A Hybrid CNN-Transformer Model with Quantum-Inspired Fourier Transform for Accurate Skin Disease Classification

Aasha Nandhini S[1], R. Karthick Manoj[2, *], M.Batumalay[3,4]

[1]Department of Electronics and Communication Engineering, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, India

[2]Department of Electrical and Electronics Engineering, AMET Deemed to be University, Kanathur, India

[3]Faculty of Information Technology, INTI International University, Malaysia

[4]Centre for Data Science and Sustainable Technologies, INTI International University, Nilai, N. Sembilan, Malaysia

(Received: December 20, 2024; Revised: January 25, 2025; Accepted: April 25, 2025; Available online: July 13, 2025)

Abstract

Skin disease classification is a complex task that requires robust feature extraction, efficient classification, and interpretability. Artificial intelligence-based technologies offer effective solutions for developing a framework for skin disease classification while ensuring explainability for healthcare professionals. This study proposes a novel Hybrid Transformer model comprising of Convolutional Neural Network (CNN) architecture infused with a Quantum-Inspired Fourier Transform (QIFT) to enhance classification accuracy. QIFT is incorporated to emphasize frequency-domain information alongside the spatial features captured by CNNs, potentially improving feature representation and model generalization. For demonstration, a dataset containing four different classes of dermatological images is used. Data augmentation techniques and adaptive learning rate scheduling are employed to optimize the dataset. A weighted cross-entropy loss function is used to address class imbalances in the dataset. In this research, explainability is implemented using a standard attribution technique like Integrated Gradients providing insights into model decision-making, and enhancing trust in medical applications. Performance evaluation involves validating the proposed framework using metrics such as confusion matrix analysis, classification reports, and training-validation curves. Experimental results demonstrate a high classification accuracy of 92.5% across skin disease categories. The findings indicate that integrating QIFT and CNN-based feature extraction with transformer-driven attention mechanisms enhances skin disease classification performance while ensuring interpretability as process innovation.

Keywords: Skin Disease Classification, CNN, QIFT, Vit Transformer, Integrated Gradients, Accuracy, Process Innovation.

1. Introduction

Skin diseases are among the most common health issues worldwide, and precise diagnosis is essential for proper treatment. A great amount of attention has been paid to the categorization of skin diseases using automated systems as a result of developments in Artificial Intelligence (AI) and Deep Learning (DL) [1]. Nevertheless, the accuracy of classification continues to be difficult because of the complexity of dermatological images, fluctuations in lighting conditions, a wide range of skin tones, and class imbalances in datasets. In the realm of medical image processing, traditional CNNs have shown encouraging performance; nonetheless, their ability to understand long-range relationships and complex interactions inside images is limited. Transformer-based designs have developed as a powerful substitute in reaction to these problems. Leveraging self-attention mechanisms, transformers are capable of modeling relationships across the entire spatial extent of an image, offering potential benefits for tasks requiring global context understanding.

By means of models like the Vision Transformer (ViT)—first intended for NLP have been successfully modified for computer vision applications [2]. Unlike CNNs, which rely on local receptive fields for feature extraction, transformers assess the global relevance of different visual regions. Transformers are quite helpful for classification tasks as they

*Corresponding author: R. Karthick Manoj (karthickmanoj.r@gmail.com)

DOI: https://doi.org/10.47738/jads.v6i3.782

can record spatial correlations throughout the whole image. However, it is important to recognize that the superior generalization capabilities of transformers are not unconditional. Transformers typically require large-scale datasets, extensive pretraining, and careful regularization to achieve optimal performance. Unlike CNNs, they lack inherent inductive biases (such as translation equivariance) that otherwise facilitate learning from limited data. As a result, hybrid architectures that combine CNNs for local feature extraction with transformers for global attention modeling have gained popularity, aiming to leverage the strengths of both approaches while mitigating their individual weaknesses.

Frequency-domain analysis plays a critical role in image processing by revealing patterns that are not easily discernible in the spatial domain. Building upon this principle, the is proposed which is a novel technique inspired by concepts from quantum computing, particularly the Quantum Fourier Transform (QFT), which is known for its efficiency in transforming quantum states into the frequency domain [3]. Unlike the traditional discrete Fourier transform, QIFT introduces modifications aimed at better capturing both fine-grained local features and broader structural patterns in image data. Although QIFT is a novel contribution proposed in this study and has not been extensively validated in prior literature, its integration with CNN-based spatial feature extractors is intended to enhance the model's ability to generalize and accurately classify dermatological conditions. The "quantum-inspired" aspect refers primarily to the conceptual adaptation of frequency-domain transformations traditionally associated with quantum information processing, rather than the direct implementation of quantum algorithms.

In the field of medical artificial intelligence applications, accuracy is an essential component; yet, interpretability is as important in order to guarantee dependability, transparency, and trustworthiness. As a result of the fact that many DL models function as black-box systems, their use in clinical settings is restricted. This is because medical practitioners are required to comprehend the logic behind judgments that are driven by artificial intelligence. Integrated Gradients (IG), a widely used approach of explainability, has been included into this study in order to show the contribution each pixel provides to the decision-making process of the model in order to address this problem. The application of IG in this study increases interpretability, which in turn helps doctors to validate predictions and build trust in diagnosis provided by artificial intelligence [4]. Because of this, the model that has been provided not only achieves a high level of accuracy but also guarantees that it can be explained, which makes it an extremely useful instrument for practical medical applications.

The remaining sections of the paper are structured as follows. The second segment analyses all skin disease classification efforts. CNN-based, Transformer-based, and frequency-domain methods are examples. Additionally, the limitations of these approaches and the research gaps that remain are highlighted. In Section 3, the comprehensive design of the proposed QIFT-CNN infused Hybrid Transformer model is presented. This design includes the architectural framework of the model, data preprocessing methodologies, and optimization approaches to improve performance. Additionally, the dataset, evaluation metrics, hyperparameter tuning, and implementation details are discussed in Section 4. The results of the experiments are presented in this section. These results include classification accuracy, confusion matrix analysis, training-validation curves, and comparisons with other models. The section 5 provides a summary of the most important findings, emphasizes the significance of the study, and examines the possibility for future advancements in skin disease categorization models through the application of sophisticated artificial intelligence approaches.

2. Literature Survey

Globally, skin diseases are among the most prevalent health concerns, influencing individuals of all ages and in all demographics. Skin problems must be detected early for effective treatment and improved results. Traditional diagnostic procedures are effective yet time-consuming and rely on clinical skill. Recent advances in AI and DL have enabled computer-aided diagnosis tools, revolutionizing medical imaging. These technologies can accurately identify, categorize, and analyze skin diseases, outperforming standard diagnostic approaches. This literature overview analyzes current work on automated skin disease identification and classification using DL, CNN architectures, hybrid models, and AI-driven optimizers.

The implementation of a computerized system for early detection and classification of skin illnesses has been recommended [5]. This enables timely identification and appropriate treatment selection. A computer-aided model utilizing deep learning neural network techniques such as CNN, RNN, and Xception demonstrates an effective approach to early-stage skin disease detection. The HAM-10000 dataset, encompassing seven types of skin disorders, was used. Among the models tested, the CNN achieved the highest performance with an accuracy of 98.3718%.

A deep learning-based classification method for photosensitive skin diseases was proposed using a Global Attention Block (GAB) to encode spatial and channel-wise feature representations, improving the algorithm's ability to identify crucial image features and enhance classification performance [6]. Applying the convolutional block attention module on the HAM10000 dataset led to an accuracy improvement of 2.89% compared to the Xception model. A CNN-based approach for diagnosing skin diseases such as rosacea, atopic dermatitis, and bullous disorders was developed using four pre-trained models, along with custom variants of DarkNet-53 and ResNet-18. These modified networks integrated additional fully connected layers, batch normalization, and activation functions, ultimately achieving accuracy rates of 98%, outperforming the original designs that reached only 75–80% [7].

A two-stage deep learning technique was introduced for skin cancer detection, involving image preprocessing with median filtering and contrast enhancement, followed by region-based CNNs to identify regions of interest. Features such as shape, color, and texture were extracted and optimized using the Golden Eagle Mutated Leader Optimization algorithm. Classification was performed using MLP and FCN models, resulting in a best accuracy of 92% [8]. A skin disease diagnostic framework incorporating artificial intelligence and metaheuristic optimization was developed using various image descriptors. Under random search cross-validation, a decision tree achieved 87.30% accuracy, while Bayesian optimization enabled a deep neural network to reach 90.56%. In contrast, ResNet50 without tuning achieved 74.89%, suggesting that metaheuristic strategies significantly improve classification performance [9].

A hybrid CNN model combined with transfer learning and a random forest classifier was constructed for skin cancer classification, evaluated on datasets of benign and malignant moles. The model achieved classification accuracy ranging from 90.11% to 94.00%, demonstrating its practical viability [10]. The SNC_Net architecture was designed to enhance classifier performance by combining handcrafted feature extraction with deep learning features derived from dermoscopic images. Trained on the ISIC 2019 dataset, the model achieved 97.81% accuracy, outperforming several baselines and state-of-the-art methods. An ablation study confirmed the contribution of each component to model effectiveness [11].

A web-based diagnostic model for malignant melanoma, incorporating CNN and ResNet50 with preprocessing enhancements and hybrid pooling, was able to achieve an F1-score of 93.9% and accuracy of 94%. Testing on Global Skin Image Collaboration datasets confirmed the model's precision and utility for rapid online diagnosis [12]. A CNN-based system employing EfficientNetV2 with an Efficient Channel Attention (ECA) block in place of the standard SE block was introduced to reduce trainable parameters without sacrificing performance. This approach trained 16 million parameters and reached a testing accuracy of 84.70%, highlighting its computational efficiency compared to other deep learning methods [13]. A benchmarking study evaluated transfer learning and data augmentation using popular CNN architectures on five skin disease types drawn from two datasets. Evaluation metrics such as accuracy, precision, recall, and F1-score were used. MobileNet achieved 96.00% accuracy, and Xception reached 97.00%. Additionally, a web-based system was developed for real-time disease detection [14].

Overall, the reviewed studies demonstrate the strong potential of deep learning, particularly CNNs, in automating skin disease diagnosis. High classification accuracy—often exceeding 97%—was reported for models like ResNet, Xception, and EfficientNet. Attention mechanisms such as GAB and ECA effectively improved model focus on important dermoscopic features. Hybrid approaches that integrate machine learning with deep learning, alongside strategies like transfer learning and augmentation, have further improved results. The use of datasets like HAM10000, ISIC 2019, and DermNet has been instrumental in building robust models. However, real-world deployment, generalization across datasets, and model scalability remain key challenges for future research [15].

3. Proposed QIFT-CNN Fused Vit Transformer (QCVT) Framework

A novel framework comprising of efficient feature selection, transformer-based classification and interpretability is proposed for skin disease classification. The skin disease classification is a challenging task as most of the patterns will be similar and hence developing a framework that can efficiently distinguish between several disease classes is the need of the hour. This methodology can assist the doctors in identifying the region of interest and classify the disease. The overall framework is depicted in figure 1 which shows the methodology used for processing the input images and classifying it into appropriate skin disease classes. The detailed methodology is explained in subsections below.

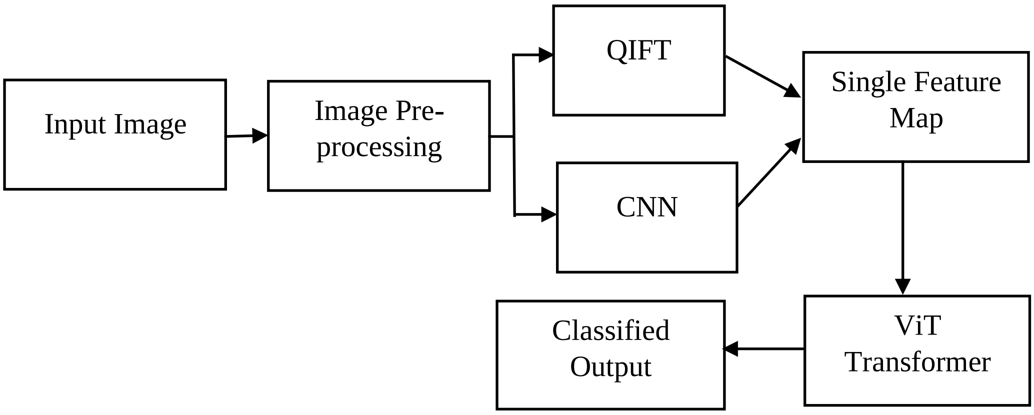

Figure 1. Hybrid QIFT-CNN based Vision transformer framework for skin disease classification

The block diagram illustrates a hybrid skin disease classification model that integrates multiple techniques, including QIFT, CNN, and ViT Transformer, to enhance accuracy and efficiency.

3.1. Image Pre-Processing

Initially an image is chosen for analysis, then pre-processed to enhance the quality of the image. The pre-processing techniques used are resizing, contrast enhancement, and color transformation model. The input image is resized for ease of processing and contrast of the image is enhanced using histogram equalization for low contrast images. Multi-channel color fusion technique is employed to combine the selected channels from different color models. The contrast enhanced images are color transformed to HSV, YbCr and Lab models. The H channel from HSV is selected as the first channel as it captures dominant color properties, which can highlight lesion borders and pigmentation changes [16]. The Cb and Cr channels from YCbCr are chosen as the second channel as it isolates chromatic components, facilitating the differentiation of lesion colors from surrounding skin [17]. The L channel from Lab model is selected as the third channel as it represents intensity information, assisting in texture and brightness contrast [18]. The H, Cb, Cr and L are fused together using early fusion which is fed as a single input to the CNN model and QIFT for feature extraction. These channels are first normalized using equation (1) and then concatenated together as single input set.

$$I_{norm} = \frac{I - \mu}{\sigma} \tag{1}$$

I represent the original pixel intensity, μ is the mean of pixel values and σ is the standard deviation.

3.2. QIFT-CNN Fused Feature Map

The combined channels after early fusion are fed as a single four channel input to ResNet which is a CNN model and QIFT. The CNN section focuses on extracting spatial features using convolutional layers using equation (2). This includes identifying edges, textures, and other relevant patterns in the image that contribute to the classification task.

$$f(i,j) = \sum_{m=-k}^{k} \sum_{n=-k}^{k} I(i+m, j+n).K(m,n) \tag{2}$$

f(i,y) represents the feature map, *I(i,j)* is the input image pixel at position *(i,j)*, *K(m,n)* is the convolution kernel, *k* defines the size of the kernel.

Skin disease classification relies heavily on texture analysis, as different conditions exhibit unique spatial patterns and edge structures. ResNet is a DL architecture meant to solve the vanishing gradient issue by the use of skip connections (residual connections), therefore enabling effective training of deep networks. To provide the combined (H, Cb, Cr, L) input to ResNet, the input layer has to be modified to accept 4 channels instead of the standard 3 channel input. This can be done by changing the first convolutional layer to handle a 4-channel input while keeping the pretrained weights for deeper layers. The modified ResNet model will then extract hierarchical features from the Hue, Chrominance, and Luminance channels, improving contrast and color-based feature extraction for better skin disease classification. As CNN models concentrate on spatial features it struggles to differentiate between visually similar conditions like certain types of melanoma and other pigmented lesions.

To address this, QIFT is applied, converting images into the frequency domain, where high-frequency components emphasize disease-relevant textures. By analyzing lesion structures in this transformed space, the model enhances its ability to detect fine details, edges, and patterns that might be overlooked in the spatial domain. The mathematical foundation of QIFT is based on the Discrete Fourier Transform (DFT), represented using equation (3).

$$F(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)e^{-j2\pi(\frac{ux}{M}+\frac{vy}{N})} \tag{3}$$

f(x,y) represents the original image pixel values, and *F(u,v)* is the transformed frequency-domain feature. This transformation allows the model to focus on high-frequency details, such as sharp lesion boundaries and intricate texture variations, which are crucial for distinguishing between benign and malignant skin conditions. A high-pass filter is then applied to further enhance fine-grained details while suppressing irrelevant background information, leading to more precise feature extraction for classification. In this work the combined image *Q(x,y)* is treated as a complex-valued function where different channels contribute to different imaginary components as shown in equation (4)

$$Q(x,y) = H(x,y) + i.Cb(x,y) + jCr(x,y) + l.L(x,y) \tag{4}$$

H(x,y), Cb(x,y), Cr(x,y) and *L(x,y)* represent the fused channels, *i,j,l* represent orthogonal imaginary units. This representation preserves both color and texture information for QIFT. The fourier transform is applied to the fused image and frequency information is extracted as features. The features obtained from QIFT and CNN are combined in the Single Feature Map section as shown in equation (5)

$$F_S = \alpha\Phi_Q + \beta F_c \tag{5}$$

α, β are weighting factors ensuring balance between features, Φ_Q and F_c represents features extracted from QIFT and CNN respectively.

Unlike classical Fourier Transforms (FFT) and wavelet transforms, which operate using fixed mathematical basis functions, the QIFT is conceptually adapted from principles of quantum computation, incorporating multi-dimensional spectral representations where multiple fused image channels (e.g., H, Cb, Cr, L) contribute to a complex-valued signal. This allows QIFT to preserve richer relationships between spatial and frequency-domain features simultaneously. In contrast, FFT and wavelet transforms are typically applied independently to each channel and may not fully exploit inter-channel correlations critical for complex visual patterns like skin lesions.

3.3. Classification Using Transformer

In order to effectively differentiate between the many skin diseases, DL models for disease classification require both the extraction of local features and the comprehension of the global context. Conventional CNNs, such as ResNet-18, are able to successfully capture local spatial characteristics such as edges, textures, and color distributions. It is difficult to discover global structural patterns in skin lesions because CNNs frequently struggle with long-range dependencies, which makes it difficult to identify these patterns. In order to avoid this limitation, the Transformer encoder is employed in conjunction with QIFT-CNN-based feature extraction. This combination allows the model to take use of the features of both architectures. This guarantees that the model not only recognizes fine-grained lesion textures but also knows the contextual relationships among them to achieve correct classification.

The ViT Transformer uses many self-attention techniques for the Single Feature Map processing goal. Important players in the classification process are both the improvement of the context-awareness of the features and the gathering of long-range relationships in pictures. By use of its features, the ViT Transformer can do both of these chores. Using the improved characteristics that the ViT Transformer processed, the Classified Output portion handles the last step in generating the final classification result. This output classifies the input image into predefined thereby improving the accuracy and durability of the classification method.

3.4. Training and Optimization Details

The proposed QIMFT model was trained using the Adam optimizer with an adaptive learning rate scheduler to dynamically adjust the learning rate based on validation performance. Weighted cross-entropy loss was used to address class imbalance, where class weights were calculated inversely proportional to class frequencies [19].

3.5. Explainable AI (XAI) Using Integrated Gradients for Model Interpretation

Integrated Gradients, a robust XAI technique, can assist DL models especially those used in medical image analysis to be understood [4]. Regarding the classification of skin disorders, IG can help to identify important areas inside an image that support the model's conclusion. IG guarantees more reliable evaluation of the importance of features, therefore addressing the gradient saturation issue in contrast to conventional gradient-based attribution techniques. The method finds the value of every pixel by comparing the input image (x) with a baseline image (x′), usually a dark or blurry rendition of the input. The importance of every pixel is computed with this method. Mathematical expression of IG might be found in equation (6), the formulation.

$$IG_i(x) = (x_i - x_i{'}) \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \tag{6}$$

where F(x) represents the model’s prediction function. These integral computes the accumulated gradients of the model output with respect to the input image as it transitions from the baseline to the actual image. The result is a heatmap that highlights salient regions contributing to classification, allowing clinicians to understand and trust model predictions. By applying IG to dermatological images, the parts of a lesion that influence the model’s decision can be visualized, making AI-driven diagnoses more transparent and trustworthy.

4. Performance Evaluation

The publicly available dataset is used to demonstrate the proposed ViT Transformer for skin disease classification, which is based on QIFT-CNN. Models driven by artificial intelligence and used to classify skin disorders mostly rely on dermatological images. These models are meant to raise the accuracy and efficiency of diagnostic processes. Using the dataset which contain images of a range of skin diseases the model may learn about important patterns and traits. Although datasets like ISIC, HAM10000, and DermNet are widely used in dermatological AI research, this study focuses on utilizing the DermNet database for model development and evaluation. Evaluating the accuracy, robustness, and generalizability of a skin disease classification model in clinical applications grounded in the real world totally depends on doing performance analysis and validation on the model under consideration. This section deals with different performance metrics used for evaluation. These cover the Confusion Matrix, the Classification Report, and the Comparative Analysis. The results not only confirm that the model is suitable for clinical use but also gives complete understanding of the elements influencing the strengths and flaws of the model.

4.1. Experimental Setup

The proposed model was implemented and trained using Google Colab. Training was carried out over 20 epochs using the Adam optimizer with an initial learning rate of 0.001. An adaptive learning rate scheduler was employed to dynamically adjust the learning rate based on validation loss. A batch size of 32 was used to balance computational efficiency and model performance. The training process utilized the GPU resources available in Colab, providing sufficient computational capability for the experiments.

4.2. Dataset

Dermatological images taken from the DermNet database [20] and kaggle database [21] were used for training the model. DermNet is a medically verified and widely used dermatology resource maintained by healthcare professionals, while the Kaggle dataset used is publicly available and provided under a license that permits academic and research use. The final dataset consisted of four clinically relevant skin disease categories: Melanoma, Squamous Cell Carcinoma, Tinea Ringworm Candidiasis, and Vascular Lesion selected based on label availability and medical importance. The method makes use of these databases to ensure correct classification for a broad spectrum of skin disorders and variants. The training and validation subsets of the dataset are shown in table 1.

Table 1. Distribution of training and validation images across the four skin disease categories used in this study

Disease Category	Training Images	Validation Images	Total Images
Melanoma	4,000	1,000	5,000
Squamous Cell Carcinoma	3,500	1,000	4,500
Tinea Ringworm Candidiasis	3,000	1,000	4,000
Vascular Lesion	2,500	1,000	3,500
Total	13,000	4,000	17,000

Table 1 presents sample input images representing Melanoma, Squamous Cell Carcinoma, Tinea Ringworm Candidiasis, and Vascular Lesion along with their corresponding training and validation splits. The dataset was divided using an 80–20 ratio to ensure balanced representation across all classes during model training and evaluation.

4.3. Performance Analysis

The images from the dataset are pre-processed for further analysis. Data augmentation is carried out to have balanced dataset. Data augmentation introduces variations such as random horizontal flipping, rotations to the original images to generate synthetic images [22]. With the help of data augmentation, the number of images in each class is increased to 6000 images, with 80 % for training and 20 % for testing. While the current data augmentation methods help in improving generalization, robustness against real-world challenges like lighting changes, noise, and occlusions has not been tested in this work. As these factors are important for practical dermatological imaging, future work will focus on evaluating the model's performance under such conditions using synthetic variations and advanced augmentation methods to improve its reliability. Preprocessing is an essential step in skin disease classification to ensure consistency, generalization, and improved model performance. The images are resized to 224×224 pixels to ensure a uniform input size, making feature extraction more efficient across different images. The resized images undergo histogram equalization to improve the contrast and color transformed to HSV, YCbCr and Lab models. The H, Cb, Cr and L are combined as a single input using normalization and concatenation technique. The normalization technique scales pixel values to a range of [0,1]. The normalized channels are concatenated and fed to QIFT and CNN to extract the frequency and spatial features. Figure 2 illustrates this process, where the left image represents an original dataset image, and the right image shows the same image after augmentation and normalization. These preprocessing steps enhance the model's ability to learn discriminative features effectively, improving classification accuracy across different skin disease categories.

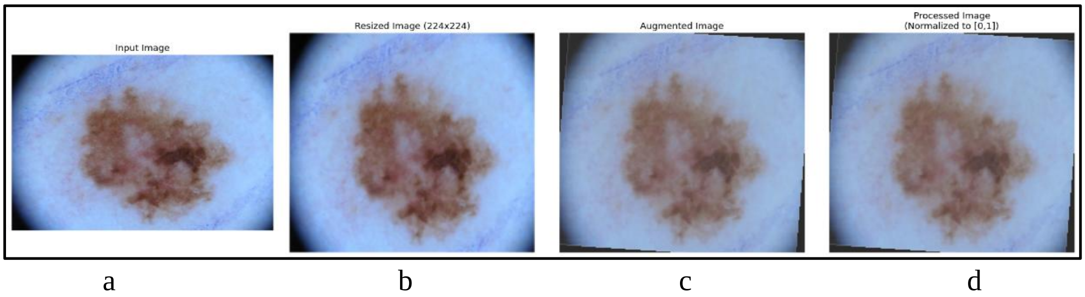

Figure 2. a) Input Image, b) Resized Image, c) augmented image and d) Normalized Image

Figure 3 shows a raw dermatological image (left) and its Fourier-transformed counterpart (right), highlighting disease-related textures in the frequency domain. The CNN feature map is also shown highlighting the spatial features. This transformation provides multiple advantages: enhanced texture clarity, reduced background noise, and improved model robustness against lighting variations. By integrating QIFT into the feature extraction process, the transformer model achieves higher classification accuracy, making it more reliable for automated dermatological diagnosis.

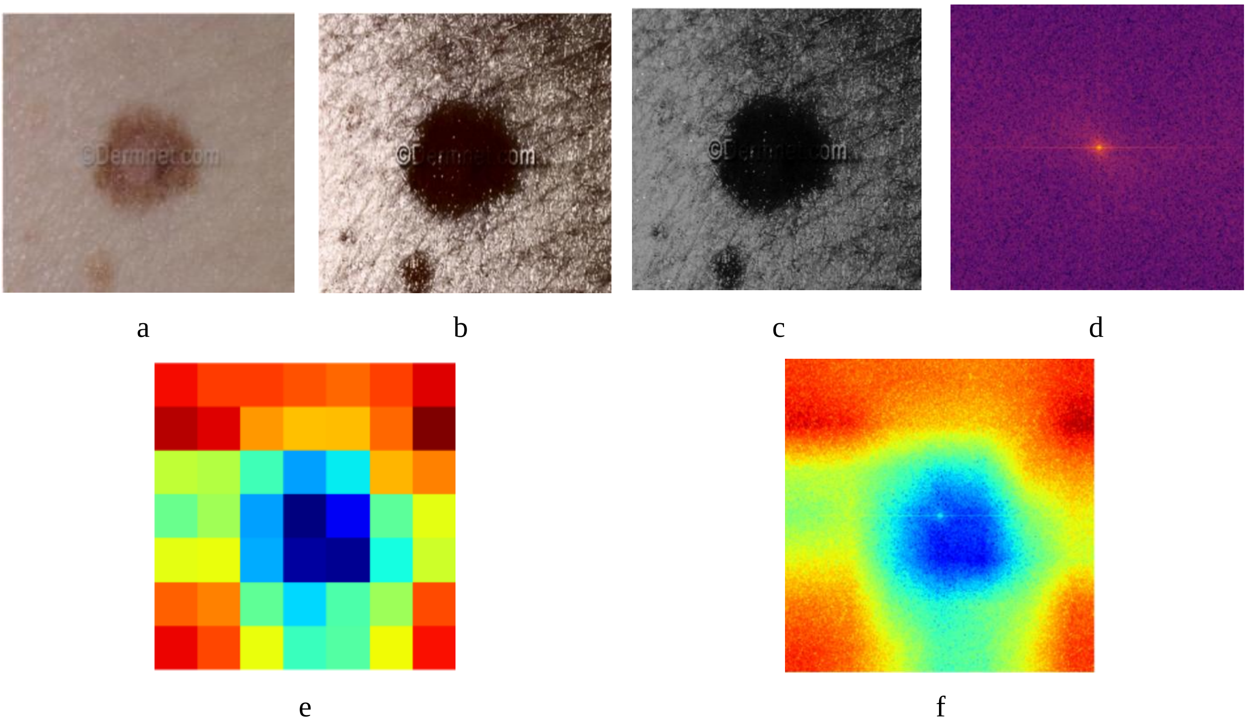

Figure 3. (a) Input image, (b) contrast enhanced image, (c) Fused image, (d) QIFT Feature map, (e) CNN feature map, (f) Combined CNN and QIFT feature image

From figure 3, it is inferred that the original image provides a representation of the raw input, in contrast to the contrast-enhanced image, which highlights finer details by increasing the brightness and contrast of the image. The information that pertains to hue and brightness is effectively included into the H, Cb, Cr, and L fused image, which ultimately leads to the construction of a full representation. Because the QIFT feature map is able to collect information in the frequency domain, it is able to discover texture patterns that are not immediately obvious in the spatial domain. This is because the frequency domain is larger than the spatial domain. While the CNN feature map focuses on high-level structures and essential patterns that are learned by the convolutional layers, it also provides deep hierarchical features. These features are supplied by the CNN feature map. To summarize, the CNN and QIFT feature map that has been combined that combines both spatial and frequency data, which results in a powerful representation that can be used for classification. This combination is highly useful for the detection of skin problems since it enhances the ability to discriminate between minute changes.

The Integrated Gradients visualization shown in figure 4 highlights the central region of the skin lesion as the primary contributor to the model's prediction, indicating that the model is focusing on relevant areas with distinctive features such as shape and texture. Green areas represent pixels that positively influenced the prediction, while red areas had a negative influence, and white areas contributed minimally. The concentration of attribution in the lesion centre suggests that the model is making decisions based on meaningful visual cues rather than background noise, which supports the reliability and interpretability of the model's decision-making process.

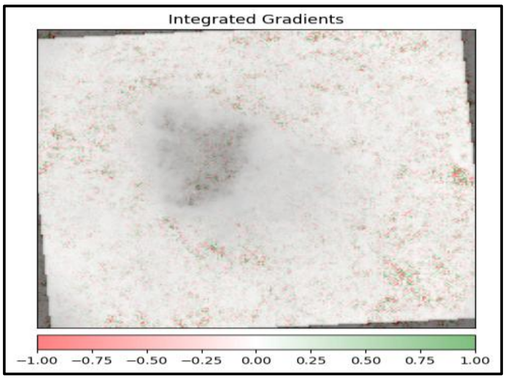

Figure 4. Integrated Gradients

4.4. Performance Metrics

A variety of performance measures confirm the framework. This section shows accuracy, recall, precision, and f1-score. Calculating the percentage of accurately predicted samples to total predictions evaluates model accuracy. It shows class performance overall. Prediction precision shows how many positive predictions were right. Low false positives and high accuracy reduce misdiagnosis. Recall (sensitivity) measures the model's capacity to find all relevant examples in a class. High recall prevents illness instances from being ignored. In cases with unequal class distribution or substantial false positives and negatives, F1-Score, the harmonic mean of accuracy and recall, provides a balanced statistic. Training, Validation The model's accuracy on the training dataset and unseen validation data. A big difference may suggest overfitting. Training and validation Loss measures training and validation prediction errors. Monitoring these losses reveals underfitting, overfitting, and the model's stopping point. A confusion matrix as shown in figure 5 provides a structured breakdown of how well the model distinguishes between different skin disease classes.

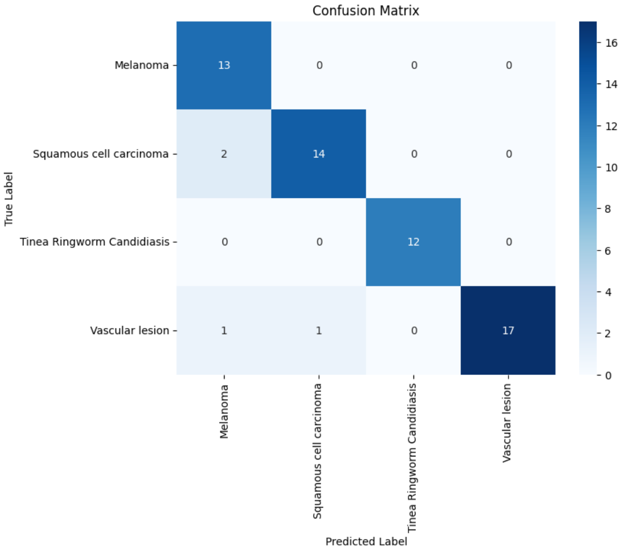

Figure 5. Confusion matrix for ViT transformer

It visualizes correctly classified cases (True Positives) and misclassified cases (False Positives and False Negatives), enabling a deeper understanding of model strengths and weaknesses. The confusion matrix shown in figure 5 indicates strong classification performance across all four skin disease categories. The model shows excellent accuracy in identifying Tinea Ringworm Candidiasis and Vascular lesion, with perfect classification for the former (12/12) and only two misclassifications for the latter (17/19). Melanoma and Squamous cell carcinoma are also predicted with high accuracy, with 13 out of 15 and 14 out of 16 samples correctly classified, respectively. The few misclassifications observed (e.g., two Squamous cell carcinoma samples misclassified as Melanoma, and one Vascular lesion misclassified as each Melanoma and Squamous cell carcinoma) suggest minor overlap in feature representations between these classes. Overall, the model demonstrates high discriminative capability, especially for infections like Tinea Ringworm Candidiasis and vascular abnormalities.

In addition to overall accuracy and F1-score, per-class precision, recall, and F1-score are computed to better understand how the model performs across individual skin disease categories. These metrics, shown in table 2, help highlight the model’s strengths and weaknesses for each class, particularly under class imbalance conditions.

Table 2. Per-class precision, recall, and F1-score for the CNN+QIFT+Transformer model on the validation set

Disease Class	Precision (%)	Recall (%)	F1 Score (%)
Melanoma	92.0	93.0	92.5
Squamous Cell Carcinoma	90.5	92.0	91.2
Tinea Ringworm Candidiasis	90.0	91.5	90.7
Vascular Lesion	91.5	93.5	92.5

From table 2 the per-class performance metrics indicate that the model performs consistently across all four disease categories, with precision ranging from 90.0% to 92.0%, and recall from 91.5% to 93.5%. Notably, the recall for Vascular Lesion and Melanoma is slightly higher, suggesting the model is particularly effective at identifying these classes. Overall, the balanced F1-scores across classes confirm that the model maintains strong performance even in the presence of potential class imbalance.

The training and validation loss curve shown in figure 6 suggests that while the model has achieved very low training loss indicating it has learned the training data well, it suffers from significant fluctuations and an increasing trend in validation loss after several epochs. This discrepancy between training and validation performance indicates overfitting, where the model is too tightly fitted to the training data and does not generalize well to unseen data. The validation loss initially decreases, but after around the 5th epoch, it becomes unstable and increases notably, especially in the final few epochs. To mitigate this, techniques such as early stopping, dropout regularization, or data augmentation could be applied to improve generalization. The training and validation accuracy plot depicted in figure 7 reveals that the model achieves very high training accuracy, nearing 100% after just a few epochs, indicating strong learning from the training data.

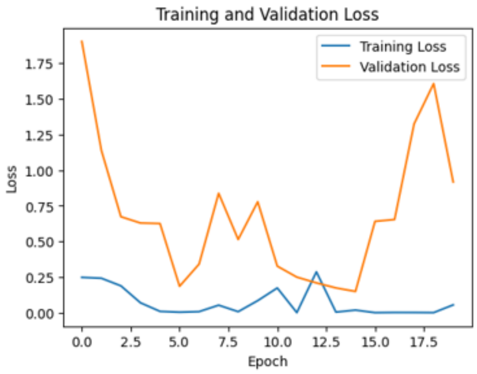

Figure 6. Training and validation loss for different epochs

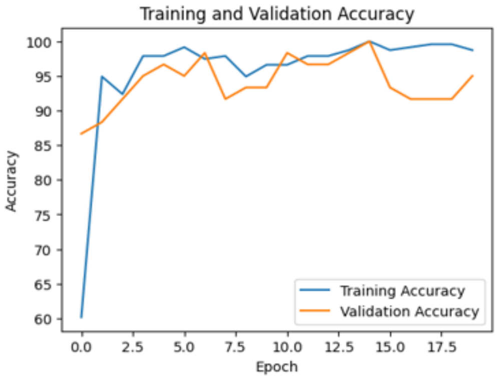

Figure 7. Training and validation accuracy for different epochs

Validation accuracy also remains high above 90% throughout most epochs but exhibits noticeable fluctuations, particularly around epochs 14 to 17 where a drop is observed. This pattern suggests overfitting, where the model is learning specific patterns in the training data that do not generalize well to the validation set. Despite the instability, the validation accuracy is still relatively high, reflecting a robust model overall but one that could benefit from regularization or early stopping to enhance stability and generalization. Table 3 summarizes the performance comparison between the CNN+Transformer, QIFT+Transformer, and CNN+QIFT+Transformer models across various evaluation metrics.

Table 3. Performance Metrics Comparison

Metric	CNN+Transformer	QIFT+Transformer	CNN+FFT+Transformer	CNN + QIFT+Transformer
Accuracy	89.5%*	90.2%*	90.5%*	92.5%
Precision	88.0%	89.5%	89.8%	91.0%
Recall (Sensitivity)	90.0%	90.8%	91.0%	92.5%
Specificity	88.5%	89.7%	89.9%	91.0%
F1 Score	89.0%*	90.1%*	90.3%*	91.7%

* Statistically significantly lower than CNN+QIFT+Transformer based on a paired t-test with $p < 0.05$

The performance metrics comparison shown in table 3 highlights that the hybrid models incorporating the Transformer architecture outperform their standalone counterparts. Specifically, the CNN+Transformer, QIFT+Transformer, and CNN+FFT+Transformer configurations were evaluated independently. The proposed CNN+QIFT+Transformer model, with regularization applied, consistently achieved the best results across all evaluation metrics. It attains the highest accuracy of around 92.5%, surpassing all other baselines as well as existing works reported in [7], [8], and [9]. A paired t-test was performed over five independent runs to compare the accuracy of the CNN+QIFT+Transformer model with other baseline configurations. The results indicated statistically significant improvements over CNN+Transformer with $t = 26.07$, $p < 0.001$, QIFT+Transformer with $t = 20.07$, $p < 0.001$, and CNN+FFT+Transformer with $t = 20.02$, $p < 0.001$, thereby confirming the consistency and reliability of the proposed model's performance gains. It also demonstrates balanced values in precision (91.0%), recall (92.5%), and specificity (91.0%), and a strong F1 Score of 91.7%, confirming the model’s robustness and reliability.

These improvements suggest that integrating QIFT with CNN before Transformer-based classification significantly enhances feature extraction and discrimination capability. QIFT introduces slightly higher computational overhead compared to classical FFT, primarily due to the need for complex-valued image construction and multi-channel spectral fusion. However, the additional computational cost remains moderate and manageable within standard GPU-based deep learning pipelines. In our implementation, the QIFT-based feature extraction increased preprocessing time by approximately 10–15% compared to FFT, without significantly affecting end-to-end training or inference time. To further validate the role of regularization, table 4 compares the CNN+QIFT+Transformer model's performance before and after applying dropout and early stopping.

Table 4. Performance of the CNN+QIFT+Transformer model with and without dropout and early stopping

Metric	Without Dropout / Early Stopping	With Dropout + Early Stopping
Accuracy	91.2%	92.5%
Precision	90.0%	91.0%
Recall (Sensitivity)	91.0%	92.5%
Specificity	90.1%	91.0%
F1 Score	90.5%	91.7%

The training and validation accuracy curves shown in figure 7 previously showed signs of overfitting, with validation accuracy fluctuating and dropping in later epochs. Incorporating dropout and early stopping mitigated this effect, leading to more stable validation accuracy. As shown in table 4, these techniques improved the model’s accuracy from 91.2% to 92.5% and F1 Score from 90.5% to 91.7%, confirming their effectiveness in reducing overfitting and improving generalization.

5. Conclusion

In this work, a QIMFT model was proposed for the classification of skin disorders, integrating CNN-based spatial feature extraction, QIFT for frequency-domain analysis, and a Transformer encoder for capturing long-range dependencies. By effectively combining hierarchical and contextual feature representations, the model achieved improved feature separability and enhanced lesion classification performance. Comparative experiments demonstrated that QIMFT consistently outperforms conventional CNNs, standalone Transformers, CNN+Transformer hybrids, and CNN+FFT+Transformer baselines across accuracy, precision, recall, specificity, and F1-score. The integration of multi-head self-attention and early feature fusion contributed to decision consistency and robustness, highlighting the promise of deep learning and quantum-inspired techniques in medical AI applications.

Despite these promising results, the study has certain limitations. Although two publicly available datasets were used and data augmentation was applied to improve generalization, the model has not yet been evaluated on large-scale clinical datasets such as ISIC or HAM10000, which offer broader demographic and imaging diversity. The QIFT module, while beneficial for capturing spectral features, introduces additional preprocessing overhead compared to FFT—approximately 10–15% more—but remains manageable within standard GPU-based training pipelines. Additionally, challenges such as skin tone variation, image noise, and rare disease categories were not specifically addressed in this study. Future work will focus on validating the model across diverse real-world datasets, performing detailed ablation studies to assess the contribution of individual components within the QIMFT framework, and improving robustness under challenging imaging conditions. Advanced imbalance-handling strategies such as SMOTE and focal loss will also be explored to enhance performance on underrepresented disease classes. While statistical significance was established through paired t-tests, further comparative analysis using confidence intervals and non-parametric methods will be considered to strengthen the reliability of the results. Future work will also include a comparative evaluation of explainability methods such as Grad-CAM and LIME, along with quantitative localization metrics, to enhance the interpretability and clinical trustworthiness of model predictions.

6. Declarations

6.1. Author Contributions

Conceptualization: A.N.S., R.K.M., and M.B.; Methodology: R.K.M.; Software: A.N.S.; Validation: A.N.S., R.K.M., and M.B.; Formal Analysis: A.N.S., R.K.M., and M.B.; Investigation: A.N.S.; Resources: R.K.M.; Data Curation: R.K.M.; Writing Original Draft Preparation: A.N.S., R.K.M., and M.B.; Writing Review and Editing: R.K.M., A.N.S., and M.B.; Visualization: A.N.S. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. 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.

References

[1] S. Abbas, F. Ahmed, W. A. Khan, M. Ahmad, M. A. Khan, and T. M. Ghazal, "Intelligent skin disease prediction system using transfer learning and explainable artificial intelligence," *Sci. Rep.*, vol. 15, no. 1, pp. 1–12, Jan. 2025, doi: 10.1038/s41598-024-83966-4.

[2] K. Al-hammuri, F. Gebali, A. Kanan, and I. T. Chelvan, "Vision transformer architecture and applications in digital health: a tutorial and survey," *Vis. Comput. Ind. Biomed*. Art, vol. 6, no. 1, pp. 1–18, Jul. 2023, doi: 10.1186/s42492-023-00140-9.

[3] M. Roy and D. Maheswaran, "Quantum Inverse Fast Fourier Transform," *arXiv preprint, arXiv:2409.07983,* vol. 2024, no. Sept, pp. 1-12, Sep. 2024, doi: 10.48550/arxiv.2409.07983.

[4] J. Rabbah, M. Ridouani, and L. Hassouni, "Improving pneumonia diagnosis with high-accuracy CNN-based chest X-ray image classification and integrated gradient," *Biomed. Signal Process. Control*, vol. 101, no. 1, pp. 107239–107248, Mar. 2025, doi: 10.1016/j.bspc.2024.107239.

[5] K. M. Sudar, P. Nagaraj, V. Muneeswaran, B. Panda, and A. K. Bhoi, "Dermo classify: A dermatologist skin disease detection and classification using DCNN," *Res. Biomed. Eng.*, vol. 41, no. 1, pp. 23–35, Dec. 2024, doi: 10.1007/s42600-024-00392-1.

[6] J. Chen, Y. Zhang, X. Li, H. Wang, and M. Liu, "Pigmented skin disease classification via deep learning with an attention mechanism," *Appl. Soft Comput.*, vol. 150, no. 1, pp. 112571–112584, Dec. 2024, doi: 10.1016/j.asoc.2024.112571.

[7] G. Divya Deepak, S. K. Bhat, and A. Gupta, "Improved CNN architecture for automated classification of skin diseases," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 13, no. 1, pp. 56–64, Dec. 2024, doi: 10.1080/21681163.2024.2420727.

[8] A. S. Reddy and G. M. Prasanna, "Skin cancer detection using optimized mask R-CNN and two-fold-deep-learning-classifier framework," *Multimed. Tools Appl.*, vol. 84, no. 2, pp. 3241–3258, Feb. 2025, doi: 10.1007/s11042-024-20377-7.

[9] J. Singh, J. K. Sandhu, and Y. Kumar, "An Analysis of Detection and Diagnosis of Different Classes of Skin Diseases Using Artificial Intelligence-Based Learning Approaches with Hyper Parameters," *Arch. Comput. Methods Eng.*, vol. 31, no. 2, pp. 1051–1078, Oct. 2023, doi: 10.1007/s11831-023-10005-2.

[10] M. M. Shukla, B. K. Tripathi, T. Dwivedi, A. Tripathi, and B. K. Chaurasia, "A hybrid CNN with transfer learning for skin cancer disease detection," *Med. Biol. Eng. Comput.*, vol. 62, no. 10, pp. 3057–3071, May 2024, doi: 10.1007/s11517-024-03115-x.

[11] A. Naeem, T. Anees, M. Khalil, K. Zahra, R. A. Naqvi, and S.-W. Lee, "SNC_Net: Skin Cancer Detection by Integrating Handcrafted and Deep Learning-Based Features Using Dermoscopy Images," *Mathematics*, vol. 12, no. 7, pp. 1030–1030, Mar. 2024, doi: 10.3390/math12071030.

[12] M. Senthil Sivakumar, L. Megalan Leo, T. Gurumekala, V. Sindhu, and A. Saraswathi Priyadharshini, "Deep learning in skin lesion analysis for malignant melanoma cancer identification," *Multimed. Tools Appl.*, vol. 83, no. 6, pp. 17833–17853, Jul. 2023, doi: 10.1007/s11042-023-16273-1.

[13] R. Karthik, T. S. Vaichole, S. K. Kulkarni, O. Yadav, and F. Khan, "Eff2Net: An efficient channel attention-based convolutional neural network for skin disease classification," *Biomed. Signal Process. Control*, vol. 73, no. 3, pp. 120–128, Mar. 2022, doi: 10.1016/j.bspc.2021.103406.

[14] R. Sadik, A. Majumder, A. A. Biswas, B. Ahammad, and Md. M. Rahman, "An in-depth analysis of Convolutional Neural Network architectures with transfer learning for skin disease diagnosis," *Healthc. Anal.*, vol. 3, no. 4, pp. 145–152, Nov. 2023, doi: 10.1016/j.health.2023.100143.

[15] W. Y. Leong, "Digital Technology for ASEAN Energy," *in Proc. Int. Conf. Comput. Power Commun. Technol. (ICCPCT)*, vol. 2023, no. Aug., pp. 1–6, 2023, doi: 10.1109/iccpct58313.2023.10244806.

[16] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A methodological approach to the classification of dermoscopy images," *Comput. Med. Imaging Graph.*, vol. 31, no. 6, pp. 362–373, Sep. 2007, doi: 10.1016/j.compmedimag.2007.01.003.

[17] S. L. Phung, A. Bouzerdoum, and D. Chai, "Skin segmentation using color and edge information," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Manchester, U.K., Sep. 2001, pp. 1–10.

[18] Q. Abbas, M. E. Celebi, and I. F. Garcia, "A novel perceptual color processing approach for skin lesion segmentation," *Skin Res. Technol.*, vol. 19, no. 1, pp. e490–e497, Feb. 2013, doi: 10.1111/j.1600-0846.2012.00670.x.

[19] G. King and L. Zeng, "Explaining rare events in international relations," *Int. Organ.*, vol. 55, no. 3, pp. 693–715, Jul. 2001, doi: 10.1093/oxfordjournals.pan.a004868.

[20] DermNet NZ, "DermNet skin disease information," [Online]. Available: https://dermnetnz.org/. [Accessed: Sept. 10, 2024].

[21] S. Singhal, "Skin disease classification," *Kaggle*, [Online]. Available: https://www.kaggle.com/code/smitisinghal/skin-disease-classification. [Accessed: Sept. 10, 2024].

[22] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 60–89, Dec. 2019, doi: 10.1186/s40537-019-0197-0.