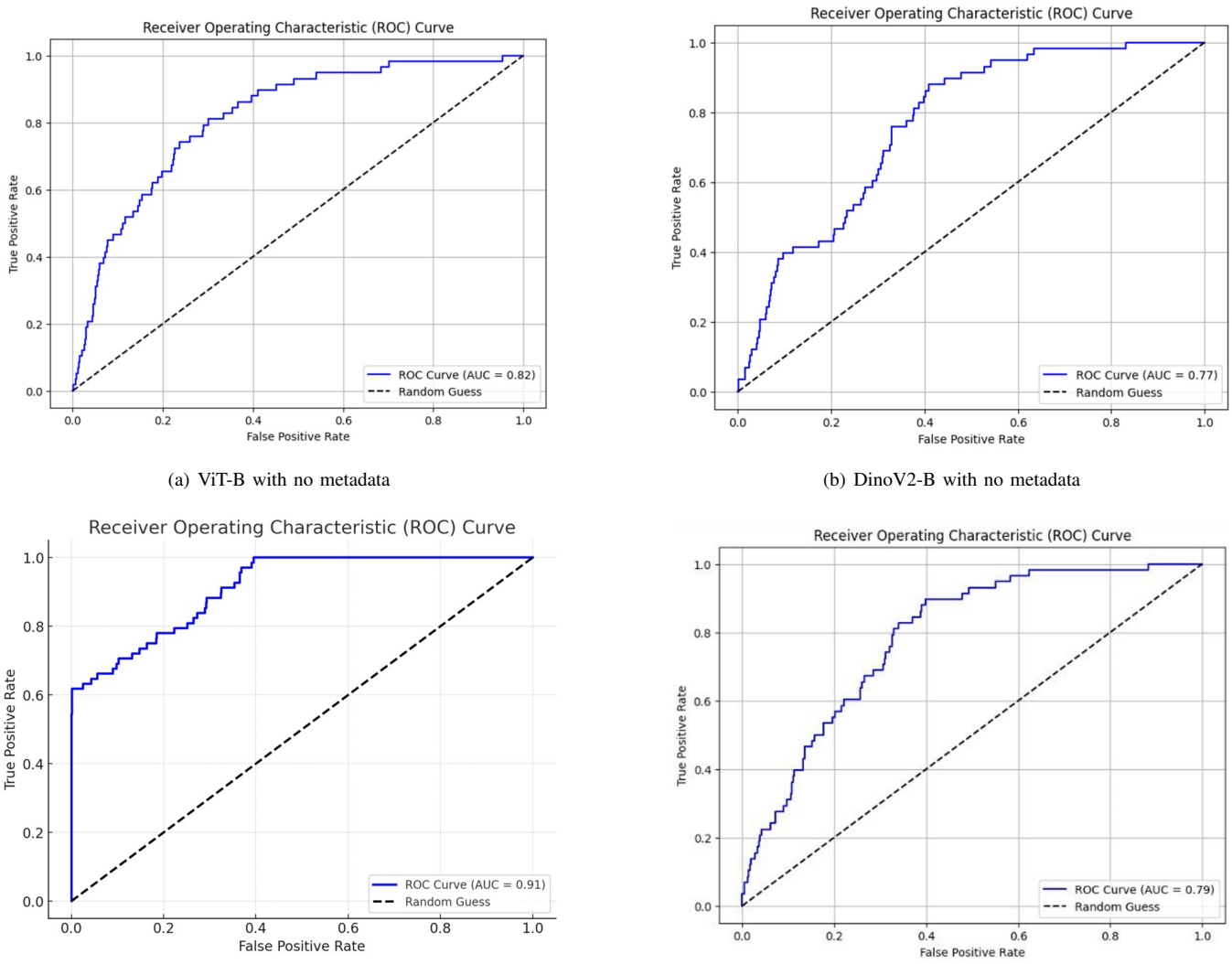
(d) DinoV2-B with metadata

Fig. 12: Confusion matrices for various models and metadata configurations.

- **Augmentation Techniques:** Hair removal and geometric transformations were critical in reducing noise and enhancing model generalization.
- **Class Imbalance:** Weighted loss functions effectively addressed the dataset's imbalance, improving the model's sensitivity to malignant cases.

Figure 14 elucidates the results of the state-of-the art models as mentioned by [19]. In general, it could be infered that the transforms integrate with metadata perform better than the results in the melanoma detection task. The results across the ROC curves, confusion matrices, and performance table show that incorporating metadata improves the performance of ViT (Vision Transformer) models significantly. For ViT-B, the ROC curve jumps from an AUC of 0.82 (no metadata) to 0.91 (with metadata), indicating better discrimination between positive and negative classes. The confusion matrix highlights this improvement, showing an increase in True Positives (TP) and a reduction in False Positives (FP) and False Negatives (FN) when metadata is included. These results are reflected in the performance metrics, where ViT-B with metadata achieves the highest precision (0.8798), recall (0.8048), and F1-score (0.8383).

In contrast, DinoV2 models show only modest improvements with metadata. While the accuracy remains high (0.9728 for the baseline), precision and recall do not improve significantly, with DinoV2 models continuing to struggle with identifying positive instances. The ROC curve and confusion matrix for DinoV2 with metadata indicate limited gains, as the model remains conservative in predicting positives, resulting in low F1-scores and recall. Overall, ViT-B with metadata stands out as the most effective configuration, highlighting that metadata significantly boosts performance for some architectures like ViT, but not equally for all models, such as DinoV2. This is possibly because the DinoV2 model is trained using a SSL approach and it could not handle the sever class imbalance of the ISIC 2020 data.



(a) ViT-B with no metadata

(b) DinoV2-B with no metadata

(c) ViT-B with metadata

(d) DinoV2-B with metadata

Fig. 13: ROC-AUC curves for various models and metadata configurations.

Table 10
A performance comparison of the baseline models on the ISIC 2020 testing set without the use of a pre-trained model, results are reported on their best epoch.

Model	Params	Best epoch	AUC
DenseNet121	7,039,554	25	0.77
DenseNet169	12,646,210	47	0.66
DenseNet201	18,325,826	5	0.63
EfficientNetB0	4,052,133	37	0.75
EfficientNetB1	6,577,801	30	0.76
EfficientNetB2	7,771,387	67	0.73
EfficientNetB3	10,786,609	8	0.75
EfficientNetB4	17,677,409	45	0.65
InceptionResNetV2	54,339,810	10	0.64
InceptionV3	21,806,882	47	0.50
ResNet50	23,591,810	27	0.71
ResNet50V2	23,568,898	41	0.73
ResNet101	42,662,274	21	0.70
ResNet101V2	42,630,658	37	0.76
ResNet152	58,375,042	18	0.67
ResNet152V2	58,335,746	25	0.65
VGG16	134,268,738	21	0.77
VGG19	139,578,434	30	**0.80**
Xception	20,865,578	16	0.75

Fig. 14: State of the art results on the ISIC 2020 dataset

VII. Conclusion and Future Work

This study highlights the transformative potential of Vision Transformers (ViT) and self-supervised models like DinoV2 in advancing melanoma detection through automated classification of dermoscopic images. By leveraging metadata such as age, gender, and anatomical site, we demonstrated significant improvements in the sensitivity and precision of the models, particularly for the minority malignant class.

ViT emerged as the superior model, achieving an accuracy of 99.03% and an F1-score of 83.83%, outperforming both DinoV2 and previously reported CNN-based approaches. The integration of metadata proved pivotal, enhancing the model's contextual understanding and improving its ability to detect malignant lesions. Augmentation techniques such as hair removal, geometric transformations, and color jittering further contributed to the robustness and generalizability of the models.

While the results are promising, the study also revealed limitations, such as the challenges DinoV2 faced with class imbalance and the dependency on metadata availability. These findings provide a strong foundation for future research aimed

at addressing these challenges and further improving automated skin cancer detection systems.

Building upon the findings of this study, the following avenues for future research are proposed:

- **Cross-Dataset Validation:** Evaluate the models on external datasets such as HAM10000 and DermNet to assess generalizability across diverse populations and imaging conditions.
- **Explainable AI:** Develop interpretable frameworks using techniques like SHAP, Grad-CAM, and LIME to provide clinicians with insights into model predictions and build trust in AI-driven diagnostics.
- **Few-Shot Learning:** Explore methods to improve model performance on underrepresented classes by incorporating few-shot learning or advanced oversampling techniques like SMOTE or ADASYN.
- **Multimodal Fusion:** Investigate hybrid models that combine image-based embeddings with text or electronic health record (EHR) data to enhance diagnostic accuracy further.
- **Real-Time Deployment:** Optimize model architectures for real-time deployment in clinical settings, focusing on computational efficiency and robustness to noisy inputs.

FUNDING

No funding was necessary to conduct this work.

CONFLICT OF INTEREST

Few of the contents presented in this report the experiences dealing with the 31 disease multiclass classification are directly taken from experiments conducted by team member Arrun Sivasubramanian in his previous institution (Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India), and the work is under review in Computers in Biology and Medicine (https://doi.org/10.48550/arXiv.2407.14757). However, all experiments with the binary classification of melanoma on the ISIC2020 dataset are novel.

REFERENCES

[1] James, W., Elston, D. & Berger, T. Andrew's diseases of the skin E-book: clinical dermatology. (Elsevier Health Sciences,2011)

[2] Roky, Amdad Hossain, et al. "Overview of skin cancer types and prevalence rates across continents." Cancer Pathogenesis and Therapy 2 (2024)

[3] Hay, R., Augustin, M., Griffiths, C., Sterry, W., Dermatological Societies, B., Grand Challenges Consultation groups, Abuabara, K., Airoldi, M., Ajose, F., Albert, S., Armstrong, A. & Others The global challenge for skin health. *British Journal Of Dermatology*. **172** pp. 1469-1472 (2015)

[4] Langemo, D. & Brown, G. Skin fails too: acute, chronic, and end-stage skin failure. *Advances In Skin & Wound Care*. **19**, 206-212 (2006)

[5] Xu, H. & Li, H. Acne, the skin microbiome, and antibiotic treatment. *American Journal Of Clinical Dermatology*. **20**, 335-344 (2019)

[6] Inthiyaz, S., Altahan, B., Ahammad, S., Rajesh, V., Kalangi, R., Smirani, L., Hossain, M. & Rashed, A. Skin disease detection using deep learning. *Advances In Engineering Software*. **175** pp. 103361 (2023)

[7] Yann LeCun, Yoshua Bengio, Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[8] Yue, W., Liu, S. & Li, Y. Eff-PCNet: An Efficient Pure CNN Network for Medical Image Classification. *Applied Sciences*. **13**, 9226 (2023)

[9] Zhou, Q., Huang, Z., Ding, M. & Zhang, X. Medical image classification using light-weight CNN with spiking cortical model based attention module. *IEEE Journal Of Biomedical And Health Informatics*. **27**, 1991-2002 (2023)

[10] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H. & Ni, B. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*. **10**, 41 (2023)

[11] Rafay, A. & Hussain, W. EfficientSkinDis: An EfficientNet-based classification model for a large manually curated dataset of 31 skin diseases. *Biomedical Signal Processing And Control*. **85** pp. 104869 (2023)

[12] Mohan, Jayanth, et al. "Enhancing skin disease classification leveraging transformer-based deep learning architectures and explainable ai." arXiv preprint arXiv:2407.14757 (2024).

[13] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M. & Others Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *ArXiv Preprint ArXiv:1902.03368*. (2019)

[14] Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*. **5**, 1-9 (2018)

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

[16] Gourav Ganesh and K. Somasundaram,*Detect Melanoma Skin Cancer Using An Improved Deep Learning CNN Model With Improved Computational Costs*, 2023.

[17] Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." arXiv preprint arXiv:2304.07193 (2023).

[18] Bajwa, M. N., Muta, K., Malik, M. I., Siddiqui, S. A., Braun, S. A., Homey, B., Dengel, A., & Ahmed, S. (2020). Computer-aided diagnosis of skin diseases using deep neural networks. Applied Sciences, 10(7), 2488.

[19] Cassidy, Bill, et al. "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations." Medical image analysis 75 (2022): 102305.