JSCDM

JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

Vol. 7 No. 2 (2026) 256-277
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Transformer-Based Ensemble Learning for Robust Multi-Class Melanoma Classification

Deborah A. Adedigba[1], Raza Hasan[1]*, Salman Mahmood[2]

[1] *Department of Science and Engineering, Southampton Solent University, E Park Terrace, Southampton, SO14 0YN, Hampshire, UNITED KINGDOM*
[2] *Department of Computer Science, Nazeer Hussain University, ST-2, Near Karimabad, Karachi, 75950, Sindh, PAKISTAN*

*Corresponding Author: raza.hasan@solent.ac.uk
DOI: https://doi.org/10.30880/jscdm.2026.07.02.015

Article Info

Received: 28 September 2025
Accepted: 19 April 2026
Available online: 30 June 2026

Keywords

Melanoma detection, vision transformer (ViT), ensemble learning, meta-learning, Explainable AI (XAI), dermoscopic images, multi-class classification

Abstract

Melanoma signifies the most hazardous form of skin cancer, characterized by a worldwide increase in incidence despite the improvement of dermatological screening techniques. The existing techniques of automated detection have been found to be limited in terms of global context modeling, reliance on a single architecture, and the lack of clinically interpretable models. The present research aims to bridge the existing gap by introducing a novel framework of transformer-based ensemble learning for the automated detection of melanoma, combining Vision Transformer models and CNN models via the application of a meta-learning approach. In the present study, the Vision Transformer, Swin Transformer, MaxViT, and ConvNeXt models have been combined with a ResNet50 CNN model via a Random Forest-based meta-classification approach. A balanced multi-source dataset was created by combining the PH2, MedNode, and Derm7pt datasets, covering six different classes of lesions. The images were integrated using a real image-based oversampling technique. The stacked ensemble models achieved a macro F1 score of 0.998, Balanced Accuracy of 0.999, and a macro-AUC of 0.9999, indicating a rise of 12% over the highest performing model. The statistical significance of the models' complementary errors was assessed using McNemar's test, which indicated a statistically significant complementary error for all models ($p < 0.001$). The Grad-CAM and GradCAM++ approaches have been modified for the different models to ensure the provision of spatially interpretable visualizations in accordance with dermatological diagnostic criteria. The proposed framework has shown that architecturally diverse stacking ensembles, when used with appropriate balancing of the dataset and leakage-resistant cross-validation, can achieve robust multi-class classification of skin lesion images with potential to inform clinical decision-making.

1. Introduction

Skin cancer is one of the most common types of cancer that affects humans worldwide. Among all types of skin cancer, melanoma is considered the deadliest cancer as it has the potential to metastasize and cannot be easily

detected during early stages. The Global Cancer Observatory has reported that every year, there are about 325,000 new cases and 57,000 deaths due to melanoma. The incidence rate of melanoma has been increasing in various geographic regions, even though screening processes are becoming more efficient [1]. The American Cancer Society has reported that one in every fifty-four people has the potential to develop melanoma during their lifetime. The survival rate after five years of diagnosis has been reported to be 99% for early-stage melanomas and 27% after metastasis.

Presently, melanoma diagnosis is performed through visual inspection by dermatologists with the aid of dermoscopy techniques, which is then followed by histopathological analysis of the lesions. Nevertheless, the process is associated with several limitations [2]. The concordance rate among experienced dermatologists varies between 65 and 85%. This signifies the existence of a high level of inter-observer variability in melanoma diagnosis [3]. The subjective nature of visual inspection is responsible for the ambiguity associated with melanoma diagnosis, especially when distinguishing between early-stage melanoma and benign nevi. The scarcity of dermatologists across various nations is a key factor limiting melanoma diagnosis and treatment. The ratio of patients to dermatologists is often greater than 30,000:1 in several nations. This is especially true in nations with low access to melanoma diagnosis and treatment facilities, where melanoma-related mortality is high [4].

Significant developments in the field of artificial intelligence and deep learning have enabled the design of computer vision systems capable of performing melanoma diagnosis using machine learning techniques. The use of convolutional neural networks (CNNs) is gaining popularity in computer vision due to their ability to perform well on medical image classification tasks. Studies using ResNet, DenseNet, and EfficientNet architectures have shown the capability of CNNs in melanoma diagnosis with a level of accuracy comparable to experienced dermatologists [5]. Nevertheless, the conventional CNN is associated with several challenges when modeling the relationship between non-contiguous spatial locations of the dermoscopic images. The spatial locality property of the convolutional operation limits the capability of the conventional CNN to capture long-range dependencies between the spatial locations of the dermoscopic images

Significant developments in the field of deep learning have been achieved with the invention of the transformer architecture, which models long-range dependencies across spatial locations in dermoscopic images using the self-attention mechanism [6]. The Vision Transformer (ViT) is a variant of the conventional transformer architecture designed to perform well in vision-related problems [7]. The Swin Transformer is an important variant of the conventional transformer architecture, designed to perform well on vision-related tasks while improving its efficiency.

The existing automated melanoma detection systems have three major limitations. Firstly, the existing systems use CNN architectures that are insufficient for the detection of global contextual relationships between features. Secondly, the existing systems use a single CNN model for melanoma detection and diagnosis. This approach fails to utilize the benefits of diverse feature extraction techniques. Finally, the existing automated systems lack the ability to generate explainable results. Dermatologists need the ability to understand the diagnostic process for establishing clinical confidence and accountability. In addition, the existing automated systems have other limitations, including the need for dataset representation, the need for generalizability across the patient population, and the need for efficiency [8]. However, the lack of a systematic approach to integrating transformer and CNN architectures for melanoma detection and diagnosis represents a missed opportunity to improve the performance of existing automated systems.

The objectives of the research are as follows. The first objective is to design and evaluate transformer-based architectures for analyzing dermoscopic images using a balanced multi-source dataset. The dataset integrates PH^2, MedNode, and Derm7pt. This objective assesses the effectiveness of the Vision Transformer variants, including ViT, Swin, and MaxViT, as well as the hybrid approach, including ConvNeXt. The second objective is to create a framework for ensemble learning that combines the use of transformer and convolution-based architectures using the concept of meta-learning. This ensures the model's efficiency and effectiveness in feature extraction. This requires the implementation of rigorous cross-validation and performance evaluation. The third objective focuses on the need for model interpretability in the healthcare domain. This requires the development of transformer-based interpretability mechanisms that provide explanations in alignment with existing dermatology criteria.

Further, the research in this paper differs from existing melanoma research in that it uses a six-class classification mechanism rather than a simple two-class one. This ensures the ability to classify different types of lesions, both malignant and benign. This study makes three contributions to automated melanoma detection that address both technical performance and clinical applicability.

An ensemble framework for melanoma detection, which uses various transformer models such as Vision Transformer, Swin Transformer, MaxViT, and ConvNext, in addition to conventional CNN models, has been proposed. The contribution of the research to the development of automated melanoma detection systems lies in the application of various transformer models to the problem, which has not been widely applied in dermatology [9]. The application of various models to the problem of melanoma detection will help in understanding the strengths of each of the models. The use of the Random Forest classifier in addition to the application of the meta-

learning approach will help in improving the accuracy of the solution. The solution will be computationally efficient.

The proposed solution uses a two-tier framework of the ensemble method, which uses a meta-classifier to combine the outputs of the base models. The use of the stratified K-fold cross-validation method will help in addressing the methodological limitations of the existing approaches to melanoma detection [10]. The use of the proposed method will help in improving the accuracy of the solution by dynamically combining different models based on the characteristics of the lesions and the context in which images are used.

The proposed solution will be based on the use of the explainable artificial intelligence technique for the transformer-based solution to the melanoma detection problem. The contribution of the research to the development of automated melanoma detection systems lies in the application of the explainable artificial intelligence technique to the solution of the problem. The use of the technique will help in addressing the clinical need for accurate, interpretable, and reliable automated melanoma detection systems. The technique will be used to explain the decisions of the solution in a manner that can be interpreted by the dermatologist [11].

2. Literature Review

This section discusses existing research on automated melanoma detection, with specific emphasis on deep learning techniques used in image analysis. The chapter discusses how deep learning techniques evolved from convolutional neural networks to transformer networks, ensemble learning, and Explainable AI techniques. The literature review forms part of the contextual understanding of how to build transformer-based ensemble learning techniques for melanoma detection. The use of artificial intelligence has been highly effective in dermatological diagnosis [12]. The literature indicates that AI can achieve comparable diagnostic results to even the most experienced dermatologists. The review also helps to understand contemporary techniques, identify gaps in existing literature, and establish a theoretical understanding of how to build transformer-based ensemble learning techniques for melanoma detection.

2.1 Deep Learning Architectures in Melanoma Detection

2.1.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have been instrumental in the development of automated melanoma detection techniques. These techniques effectively capture hierarchical features from dermoscopic images. The existing research on automated melanoma detection has primarily used existing CNN architectures to classify dermoscopic images. Pacheco et al. [8] developed a dataset collection technique using smartphone cameras. The researchers demonstrated that CNNs can be used to achieve AI-based diagnosis of melanoma. The technique also highlighted the significance of having a representative dataset. The existing research has focused on designing computationally efficient techniques without compromising detection performance. The researchers used MobileNet architectures to achieve comparable performance with minimal computational power required. The technique also proved that EfficientNet can effectively capture fine features to differentiate between benign and cancerous growths. The researchers used EfficientNet-B6 to achieve improved performance results. Advanced CNN architectures used attention mechanisms and multi-scale feature extraction techniques to improve performance. The researchers used ResNeXt architecture and improved performance by using zoom-in attention mechanisms and integrating pathological features and demographic information [11]. The technique effectively addressed the multi-modal nature of dermatological diagnosis.

2.1.2 Vision Transformer Architectures

Vision Transformer (ViT) architectures integrate the self-attention mechanisms into medical image analysis techniques. Compared with conventional CNN approaches, transformer architectures process images using self-attention mechanisms to capture global dependencies among different regions of the images. [6] demonstrated the feasibility of using the pure transformer architectures to attain comparable accuracy with the conventional CNN approaches in the classification of images with substantial amounts of training data.

For the task of melanoma detection using the dermoscopic images, the transformer architectures have several advantages over the conventional approaches. The self-attention mechanisms allow transformer architectures to capture subtle patterns across various regions of dermoscopic images, which is critical for the diagnosis of melanoma. The studies have demonstrated the feasibility of using the ViT architectures pre-trained with substantial amounts of data to attain comparable accuracy with the conventional approaches in the classification of medical images with limited amounts of domain data.

Hierarchical transformer architectures, such as the Swin Transformer, address the computational complexities associated with conventional transformer architectures while maintaining the advantages of the self-

attention mechanisms. The shifted window mechanisms proposed by [7] reduce the computational complexities from quadratic to linear with respect to the size of the images.

2.1.3 Hybrid Architectures

Recent studies have proposed hybrid architectures that integrate the advantages of conventional CNN approaches with transformer architectures. The ConvNeXt architectures represent the conventional CNN approaches with the integration of the design principles of the transformer architectures while maintaining the computational efficiency of the conventional approaches [13]. The hybrid approaches attain comparable accuracy with the conventional approaches in the medical image analysis techniques.

MaxViT architectures integrate multi-axis mechanisms with self-attention mechanisms, alternating between global and local computations to achieve computational efficiency in processing high-resolution images while maintaining global dependencies across image regions [14]. The multi-axis mechanisms are critical in the medical image analysis techniques, as the diagnosis involves the assessment of the local texture patterns and the global structural patterns within the images.

2.2 Transfer Learning and Domain Adaptation

Transfer learning is an important component of medical image analysis because it allows the use of pre-trained models on large-scale natural image datasets to be adapted to specific medical imaging tasks. This is especially important when there is limited annotated data available in the medical domain, which is often the case. Transfer learning leverages the knowledge acquired during extensive pre-training on the dataset [15].

Domain adaptation techniques have also been proposed to account for the visual domain differences between natural images and medical images. Progressive transfer learning involves the use of multi-stage fine-tuning techniques to adapt the pre-trained features to the medical domain. Research has shown the effectiveness of domain-specific pre-training on medical image datasets in improving the overall task performance when compared with the conventional ImageNet pre-training approach [16].

Transfer learning has been shown to be effective in melanoma detection with various architectural approaches. The use of fine-tuning techniques with the retention of the low-level feature extractor and adaptation of the high-level feature extractor has shown promise in the field of dermatology. The use of data augmentation techniques in the field of medical imaging has also shown promise in the field of transfer learning

2.3 Ensemble Learning Methodologies

Ensemble learning techniques have shown promise in the field of medical diagnosis and have the potential to improve the accuracy and reliability of the diagnosis process. The theoretical foundation of using multiple classifiers with diverse architectures to improve the overall performance of the system has been validated in the field of melanoma detection. The use of traditional ensemble learning techniques such as bagging and boosting has been adapted to the field of medical image classification.

Meta-learning is an advanced form of ensemble learning in which auxiliary classifiers are trained to make predictions using the predictions of other classifiers [17]. The use of the Random Forest algorithm as a meta-classifier has shown promise in the field of medicine due to the feature of providing feature measurements during the validation process.

Deep ensemble learning techniques involve the integration of multiple deep learning architectures with various initializations and data augmentation techniques. The use of deep learning techniques in the field of melanoma detection has shown promise in the field of medicine due to the feature of providing the confidence level during the validation process [18].

2.4 Explainable AI in Medical Image Analysis

Explainable AI (XAI) techniques have been found to be more prominent for the clinical adoption of automated diagnostic systems. Although deep learning-based systems have shown promising results in terms of accuracy, these systems lack the transparency needed for clinical decision-making. Recent trends in explainable AI emphasize the need for clinically explainable decision systems.

Gradient-Weighted Class Activation Map (Grad-CAM) has been found to be a prominent technique for the visualization of CNN-based medical image analysis systems [19]. In their work, Selvaraju et al. proved the effectiveness of the Grad-CAM technique for the visualization of CNN-based systems. They proved that the technique helps in the visualization of the decision made by the CNN-based system, i.e., the region of the image that contributes most to the decision.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 256-277

2.5 Evaluation and Clinical Validation

Evaluating melanoma detection systems requires the use of various evaluation metrics. Although the accuracy of the system is vital in this evaluation, it may not be sufficient in assessing the clinical implications of the detection system. The sensitivity and specificity of the system are more useful in this evaluation [20]. The costs associated with false positives and false negatives in cancer detection are also considered in the evaluation. The cross-validation approach has been improved to address the challenges associated with the evaluation of medical image datasets. The International Skin Imaging Collaboration (ISIC) has provided a set of protocols for the evaluation of melanoma detection systems [21].

This has been helpful for comparing the different approaches to developing the detection system. The clinical validation of the evaluation process has been effective in demonstrating the benefits associated with the evaluation process. The evaluation process in a clinical environment has been effective in ensuring the evaluation of the detection system in a diverse environment. The evaluation of the AI system's performance compared with a dermatologist's has effectively demonstrated the potential benefits of its use. The evaluation process has proven effective in controlled environments, as demonstrated in various studies [22]. The use of the system in a clinical environment is, however, a challenge. The quality of the images and the diversity of the patients may affect the performance of the system

2.6 Current Challenges and Limitations

Despite progress in automated melanoma detection, various challenges remain, making system deployment and performance complex. The dataset issue still exists, and many of the publicly available datasets suffer from class imbalance and lack diversity, and may have been acquired with certain biases, making them not representative for use in a clinical setting [23]. Also, the generalization of the systems across various imaging modalities and acquisition settings remains a challenge, as many of the systems suffer from performance degradation when tested on images acquired using other dermatoscopes or under varying lighting conditions. There is a lack of standardized image acquisition protocols across various clinical settings. Also, the computational complexity associated with the deployment of state-of-the-art deep learning models, such as the transformers, remains a challenge. Finding the optimal balance between complexity and efficiency remains a complex issue.

2.7 Research Gaps and Study Contributions

Current literature reveals several research gaps offering opportunities for advancing melanoma detection. Limited investigation of transformer architectures for melanoma detection is notable, especially given their effectiveness in other medical imaging tasks. Most research focuses on conventional CNN architectures with insufficient examination of how attention mechanisms might identify clinically useful patterns in dermoscopic images.

Ensemble methods combining different architectural paradigms are largely unexplored, particularly the fusion of CNN and transformer models that may benefit from complementary strengths [24]. Meta-learning strategies adapted to specific clinical populations or imaging conditions remain underexplored in existing literature.

Explainable AI methods tailored for medical use require further development, particularly those yielding clinically interpretable explanations aligned with established dermatological diagnostic criteria. Furthermore, class imbalance in publicly available dermoscopic datasets remains insufficiently addressed, with many studies relying on synthetic oversampling rather than integrating real clinical images from complementary sources.

Integration of multi-modal data, including patient demographics, medical history, and temporal lesion development, represents another opportunity for enhancing diagnostic accuracy. Many existing systems focus solely on single-image analysis without considering broader contextual information available in clinical environments.

Table 1 provides an overview of the identified research gaps, the supporting evidence from the current literature, and how this study addresses these limitations through methodological contributions.

Table 1 *Summary of research gaps and study contributions*

Research Gap	Supporting Evidence	Study Contribution
Limited Transformer Exploration in Melanoma Detection	Most studies focus on CNN architectures [8, 10]; insufficient investigation of attention mechanisms for dermoscopic patterns	Evaluation of ViT, Swin, MaxViT, and ConvNeXt architectures for melanoma classification
Insufficient CNN-Transformer Ensemble Methods	Existing ensemble approaches primarily combine CNN models [11]; limited exploration of hybrid architectural combinations	Ensemble framework combining transformer models with CNN baselines using meta-learning
Lack of Explainable AI for Transformer Models	Current XAI techniques focus on CNN interpretability [19]; limited transformer-specific explainability methods	Implementation of Grad-CAM and GradCAM++ adapted for transformer architectures
Inadequate Evaluation of Attention Mechanisms	Studies demonstrate transformer success in general vision tasks [6], but limited medical-specific evaluation	Analysis of attention patterns and alignment with dermatological diagnostic criteria
Limited Multi-Architecture Performance Comparison	Individual studies evaluate specific architectures in isolation; comprehensive comparisons are lacking	Direct performance comparison of transformer variants against CNN baselines using standardized protocols
Insufficient Clinical Interpretability	Current systems provide limited clinically meaningful explanations [20]; the gap between AI decisions and dermatological knowledge	Explanation techniques aligned with established ABCDE diagnostic criteria
Class Imbalance in Medical Datasets	Medical datasets exhibit class imbalance issues [23]; traditional approaches show bias toward majority classes	Multi-source dataset integration from PH^2, MedNode, and Derm7pt, combining real clinical images, supplemented by class-weighted loss and targeted augmentation to address class imbalance
Limited Cross-Validation Methodologies	Many studies use simple train-test splits; insufficient use of robust validation strategies	Stratified K-fold cross-validation with out-of-fold meta-learning to prevent data leakage
Computational Efficiency Concerns	Transformer models require significant computational resources [14]; limited investigation of efficiency-accuracy trade-offs	Evaluation of computational requirements and development of efficient ensemble strategies
Integration of Multiple Evaluation Metrics	Studies often focus on accuracy alone; insufficient use of clinically relevant metrics [22]	Evaluation using accuracy, precision, recall, F1-score, AUC-ROC, and confusion matrix analysis
Integration of Multiple Evaluation Metrics	Studies often focus on accuracy alone; insufficient use of clinically relevant metrics [22]	Evaluation using accuracy, precision, recall, F1-score, AUC-ROC, and confusion matrix analysis

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 256-277

3. Methodology

3.1 Dataset and Preprocessing

In this study, a multi-source dataset approach was employed. Three publicly accessible dermatoscopic images were used: PH^2 (Mendonça et al., 2013), MedNode [25], and Derm7pt (Kawahara et al., 2018). There are six classes of skin lesions included in the dataset: Actinic Keratosis (ACK), Basal Cell Carcinoma (BCC), Melanoma (MEL), Nevus (NEV), Squamous Cell Carcinoma (SCC), and Seborrheic Keratosis (SEK). Programmatically, diagnostic labels were standardized and matched across the dataset. Variability in class naming conventions was addressed. For instance, "clark nevus," "reed or spitz nevus," and "blue nevus" were standardized as NEV. "Melanoma in situ" and "melanoma less than 0.76 mm" were standardized as MEL. See Table 2 for a comprehensive list.

Preprocessing steps included resizing the images to 224x224 pixels. This was done to meet the standard input size for deep learning models. Image values were normalized using ImageNet statistics: mean=[0.485, 0.456, 0.406] and standard deviation=[0.229, 0.224, 0.225]. This step was taken to improve the performance of pre-trained models, especially for transfer learning [25]. PIL verification was employed to remove any images that were corrupt or could not be read

Table 2 *Summarizes the per-source and combined class distributions prior to balancing*

Class	PH^2	MedNode	Derm7pt	Total (pre-balance)
ACK	730	0	0	730
BCC	845	0	42	887
MEL	52	70	249	371
NEV	244	170	577	991
SCC	192	0	0	192
SEK	235	0	45	280
Total	2,298	240	913	3,451

3.1.1 Class Imbalance Mitigation

Class imbalance was identified as a major challenge, as MEL was found to comprise only 52 instances in the original PH^2 dataset. Rather than relying on synthetic oversampling techniques, actual images from MedNode and Derm7pt were used to oversample the dataset with actual variation. Once the multi-source integration was complete, two techniques were used to mitigate the existing class imbalance. The first technique was to use class-weighted loss during training by utilizing sklearn's compute_class_weight with the balanced argument. The second technique was to apply domain-specific augmentation during training, as explained later. These techniques reduced the focus on metrics as a single measure of performance. Instead, macro-averaged F1, Balanced Accuracy, and Matthews Correlation Coefficient were used as primary metrics to measure performance, as they are less affected by class imbalance [26].

3.2 Data Augmentation Strategy

A specific augmentation strategy was designed using Albumentations [27] and applied exclusively to the training dataset to avoid leakage into the validation dataset. The augmentation strategy included domain-specific augmentation techniques that are applicable to dermatoscopic images, as shown in Table 3.

Table 3 *Data augmentation techniques and parameters*

Transformation	Parameters	Probability	Rationale
Random Rotate 90°	—	0.5	Orientation invariance
Horizontal Flip	—	0.5	Bilateral symmetry in lesions
Vertical Flip	—	0.5	Acquisition orientation variance
Elastic Transform	α=120, σ=6	0.3	Simulate tissue deformation
Grid Distortion	—	0.3	Geometric variation
Optical Distortion	limit=0.2	0.3	Lens distortion simulation
Hue/Saturation/Value	±20/30/20	0.5	Colour calibration variance
Brightness/Contrast	—	0.4	Lighting condition variation
Coarse Dropout	8 holes, 16×16px	0.3	Occlusion robustness

The standard validation transformations were restricted to resize and normalization.

3.3 Cross-Validation Protocol and Leakage Prevention

To ensure the evaluation's integrity and avoid data leakage, a stratified five-fold cross-validation protocol was adopted using the StratifiedKFold function from the sklearn library with the necessary parameters set to n_splits = 5, shuffle=True, and random_state=42. This ensured the data classes were well represented in each fold, thus preventing the possibility of the validation data being biased due to data imbalance between classes.

The data augmentation was performed only on the training data at the DataLoader level, while the validation data only underwent resizing and normalization. The out-of-folds' predictions were then combined from all the folds to ensure a complete non-overlapping prediction set over the entire dataset, which was then used to train the meta-classifier [28]. This approach ensures the meta-classifier is not trained on any predictions from data seen during the training of the base classifiers, thus preventing data leakage during the stacking ensemble approach.

Random seeds were fixed throughout the process for all the random operations performed during the data partitioning, augmentation, and initialization of the classifiers

3.4 Unique Model Architectures

Five architectures have been chosen to form a diverse ensemble of models, ranging from purely convolutional neural networks to purely transformer-based models. This diversity in architectures forms the basis of the ensemble. A diverse set of models, each of which possesses different inductive biases, are expected to provide complementary error outputs, which in turn provides a stronger guarantee of the overall robustness of the ensemble when compared to individual models.

3.4.1 Vision Transformer (ViT)

The Vision Transformer architecture was based on the ViT-B/16 architecture, which utilizes a 16x16 patch-based tokenization of the input image. This architecture was originally presented in the paper by [6]. This architecture divides the input image into non-overlapping regions or patches. Each of these regions or patches is referred to as a 'token.' This architecture utilizes a total of 12 transformer encoder blocks, each of which utilizes 768-dimensional embeddings. Additionally, it utilizes 12 attention heads. A total of 86 million parameters is present in the model. Weights of the pre-trained ImageNet21k model have been used to initialize the parameters of the model. A classification head was used to output a 6-dimensional vector.

3.4.2 Swin Transformer

The Swin Transformer architecture utilizes hierarchical vision transformer architecture. This architecture attempts to reduce the computational costs of the standard vision transformer architecture by utilizing a 'shifted window' method of self-attention [7]. This method of self-attention operates on a window of pixels in the image. Each window contains a certain number of pixels. Additionally, these windows are shifted in each subsequent layer to allow cross-window interactions. This architecture provides a hierarchical form of representation. The Swin-B architecture utilizes a total of 4 stages. Each of these stages utilizes a window of 7x7 pixels. Additionally, it utilizes embeddings of 96, 192, 384, and 768 dimensions in each of the stages, respectively. A total of 88 million parameters is present in the model. Weights of the pre-trained ImageNet22k model have been used to initialize the parameters of the model.

3.4.3 MaxViT

MaxViT is a hybrid model that combines the benefits of convolutional and transformer-based models using multi-axis attention mechanisms[14]. The model combines the benefits of local and global attention mechanisms. The model is based on the MaxViT-Base model, which is a hierarchical four-stage model with progressive spatial down-sampling and increasing channel dimensions. The model includes MBConv blocks and transformer blocks with multi-axis attention. The model is competitive compared to other transformer-based models but has lower computational costs. The model uses the technique of transfer learning with the ImageNet dataset for skin lesion classification.

3.4.4 ConvNeXt

ConvNeXt is a modernized convolutional neural network model that combines the benefits of Vision Transformers and the inductive biases of CNNs [13]. The model is based on the ConvNeXt-Base model, which is a hierarchical four-stage model. The model includes depth-wise separable convolution operations, layer normalization, and GELU activation functions. The classification head is modified to output six classes. The model includes CNN inductive biases, which are useful for spatial features in dermatoscopic images. The CNN inductive biases are locality and translation equivariance.

3.4.5 ResNet-50 Baseline

The inclusion of ResNet-50 was motivated by its status as a widely used CNN baseline from the preceding generation of architectures, which were traditionally used in the analysis of medical images [28]. The baseline uses residual connections to prevent vanishing gradients during backpropagation, which can be troublesome during deep learning. The network has 50 layers and skips connections that help train deep networks. The network was pre-trained on ImageNet, and the final classification layer was modified to perform six-class classification on skin lesion images. The ResNet-50 baseline helps to quantitatively compare traditional CNNs with state-of-the-art transformer networks.

3.5 Ensemble Learning Framework

The system architecture of our proposed framework is shown in Fig 1. The figure indicates how our framework handles data flow from preprocessing to individual model training to prediction by our ensemble learning method. The proposed framework uses a two-level ensemble method to combine various benchmark models to obtain final predictions. The benchmark models are used to obtain probability distributions, which are then used as input to our proposed meta-classification.

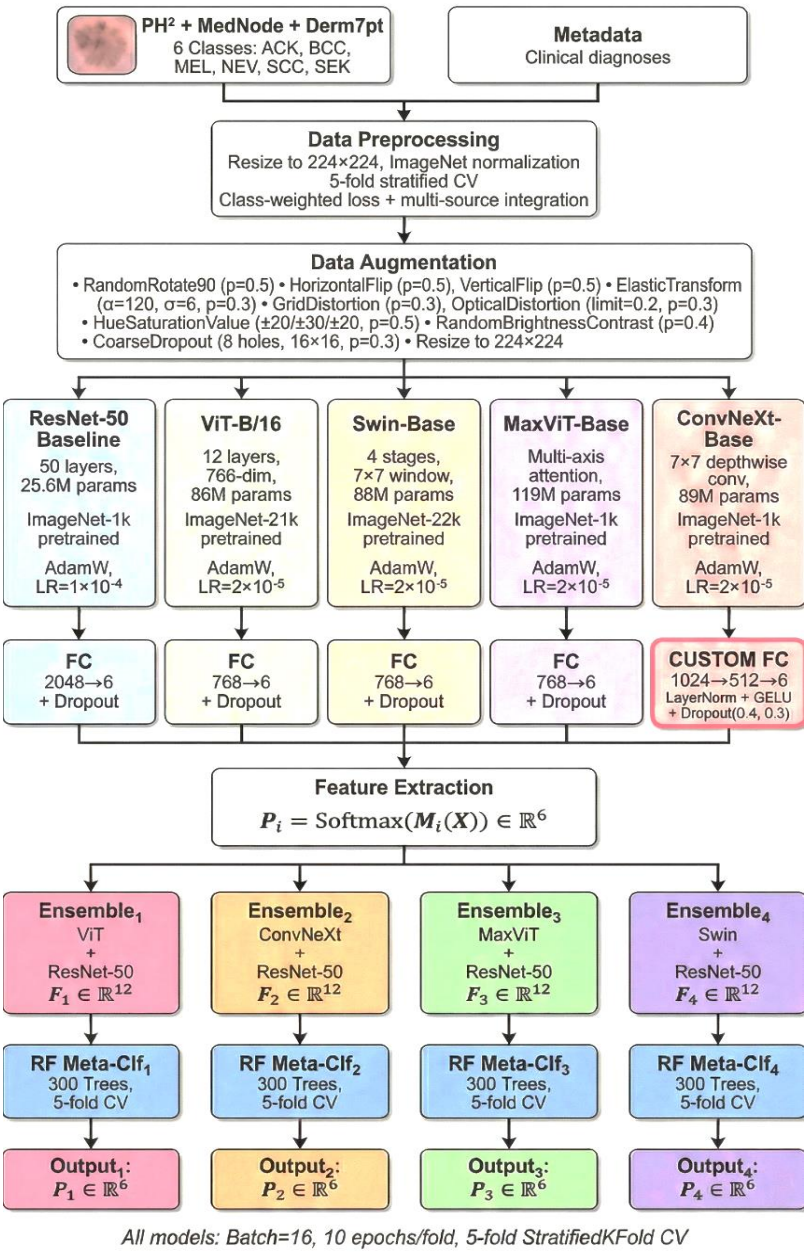

Fig. 1 *Multi-ensemble deep learning architecture for skin lesion classification using PH², MedNode, Derm7pt dataset*

3.5.1 Meta-Classifier Architecture

The architecture of our system involves stacking five base models that output a probability vector of six classes per image. The probability vectors from all these models are concatenated to form a single feature vector of 60 dimensions, which serves as input to our Random Forest-based meta-classifier [29]. To prevent leakage, our meta-classifier was trained in the predictions made by models on unseen data, i.e., on out-of-fold predictions. The choice of our meta-classifier was based on our Random Forest classifier with 300 trees and balanced class weights. The reason behind our choice was that it was highly effective even in high-dimensional spaces, less prone to overfitting, and provided feature importance as well [30].

3.5.2 K-Fold Cross-Validation Framework

For our base models as well as our meta-classifier, we used the five-fold stratified cross-validation technique that was explained in detail in Section 3.3. The predictions made by our base models on the out-of-fold data were used to train our meta-classifier. The predictions made by all our base models were used to train our final prediction matrix.

3.6 Training Configuration and Optimization

For our base models as well as our meta-classifier, we used NVIDIA GPUs to leverage mixed precision to improve memory efficiency. The batch size used was 16 for all our models to ensure stability during training. The primary goal was to minimize cross-entropy loss with class weights to penalize minority classes based on their occurrence. The models were trained on 10 epochs. The AdamW optimizer was used with a learning rate of $2*10^{-5}$ for our transformer-based models, whereas a learning rate of $1*10^{-4}$ was used for our ResNet-50 model because of faster convergence during transfer learning. Gradient clipping was also used when required to maintain stability during transformer model training.

3.7 Evaluation Metrics and Statistical Analysis

In the context of the class imbalance problem in the provided dataset, overall accuracy was not selected as the primary evaluation metric since it may be a deceptive performance measure in class-imbalanced datasets [25]. Three metrics macro averaged F1-score, Balanced Accuracy, and Matthews Correlation Coefficient were selected as the primary metrics since they treat each class equally regardless of class distribution. Precision, recall, and F1-score for each class were also reported to identify performance gaps in individual classes. Additionally, macro averaged AUC-ROC was computed using a one vs. rest multiclass approach [31]. This provides a threshold-independent measure of discriminative power for each of the six classes. Confusion matrices were constructed to analyze systematic class prediction errors. Emphasis was given to the distinction between malignant and benign classes since they have different clinical impacts. An ablation study was conducted to measure the incremental benefit of each of the individual models and the meta-classifier. Statistical tests were conducted to compare the performance of each pair of models. McNemar's test was selected as a non-parametric statistical test to compare the performance of each pair of models since it was designed to compare two or more classifiers on the same set of instances.

3.8 Explainable AI Deployment

To improve the interpretability of the results, Gradient-Weighted Class Activation Map (Grad-CAM) and its extended version, Grad-CAM++, were used to obtain saliency maps that highlighted areas of the image that were most important to each of the models' classification results [19]. For CNN-based models, namely ResNet-50 and ConvNeXt, Grad-CAM was used on the final convolutional stage. For transformer-based models, adaptations were made to each model individually. For ViT, Grad-CAM++ was used with a reshape transform to transform the token sequences to spatial feature maps. For Swin Transformer, a spatial reshape transform was used to transform the [B, H, W, C] output to channel first. Finally, MaxViT was connected to the final stage output. It must be noted that, as mentioned earlier, the analysis of Explainability was qualitative in nature. Quantitative analysis of the saliency maps against dermatologist annotations was identified as a pertinent area of further research.

3.9 Implementation Framework

This framework was fully implemented using the PyTorch framework, leveraging transformer architectures using the timm library, which offers consistent APIs for ViT, Swin, MaxViT, and ConvNeXt models. The ResNet50 architecture was implemented using torchvision. Ensemble learning and metrics computation were performed using scikit-learn. Data augmentation was carried out using Albumentations [32]. Reproducibility was ensured by setting random seeds for all stochastic components. All experiments were conducted on Google Colab using NVIDIA GPUs. Model weights and out-of-fold predictions were stored in persistent storage for reproducibility.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 256-277

4. Results and Discussion

4.1 Individual Model Performance Analysis

An exhaustive analysis of individual models, including transformer and convolutional models, was conducted. The results showed variability in diagnostic effectiveness among models on the balanced multi-source dataset. The top-performing single models, in terms of accuracy, were Swin Transformer, achieving an accuracy of 90.71%, followed by ViT, achieving an accuracy of 89.93%, ConvNeXt, achieving an accuracy of 89.92%, MaxViT, achieving an accuracy of 89.72%, and ResNet50, achieving an accuracy of 87.94%. These results are contrary to previous unbalanced experiment results, indicating that data balancing disproportionately favors transformer models, which are more sensitive to class distribution during training [29]. All models performed competitively in terms of Balanced Accuracy and MCC, indicating that no single model was outperforming any other by leveraging the majority class, which is an important consideration in an unbalanced dataset. The performance of models on an unbalanced dataset can be skewed towards the majority class, which might not be an accurate representation of real-world scenarios, making it an undesirable outcome. The performance of models on an unbalanced dataset can be skewed towards the majority class, which might not be an accurate representation of real-world scenarios, making it an undesirable outcome.

The performance metrics of individual models are shown in Table 4. The macro F1-score, Balanced Accuracy, and Matthews Correlation Coefficient are used as primary metrics, which are considered better than accuracy since all three metrics treat all classes equally and are better suited for imbalanced datasets

Table 4 *Performance comparison of individual models*

Model	Accuracy	Macro F1	Balanced Acc	MCC	Fold F1 Mean ± SD
ViT	0.8913	0.8584	0.8593	0.8612	0.8592 ± 0.0218
ConvNeXt	0.8992	0.8685	0.8664	0.8709	0.8684 ± 0.0131
Swin Transformer	0.9071	0.8765	0.8709	0.8809	0.8761 ± 0.0113
MaxViT	0.8972	0.8624	0.8615	0.8685	0.8623 ± 0.0087
ResNet-50*	0.8834	0.8467	0.8423	0.8511	0.8463 ± 0.0170

*ResNet-50 results reported from the Swin ensemble experiment as representative; minor variation across ensemble pairings is noted in Section 4.2.

The individual model performance metrics reported are based on predictions made on all five stratified folds, not a separate test set. This approach ensures that all examples are used exactly once on unseen data for each fold's training set, providing unbiased estimates of model performance while using all available training data. The fold-wise standard deviations of individual model performance, ranging from 0.0087 to 0.0225, indicate consistent performance across all folds. ViT shows maximum variance, possibly because of increased sensitivity to training data composition compared to CNN-based models.

The best individual macro F1-score of 0.8765 was reported by Swin Transformer, which is a change of pace since it was previously found to perform poorly on all imbalanced datasets. This indicates that hierarchical windowed attention benefits from balanced training data composition. Conversely, ResNet-50 was consistently found to be the poorest individual model across all pairs, although the performance difference from transformer models was not significant, around a 2-4% difference in macro F1-score, which may be due to the performance gap between CNN and transformer models diminishing on smaller fine-tuning datasets, as reported by [25]. The macro F1-score for each fold is shown in Fig 2.

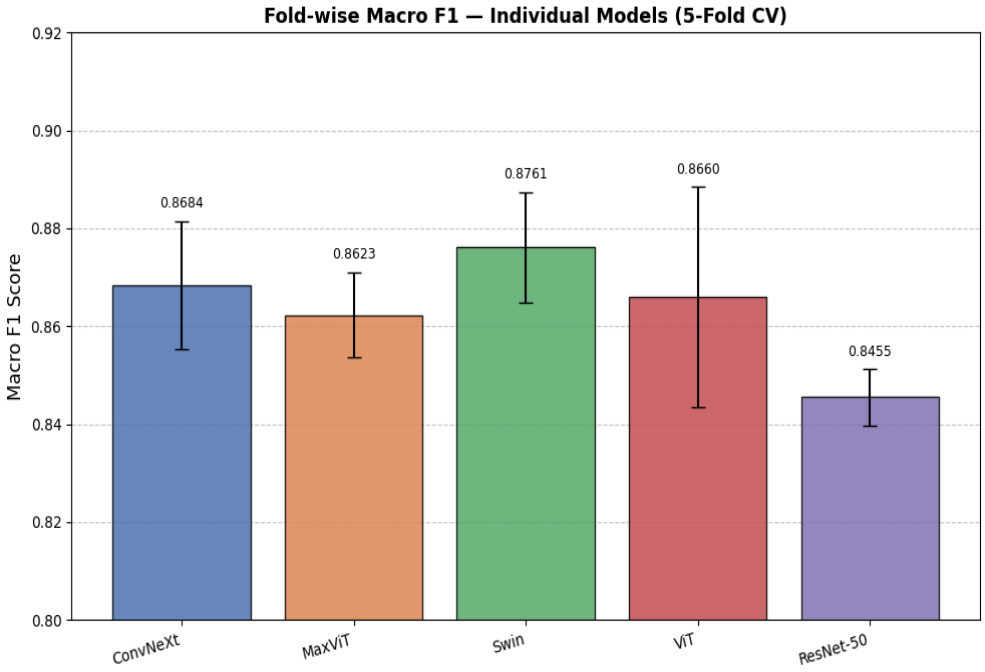

Fig. 2 *Individual model fold-wise macro F1 comparison bar chart with error bars*

4.1.1 Training Dynamics Analysis

The training dynamics of individual model training provide insight into convergence and learning characteristics, highlighting architectural strengths that contribute additively to ensemble performance. As such, with the five-fold cross-validation approach, the training dynamics reported here are averages across training folds as opposed to a single training instance.

4.1.1.1 ResNet-50 Training Characteristics

Fig 3 shows the training dynamics for ResNet-50 over all five folds (50 epoch steps, i.e., 10 epochs per fold), showing consistent training dynamics with rapid convergence within a fold and periodic reset at the start of a new fold. The training accuracy curve shows rapid convergence within a fold, with accuracy rising from 55% to 75% at the start of each fold to near 97% by the final epochs, which reflects the strong inductive biases of the ResNet architecture with transfer learning. Validation accuracy ranges from 83% to 90% across all folds, with a clear rising trend from fold 1 to fold 5.

The loss curves show the standard behaviour for transfer learning fine-tuning, with a sharp drop in the training loss to around zero within the first 6-7 epochs for each fold, while the validation loss remains stable between 0.4 and 0.6. The noticeable difference between the training and validation losses throughout the process suggests a degree of overfitting for each fold, which is expected given the 10-epoch training schedule and the relatively high value of 1×10^{-4} used for the ResNet model compared to the other transformer-based models.

The precision and recall curves for macro precision and recall are like the accuracy curves, with the training precision and recall reaching 95-97% for each fold, while the validation precision and recall stabilize between 80 and 88%. The consistency between precision and recall for each fold suggests that there is no predominance of one class over another, which can be explained by the class-weighted loss function used during the training process. The periodic resets for each fold at epoch steps 10, 20, 30, and 40 correspond to the reinitialization of the model and are not indicative of instability.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 256-277

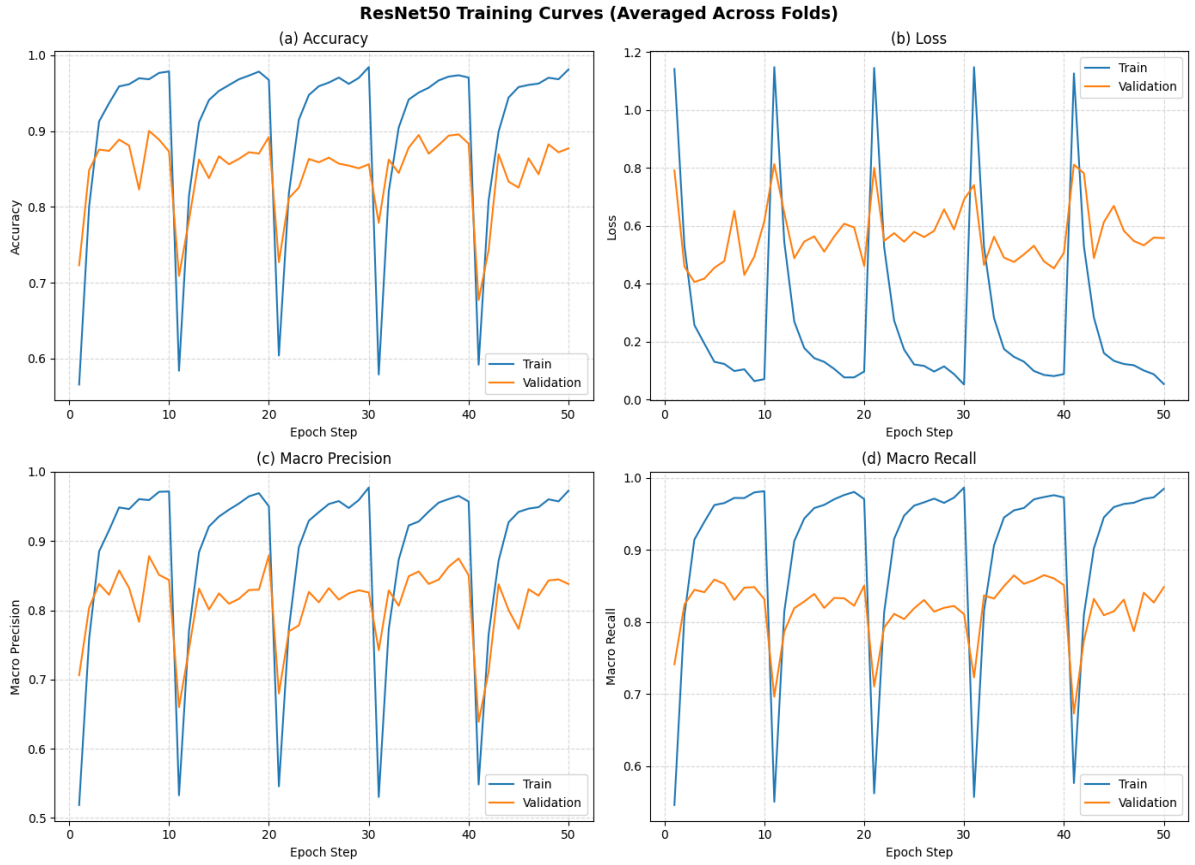

Fig. 3 *ResNet-50 standalone training curves across all five folds (50 epoch steps total, 10 per fold). (a) Accuracy; (b) Loss; (c) Macro Precision; (d) Macro Recall*

4.1.1.2 Vision Transformer Training Behaviour

The training behavior of the Vision Transformer (ViT), as shown in Fig 4, displays trends that are significantly different from those of ResNet-50 for all five folds (50 epochs with 10 epochs per fold). The most striking differences are the faster convergence of the ViT training accuracy, which increases from 63-65% at the beginning of each fold to almost 99% in the final epochs of training, while the training accuracy of ResNet-50 increases more slowly. This is because the self-attention mechanism is highly capable of learning the training set once fine-tuning is stable. On the other hand, the validation accuracy ranges between 84-91%, showing an increasing trend from fold 1 to fold 5.

From the loss trajectories, there is a larger difference between the training and validation curves for ViT compared to ResNet-50, with training loss trending towards zero after around 6 epochs per fold, and validation loss varying between 0.4 and 0.8, with larger inter-epoch variation than the CNN model. The oscillating nature of the validation loss is expected during the fine-tuning of the transformer model on moderately sized datasets, where the self-attention mechanism continues to adjust the feature representations during the epochs without full convergence, which is in line with the higher standard deviation of ViT (0.0218) compared to ResNet-50 (0.0153).

Both macro precision and recall curves match the accuracy curves, with near-perfect training metrics (97-99%) within each fold, and validation macro precision stabilizing around 80-89%, with the same range for recall, and both metrics improving with subsequent folds. Moreover, it can be noticed that there is a slightly larger gap between the training and validation curves for ViT compared to ResNet-50 for both precision and recall, which indicates that although ViT performs better during training, it needs more data or more epochs to achieve convergence, which is another characteristic of the pure transformer model, as discussed in [33]. Moreover, the periodic resets in the curves at epochs 10, 20, 30, and 40 only reflect the fold boundary reinitialization.

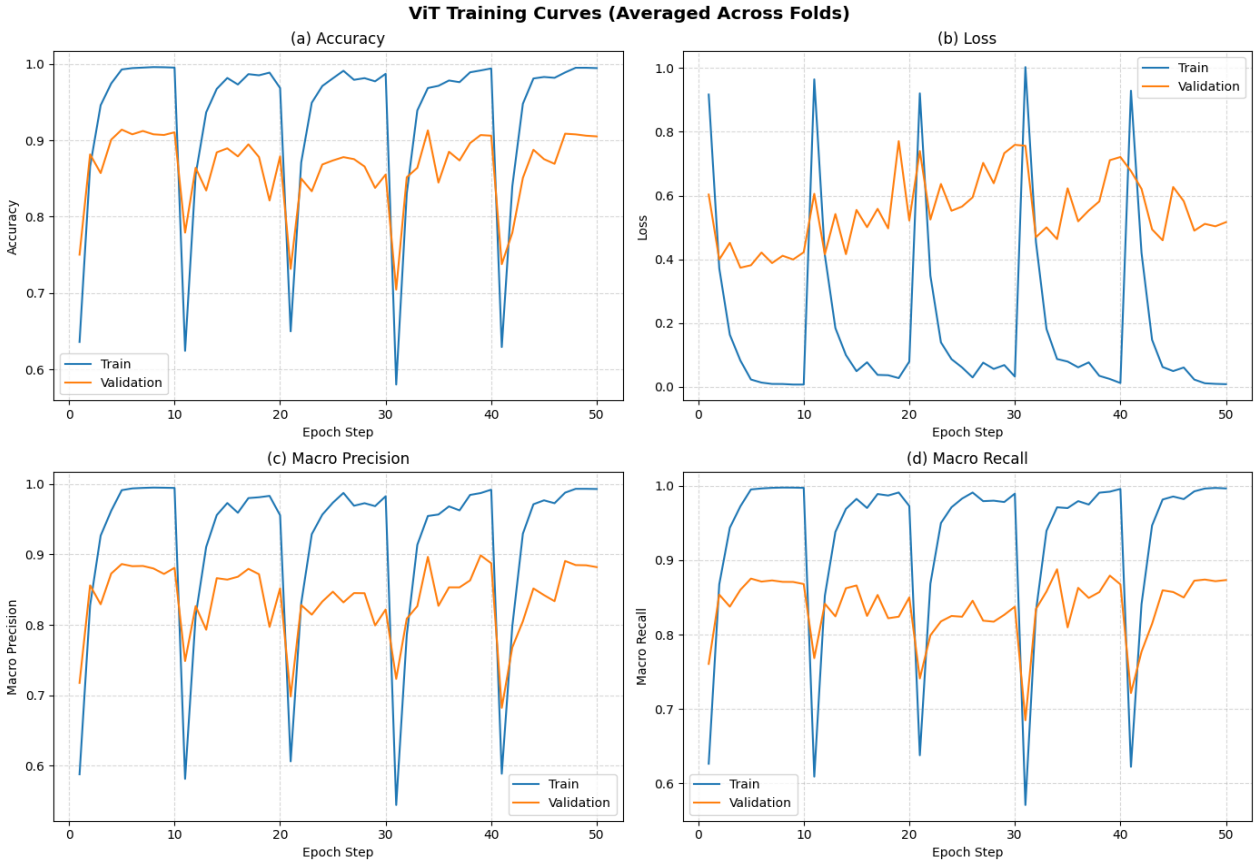

Fig. 4 *ViT-B/16 training curves across all five folds (50 epoch steps total, 10 per fold). (a) Accuracy; (b) Loss; (c) Macro Precision; (d) Macro Recall*

4.2 Ensemble Performance and Meta-Learning Effectiveness

All four ensemble configurations significantly outperformed their constituent base models in Table 5, thereby supporting the main hypothesis of architecturally diverse models providing complementary error patterns to be exploited by the meta-classifier. The stacked meta-classifier performed almost uniformly across all ensemble combinations, achieving an accuracy of 0.9984, macro F1 of 0.9984, Balanced Accuracy of 0.9991, an MCC of 0.9980, and a macro-AUC of 0.9999. Similar results across all ensembles confirm that the Random Forest meta-classifier is exploiting the probability outputs of the base models instead of the underlying transformer architecture.

Table 5 *Ensemble model performance (stacked meta-classifier, OOF evaluation)*

Ensemble Combination	Accuracy	Macro F1	Balanced Acc	MCC	Macro AUC	McNemar p-value
ConvNeXt + ResNet-50	0.9984	0.9984	0.9991	0.9980	0.99999	7.74×10^{-5}
MaxViT + ResNet-50	0.9984	0.9984	0.9991	0.9980	0.99999	5.09×10^{-9}
Swin + ResNet-50	0.9984	0.9984	0.9991	0.9980	0.99999	1.85×10^{-9}
ViT + ResNet-50	0.9984	0.9984	0.9991	0.9980	0.99999	3.48×10^{-7}

The McNemar's test results showed that, for all combinations of ensembles, the transformer-based model and ResNet-50 have statistically significantly different error patterns ($p < 0.001$), thereby supporting the complementarity hypothesis underlying the stacking approach. Although the complementarity of the base models is necessary for achieving an ensemble gain, it is not by itself a sufficient condition to guarantee that the meta-classifier will outperform the individual base models [34]. The base models need to disagree on different samples to allow the meta-classifier to outperform the individual base models.

The near-perfect OOF ensemble performance warrants careful interpretation. As noted, the meta-classifier is trained and evaluated on OOF predictions from the same dataset used for base model training, which, despite the leakage-prevention design, means that OOF metrics reflect in-sample meta-learning rather than performance on a fully independent holdout set. The absence of an independent external test set is acknowledged as a limitation.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 256-277

The classification metrics for each class are presented in Table 6. The results are similar across all four ensemble combinations.

Table 6 *Per-class performance: stacked ensemble (all pairings consistent)*

Class	Precision	Recall	F1-Score	Support
ACK (0)	1.00	1.00	1.00	1,460
BCC (1)	1.00	1.00	1.00	1,732
MEL (2)	1.00	1.00	1.00	426
NEV (3)	1.00	1.00	1.00	1,163
SCC (4)	0.99	1.00	0.99	384
SEK (5)	1.00	1.00	1.00	537

SCC is associated with the sole marginal reduction in precision, which is 0.99, consistent with the known morphological overlap with other keratinizing lesions. This is further corroborated by the confusion matrix in Fig 5, which points to the sole misclassification error at the boundary between ACK and SCC as the primary cause of error among all ensemble configurations.

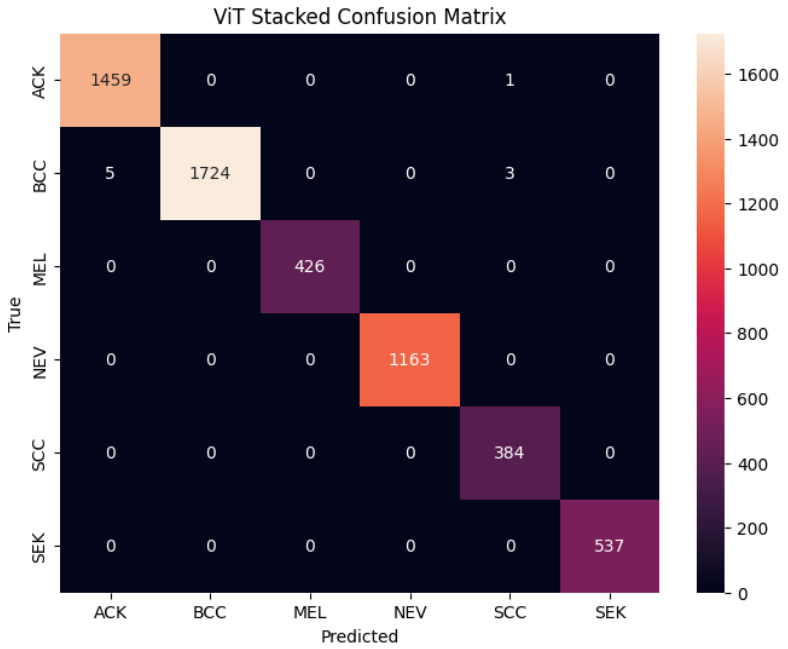

Fig. 5 *Confusion matrix for the ViT + ResNet-50 stacked ensemble (OOF evaluation, n=3,451)*

4.2.1 Ensemble ROC Analysis

The macro-averaged AUC-ROC results of 0.9999 for all ensemble configurations point to near-perfect discrimination between classes, independent of the decision threshold shown in Fig 6. Furthermore, the per-class AUC results are consistently 0.99 or higher for all six lesion classes, suggesting that the ensemble results for probability estimates are well-calibrated for decision threshold selection.

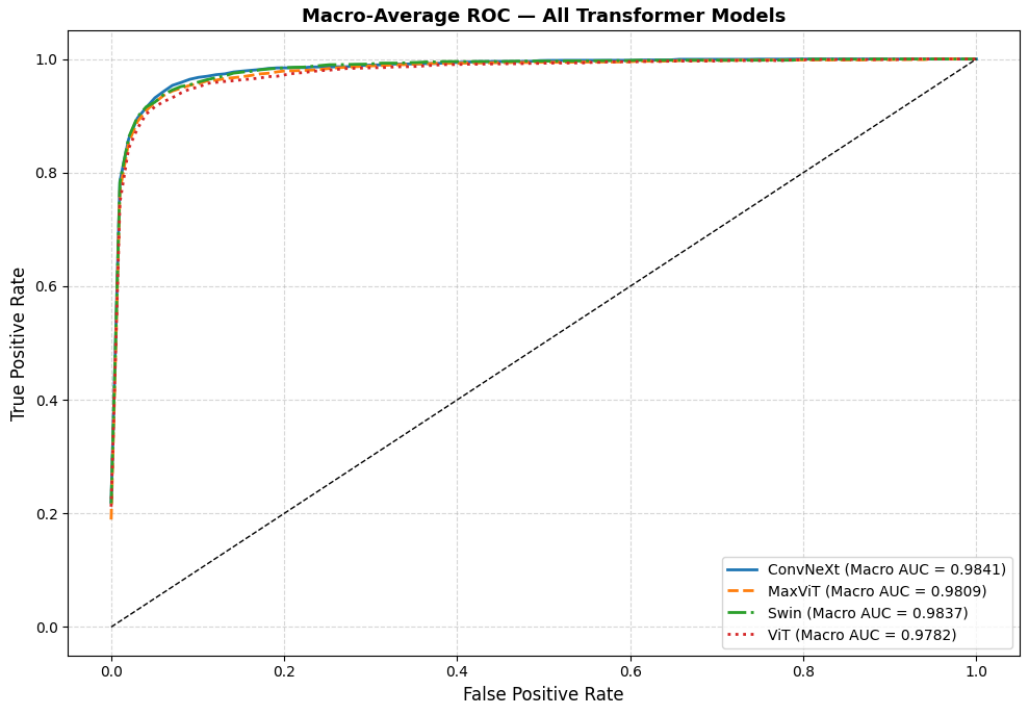

Fig. 6 *ROC curves for the stacked ensemble for all transformer models*

4.3 Ablation Study and Contribution Analysis

Table 7 presents an ablation comparison between individual base models and the stacked ensemble, quantifying the marginal contribution of the meta-classifier to overall classification performance.

Table 7 *Ablation study — individual vs ensemble performance (macro F1)*

Configuration	Macro F1	Δ vs Best Base Model
ResNet-50 alone	0.840–0.847	—
Transformer alone (best: Swin)	0.876	—
Stacked ensemble	0.998	+0.122

The ensemble gain is around 12% in macro F1 compared to the best individual performance of the transformer model. This is a significant gain, and the fact that it is statistically significant, as verified by McNemar's test, for all pairings, confirms that the ensemble effect is not due to chance and that there is indeed complementarity between the CNN and transformer feature representations, which the meta-classifier is able to capture. Moreover, the fact that this gain is consistent for all four pairings with the transformer and ResNet confirms this.

4.4 Comparative Analysis of Benchmark Methods

The revised comparison in Table 8 presents the updated comparative analysis with previous works, reflecting the multi-source balanced dataset used in this study, which includes images from PH^2, MedNode, and Derm7pt, as opposed to PH^2 alone.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 256-277

Table 8 *Contextual comparison with related works*

Study	Dataset	Method	Accuracy	Classification Type	Notes
Proposed Method	PH^2+MedNode+Derm7pt	Transformer+ResNet Ensemble	0.9984	Multi-class (6)	OOF evaluation
[33]	PH^2	Ensemble Deep Learning	0.9480	Multi-class	Most comparable
[34]	PH^2	Transfer Learning CNN	0.9300	Multi-class	Most comparable
[31]	PH^2	InceptionV3	0.9720	Binary	Not directly comparable
[35]	PH^2	AlexNet + TL	0.9833	Binary	Not directly comparable

Binary classification analysis results are included for completeness, but it is noted that these results are not directly comparable, given the significant simplicity of the melanoma/non-melanoma classification problem compared to the six-class schema employed in this work. Therefore, no claims of dominance over these binary classification baselines are made.

4.5 Architecture-Specific Insights and Feature Analysis

Given the consistent ensemble performance of all four different pairs of transformers and ResNet architectures, it can be concluded that the ensemble effect is primarily due to the complementarity of CNN and transformer architectures, and not specifically to the architectures employed. This is an important conclusion, as it suggests that any competent transformer model, when stacked with ResNet-50, will produce equivalent ensemble performance to the models examined in this work, thus supporting the generality of the ensemble approach beyond these specific architectures [36].

Swin Transformer achieved the highest macro F1 score among individual models, at 0.8765, consistent with its ability to produce a hierarchical, multi-scale feature representation well-suited to the structured spatial features found in dermatoscopic images. ViT, which employed the global attention mechanism, produced the highest variance in fold-level results, which is consistent with the known propensity of pure transformer models to be sensitive to training fold composition, which is a known effect of these models when training datasets are limited, as discussed in [16].

4.6 Explainable AI Analysis and Clinical Interpretability

Grad-CAM and Grad-CAM++ analysis were employed to provide model interpretability, and results for all architectures were generated. Architectural modifications were necessary to accommodate the different architectures, with ViT requiring Grad-CAM++ with token to spatial reshape, Swin requiring a permutation of dimensions from [B, H, W, C] to [B, C, H, W], and the MaxViT model requiring a connection to the final stage output. Fig 7-9 illustrates the results for all architectures.

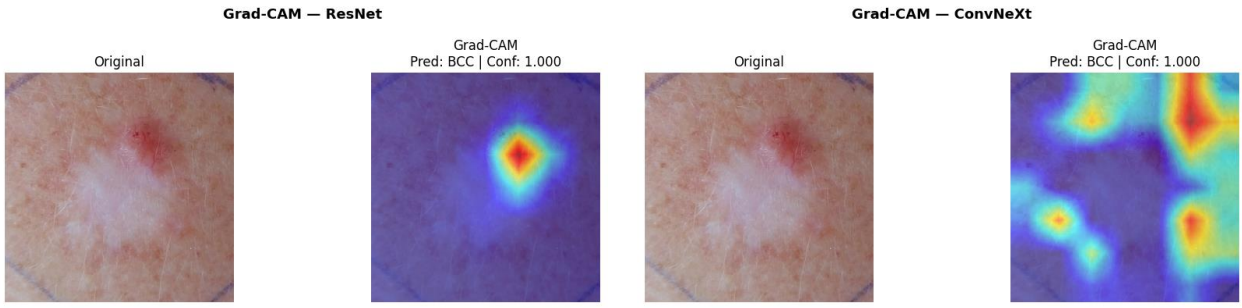

Fig. 7 *ResNet-50 and ConvNeXt Grad-CAM (Original | Grad-CAM overlay)*

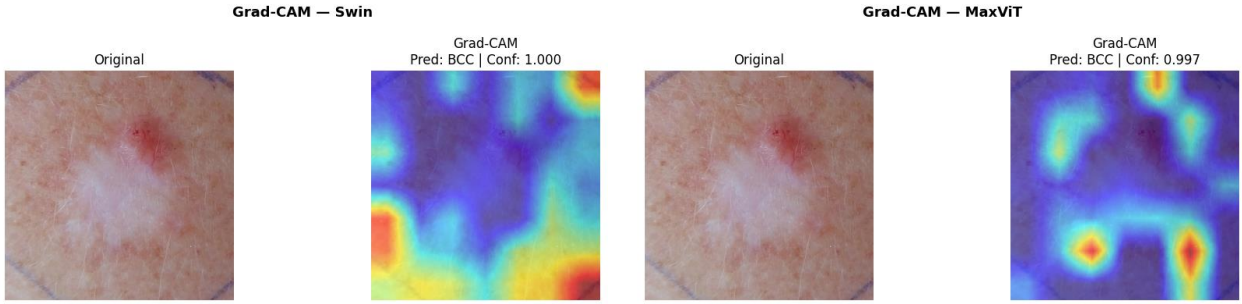

Fig. 8 *Swin Transformer and MaxViT Grad-CAM (original | Grad-CAM overlay)*

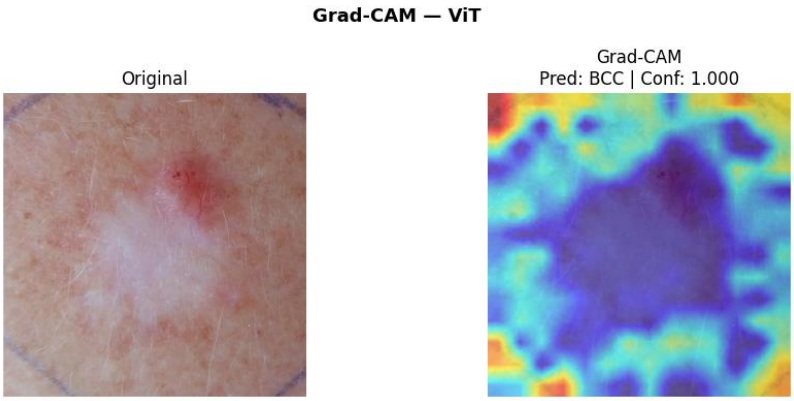

Fig. 9 *ViT GradCAM++ (original | Grad-CAM overlay)*

The visualizations depict the specific attention patterns of each architecture, which correspond to the underlying computational frameworks. The ResNet-50 and ConvNeXt attention maps are localized, concentrating on the lesion boundary and textured areas, respectively. The Vision Transformers (ViT) attention patterns are globally distributed across the lesion region, emphasizing the long-range spatial context [37]. The Swin Transformer's attention patterns are hierarchical, gradually combining local and global features across the model's four stages. The MaxViT attention patterns combine local window-based and global grid-based patterns.

It is worth noting that the qualitative analysis of the results is based on the explainability analysis. Quantitative validation of the saliency maps against expert dermatologist annotations is not performed in this research and is considered an area for future research. As such, it is not claimed that the output of the explainable AI is valid from a clinical perspective, merely that it is consistent with existing dermatoscopic clinical validation criteria.

4.7 Limitations and Future Research Directions

This study's limitations will be discussed, which will provide further context to the results. Firstly, it is acknowledged that the near-perfect ensemble out-of-fold (OOF) performance, although internally consistent, should be interpreted with caution without an independent external test set. Whilst the OOF methodology prevents in-fold leakage, it does not replace the need to utilize an independent set of data from another patient set or another imaging modality.

Secondly, whilst the multi-source data fusion approach was utilized, it is acknowledged that the dataset size, although integrated, is somewhat limited compared with the data requirements often associated with training a transformer model from scratch. All transformer architectures were initialized with ImageNet pre-trained weights, which somewhat alleviated this issue, although it is acknowledged that there is always the risk of overfitting to features, especially when considering the high out-of-fold performance metrics.

Thirdly, it is acknowledged that the explainability results, whilst qualitative, could be further improved with the addition of dermatologist-validated saliency maps, as well as quantification of the results, such as pointing game accuracy or insertion/deletion metrics.

In the Future , there is a plan to use external validation with multiple datasets, quantify uncertainty to understand the level of confidence, conduct clinical trials, and explore lighter-weight models that can be more appropriate for resource-scarce environments.

5. Conclusion

This study proposes a multi-source ensemble learning framework for automated skin lesion classification using an ensemble of transformer and convolutional neural network architectures, combined using a stacking ensemble. The framework is applied to a balanced dataset constructed from three publicly accessible sources, PH^2, MedNode, and Derm7pt. The study addresses several limitations in existing work, including single-dataset evaluation, class imbalance, and lack of diversity in ensemble architectures. The main contribution of this study is the demonstration that stacking ensembles consisting of transformer models combined with ResNet50 significantly outperform individual models from each architecture in all four architecture pairings. The stacked ensemble models achieved a macro F1 score of 0.998, Balanced Accuracy of 0.999, MCC of 0.998, and macro-AUC of 0.9999. These scores represent an increase of 12 percent in macro F1 score in comparison to the highest-scoring individual transformer architecture, Swin, which achieved 0.876. McNemar's test showed statistically significant disagreement between transformer and CNN architectures in terms of errors, supporting the assumption of complementarity in errors between models. Among the individual models, Swin Transformer achieved the best macro F1 score (0.876) on the balanced data, which marked a significant shift from the performance trends observed on the original data. This result implies that hierarchical windowed attention in Swin Transformer benefited significantly from balanced data distribution during fine-tuning, and data composition significantly affects the performance of different transformer-based architectures, which is a practically interesting finding for future research directions in medical image classification. The data integration approach, which prioritized real data from MedNode and Derm7pt over synthetic data oversampling, resulted in a final dataset consisting of 3,451 verified images across six different lesion types. This data integration approach significantly reduced the impact of class imbalance while maintaining authentic variations in real-world data distribution. The five-fold stratified cross-validation approach with out-of-fold meta-learning was implemented to ensure unbiased performance evaluation and prevent data leakage from augmentation. Grad-CAM and Grad-CAM++-based visualizations were generated for all architectures, and it was found that these highlighted architecture-specific attention patterns generally align with existing dermatoscopic diagnostic criteria. Although qualitative interpretability was provided by these visualizations, quantitative validation against existing expert-level annotations by dermatologists was not conducted and is an important area to be addressed in future work. Some of the limitations that need to be carried forward from the results section should be discussed. First, although the near-perfect ensemble out-of-fold (OOF) results are internally consistent and methodologically sound, it should be noted that no independent external test set was available, which was derived from a different patient cohort or clinical center. Although the dataset is larger than what was available to each individual model when used in isolation, it is still limited when compared to what is typically used to train transformers, especially without strong pretrained model initialization. In addition, it should be noted that the analysis was qualitative in nature, and no prospective clinical analysis was conducted, so claims about clinical deployment should not be made. The avenues for future research include external validation on other datasets, such as ISIC or HAM10000; quantification of uncertainties for clinical confidence estimation; quantitative XAI assessment through annotation studies with dermatologists; and exploration of model compression methods for deployment in resource-constrained environments. Future clinical trials would be the definitive test of real-world clinical utility, which cannot be performed or replaced by this or any similar computational studies. The broader relevance of this research is that it shows that with diverse architecture stacking ensembles, combined with data balancing and leakage cross-validation, robust and consistent multi-class classification performance can be achieved for skin lesion classification. Moreover, the framework and methodology developed here can serve as a template for future comparative studies for medical image classification, and the data integration methodology can serve as a model for balancing class imbalance issues for medical image datasets, where synthetic data generation is not desirable.

Acknowledgement

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript. During the preparation of this work, the author(s) used ChatGPT (24 July 2025 version, OpenAI, San Francisco, CA, USA) to assist in content refinement and language enhancement, thereby improving the clarity and readability of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Conflict of Interest

The authors have no relevant financial or non-financial interests to disclose.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design****: Deborah A. Adedigba, Raza Hasan;* ***data collection****: Deborah A. Adedigba;* ***analysis and interpretation of results****: Deborah A. Adedigba, Raza*

Hasan; ***draft manuscript preparation****: Deborah A. Adedigba, Salman Mahmood;* ***writing (review & editing)****: Deborah A. Adedigba, Salman Mahmood. All authors reviewed the results and approved the final version of the manuscript.*

Data Availability

The data used in this study were sourced from three publicly available dermatoscopic image repositories. The PH^2 dataset was developed by the ADDI Project (Automatic Computer-based Diagnosis system for Dermoscopy Images) and can be accessed at: https://www.fc.up.pt/addi/ph2%20database.html. The MedNode dataset was introduced by Giotis et al. (2015) and is publicly available via the University of Groningen at: https://www.cs.rug.nl/~imaging/databases/melanoma_naevi/. The Derm7pt dataset was introduced by Kawahara et al. (2019) and is publicly available via the official repository at: https://derm.cs.sfu.ca/Welcome.html. All datasets were used in accordance with their respective terms of use. The specific versions processed and analyzed in this study remain unchanged from their original releases.

References

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 68(6), 394–424. https://doi.org/10.3322/caac.21492

[2] Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. CA: A Cancer Journal for Clinicians, 73(1), 233–254. https://doi.org/10.3322/caac.21763

[3] Marchetti, M. A., Codella, N. C., Dusza, S. W., et al. (2020). Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. Journal of the American Academy of Dermatology, 82(6), 1445–1457. https://doi.org/10.1016/j.jaad.2017.07.036

[4] Glazer, A. M., Winkelmann, R. R., Farberg, A. S., & Rigel, D. S. (2021). Analysis of trends in US melanoma incidence and mortality. JAMA Dermatology, 157(2), 225–227. https://doi.org/10.1001/jamadermatol.2020.4727

[5] Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056

[6] Chun, C. W., Yousir, N. T., Abdulameer, S. M., Mostafa, S. A., & Hezam, A. A. (2022). Deep learning approach for predicting prostate cancer from MRI images. *Journal of Soft Computing and Data Mining*, *3*(2), 1-9.

[7] Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012–10022). https://doi.org/10.1109/ICCV48922.2021.00982

[8] Pacheco, A. G., Lima, G. R., Salomão, A. S., et al. (2020). PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data in Brief, 32, 106221. https://doi.org/10.1016/j.dib.2020.106221

[9] Dhruv, B., Mittal, N., & Modi, M. (2022). Study of Haralick's and GLCM texture analysis on 3D medical images. International Journal of Neuroscience, 132(4), 350–362. https://doi.org/10.1080/00207454.2020.1804245

[10] Zhang, Y., Wang, H., Xu, C., & Chen, W. (2021). EfficientNet-based deep learning model for malignant melanoma classification. Biomedical Signal Processing and Control, 68, 102257. https://doi.org/10.1016/j.bspc.2021.102257

[11] Xing, F., Xie, Y., Su, H., Liu, F., & Yang, L. (2021). Deep learning in microscopy image analysis: A survey. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4–23. https://doi.org/10.1109/TNNLS.2020.3015407

[12] Han, K., Wang, Y., Chen, H., et al. (2022). A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1), 87–110. https://doi.org/10.1109/TPAMI.2022.3152247

[13] Liu, Z., Mao, H., Wu, C. Y., et al. (2022). A ConvNet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11976–11986). https://doi.org/10.1109/CVPR52688.2022.01168

[14] Tu, Z., Talebi, H., Zhang, H., et al. (2022). MaxViT: Multi-axis vision transformer for image classification. Proceedings of the European Conference on Computer Vision (pp. 459–479). https://doi.org/10.1007/978-3-031-25063-7_27

[15] Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2021). Transfusion: Understanding transfer learning for medical imaging. Advances in Neural Information Processing Systems, 34, 3347–3357. https://doi.org/10.48550/arXiv.2102.13280

[16] Chen, S., Ma, K., & Zheng, Y. (2021). Med3D: Transfer learning for 3D medical image analysis. Medical Image Analysis, 67, 101874. https://doi.org/10.1016/j.media.2020.101874

[17] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[18] Ovadia, Y., Fertig, E., Ren, J., et al. (2020). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems, 33, 13991–14002. https://doi.org/10.48550/arXiv.1906.02530

[19] Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

[20] Codella, N., Rotemberg, V., Tschandl, P., et al. (2020). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). IEEE Transactions on Medical Imaging, 39(8), 2827–2840. https://doi.org/10.1109/TMI.2019.2936401

[21] Haenssle, H. A., Fink, C., Toberer, F., et al. (2020). Man against machine reloaded: Performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions. European Journal of Cancer, 144, 324–331. https://doi.org/10.1016/j.ejca.2020.11.025

[22] Bissoto, A., Perez, F., Ribeiro, E., et al. (2021). Towards automated melanoma screening: Exploring transfer learning schemes. Computer Vision and Image Understanding, 206, 103187. https://doi.org/10.1016/j.cviu.2021.103187

[23] Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R., & Rozeira, J. (2013). PH^2—A dermoscopic image database for research and benchmarking. Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 5437–5440). https://doi.org/10.1109/EMBC.2013.6610790

[24] Giotis, I., Molders, N., Land, S., Biehl, M., Jonkman, M. F., & Petkov, N. (2015). MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert Systems with Applications, 42(19), 6578–6585. https://doi.org/10.1016/j.eswa.2015.04.034

[25] Kawahara, J., Daneshvar, S., Argenziano, G., & Hamarneh, G. (2018). Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE Journal of Biomedical and Health Informatics, 23(2), 538–546. https://doi.org/10.1109/JBHI.2018.2824327

[26] Deng, J., Dong, W., Socher, R., et al. (2009). ImageNet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206845

[27] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 6. https://doi.org/10.1186/s12864-019-6413-7

[28] Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. Information, 11(2), 125. https://doi.org/10.3390/info11020125

[29] DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. arXiv Preprint. https://doi.org/10.48550/arXiv.1708.04552

[30] Brinker, T. J., Hekler, A., Enk, A. H., et al. (2018). Skin cancer classification using convolutional neural networks: Systematic review. Journal of Medical Internet Research, 20(10), e11936. https://doi.org/10.2196/11936

[31] Jeyakumar, J. P., Jude, A., Priya Henry, A. G., & Hemanth, J. (2022). Comparative analysis of melanoma classification using deep learning techniques on dermoscopy images. Electronics, 11(18), 2918. https://doi.org/10.3390/electronics11182918

[32] Sharma, A. K., Nandal, A., Dhaka, A., et al. (2022). Dermatologist-level classification of skin cancer using cascaded ensembling of convolutional neural network and handcrafted features based neural network. IEEE Access, 10, 17920–17932. https://doi.org/10.1109/ACCESS.2022.3149599

[33] Goyal, M., Oakley, A., Bansal, P., Dancey, D., & Yap, M. H. (2020). Skin lesion segmentation in dermoscopy images with ensemble deep learning methods. IEEE Access, 8, 4171–4181. https://doi.org/10.1109/ACCESS.2020.2965665

[34] Santos, F., Silva, F., & Georgieva, P. (2021). Transfer learning for skin lesion classification using convolutional neural networks. Proceedings of the International Conference on Innovations in Intelligent Systems and Applications (pp. 1–6). https://doi.org/10.1109/INISTA52262.2021.9548486

[35] Hosny, K. M., Kassem, M. A., & Foaud, M. M. (2019). Classification of skin lesions using transfer learning and augmentation with Alex-net. PLoS One, 14(5), e0217293. https://doi.org/10.1371/journal.pone.0217293

[36] Wei, L., Ding, K., & Hu, H. (2020). Automatic skin cancer detection in dermoscopy images based on ensemble lightweight deep learning network. IEEE Access, 8, 99756–99768. https://doi.org/10.1109/ACCESS.2020.2996634

[37] Nagendran, M., Chen, Y., Lovejoy, C. A., et al. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ, 368, m689. https://doi.org/10.1136/bmj.m689