Computers in Biology and Medicine 190 (2025) 110007

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Enhancing skin disease classification leveraging transformer-based deep learning architectures and explainable AI

Jayanth Mohan [a,1], Arrun Sivasubramanian [a,1,*], Sowmya V. [a], Vinayakumar Ravi [b]

[a] Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India
[b] Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia

ARTICLE INFO

Keywords:
Skin disease classification
Vision transformers
Swin transformers
DinoV2
GradCAM
SHAP

ABSTRACT

Skin diseases affect over a third of the global population, yet their impact is often underestimated. Automating the classification of these diseases is essential for supporting timely and accurate diagnoses. This study leverages Vision Transformers, Swin Transformers, and DinoV2, introducing DinoV2 for the first time in dermatology tasks. On a 31-class skin disease dataset, DinoV2 achieves state-of-the-art results with a test accuracy of 96.48 ± 0.0138% and an F1-Score of 97.27%, marking a nearly 10% improvement over existing benchmarks. The robustness of DinoV2 is further validated on the HAM10000 and Dermnet datasets, where it consistently surpasses prior models. Comparative analysis also includes ConvNeXt and other CNN architectures, underscoring the benefits of transformer models. Additionally, explainable AI techniques like GradCAM and SHAP provide global heatmaps and pixel-level correlation plots, offering detailed insights into disease localization. These complementary approaches enhance model transparency and support clinical correlations, assisting dermatologists in accurate diagnosis and treatment planning. This combination of high performance and clinical relevance highlights the potential of transformers, particularly DinoV2, in dermatological applications.

1. Introduction

Human skin serves various functions, including protecting the human body from contaminants, heat, and UV radiation [1]. Skin disorders are significantly more common than we know, with skin and subcutaneous disease impairment accounting for 4. 02% of the total cases of disability in India in 2017 [2]. These skin diseases are growing increasingly hazardous as time passes. Dermatologists believe that the injury can be addressed if it is recognized in time, but things can become tricky when they rely on manual approaches alone to identify diseases. The fundamental reason for this is that there are many types of diseases. Furthermore, physical diagnosis might be challenging because many skin diseases have similar visual characteristics that further increase difficulty in diagnosis and suggesting medical treatment [3].

The severity and symptoms of these skin problems vary greatly, with some skin diseases being hereditary while external 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.

Deep learning (DL) techniques, especially convolutional neural networks (CNNs), have been essential in unsupervised feature extraction from images in recent years [6]. Many academics have created many CNN designs to improve the performance in domains with high availability and diverse annotated data [7], and they have also played an essential part in medical image-based classification and analysis [8,9]. In the big data era, high-performance GPUs have also enabled mapping a big dataset on a network for improved CNN implementation [10]. All these factors have helped reduce human error and variability in medical diagnoses, leading to improved patient safety and satisfaction while enhancing diagnostic efficiency and accuracy,

Following extraordinary success on natural language tasks, transformer neural networks have been effectively applied to various computer vision challenges, yielding state-of-the-art results and pushing academics to reassess the dominance of CNNs [11]. Taking advantage of developments in computer vision, the medical imaging profession has seen increased interest in transformers that can capture global

* Corresponding author.
E-mail addresses: jay.thinkai@gmail.com (J. Mohan), arrun.sivasubramanian@gmail.com (A. Sivasubramanian), v_sowmya@cb.amrita.edu (Sowmya V.), vravi@pmu.edu.sa (V. Ravi).
[1] Equal Contribution

https://doi.org/10.1016/j.compbiomed.2025.110007
Received 19 July 2024; Received in revised form 27 January 2025; Accepted 5 March 2025
Available online 20 March 2025

Fig. 1. Sample images of each of the 31 classes (with abbreviations) of the SDC dataset [17].

context as opposed to CNNs with local receptive fields [12,13]. Though works [14–16] have explored transformers for SDC, their study is limited to models that classify skin diseases for a small corpus. They are also trained on data containing samples belonging to fewer classes, which limits the diversity of the diseases in the study. As demonstrated in our work, the models they use alone cannot capture diversity in the distribution of diseases. The introduction of transformer architectures such as DinoV2 in the computer vision community warrants its utilization for complex and critical dermatological tasks such as SDC, which could help the general public as well as dermatologists in terms of time and resources. Moreover, the works do not provide any insight into the extent of the spread of the disease that could further help determine factors like severity or rate of spread of the disease.

Thus, this work addresses leveraging transformer architectures, such as Vision Transformers (ViT), Swin Transformers, and DinoV2, to classify diverse skin diseases. All the variants of these models are trained and tested on a dataset containing 31 skin diseases and their augmentation to overcome regularization and assist with data-limited classes to perform a comprehensive analysis. The samples of each class of the overall dataset are shown in Fig. 1. Since the DinoV2 model was recently introduced, the model performance has also been evaluated for other benchmark SDC datasets, such as HAM10000 and Dermnet, to test the robustness of the model on smaller datasets focusing on a relatively lesser number of classes, yet widespread skin diseases. The authors believe the suggested study's practical impact is extremely valuable to doctors and the medical industry. The model's excellent best test accuracy of 96.48% and F1-Score of 97.28% (an improvement of approximately 10% in accuracy and F1-Score over existing results) can aid these organizations in improving their ability to diagnose skin problems and offer patients more effective treatments. The interpretability of the results using the explainable AI (XAI) outputs, such as GradCAM and SHAP, obtained for test samples on the top-performing models used in this work can additionally guide dermatologists to perform clinical correlations and determine all the regions of occurrence. The dermatologists and the research community can utilize the results of this study to develop a mobile application for health organizations to swiftly and correctly identify skin problems, saving time and resources while increasing patient satisfaction with improved diagnosis and treatment. The major contributions of this work are:

- Utilizing DinoV2 for the first time for a skin disease diagnosis task and other transformers and CNNs, achieve state-of-the-art classification on a 31-class SDC dataset [17] for diverse skin ailments, ensuring accurate and rapid diagnoses.
- Perform a comprehensive comparative analysis on ConvNeXt and other popular CNN architectures, alongside all variants of three transformer architectures - ViT, Swin Transformers and DinoV2 on the augmented and unaugmented datasets.
- Evaluate the robustness of the proposed methodology by validating the performance on two smaller benchmark datasets: the HAM10000 and Dermnet datasets, with fewer samples and popular skin diseases.
- Including the XAI results - SHAP and GradCAM, which help to demystify the black-box nature of AI algorithms and assist dermatologists with the accurate and early diagnosis of skin diseases. The region-level heat maps and pixel correlation coefficient plots also aid in determining the exact regions of infection, possibly giving more insights into the severity of the infection and developing efficient treatment plans.

The manuscript is structured as follows: Section 2 contains the related works done in the literature and the relevant gaps discovered

and addressed. Section 3 outlines the suggested technique, data curation, and experimental setup, whereas Section 4 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 5 concludes the work by summarizing it and describes the advantages that medical professionals could leverage. It also elucidates the limitations of the work and the future scope of improvement. Finally, sources utilized to identify literature are included in the final section of the manuscript.

2. Related works

The epidermis shields internal organs, which can get scarred or damaged due to infections or other factors such as worsening pollution and unhealthy diet. People commonly ignore the warning indications of a skin condition, and most current procedures for detecting and treating skin diseases rely on biopsies performed by a clinician. Since SDCs might be challenging to diagnose in a clinical context, the frequency of skin disorders has been growing, demanding quick and accurate detection [18]. With the introduction of large-scale datasets such as ISIC 2018, [19] HAM 10000 [20] and Dermnet [21], several works in literature utilize deep learning models that can capture accurate features for feature classification with convolution and transformers. A proper diagnosis, assisted by these model predictions, can aid in the recovery from such ailments.

Karthik et al. [22] developed Eff2Net, a CNN that employs a channel attention block called ECA rather than the typical module to identify skin diseases. They evaluated the model on four diseases to obtain an accuracy of 84.70%. Hossen et al. [23] built a unique dataset of four dermatological diseases and compared a novel CNN with previous benchmark techniques. Image augmentation was also used to increase the size of the database and the model's scope. The model demonstrated good accuracy for the diseases. The combination of CNN-based SDC and a federated learning methodology provides an efficient way to classify skin diseases while protecting data. It motivated us to determine if augmentation additionally boosts results for the primary dataset.

Andre Esteva et al. [24] fine-tuned all the layers of InceptionV3 on a composite dataset to report a 72.1% accuracy on the HAM10000 dataset. Kshirsagar et al. [25] created a cutting-edge solution identifying skin problems with LSTM and MobileNetV2. The main goal of this research was to perform SDC correctly and determine if a hybrid technique can aid in preventing people. Though Saket S. Chaturvedi et al. [26] initially attempted to classify the HAM 10000 dataset using the ResNet101 backbone for feature extraction, they yielded better results, with an accuracy of 91.47%. An improvisation was suggested by Anand et al. [27], who suggested a pre-trained Xception model with transfer learning capability. The model was trained and tested on the HAM10000 dataset, classifying skin disorders with an accuracy of 96.40%. With an accuracy of 99%, the suggested model did exceptionally well in diagnosing Benign Keratosis. This strategy can help people and clinicians determine if medical intervention is required. Nevertheless, the authors of [28] proposed a fine-tuned Xception architecture to get high accuracy and an F1-score of 96% on the 7-class MNIST HAM 10000 dataset, with data augmentation applied to prevent the class imbalance problem prevalent in the dataset to boost the results.

Hameed et al. [29] proposed an intelligent diagnostic technique for a more attractive cutaneous lesions class. The proposed approach was realized through hybrid techniques: error-correcting CNN and outcome codes based on a usable support vector machine. The study makes use of 9144 images acquired from public sources. AlexNet, a CNN-approved approach, was used to extract the feature. Filali et al. [30] used the PH2 dataset to detect melanoma using pre-trained and trained-from-scratch CNN models. They also applied preprocessing on the input image fed to the CNN using the Otsu algorithm to report an accuracy of 87.8%. A similar study was carried out by Ly et al. [31] with a model trained from scratch with a balanced PHDB dataset for classifying malignant skin cancer, with a reported accuracy of 86% even without a publicly available HAM 10000 during their experimentation. There are works using ResNet [32] and ResUNet [33], which get satisfactory results for the SDC task.

Some studies utilize private and custom datasets for SDC. Velasco et al. [34] introduced a model utilizing MobileNet for finding skin lesions with accuracy enhanced by using novel sampling strategies and preprocessing of input data. It was 84.28% accurate using basic sampling methods. The accuracy was 93.6%, with a skewed dataset and typical input record preparation. When oversampling in the dataset was found, the model's accuracy climbed to 91.8%. Voggu and Rao [35] suggested research in which three separate skin diseases would be detected using a novel approach. In this methodology, images of the skin are first preprocessed using filtering and alteration to reduce noise and undesired heredity.

Since it was required to create automated methods for boosting analysis accuracy for various skin types and psoriasis symptoms, deep neural algorithms have been used to detect skin problems automatically. Bhavani et al. [36] suggested a method for identifying various skin problems. Three examples, Mobile Net, Inception V3, and ResNet, are trained on a collection of machine learning features, notably logistic regression. Integrating the three CNNs in a hybrid architecture can result in excellent performance, though it reduces the evaluation's space and time complexity. The authors of [17] were among the pioneers to do diverse work on a combined 31-class dataset, obtained by merging the non-overlapping and high sample quantity classes of two SDC datasets: Atlas Dermatology and ISIC 2018, containing 26 and 8 classes, respectively. The authors claim that the class count in the combined dataset is much higher than the benchmark datasets proposed in the literature. The results show that the EfficientNetB2 model performed the best with an 87.15% accuracy for 31 classes of the augmented dataset.

Transformers have shown to be quite adept at handling complicated visual data. Their superior performance over CNNs in various visual tasks has been the driving force behind this revolution. They have become a potent substitute, processing image patches through self-attentional processes. There have been a great deal of studies on improving transformer topologies due to their effectiveness in tasks like image classification, including skin diseases, as evidenced by the literature. Cai et al. [14] demonstrated a multimodal transformer for categorizing skin disorders. The architecture comprises dual encoders for images and metadata and a decoder for fusing the multimodal data. The proposed network employs a Vision Transformer model to extract deep features from images and incorporate metadata that serves as soft classification labels. In the decoder, the attention mechanism aids in the fusion of image and metadata characteristics. It performs well with an accuracy of 93.81%, an improvement over state-of-the-art methods by 1% on the ISIC dataset, making it a viable method for identifying skin diseases.

Aladhadh, Suliman, et al. [37] were among the first to suggest Medical-VIT for Skin disease classification. Mild geometric and brightness-contrast-based augmentations helped their model fetch a test accuracy of 96.14%. However, inspired by the work of [38], LesionAid [15], a novel multiclass prediction framework that classifies skin lesions based on ViT-GAN used GAN was used as an up-sampling algorithm to extract the genuine representation of the data from the raw images and synthesize new images to tackle the class imbalance problem. A model fine-tuned on such a synthetically up-sampled task yielded an immaculate validation accuracy of 97.4% for classifying the HAM 10000 dataset using Vision transformers. The results were also closely followed by the one trained on Swin Transformers and its variants for the ISIS 2018 dataset by Selen Ayas [16] to get an accuracy of 97.2% using a weighted CCE loss in the Large22K model.

Despite the improvement in the results of transformers on benchmark datasets and a few works using XAI to prove their efficiency, the

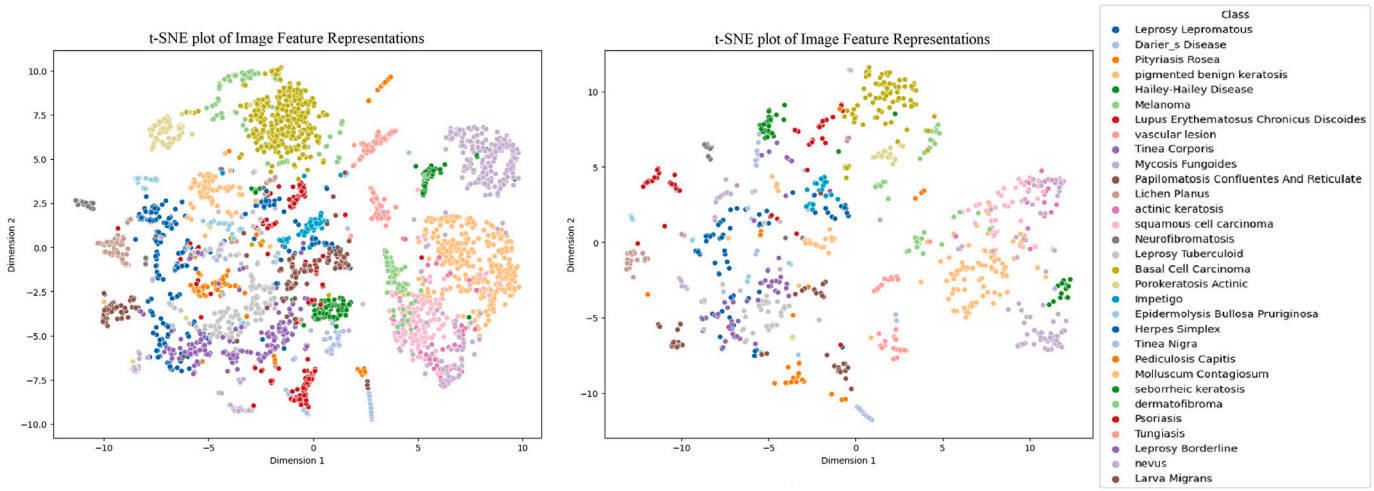

Fig. 2. t-SNE plot of the train (left) and test (right) data.

models have been trained on smaller benchmark datasets to perform SDC. These datasets do not capture all regions in which diseases occur in the human body and different geographical areas of occurrence of these diseases and focus only on prominently occurring diseases. This may lead to diagnosing a rare disease as a popularly known disease that exhibits the same visible symptoms. With the growing number of skin disease cases belonging to a diverse category of infections, it is quintessential to accurately classify a much larger number of diseases containing more samples per class with a single transformer model. To the best of our knowledge, no study in the literature has used a complex transformer architecture like DinoV2 for a dermatology task. Thus, this work utilizes state-of-the-art transformers and performs transfer learning to improve prediction accuracy for a diverse 31-class dataset to improve the quality of diagnosis and prognosis of dermatological diseases. These results are compared with the dataset's benchmark results produced by CNN architectures. The robustness of the proposed methodology is also tested by fine-tuning the model on other smaller datasets focusing on prominent dermatological problems.

In addition to experiments with state-of-the-art models, the black-box nature of the trained models is unraveled with the help of GradCAM and SHAP—two XAI frameworks that help dermatologists, doctors, and medical experts understand and visualize the regions of the image prioritized by each transformer to automate the diagnosis. This would assist them in diagnosing the disease more accurately and assist dermatologists with additional information like regions of occurrence that could be neglected because of human error. Additionally, getting information on severity using heatmaps and the extent and rate of spread can aid in administering treatment after cross-validating patient clinical records.

3. Methodology

3.1. Dataset description

Abdul Rafay and Waqar Hussain [17] initially curated the dataset by combining the majority classes (categories with more than 80 samples) of the Atlas Dermatology and ISIC 2018 datasets, containing 3399 and 561 images, respectively, to obtain a total of 4910 samples. The dataset was split into an 80:20 train-test split. In our study, the train data was further divided into a 90:10 split, resulting in an overall train-validation-test split of 72:8:20.

There were 561 different skin conditions listed in Atlas Dermatology, some of which lacked inadequate data to train and construct a deep model due to the scarcity of data. Even yet, just 9 to 10 samples were available for several classes. As a result, a threshold was

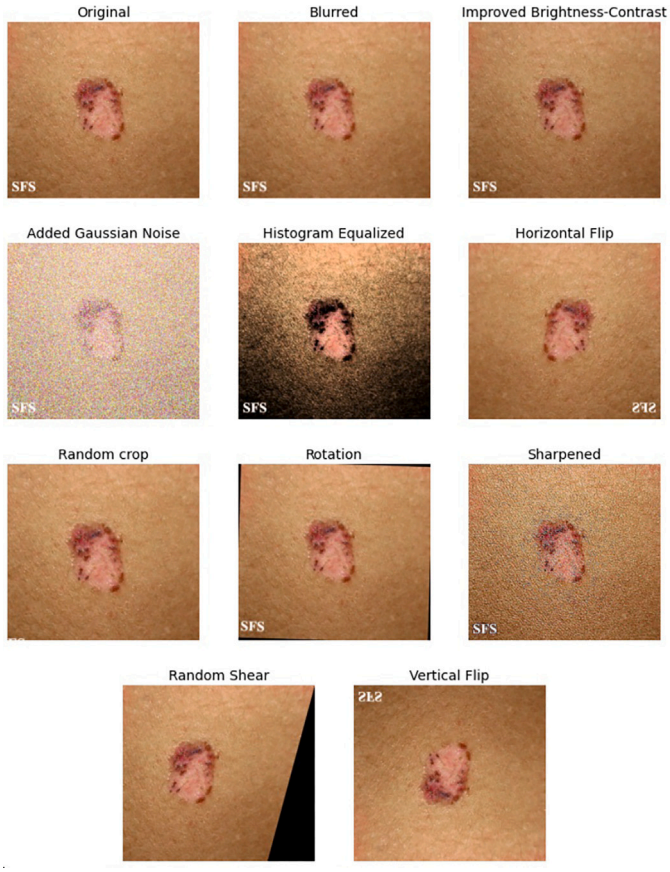

Fig. 3. Geometric augmentations used to upsample the dataset.

established to curate the dataset manually, collecting data from classes with at least 80 examples. The dataset had 24 classes containing 3399 samples after the filtration process. The second source, ISIC 2018, listed nine types of skin ailments. However, two of these nine classifications previously existed in the Atlas Dermatology dataset. Following filtration results, the two classes were omitted from the nine before the merger. The data from both sources was combined into a single dataset, and the resultant dataset had 31 classes and 4910 samples in total.

However, the distributions followed by these data samples are slightly different, as evident from the train and test T-Stochastic Nearest

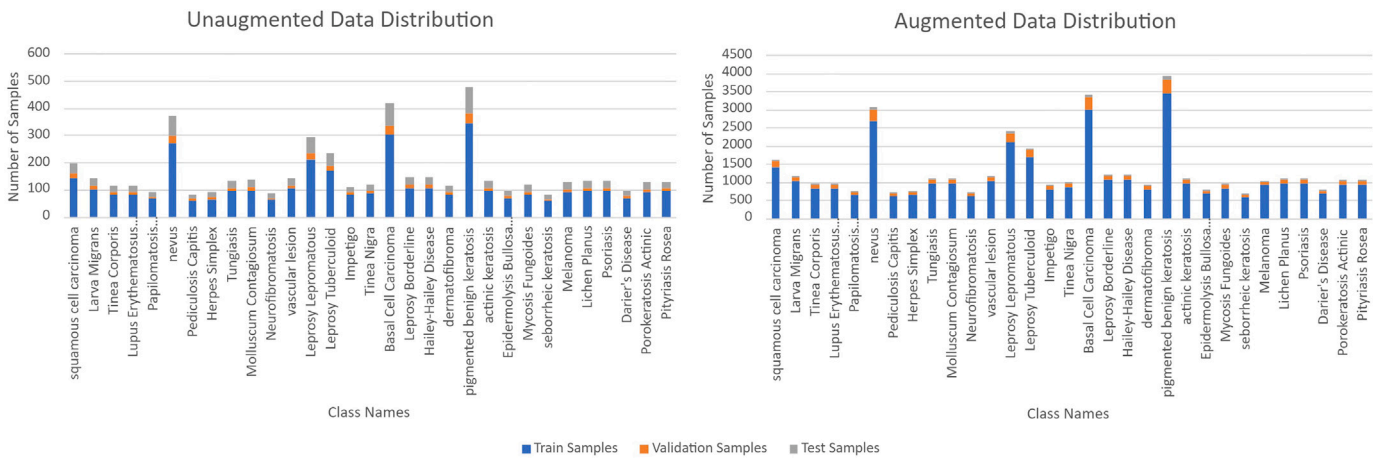

Fig. 4. Train-Validation-Test data distribution for the unaugmented/raw and augmented datasets.

Table 1
Sample distribution of the main dataset.

	Train	Validation	Test	Total
Raw data	3524	392	994	4910
Augmented data	35,240	3920	994	49,100

Table 2
Sample distribution of the additional datasets.

Dataset	Train	Validation	Test	Total
HAM10000	7211	801	2003	10,015
Dermnet	13,950	1550	4000	19,500

Embedding (t-SNE) plots in Fig. 2. The train distribution has samples of the same class that are more cluttered together, indicating that training a model to classify samples from different classes would not be difficult. However, the test dataset t-SNE plot shows samples more distributed in space, indicating a difficult linear separability. The dataset can be oversampled using augmentations to make the model more robust. Thus, ablation experiments with ten different types of augmentations: Vertical and Horizontal Flipping, Random Shear, Sharpening, Random Rotation, Center Crop, Brightness, and Contrast variation, Histogram Equalization, Gaussian Noise, and blurring were used to up-sample the training dataset to oversample the train data, to determine if the attempt improves the overall test accuracy. The appearance of samples post data augmentation for a randomly chosen sample belonging to the "Basal Cell Carcinoma" class is shown in Fig. 3. The insights into the number of samples in different partitions of the data are mentioned in Table 1, and the data distribution for each class for the split is mentioned in Fig. 4.

Apart from this dataset, two smaller benchmark datasets have also been considered for analyzing the robustness of transformer architectures for the SDC task that contain images of popular skin diseases. The HAM10000 dataset, which comprises image samples covering important diagnostic categories like actinic keratoses and other pigmented lesions, is a large collection of multi-source dermatoscopic images of common pigmented skin lesions, providing valuable resources for research and classification purposes. It contains 10,015 images belonging to 7 classes. Another dataset called Dermnet is a collection of images used for the localization and classification of various skin diseases. A diverse group of dermatologists maintains a 23-class dataset with 19,500 images and contains images representing different skin conditions for research and diagnostic purposes. Table 2 includes the number of samples present in the additional datasets that are benchmarked in this work.

3.2. Transformer networks used

Transformers have outperformed classic CNNs in image classification, object identification, and other computer vision tasks, opening the way for integrating textual and visual information in multimodal applications. As they continue to impact the computer vision environment, research focuses on refining their design, scaling them to bigger datasets, and examining their potential for tackling various visual comprehension difficulties, including essential biomedical applications. What makes the proposed study unusual is no previous research has been undertaken utilizing transformers such as DinoV2 on a dermatology task, to our knowledge. Moreover, this dataset helped us comprehensively analyze SDC with other popular transformers on the biggest SDC dataset. Thus, in addition to the benchmark convolution networks used in the literature, the following transformers were trained on the three datasets to validate their performance on the test data split and use the metrics for the comparative analysis.

3.2.1. Vision transformers

Because of their exceptional performance and scalability, ViTs [39] have received much interest in image classification. Unlike typical CNNs that excel at capturing local features through hierarchical convolution layers, ViTs can capture regional and global dependencies in a single attention mechanism as they divide an image into non-overlapping patches and embed them linearly into a series of tokens, which are subsequently processed by transformer layers after combining with the corresponding position embeddings of the tokens. Fig. 5 explains the Vision Transformers architecture as initially proposed. The equation of the output computed by the multi-head self-attention block on the embeddings is given in Eqs. (1) and (2). It enables ViTs to record long-term relationships and contextual information over the whole image, making them helpful in dealing with complicated visual patterns.

The architecture has demonstrated the ability to handle changing-size visuals without requiring substantial architectural adjustments. ViT models pre-trained on large-scale datasets have demonstrated high transfer learning capabilities, allowing for fine-tuning on smaller datasets for specialized image classification tasks. On the other hand, they may be computationally costly and require a large quantity of training data to work well. Despite their benefits, ViTs tend to have higher memory and computational costs due to their reliance on dense attention layers, and their performance often hinges on the availability of large-scale pre-training datasets to mitigate overfitting, especially when working with smaller or more specific datasets. Nonetheless, they are a promising trend in image classification and are constantly improving, with researchers investigating different architectural enhancements and training strategies to increase their performance.

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

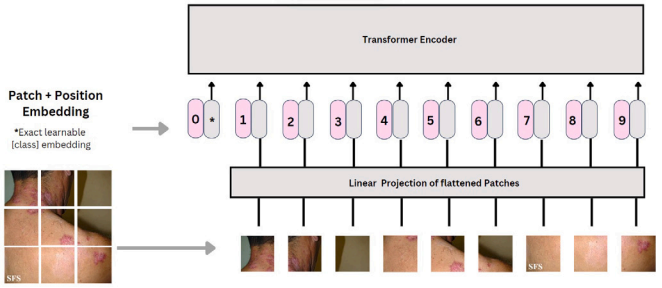

Fig. 5. Architecture diagram of Vision Transformers [39] for SDC.

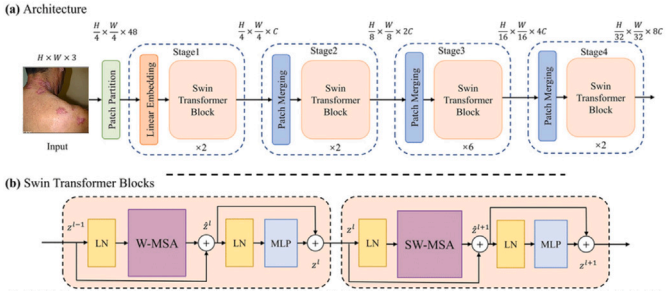

Fig. 6. Architecture diagram of Swin Transformers [40] for SDC.

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

3.2.2. Swin transformers

Swin Transformers [40] is yet another novel way of image categorization that has shown to be quite effective. Swin Transformers overcome some of the limitations of classic CNNs and ViTs by employing a hierarchical design that effectively gathers local and global information. Like ViTs, Swin Transformers also divide the image into non-overlapping patches, but they employ a hierarchical design with many stages like convolution. Each stage comprises a series of transformer layers that analyze data at various spatial resolutions, and they incorporate a sliding window mechanism to enhance their performance further, using shifted windows to capture more granular details. Specifically, after processing non-overlapping windows of patches in one layer, the windows are shifted by a certain amount in the next layer, allowing information from neighboring patches to be aggregated across layers, effectively increasing the receptive field and improving the model's ability to capture both local and global context, without missing features in kernels edges. Fig. 6 explains the Swin Transformers architecture as originally proposed.

One of Swin Transformers' primary advantages is its computational efficiency. Its linear complexity, compared to ViT's quadratic complexity, lowers total computing costs while retaining comparable performance by processing information hierarchically. As a result, it is more suitable for real applications requiring minimal processing resources. The model has demonstrated outstanding performance on various image classification standards and remains an active field of study in medical image classification due to its ability to balance efficiency with efficacy.

3.2.3. DinoV2

The self-DIstillation with NO labels (DINO) [41] is a sophisticated self-supervised learning approach for training models that improves

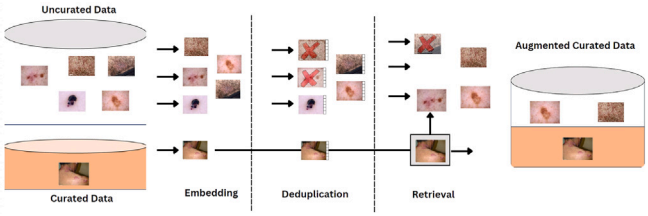

Fig. 7. Data curation for the semi-supervised learning mechanism of DinoV2.

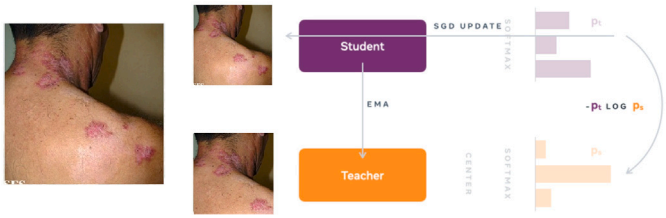

Fig. 8. Teacher student model training approach of DinoV2.

computer vision by reliably detecting specific objects inside 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 labeled dataset and estimating part of the data from that batch based on the characteristics. Some SSL systems circumvent these issues by employing DINO, which uses SSL and knowledge distillation methods. It enables extraordinary features to develop, such as robust object component recognition and robust semantic and low-level image understanding. Fig. 7 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 Fig. 8. It leads to quicker, more memory-efficient 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-neighbor search task), and the L2-norm for normalizing the embeddings are some of the strategies DINOv2 adopted to improve their results.

3.3. XAI for explainability

Explainable Artificial Intelligence, or XAI, is an important AI research and development topic as it tries to improve the transparency and interpretability of AI systems, allowing people to comprehend their decision-making processes. XAI solves the black box issue that frequently afflicts complicated machine learning models such as deep neural networks. XAI increases trust and responsibility by offering insights into why AI systems make certain predictions or conclusions. Still, it also helps users uncover and minimize biases, mistakes, and unexpected behaviors in AI applications. XAI employs various approaches and procedures, from visualization to feature attribution, aiming to make AI systems more interpretable and accessible to professionals and non-experts.

GradCAM, a computer vision algorithm, stands for Gradient-weighted Class Activation Mapping. It creates heatmaps emphasizing parts of an input image that contribute the most to a deep neural network's classification judgment. This graphic explanation explains

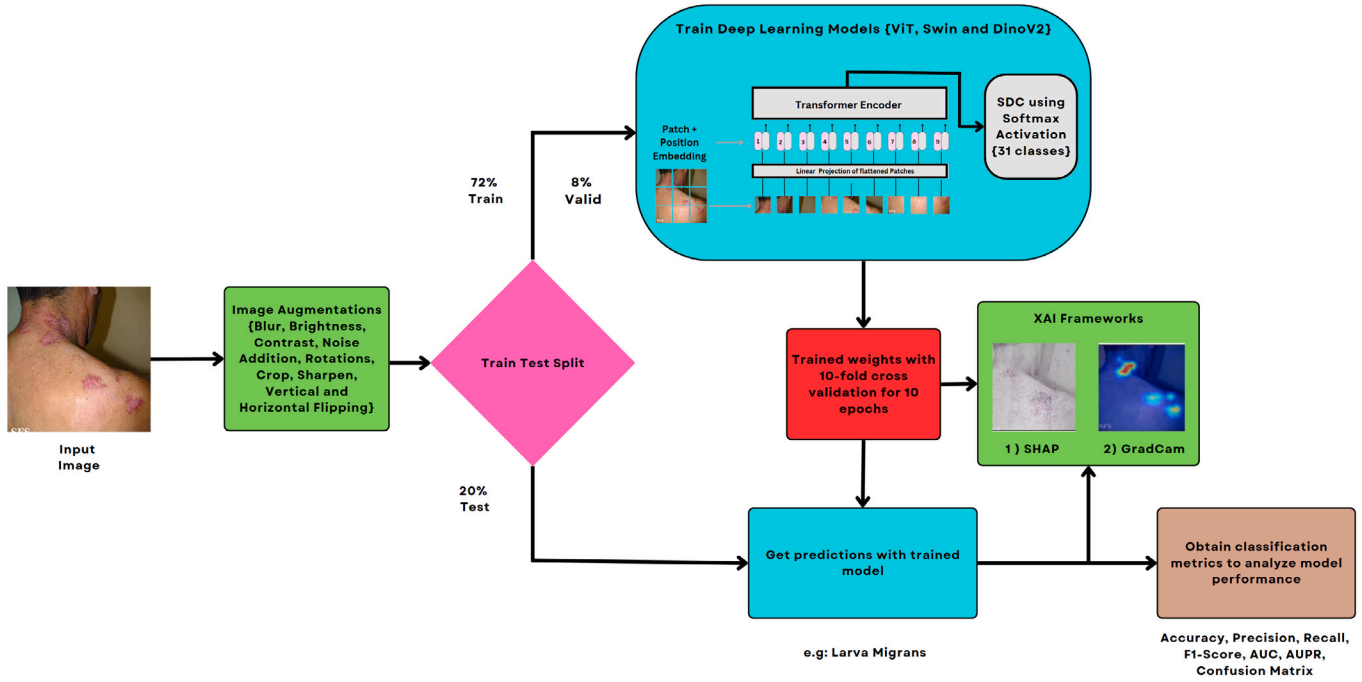

Fig. 9. Overall methodology proposed in this work.

which components of an image were important in the model's decision-making process. SHAP is a larger technique that may be used in a wide range of machine learning models, including some unrelated to computer vision. SHAP values are based on cooperative game theory and give a mechanism to ascribe the contribution of each characteristic to a certain prediction or result. This method thoroughly explains how specific input features impact model output, making it useful for model interpretation and feature engineering.

While GradCAM is particularly beneficial for visualizing deep neural network judgments in image-related tasks, SHAP offers a more adaptable technique that can be applied to various machine learning models and is particularly good for determining feature significance. Both strategies contribute to the larger subject of XAI by improving AI system transparency and interoperability, as the exhaustive methods offer distinct but complementary insights.

3.4. Experimental setup

Twenty experiments - Ten different architectures belonging to four backbones and two types of SDC datasets (with and without data augmentation) were done in this study. The experiments were done in a system with an Nvidia RTX A6000 GPU with 48 GB vRAM, a Ryzen Threadriper Pro CPU with 120 GB RAM, and 24 cores. The proposed pipeline to carry out the study done in this work is shown in Fig. 9.

The models were trained for 10 epochs with 10-fold cross-validation on all backbones of the transformers used and were fed in batches of 64 to these models. The models were trained on both the unaugmented and augmented dataset to determine if augmentation leads to overfitting on train data, as deciphered from the train and test T-SNE plots shown in Fig. 1. The geometric augmentations, as explained in the methodology, were meticulously chosen to enhance the diversity and robustness of our dataset, aiming to expose models to various data distributions and real-world variations.

The PyTorch framework and the weights from the Huggingface library were used to code and perform transfer learning using pre-trained ImageNet1k weights. Categorical cross entropy (CCE) loss was used to calculate the classification error during the training backpropagation process (the equation to calculate the loss is given in Eq. (3)), and Adam was used as the optimizer for faster training. The optimal learning rate during gradient descent was calculated using the lr_find() function, which divides the data into batches and considers choices from the learning rate yielding the least loss.

$$\text{Loss}_{\text{CCE}} = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(p_{ij}) \tag{3}$$

All models are evaluated using 10-fold cross-validation, where the dataset is repeatedly split into training and testing folds, and their performance is assessed using popular classification metrics such as accuracy, precision, recall, and F1-score [42,43]. Eqs. (4)–(7) denote the formulas for calculating the classification metrics. In addition to these metrics, we plot and analyze the area under the precision–recall curve and the receiver operating characteristic curve, which is done in the literature to prove the model's discriminatory power and performance across varying thresholds. These curves provide a comprehensive view of how well the model distinguishes between classes, offering insights into its precision, recall, and overall effectiveness in real-world applications. The results were also explained visually using XAI Tools such as GradCAM and SHAP to get more insight into the features captured by the model to diagnose a disease.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{4}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{5}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{6}$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

4. Results and discussion

The experiments done in this work leverage three transformers: Vision Transformers, Swin Transformers, and DinoV2. The ConvNeXt architecture, a benchmark in convolution-based feature extraction for image classification tasks has also been trained and validated for the main 31-class dataset. Other convolution architectures that have been

Table 3
Classification metrics for the architectures trained on the unaugmented and augmented SDC dataset.

Model	Parameters	Augmented data				Unugmented data			
		Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score
ConvNeXt-B	87,598,239	83.10	85.04	83.58	83.98	31.18	17.53	15.92	12.19
ViT-B	85,822,495	94.37	95.62	95.60	95.51	92.35	93.67	93.70	93.49
ViT-L	303,333,407	88.63	90.67	90.55	90.44	87.22	88.79	88.53	88.39
Swin-T	27,543,193	39.43	42.96	28.81	29.01	36.01	49.71	26.06	26.55
Swin-S	48,861,097	84.41	86.14	86.04	85.64	40.44	42.8	29.29	29.33
Swin-B	86,774,999	90.44	92.16	92.64	92.31	93.26	94.88	95.15	94.71
Swin-L	195,043,123	88.93	91.14	90.95	90.89	75.85	79.23	75.62	76.55
DinoV2-S	**22,080,415**	87.62	89.85	89.24	89.37	60.26	62.50	58.08	58.52
DinoV2-B	86,628,127	95.57	96.81	96.72	96.71	**96.48**	**97.55**	**97.10**	**97.27**
DinoV2-L	304,432,159	90.44	92.64	92.22	92.33	88.02	89.65	89.95	89.60

Table 4
Average metrics across 10-Fold cross validation.

Model	Accuracy	AUROC	AUPR
ConvNext-B	31.18 ± 0.045	92.09 ± 0.0439	45.19 ± 0.2265
ViT-B	92.35 ± 0.016	99.84 ± 0.034	97.11 ± 0.0668
ViT-L	87.22 ± 0.038	99.62 ± 0.0701	93.61 ± 0.1136
Swin-T	36.01 ± 0.039	90.58 ± 0.0523	38.38 ± 0.2092
Swin-S	40.44 ± 0.052	92.14 ± 0.0487	47.33 ± 0.2310
Swin-B	93.26 ± 0.017	0.9987 ± 0.0036	97.45 ± 0.0801
Swin-L	75.85 ± 0.036	98.80 ± 0.0100	83.98 ± 0.1199
DinoV2-S	60.47 ± 0.047	96.77 ± 0.0213	66.09 ± 0.1783
DinoV2-B	96.48 ± 0.0103	99.94 ± 0.0017	98.75 ± 0.0498
DinoV2-L	88.02 ± 0.030	99.69 ± 0.0060	95.01 ± 0.1126

adopted as backbones for feature extraction in the literature have also been used to extend the comparative analysis. The models are interpreted using XAI frameworks to assist dermatology and unravel the black-box nature of deep learning. Additionally, to ascertain if the best-performing transformer model on the 31-class dataset is robust, the model is also trained on two smaller datasets containing more samples of prominent diseases, and the metrics are compared with those of other benchmark models proposed in the literature.

4.1. Results on the combined SDC dataset

Table 3 shows the classification metrics obtained on the augmented and raw datasets using all the official releases of the three different transformer models used in this work. From the comparative analysis of transformer-based architectures alongside the convolutional-based architecture ConvNeXt, the results reveal several key insights from the perspective of deep learning for medical image analysis. The benchmark convolution architecture ConvNeXt-B demonstrates comparatively lower performance with just 31.18% accuracy, particularly on unaugmented data, where it struggles to generalize effectively. This suggests potential limitations in ConvNeXt-B's ability to adapt to diverse datasets without augmentation. This could be due to the fact that convolution extracts local features with the assistance of a kernel, limiting its scope to that region alone for feature extraction in a particular layer, but the attention mechanism ensures correlation computation between all patches of the image feature map in a layer. ViT improves these results because it considers the relationship between the image's different fixed patch embeddings. Yet, Swin Tranformers perform better than ViT due to an improved sliding kernel attention mechanism. Nevertheless, it is the SSL approach of DinoV2 training that obtains the best metrics.

Table 4 summarizes the performance of various models across 10-fold cross-validation using accuracy, AUROC, and AUPR metrics. DinoV2 Base emerged as the best-performing model with an accuracy of 96.48% (±0.0103), AUROC of 99.94% (±0.0017), and AUPR of 98.75% (±0.0498), demonstrating exceptional consistency and reliability. Swin Base and ViT Base also achieved high performance, with accuracies of 93.26% (±0.017) and 92.35% (±0.016), respectively, complemented by AUROC values exceeding 99.8% and AUPRs above 97%. Larger variants, such as DinoV2 Large and ViT Large, maintained robust results with accuracies around 87%–88%, AUROC values above 99.6%, and AUPRs above 93%, albeit with slightly higher variability. In contrast, smaller models like ConvNext-B, Swin Tiny, and Swin Small delivered suboptimal performance, with accuracies ranging from 31.18% to 40.44% and AUPRs below 50%. DinoV2 Small showed moderate performance, achieving 60.47% accuracy and an AUPR of 66.09%. Overall, DinoV2 Base, Swin Base, and ViT Base demonstrated superior results across all metrics, underscoring their effectiveness and reliability for the task, while smaller models exhibited significant variability and limited applicability.

Furthermore, from all the classification metrics for the test results of the experimental results, it is clear that the general performance trend of the models trained on the augmented SDC dataset is better than those trained on the unaugmented data, suggesting that the train and test distributions are indeed not as different as deciphered from the t-SNE plots. An improvement in accuracy and a similar improvement in other metrics is noticeable for all backbones trained on augmented data, except for the Swin-B and DinoV2-B backbones, whose metrics deteriorate post-augmentation. Though DinoV2-B experiences a drop of 1% in all metrics, the drop is not very significant for recall, denoting a lesser increase in the number of false negative predictions. Nevertheless, all the classification metrics of DinoV2-B consistently outperform other models across all metrics, showcasing its effectiveness in both augmented and unaugmented data scenarios. This suggests that the self-supervised pre-training method utilized in DinoV2-B yields superior results compared to the supervised pre-training approach adopted by other transformer models. The slight drop in performance metrics for DinoV2-B post-augmentation indicates a potential overfitting to the augmented training data, leading to a marginal decline in test results. This overfitting suggests that while augmentations can enhance model robustness, they might also introduce noise that affects generalization, particularly in SSL models like DinoV2-B. DinoV2's consistent performance superiority, even with fewer training samples, underscores its robustness and efficiency. The self-supervised pre-training method enables the model to learn more generalized features from the data, making it less reliant on large annotated datasets. Also, the performance of Swin Transformers drops below ViT's post augmentation, with a mere 90.44% accuracy, and all metrics drop by 2% despite architectural dominance, indicating the importance of model size in achieving higher metrics.

On the other hand, one can also infer from the results of all backbones within a family that the smaller models might generalize better than larger ones, especially in cases where the dataset is diverse and representative of the target domain. Yet, larger models tend to have more parameters, making them more prone to overfitting, especially when the dataset is not large enough to fully exploit the model's capacity. For the combined unaugmented data considered for the study,

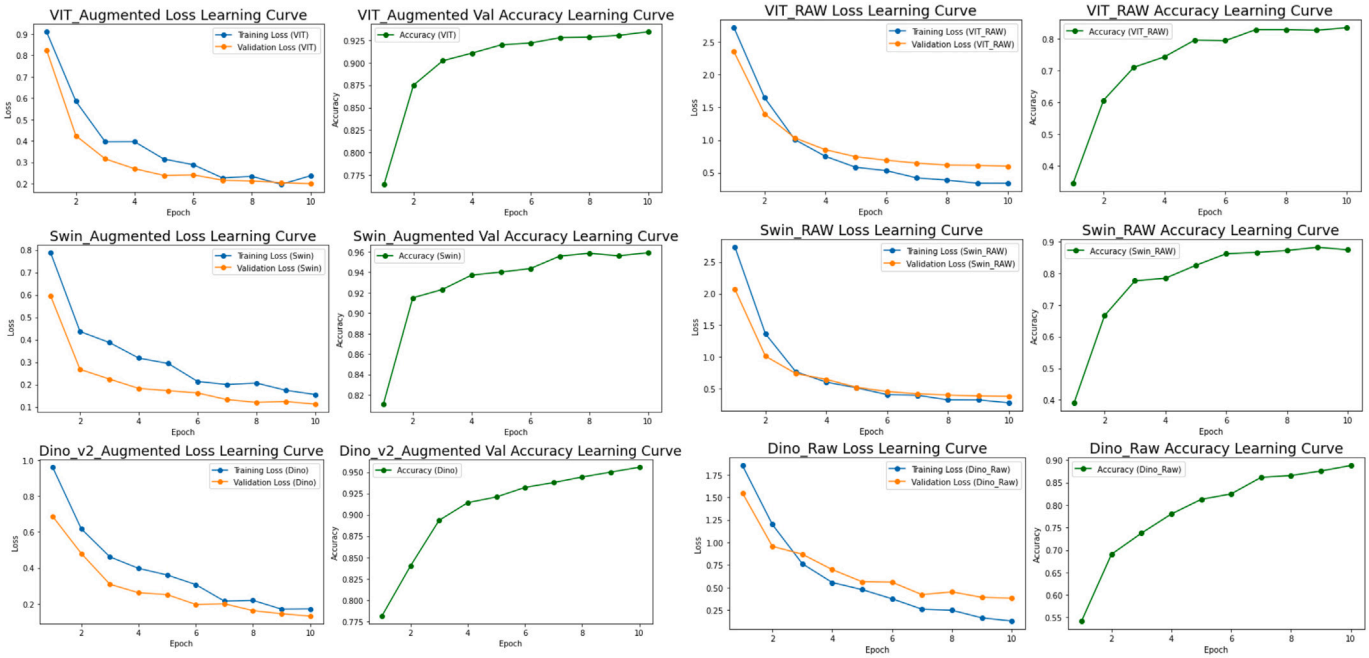

Fig. 10. Epoch vs. Loss and Accuracy curves for all trained models.

the training data per class might be limited (just like it is seen for a few classes in Fig. 4), and overfitting is more prominent, leading to a decrease in performance for larger models. This is justified by the generic trend in results across each family of transformers, where due to the size of the combined dataset and the trainable parameters of the model, the classification metrics can be easily observed to increase from tiny up to the base models of all architectures, but a small drop in the performance of the large variant is observed.

The classification metrics, in general, are better for the augmented dataset than the unaugmented data, suggesting that the augmentation strategies adapted to upsample the dataset are indeed helpful in helping the generalization of relevant features extracted by the architectures. The accuracy improvement is about 10% in cases where classification metrics have improved in general. However, this is not the case for the Swin-B and DinoV2-B models due to overfitting. The overfitting nature of the model trained on augmented data can also be substantiated by the epoch vs. loss and accuracy curves for all the best-performing variants of the chosen transformer models shown in Fig. 10. Firstly, the horizontal gap between the train and validation loss curves keeps fluctuating for the models trained with the augmented dataset. Still, a vast fluctuation is absent for the models trained on unaugmented data. This erratic fluctuation in loss curves for augmented data indicates that the models may struggle to generalize effectively, leading to increased overfitting. Upon closer examination of the loss curves on the y-axis, the values are consistently smaller for augmented data compared to the other models for the same epoch. However, the corresponding improvement in validation accuracy is not observable.

Moreover, a distinct trend emerges when assessing the validation accuracy curves. Models trained with data augmentation tend to rapidly reach high accuracy levels, often within the first few epochs, before saturating. This is evidenced by the validation accuracy higher for training with data augmentation than the data due to the validation data being a subset of the train data and the model becoming well-trained on the train data samples. However, overfitting with an augmented dataset leads to overtraining, yielding lesser classification metrics for the test data. In contrast, models trained on the raw dataset exhibit a more gradual and steady increase in accuracy over time. This phenomenon can be attributed to the overfitting observed in augmented data, where the models essentially 'memorize' the training

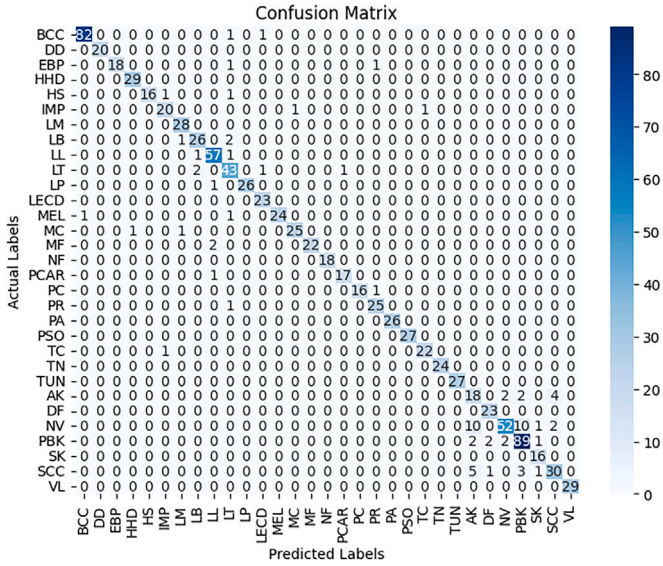

Fig. 11. Confusion matrix for the trained ViT-Base model on unaugmented data.

samples rather than learning generalized patterns. Though the models obtain almost the same quantity of false positives and false negatives, owing to a comparable precision, recall, and F1-score for their unaugmented counterparts, the lesser true positives and true negatives (as deciphered from the accuracy) make the model performance relatively poor. Thus, though augmentation strategies are helpful in general, it is not necessary for architectures like Swin transformers and DinoV2, as demonstrated by the results.

Since the best performing models were acquired with training on an unaugmented dataset, Figs. 11, 12, and 13 show the confusion matrices obtained by the models ViT-B, Swin-B, and DinoV2-B, respectively, trained on the unaugmented data. While comparing the metrics of each model, the performance of Vision Transformers is the lowest compared to Swin Transformers and DinoV2. Though ViT can theoretically extract better feature maps than CNNs using the multi-head self-attention layer

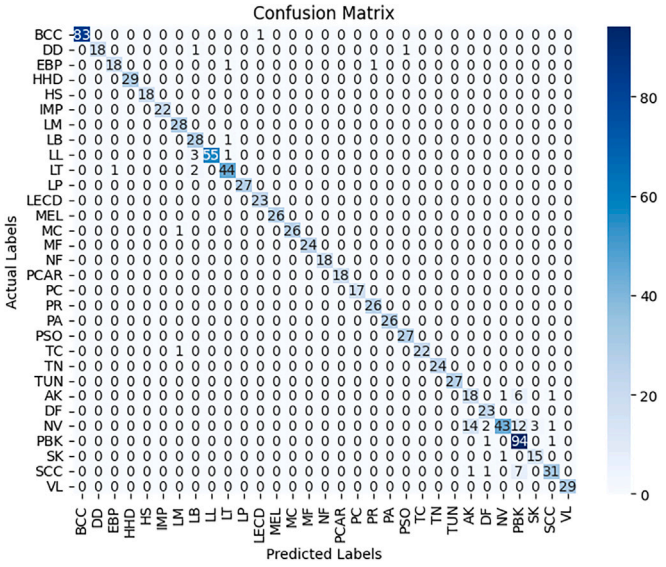

Fig. 12. Confusion matrix for the trained Swin-Base model on unaugmented data.

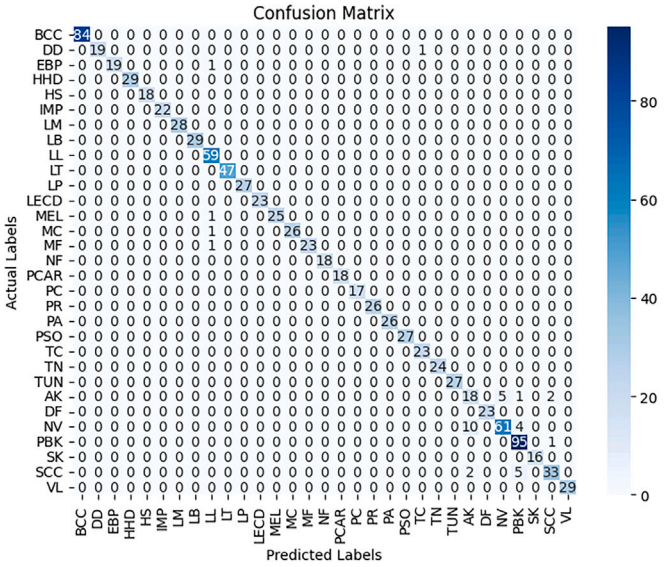

Fig. 13. Confusion matrix for the trained DinoV2-Base model on unaugmented data.

from the patch and position embeddings, due to which the model gets a test accuracy above 90%, the model is outperformed by the sliding window self-attention blocks of the Swin Transformers. Moreover, a higher classification metric of 93.26% accuracy, which is approximately a 1% improvement over the accuracy of ViT makes Swin a better architecture to perform the task. The model obtained a lesser number of false positive and false negative values. However, the standout performer in our experiments is DinoV2, a network pre-trained with semi-supervised approaches, which harnesses the Xformers framework to attain the best test accuracy of 96.48% and the least outliers (less than 40 of the 944 samples). This substantial improvement in accuracy positions DinoV2-B as the most promising architecture among the tested models, surpassing both Swin Transformers and ViT for image classification tasks. Nevertheless, the model does have a few outliers (elements not along the diagonal) in the confusion matrix, suggesting room for improvement.

Another standard inference from the confusion matrices of all three models is a significant number of samples (10 or higher) in the test set of the Nevus (N) class, being incorrectly predicted as the Actinic Keratosis (AK). This is because AK and nevus can sometimes be misdiagnosed due to overlapping clinical features. AK presents as scaly, rough patches, often on sun-exposed areas, while nevi (moles) are pigmented skin growths. However, certain types of nevi, such as dysplastic nevi, may exhibit features resembling AK, leading to misdiagnosis. Additionally, both conditions can arise from sun exposure, further complicating diagnosis. Furthermore, the differential diagnosis may be challenging, as flat pigmented lesions on sun-damaged skin, including nevi, can mimic actinic keratosis, which the transformer architectures cannot easily decipher from the training dataset.

Fig. 14 presents the ROC and precision–recall (PR) curve analysis, showing that transformer-based models like DinoV2, ViT, and Swin consistently outperform ConvNeXt in terms of generalization and managing the trade-off between precision and recall. DinoV2, particularly in its Base variant, shows strong performance across both ROC and PR curves, indicating robustness against false positives. Swin Base also performs well with balanced precision–recall trade-offs, while ConvNeXt exhibits a noticeable decline in precision as recall increases, suggesting challenges with false positives in more difficult scenarios. Notably, the Base variants of DinoV2, Swin, and ViT outperform their smaller and larger counterparts, with a favorable balance between model complexity and generalization. Although ViT, especially in its Base variant, achieves strong true positive rates, but its precision–recall curves show slightly more variability than DinoV2.

Fig. 15 shows the t-SNE plots for the feature map vectors extracted from ViT-B, Swin-B, and DinoV2-B. In all models, the class clusters of the test embedding are closer to each other, demonstrating structural similarities for samples within the same class. However, owing to inter-class similarities, some embedding projections are distributed throughout the space and overlap with closely related classes, explaining why the task itself is primarily difficult even for robust transformer-based feature extractors. Nevertheless, a model such as DinoV2, which is robustly trained well on the combined dataset, performs better, only on a dataset close to its distribution, as evident from the points being closer to the corresponding cluster centers (demonstrating low intra-class variability) and the high classification metrics obtained by the model in this work.

Table 5 highlights the model performance and classification metrics of the experiments done with the transformer-based architectures alongside the state-of-the-art models trained on the dataset used in this work. The only work on this dataset was done by the group that introduced the dataset, and they adopted convolution-based architectures from a family of architectures such as EfficientNet, VGG, and ResNet. to conclude that EfficentNetB2 achieves the best classification accuracy. Nevertheless, all transformers used in our work perform better and yield better benchmark results, with DinoV2-B improving the accuracy by approximately 10% to yield a 96.48% accuracy compared to an existing 87.15% accuracy. Thus, our study underscores the significance of transformer model architectures, pre-training strategy, and data augmentation in determining the performance of deep learning models for image classification tasks, with DinoV2-B emerging as the top-performing transformer architecture in this comparative analysis. This potentially paves the way for classifying many such medical image datasets in the future.

4.2. Explainability using XAI frameworks

In classification tasks like SDC, the additional outputs with XAI frameworks offer transparency by generating explanations highlighting the key characteristics and factors influencing a deep learning model to arrive at a specific class label prediction. It helps researchers diagnose the severity and spread of the disease by highlighting critical regions in medical images, providing deeper insights into the model's decision-making process. Understanding the rationale behind a model's choice

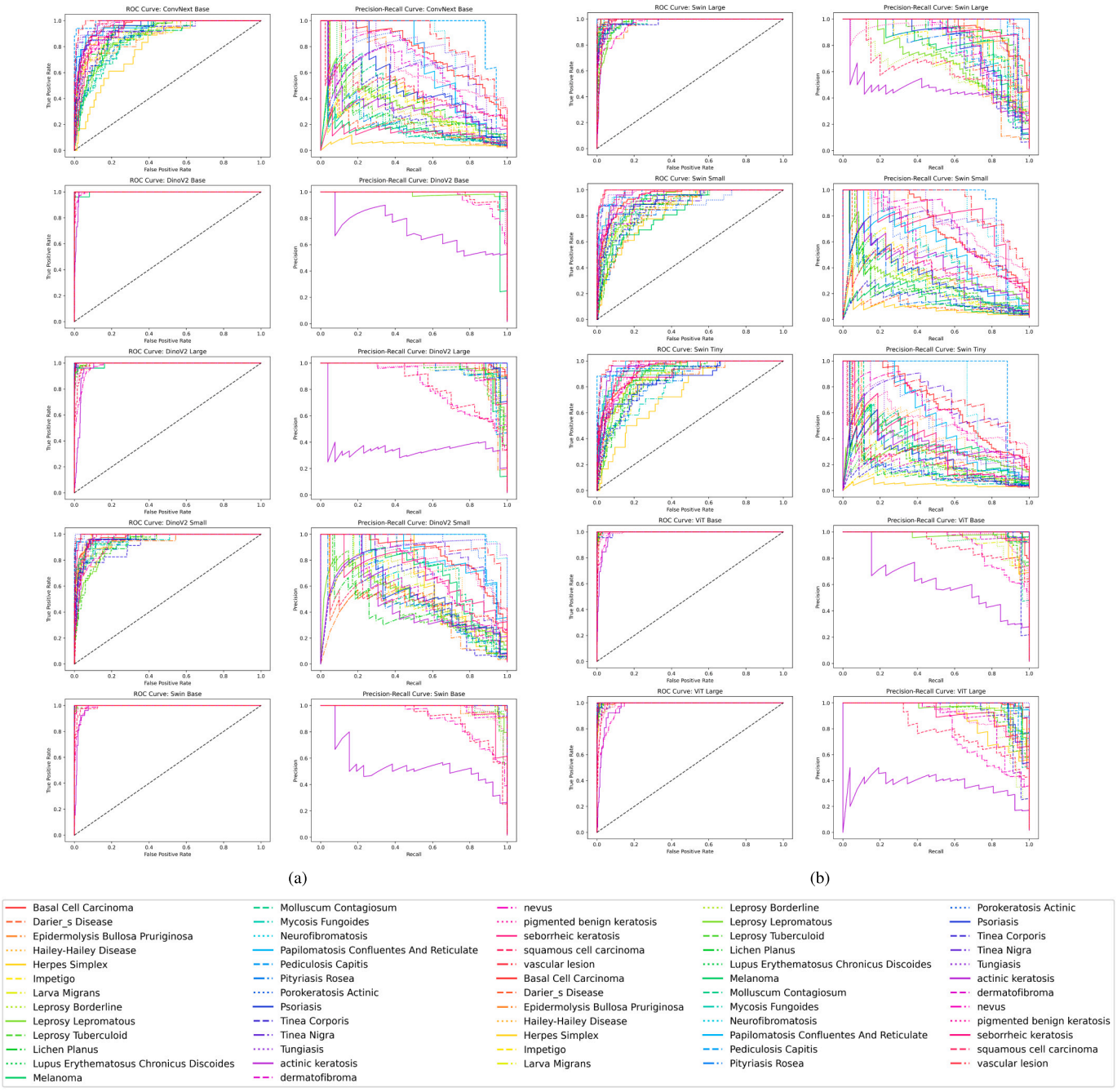

Fig. 14. Comparison of ROC Curves.

for a given input is crucial for establishing trust and ensuring accountability, particularly in medical domains like dermatology. Furthermore, the XAI framework employed in this study holds significant potential for real-world applications.

GradCAM helps model interpretability by offering useful insights into the relevance of features. It highlights the regions in an input image most influential in determining a specific classification outcome. This visualization is achieved by computing the gradient of the predicted class score with respect to the feature maps in the penultimate layer, enabling researchers to see which areas of an image contribute most to the model's decision. This information is particularly useful in medical image analysis, as it can guide practitioners toward areas requiring more attention for accurate diagnosis and treatment.

SHAP offers a complementary approach by providing a more granular understanding of the importance of features. It quantifies the contribution of different patches or tokens within an image to the model's prediction, allowing a comprehensive view of how individual parts of an input influence the output. In medical contexts, this can help identify specific regions or features critical for diagnosis and guide treatment decisions.

GradCAM and SHAP are exhaustive methods because they offer distinct but complementary insights. GradCAM generates heatmaps highlighting the most relevant regions for a model's decision, offering a coarse, localized visual explanation at the feature level. In contrast, SHAP provides pixel-level explanations, quantifying the contribution of individual input features to the output. By applying both methods together, GradCAM identifies high-impact regions, while SHAP provides pixel-level correlations within those regions. The outputs of images taken for three different classes of the test dataset for GradCAM and SHAP are shown in Figs. 16 and 17, respectively.

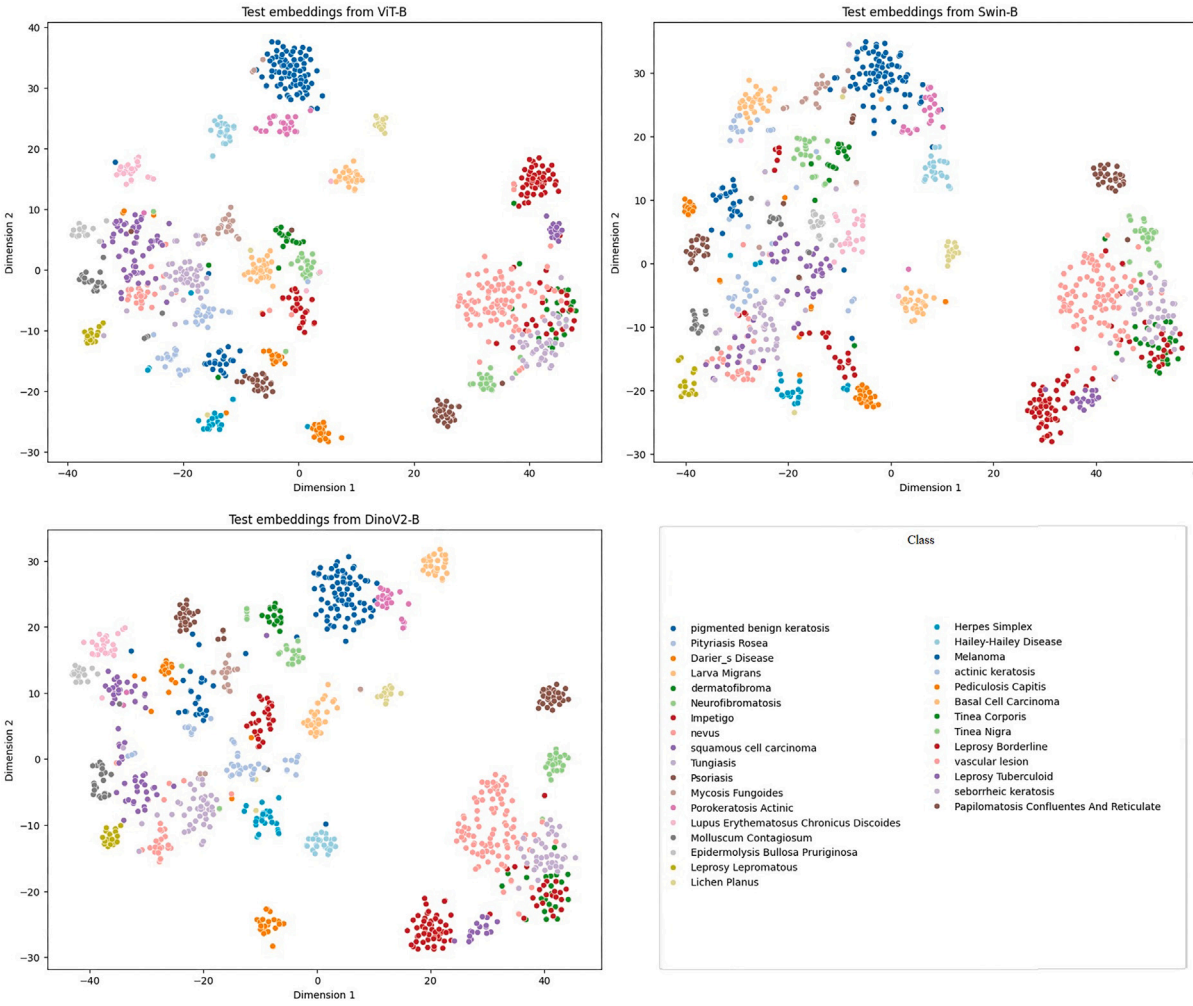

Fig. 15. T-SNE plot for the embeddings obtained from the fine-tuned transformer architectures on the unaugmented dataset.

Table 5
Comparison of SDC Models on combined data.

Authors	Year	Classes	Architecture	Accuracy	Precision	Recall	F1-Score
A. Rafay and W. Hussain [17]	2023	31	VGG-19	43.25	45.00	43.00	43.00
		31	VGG-16	57.62	58.00	58.00	58.00
		31	ResNet-152	72.13	73.00	72.00	72.00
		31	ResNet-50	75.91	76.00	76.00	76.00
		31	ResNet-101	77.50	78.00	77.00	77.00
		31	EfficientNet-B6	80.89	81.00	80.00	80.00
		31	EfficientNet-B5	81.67	82.00	82.00	82.00
		31	EfficientNet-B0	81.67	82.00	82.00	82.00
		31	EfficientNet-B1	81.80	82.00	82.00	82.00
		31	EfficientNet-B3	82.30	82.00	82.00	82.00
		31	EfficientNet-B4	84.45	84.00	84.00	84.00
		31	EfficentNetB2	87.15	87.00	87.00	87.00
Proposed work	2024	31	ViT-Base	92.35	93.67	93.70	93.49
		31	Swin-Base	93.26	94.88	95.16	94.72
		31	**DinoV2-Base**	**96.48**	**97.55**	**97.11**	**97.28**

These GradCAM and SHAP insights align with quantitative results. DinoV2-B emerges as the top performer, particularly on the raw dataset, achieving a remarkable test accuracy of 96.48%. While Swin Transformers and ViT compete closely, they fall short of DinoV2-B's performance standards. DinoV2's GradCAM heatmaps and SHAP plots on the unaugmented dataset exhibit remarkable accuracy, effectively highlighting infected regions such as the hands, neck, and ears for the three images, respectively. This precision in localization elucidates why DinoV2 surpasses other architectures in performance. In contrast, it is GradCAM heatmaps and SHAP plots on the augmented dataset show reduced accuracy due to the introduced variations, causing overfitting and impacting the efficacy of the SSL approach. The patch area is more diversified, suggesting that the model cannot narrow down to the region of infection as precisely as the model trained on the unaugmented data. These results are closely followed by the Swin Transformers model trained on the unaugmented data, with a similar area but less intense in and around the infected region. Nevertheless, Swin transformers, like DinoV2, demonstrate better GradCAM and SHAP regions on the unaugmented dataset, indicating that sliding kernel self-attention can extract relevant areas from the image.

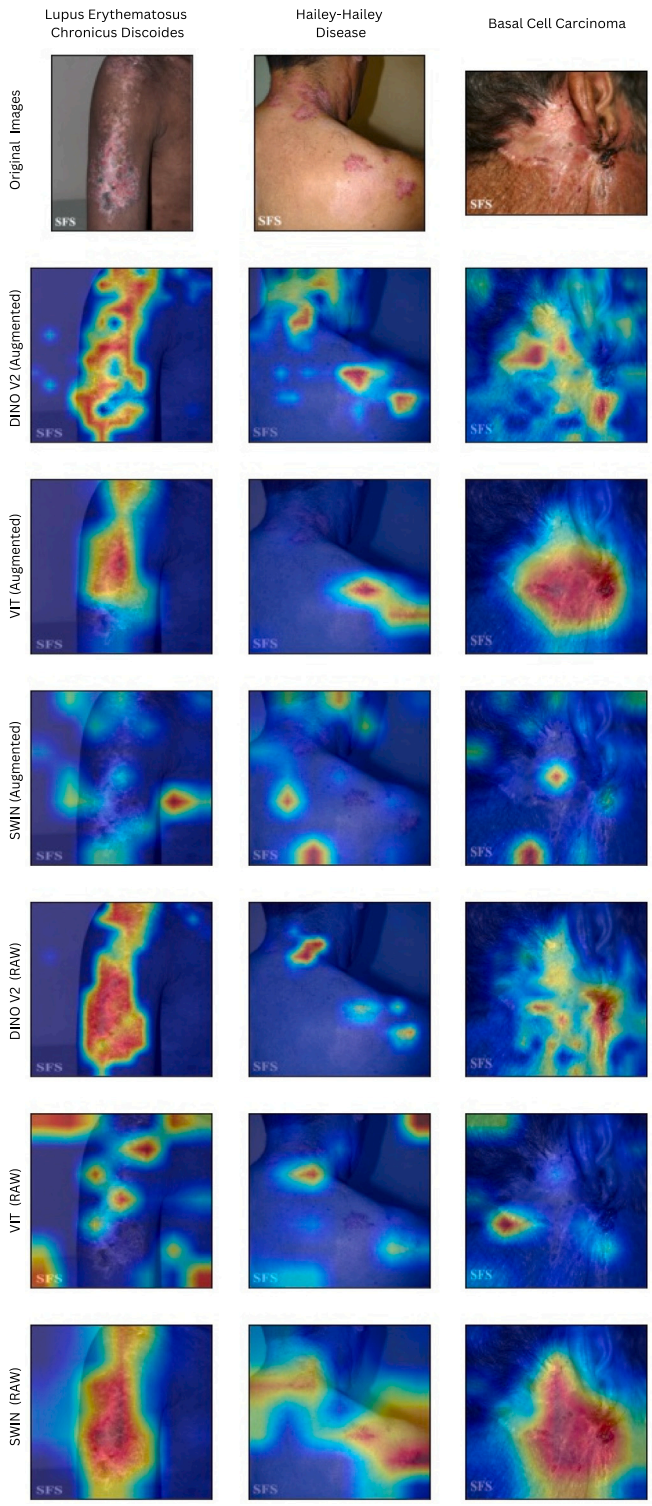

Fig. 16. GradCAM plots for Augmented & Raw Images.

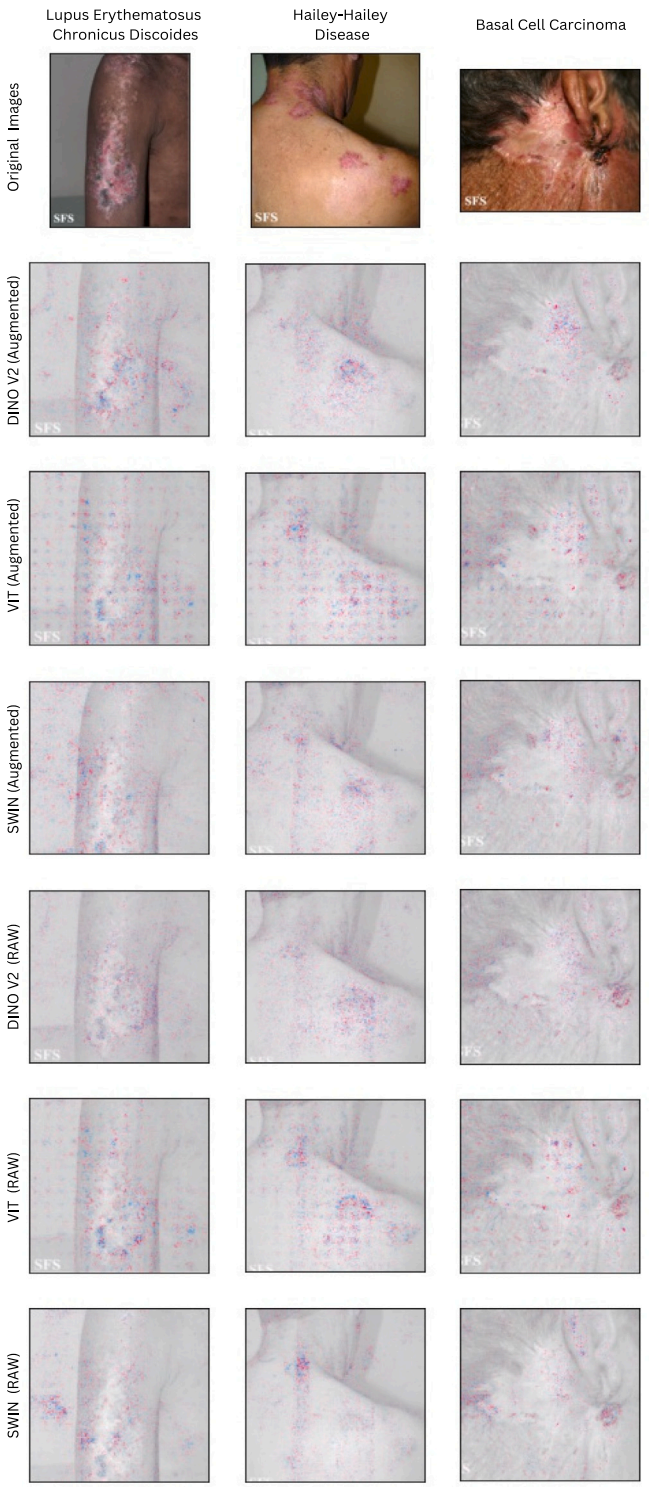

Fig. 17. SHAP plots for Augmented & Raw Images.

Notably, ViT defies the norm on the augmented dataset by displaying improved GradCAM and SHAP performance compared to the unaugmented dataset. This underscores ViT's adaptability to diverse data distributions resulting from augmentation. It successfully captures pertinent features and regions, showcasing resilience to dataset variations. Nevertheless, some of the regions of interest for the unaugmented data are outside the human body, suggesting room for improvement in performance. With the most attention to heatmap regions and SHAP points outside the skin in ViT, the model trained on the unaugmented data has the least ability to perform diagnosis, which aligns well with the performance metrics of the model.

4.3. Comparison of results on smaller benchmark datasets

To evaluate the robustness of the transformer models and showcase that the proposed pipeline can accurately automate the classification of

Table 6
Comparison of SDC Models trained on the dermnet dataset.

Authors	Year	Classes	Architecture	Accuracy	Precision	Recall	F1-Score
Aboulmira et al. [44]	2022	23	DenseNet	68.97	69.30	69.20	69.25
Sah et al. [45]	2019	23	Finetuned VGG	76.30	76.00	76.00	76.00
Bindhu et al. [46]	2023	23	FuzzyUNet+DB	95.61	**94.72**	–	–
Anurodh Kumar et al. [47]	2024	23	1D-Multiheaded CNN	88.57	88.88	88.72	88.04
Proposed work		**23**	**DinoV2-Base**	**96.23**	94.51	**94.62**	**94.54**

Table 7
Comparison of SDC Models on the HAM10000 dataset.

Authors	Year	Classes	Architecture	Accuracy	Precision	Recall	F1-Score
Saket Chaturvedi et. al. [26]	2020	7	ResNeXt101	93.20	88.00	88.00	88.00
Anand et.al. [27]	2022	7	XCeption Net	96.40	–	–	–
Aladhadh, Suliman, et al. [37]	2022	7	Medical-VIT	96.14	**96.00**	96.50	96.25
Krishna, Ghanta Sai et. al. [15]	2023	7	ViT-GAN	97.40	–	–	–
Selen Ayas [16]	2023	7	Swin-Large	97.20	85.10	**98.00**	91.10
Proposed work		**7**	**DinoV2-Base**	**97.45**	95.63	97.42	**96.46**

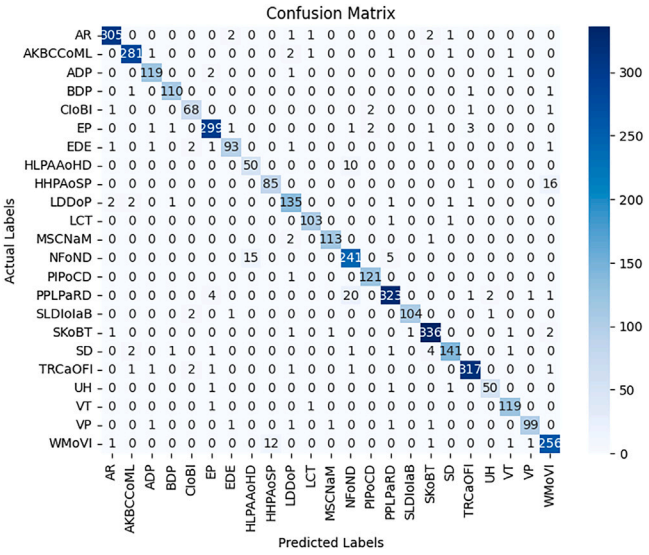

Fig. 18. Confusion Matrix for the 23-class Dermnet dataset using Dino-V2.

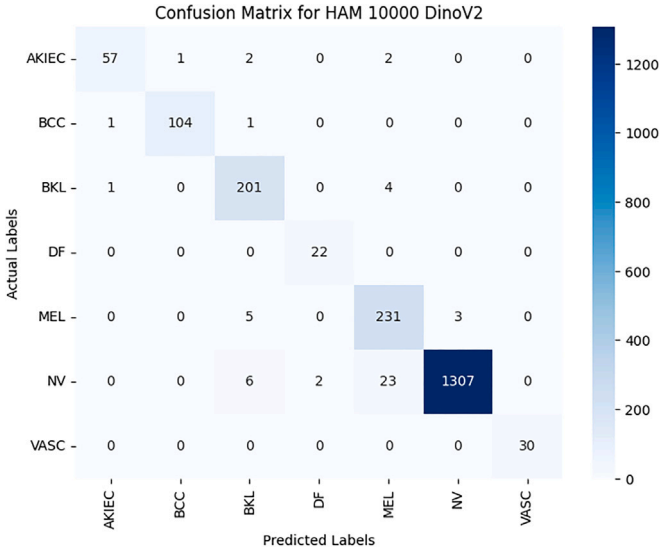

Fig. 20. Confusion matrix for the 7-class HAM10000 dataset using Dino-V2.

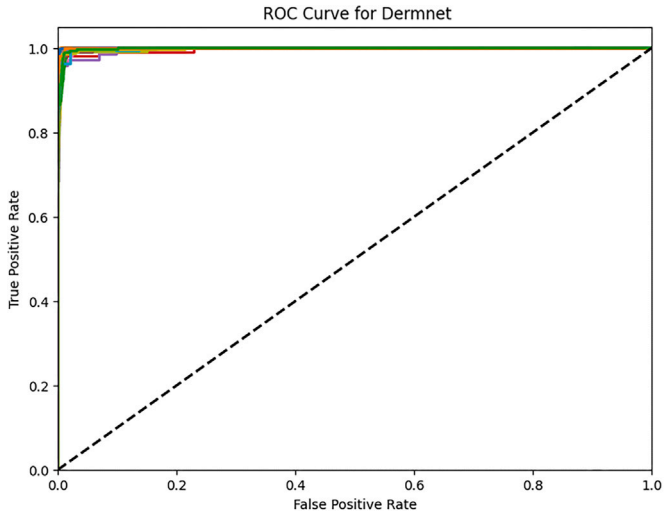

Fig. 19. ROC-AUC curve for the 23-class Dermnet dataset using Dino-V2.

a wide range of diseases by extracting relevant features from the input images, the models are also trained on smaller benchmark datasets such as Dermnet and HAM10000, which contain many samples per class for popular skin diseases.

It is evident from the confusion matrix in Fig. 18 that most of the samples are being accurately classified while only a few samples from the minority classes in the dataset, such as NFoND, WMoVI, PPLRaRD and HHPAoSP (expansions in the Appendix). Despite the anomalies, the score for ROC-AUC curves plotted for the true positive rate vs. the false positive in Fig. 19 never falls below 0.9980. The maximum samples falling on the matrix trace and the fewer false positives and false negatives clearly demonstrate the high accuracies and F1-score indicated in Table 6. The table also compares the results of DinoV2-B trained on the unaugmented Dermnet dataset to other state-of-the-art works in the literature. DinoV2-B significantly outperforms the works in the literature, with an improvement in the test accuracy compared to CNN architectures. A similar improvement can be seen in the recall and F1 scores; an overall drop in total misclassified samples improves the overall performance.

A similar trend can be seen in Table 7, which compares the results of DinoV2-B trained on the unaugmented HAM10000 dataset to other state-of-the-art works in the literature, the present benchmark accuracy being a computationally heavy ViT-GAN architecture. DinoV2-B has shown a slight increase in metrics such as F1-Score, as it maintains a good balance between the precision and recall scores, which are the current state-of-the-art ones maintained by ViT-GAN and Swin-L.

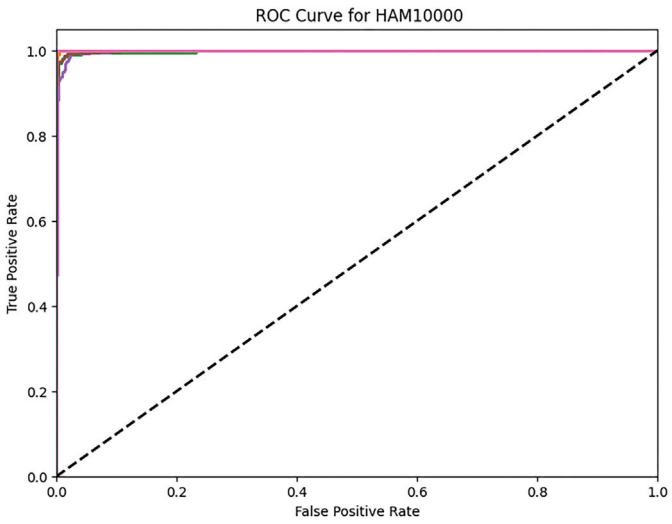

Fig. 21. ROC-AUC curves for the 7-class HAM10000 dataset using Dino-V2.

Even for highly underrepresented samples of the DF class in Fig. 20, the model robustly classifies all test samples except 2 correctly. The model accurately classifies all samples for the VASC class, once again proving its robustness. However, samples from the NV class are falsely classified as MEL due to overlapping features, which even clinicians fail to identify in extreme cases (expansions in the Appendix). Nevertheless, the scores for the ROC-AUC curves in Fig. 21 are still higher than 0.995 for all classes. These benchmark results set by DinoV2-B for the smaller datasets in the literature with fewer samples per class are the new state-of-the-art results for small datasets, probing the robustness of the transformer architectures for the SDC task.

While our proposed architecture achieved state-of-the-art results on the combined dataset, Dermnet, and HAM10000 datasets, several limitations and areas for future improvement remain to be addressed.

- Computational Complexity: One of the major limitations of our methodology is its computational intensity. The models developed in this study require significant computational resources, which may not be feasible for deployment in real-time or on resource-constrained devices. Future research efforts should focus on other lightweight architectures to reduce computational complexity while maintaining or even enhancing classification performance.
- Generalizability to Diverse Datasets: While we demonstrated the effectiveness of our methodology on Dermnet and HAM10000 datasets, its generalizability to other diverse skin disease datasets remains unexplored. Future studies should evaluate our approach on a wider range of datasets to assess its robustness and generalizability across different skin conditions.

5. Conclusion and future work

The current study presented transformers to classify a diverse set of 31 skin diseases, and the results are validated with metrics such as precision and F1 score in the data. When evaluated on the test data, the final model achieved 96.48% accuracy in detecting the condition, approximately 10% improvement over the current state-of-the-art results. Ten augmentation strategies were employed to improve the data distribution and determine any performance improvements, and augmentation helped CNNs like ConvNeXt and transformers like ViT improve their performance. However, there has been a drop in performance metrics for SSL and sliding kernel attention techniques employed in transformers like DinoV2 and Swin transformers. According to the study results and their interpretation with XAI frameworks like Grad-CAM and SHAP, the suggested model can assist society by allowing clinicians to detect skin problems more precisely and rapidly. The improvement in results using transformers and recently-introduced DinoV2-B was also compared with other state-of-the-art results in the literature, and an improvement was observed for the 23-class Dermnet and the 7-class HAM10000 dataset. An improved recall in both datasets suggests that the improved precision suggests DinoV2-B is a robust model and can yield fewer false negative diagnoses in the future. The models publically made available through this work can result in quicker and more effective therapy offered by skin specialists, enhancing patient well-being while reducing the financial burden on the healthcare system. The proposed methodology can also help the general people directly diagnose and gain an immaculate understanding of their dermatological issues without the assistance of doctors in non-complicated scenarios.

CRediT authorship contribution statement

Jayanth Mohan: Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Arrun Sivasubramanian:** Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Sowmya V.:** Writing – review & editing, Validation, Supervision. **Vinayakumar Ravi:** Writing – review & editing.

Compliance with ethical standards

None.

Availability of data and code

The data that support this study's findings are available from the first authors upon reasonable request. The best model for this dataset is deployed publically on the Hugging Face community (DinoV2-Base, Swin-Base, and VIT-Base). The codes are also available in the GitHub link: https://github.com/JAYANTH-MOHAN/Enhancing-Skin-Disease-Classification-Using-Transformers

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

See Tables 8 and 9.

Table 8
Abbreviations and their meanings for HAM10000 dataset.

Abbreviation	Meaning
AKIEC	Actinic Keratoses and Intraepithelial Carcinoma
BCC	Basal Cell Carcinoma
BKL	Benign Keratosis Like Lesions
DF	Dermatofibroma
MEL	Melanoma
NV	Melanocytic Nevi
VASC	Vascular Lesions

Table 9
Abbreviations and their meanings for dermnet dataset.

Abbreviation	Meaning
AR	Acne and Rosacea Photos
AKBCCoML	Actinic Keratosis Basal Cell Carcinoma and other Malignant Lesions
ADP	Atopic Dermatitis Photos
BDP	Bullous Disease Photos
CIoBI	Cellulitis Impetigo and other Bacterial Infections
EP	Eczema Photos
EDE	Exanthems and Drug Eruptions
HLPAAoHD	Hair Loss Photos Alopecia and other Hair Diseases
HHPAoSP	Herpes HPV and other STDs Photos
LDDoP	Light Diseases and Disorders of Pigmentation
LCT	Lupus and other Connective Tissue diseases
MSCNaM	Melanoma Skin Cancer Nevi and Moles
NFoND	Nail Fungus and other Nail Disease
PIPoCD	Poison Ivy Photos and other Contact Dermatitis
PPLPaRD	Psoriasis pictures Lichen Planus and related diseases
SLDIoIaB	Scabies Lyme Disease and other Infestations and Bites
SKoBT	Seborrheic Keratoses and other Benign Tumors
SD	Systemic Disease
TRCaOFI	Tinea Ringworm Candidiasis and other Fungal Infections
UH	Urticaria Hives
VT	Vascular Tumors
VP	Vasculitis Photos
WMoVI	Warts Molluscum and other Viral Infections

References

[1] W. James, D. Elston, T. Berger, Andrew's Diseases of the Skin E-Book: Clinical Dermatology, Elsevier Health Sciences, 2011.

[2] A. Kavita, J. Thakur, T. Narang, et al., The burden of skin diseases in India: Global burden of disease study 2017, Indian J. Dermatol. Venereol. Leprol. 89 (2023) 421–425.

[3] R. Hay, M. Augustin, C. Griffiths, W. Sterry, B. Dermatological Societies, Grand Challenges Consultation groups, K. Abuabara, M. Airoldi, F. Ajose, S. Albert, A. Armstrong, et al., The global challenge for skin health, Br. J. Dermatol. 172 (2015) 1469–1472.

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

[5] H. Xu, H. Li, Acne, the skin microbiome, and antibiotic treatment, Am. J. Clin. Dermatol. 20 (2019) 335–344.

[6] S. Inthiyaz, B. Altahan, S. Ahammad, V. Rajesh, R. Kalangi, L. Smirani, M. Hossain, A. Rashed, Skin disease detection using deep learning, Adv. Eng. Softw. 175 (2023) 103361.

[7] Yann LeCun, Yoshua Bengio, Geoffrey Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[8] W. Yue, S. Liu, Y. Li, Eff-PCNet: An efficient pure CNN network for medical image classification, Appl. Sci. 13 (2023) 9226.

[9] Q. Zhou, Z. Huang, M. Ding, X. Zhang, Medical image classification using light-weight CNN with spiking cortical model based attention module, IEEE J. Biomed. Heal. Inform. 27 (2023) 1991–2002.

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

[11] F. Shamshad, S. Khan, S. Zamir, M. Khan, M. Hayat, F. Khan, H. Fu, Transformers in medical imaging: A survey, Med. Image Anal. (2023) 102802.

[12] O. Manzari, H. Ahmadabadi, H. Kashiani, S. Shokouhi, A. Ayatollahi, MedViT: a robust vision transformer for generalized medical image classification, Comput. Biol. Med. 157 (2023) 106791.

[13] Z. Liu, Q. Lv, Z. Yang, Y. Li, C. Lee, L. Shen, Recent progress in transformer-based medical image analysis, Comput. Biol. Med. (2023) 107268.

[14] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, D. Yang, A multimodal transformer to fuse images and metadata for skin disease classification, Vis. Comput. 39 (2023) 2781–2793.

[15] G. Krishna, K. Supriya, M. Sorgile, et al., LesionAid: Vision transformers-based skin lesion generation and classification, 2023, arXiv preprint arXiv:2302.01104.

[16] S. Ayas, Multiclass skin lesion classification in dermoscopic images using swin transformer model, Neural Comput. Appl. 35 (2023) 6713–6722.

[17] A. Rafay, W. Hussain, EfficientSkinDis: An EfficientNet-based classification model for a large manually curated dataset of 31 skin diseases, Biomed. Signal Process. Control. 85 (2023) 104869.

[18] Gourav Ganesh, K. Somasundaram, Detect melanoma skin cancer using an improved deep learning CNN model with improved computational costs, 2023.

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

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

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

[22] R. Karthik, T. Vaichole, S. Kulkarni, O. Yadav, F. Khan, Eff2Net: An efficient channel attention-based convolutional neural network for skin disease classification, Biomed. Signal Process. Control. 73 (2022) 103406.

[23] M. Hossen, V. Panneerselvam, D. Koundal, K. Ahmed, F. Bui, S. Ibrahim, Federated machine learning for detection of skin diseases and enhancement of internet of medical things (IoMT) security, IEEE J. Biomed. Heal. Inform. 27 (2022) 835–841.

[24] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nat. 542 (2017) 115–118.

[25] P. Kshirsagar, H. Manoharan, S. Shitharth, A. Alshareef, N. Albishry, P. Balachandran, Deep learning approaches for prognosis of automated skin disease, Life 12 (2022) 426.

[26] S. Chaturvedi, J. Tembhurne, T. Diwan, A multi-class skin cancer classification using deep convolutional neural networks, Multimed. Tools Appl. 79 (2020) 28477–28498.

[27] V. Anand, S. Gupta, D. Koundal, S. Nayak, J. Nayak, S. Vimal, Multi-class skin disease classification using transfer learning model, Int. J. Artif. Intell. Tools 31 (2022) 2250029.

[28] L. Moataz, G. Salama, M. Abd Elazeem, Skin cancer diseases classification using deep convolutional neural network with transfer learning model, J. Phys.: Conf. Ser. 2128 (2021) 012013.

[29] N. Hameed, A. Shabut, M. Hossain, Multi-class skin diseases classification using deep convolutional neural network and support vector machine, in: 2018 12th International Conference on Software, Knowledge, Information Management & Applications, SKIMA, 2018, pp. 1–7.

[30] Y. Filali, H.E.L. Khoukhi, M. Sabri, A. Aarab, Efficient fusion of handcrafted and pre-trained CNNs features to classify melanoma skin cancer, Multimed. Tools Appl. 79 (2020) 31219–31238.

[31] P. Ly, D. Bein, A. Verma, New compact deep learning model for skin cancer recognition, in: 2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference, UEMCON, 2018, pp. 255–261.

[32] Niharika Gouda, J. Amudha, Skin cancer classification using ResNet, in: 2020 IEEE 5th International Conference on Computing Communication and Automation, ICCCA, 2020, pp. 536–541.

[33] B. Swathi, K.S. Kannan, S. Sreenivasa Chakravarthi, Getla Ruthvik, J. Avanija, C. Chandra Mohan Reddy, Skin cancer detection using VGG16, InceptionV3 and ResUNet, in: 2023 4th International Conference on Electronics and Sustainable Communication Systems, ICESC, 2023, pp. 812–818.

[34] J. Velasco, C. Pascion, J. Alberio, J. Apuang, J. Cruz, M. Gomez, B. Molina Jr., L. Tuala, A. Thio-ac, R. Jorda Jr., A smartphone-based skin disease classification using mobilenet cnn, 2019, arXiv preprint arXiv:1911.07929.

[35] S. Voggu, K. Rao, A survey on skin disease detection using deep learning techniques, J. Algebraic Stat. 13 (2022) 3916–3920.

[36] R. Bhavani, V. Prakash, R. Kumaresh, R. Srinivasan, Vision-based skin disease identification using deep learning, Int. J. Eng. Adv. Technol. 8 (2019) 3784–3788.

[37] S. Aladhadh, M. Alsanea, M. Aloraini, T. Khan, S. Habib, M. Islam, An effective skin cancer classification mechanism via medical vision transformer, Sensors 22 (2022) 4008.

[38] M. Heenaye-Mamode Khan, N. Gooda Sahib-Kaudeer, M. Dayalen, F. Mahomedaly, G. Sinha, K. Nagwanshi, A. Taylor, et al., Multi-class skin problem classification using deep generative adversarial network (DGAN), Comput. Intell. Neurosci. 2022 (2022).

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[40] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[41] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, 2023, arXiv preprint arXiv:2304.07193.

[42] Lihong Peng, et al., LDA-VGHB: identifying potential lncRNA–disease associations with singular value decomposition, variational graph auto-encoder and heterogeneous Newton boosting machine, Brief. Bioinform. 25 (1) (2024) bbad466.

[43] Lihong Peng, et al., CellDialog: A computational framework for ligand–receptor-mediated cell–cell communication analysis III, IEEE J. Biomed. Heal. Inform. (2023).
[44] A. Aboulmira, H. Hrimech, M. Lachgar, Comparative study of multiple CNN models for classification of 23 skin diseases, Int. J. Online Biomed. Eng. 18 (2022).
[45] A. Sah, S. Bhusal, S. Amatya, M. Mainali, S. Shakya, Dermatological diseases classification using image processing and deep neural network, in: 2019 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS, 2019, pp. 381–386.
[46] A. Bindhu, K. Thanammal, Multi-stage feature extraction-based classification of skin cancer detection, Soft Comput. (2023) 1–14.
[47] A. Kumar, A. Vishwakarma, V. Bajaj, S. Mishra, Novel mixed domain hand-crafted features for skin disease recognition using multi-headed CNN, IEEE Trans. Instrum. Meas. (2024).