SpectraViT: A Novel Hybrid Architecture for Enhanced Melanoma Classification

Samridhi Raj Sinha
Dept of Computer Engineering
MPSTME, NMIMS
Mumbai, India
samridhi.rajsinha53@nmims.in

Asmi Parikh
Dept of Computer Engineering
MPSTME, NMIMS
Mumbai, India
asmi.parikh80@nmims.in

Vidhi Damani
Dept of Computer Engineering
MPSTME, NMIMS
Mumbai, India
vidhi.damani52@nmims.in

***Abstract*—Melanoma is one of the most invasive types of skin cancers which requires an accurate understanding of the nature of the disease for better management. However, subtle differences in lesion morphology are one of the greatest challenges for traditional methods. This research presents SpectraViT, a hybrid architecture that combines Fourier and wavelet feature extraction techniques as part of a Vision Transformer (ViT) model. Using this technique allows transforming structures and capturing spatial and frequency and enhances information enabling the model to interpret more intricate patterns of skin lesions. In this paper we have discussed our experimental findings which have shown improvements in accuracy as compared to the existing techniques, thus validating the applicability of our framework for reliable melanoma diagnosis. We evaluate SpectraViT on a diverse melanoma dataset and demonstrate significant improvements in classification accuracy, achieving over 91% on test data.**

***Index Terms*—Melanoma classification, Vision Transformer, Fourier transforms, adaptive masking, deep learning.**

I. Introduction

Melanoma develops from melanocytes, the pigment-producing cells in the skin, and is one of the most dangerous forms of skin cancer. It is characterized by the uncontrolled growth of cells, often due to DNA damage caused by excessive ultraviolet (UV) radiation exposure. Although melanoma accounts for only 1% of all skin cancers, it is the leading cause of skin cancer-related deaths. It is also the second most common cancer among women under 30. In the United States, the American Cancer Society projected approximately 100,000 new cases of melanoma in 2023, with around 8,000 of these cases potentially leading to death if not diagnosed early [1]. This overwhelming burden on healthcare systems underscores the importance of early diagnosis for effective management, particularly for those with high UV exposure. However, melanoma diagnosis is challenging due to the difficulty of distinguishing between benign and malignant skin lesions. This makes melanoma one of the most difficult skin cancers to classify accurately.

Common diagnostic methods, such as dermoscopic examination based on the ABCDE features (Asymmetry, Border, Color, Diameter, and Evolution), are essential but often insufficient for differentiating between benign lesions and melanoma, especially for intermediate dermatologists. The subtle differences in texture, color, and structure between benign and malignant lesions complicate the identification process.

Recent advancements in deep learning, particularly through Convolutional Neural Networks (CNNs), have significantly improved the detection of medical images, including skin lesions. CNNs are particularly effective in image-based tasks due to their filters that capture local image features. However, CNNs have limitations, as they are designed to focus on small regions of an image, which may lead to overlooking large, more complex patterns indicative of melanoma. This limitation has prompted researchers to explore Vision Transformers (ViTs), which analyze images differently by treating them as a collection of patches and using attention mechanisms to capture wider, deeper relationships between image regions [2]. ViTs are better suited for identifying the subtle patterns associated with melanoma, where CNNs might miss critical features.

While ViTs have proven effective for capturing spatial patterns, they may still miss other important aspects in melanoma detection. For more comprehensive image analysis, methods such as Fourier and Wavelet Transforms are useful. Fourier Transform decomposes an image to reveal its overall frequency structure, which can help remove noise and highlight significant features. The Wavelet Transform, on the other hand, is advantageous for focusing on local details, such as edges and textures, which are crucial in melanoma detection. Combining both transforms enhances the ability to detect both macro and micro patterns, providing a more complete analysis.

In this study, we present *SpectraViT*, a novel hybrid model that leverages the Vision Transformer backbone, incorporating both Fourier and Wavelet transforms. The Fourier Transform is used to capture global features, while the Wavelet Transform focuses on local features, thereby improving the classification accuracy for melanoma detection. This hybrid approach aims to enhance the performance of existing models by combining the strengths of both frequency domain analysis and localized feature extraction.

II. Background work

Melanoma, though constituting only 5% of skin cancers, is responsible for over 75% of skin cancer-related deaths.

Identify applicable funding agency here. If none, delete this.

Early detection significantly improves survival, raising the five-year survival rate to 98% [3]. Dermatoplasty is a common diagnostic aid, yet it requires significant expertise and can lead to varied interpretations depending on the practitioner [4].

Convolutional Neural Networks (CNNs) have shown considerable promise in melanoma detection, especially on smaller datasets, as they handle image transformations (e.g., rotation and translation) efficiently. This capability allows CNNs to capture critical features without requiring extensive datasets, making them a preferred choice in resource-constrained settings [2], [5]. Vision Transformers (ViTs), while powerful, often demand large datasets and are more complex to train for medical imaging applications, leading to a greater preference for CNNs in effective and resource-efficient melanoma detection [6].

Fourier Vision Transformers (Fourier-ViTs) represent a new approach that uses Fourier Transform techniques to map spatial data into the frequency domain, enabling them to capture both high and low-frequency details for image analysis. This approach reduces the risk of overfitting in large datasets, making it especially useful in medical imaging, where accurate and precise detection is critical [7], [8].

Wavelet Vision Transformers (Wavelet-ViTs) use Wavelet Transforms to decompose images across multiple scales and orientations. This capability enhances both spatial and frequency information extraction, which is essential for capturing localized features. This characteristic allows Wavelet-ViTs to excel in tasks requiring high-resolution attention to fine details, such as medical imaging [4], [10]. By combining wavelet decomposition principles with transformer architecture, Wavelet-ViTs can improve feature extraction while maintaining computational efficiency, making them a robust choice for detailed image analysis tasks [11], [12].

Fourier and Wavelet Transforms in image processing facilitate the identification of intricate patterns by isolating specific features. Fourier Transforms, which capture overall frequency structure, are especially useful for locating patterns and textures, thereby assisting Vision Transformers in accurate melanoma detection [13]. Wavelet Transforms perform multi-resolution analyses, decomposing images into various scales and orientations, which better represents both spatial and frequency information. This dual capability gives Wavelet Transforms an advantage in detecting local features and identifying high-level, complex patterns often present in medical imaging [14], [15]. Together, these transforms create a powerful framework for enhanced image analysis and pattern recognition.

This study presents a hybrid approach, SpectraViT, that combines Fourier and Wavelet Vision Transformers, aiming to integrate the benefits of both Fourier's frequency-focused insights and Wavelet's localized feature retention. By focusing on dominant frequency components from the Fourier domain and preserving localized features, SpectraViT aims to achieve high diagnostic accuracy and computational efficiency. This approach minimizes noise and overfitting effects while enabling complex pattern recognition, making it a promising tool for medical imaging applications [6], [7], [13].

III. Dataset

The Melanoma Cancer Dataset consists of 10000 images of variety of skin lesions divided into two classes - benign (non-cancerous) and malignant (cancerous).It is further divided into training set of 9600 images used for model training and test set of 1000 images for model evaluation. The images underwent a series of preprocessing transformations to augment the dataset and improve model generalization. These transformations included random resizing, horizontal flipping, rotation, and color jittering.

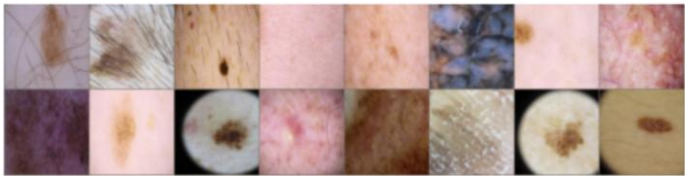

Fig. 1. Overview of the Melanoma Skin Cancer Dataset.

IV. Mathematical Transforms

A. Overview

The system proposed in this work, SpectraViT, aims to improve melanoma classification by taking advantage of the strengths of the Fourier Transform FT and Wavelet Transform WT to achieve better classification accuracy. This hybrid approach features that is of frequency as well as spatial analysis hence a robust framework for feature extraction and classification.

B. Fourier Transform

The **Fourier Transform** is one of the essential tools for converting images from the spatial domain to the frequency domain. This transformation allows for enhanced image manipulation by decomposing the image into sine and cosine components, each with specific magnitudes, frequencies, and phases. The Fourier Transform is useful for various applications, including image enhancement, filtering, and compression. In the resulting frequency domain, each point corresponds to a unique frequency from the original image, highlighting features that are important for melanoma detection.

The two-dimensional Fourier Transform of a discrete spatial function $f(m, n)$ is given by:

$$F(\omega_1, \omega_2) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f(m, n) e^{-j\omega_1 m} e^{-j\omega_2 n} \quad (1)$$

where ω_1 and ω_2 are frequency variables expressed in radians per sample. The transformed image $F(\omega_1, \omega_2)$ reveals significant frequency components of the original image, which are useful for detecting features that differentiate benign from malignant lesions.

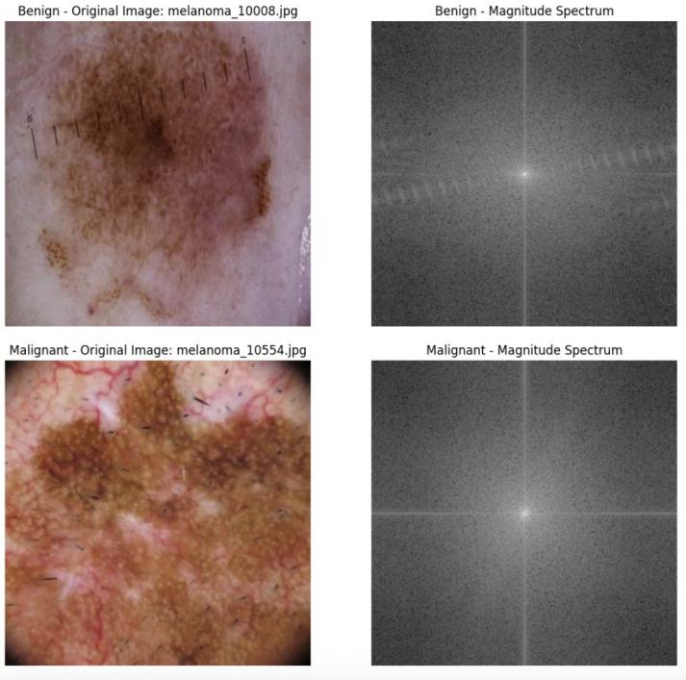

Fig. 2. Applying Fourier Transform to lesion images for analyzing frequency components

C. Wavelet Transform

The **Wavelet Transform** (WT) enables the analysis of images at multiple resolutions using wavelets, or "small waves." Unlike the Fourier Transform, which focuses solely on frequency information, the WT provides simultaneous spatial and frequency analysis, making it suitable for multiresolution approaches.

For a continuous signal $f(t)$, the Wavelet Transform with a chosen "mother wavelet" $\psi(t)$ is defined as:

$$W_f(a,b) = \int_{-\infty}^{\infty} f(t)\,\psi^*\left(\frac{t-b}{a}\right)\frac{dt}{\sqrt{|a|}} \quad (2)$$

where:

- $W_f(a,b)$ represents the wavelet coefficient at scale a and position b,
- a is the scaling factor, which adjusts the wavelet's width,
- b translates the wavelet across the signal,
- $\psi(t)$ is the mother wavelet, and
- ψ^* is the complex conjugate of ψ.

V. Proposed System

In this section, we introduce our proposed model, **SpectraViT**, which integrates Fourier and wavelet transforms within the Vision Transformer (ViT) architecture. This unique combination enhances SpectraViT's effectiveness in classifying skin cancer lesions. Here, we discuss the model architecture, data preprocessing steps, training methodologies, and performance evaluation metrics.

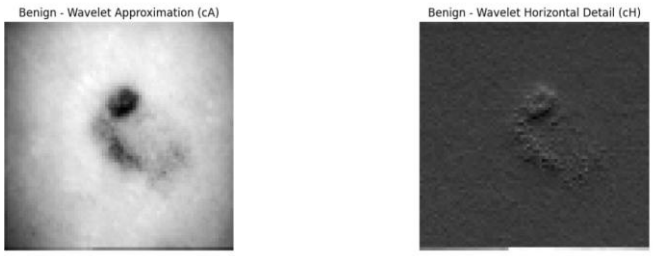

Fig. 3. Haar wavelet transform applied on lesion images for capturing spatial features

A. Model Architecture

Architecture of **SpectraViT** explores the best of the strengths of both the mathematical transforms i.e Fourier Transform (capturing global features) and Wavelet Transforms (capturing localised features).

As shown in **Figure 4**, the model consisys of the following main components:

- Fourier Layer
- Wavelet Layer
- Hybrid Pooling Layer
- Vision Transformer (ViT)

1) Fourier Layer: The **FourierLayer** transforms input images from the spatial domain to the frequency domain using the Fast Fourier Transform (FFT). This transformation captures high frequency components, which can be important in identifying patterns or trends of features responsible for melanoma.

2) Wavelet Layer: The **WaveletLayer** uses Haar wavelets to perform a a two-dimensional wavelet decomposition of the input images. This enables the model to extract features at multiple resolutions, allowing for a detailed analysis of both spatial and frequency terms.

3) Hybrid Pooling Layer: The **HybridPoolingLayer** uses max pooling to preserve features while using average pooling to reduce spatial dimension With a single pooling method, important information tends to be lost, whereas this dual pooling strategy can help the model keep important information.

4) Vision Transformer (ViT): After the feature extraction, it uses a pre-trained Vision Transformer whose weights have been trained on large datasets like ImageNet. The pretrained model acts as a robust backbone and allows the model to utilize all the learned features to help classify better. The Fourier and the wavelet outputs are concatenated, projected linearly, and out are passed through the ViT, which captures complex relationships in the data.

VI. Implementation

During each training epoch,Model was updated using mini-batches of data for each training epoch, progress monitored with tqdm library for visualisation of progress.

Next, the dataset is partitioned into training and validation subsets using an 80-20 split, facilitating the evaluation of model performance on unseen data.

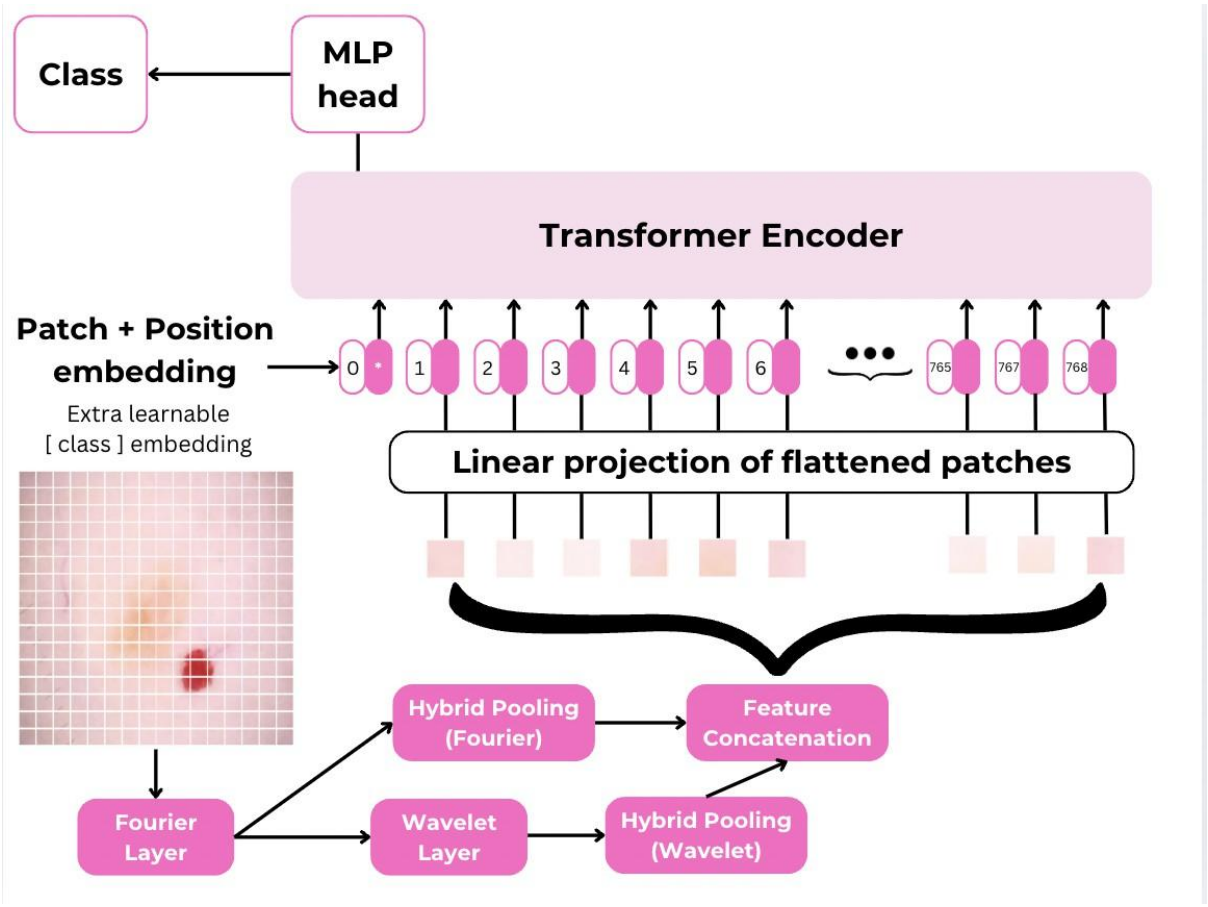

Fig. 4. Architecture diagram of SpectraViT model

TABLE I
MODEL ARCHITECTURE OVERVIEW

Layer	Description	Output Shape
Input	Input Image	[1, 3, 224, 224]
Fourier Layer 1	Fourier Transform Applied	[1, 3, 224, 224]
Fourier Layer 2	Fourier Transform Applied	[1, 3, 224, 224]
Wavelet Layer 1	Wavelet Transform Applied	[1, 3, 56, 56]
Wavelet Layer 2	Wavelet Transform Applied	[1, 3, 56, 56]
Fourier Pooling	Pooling Fourier Output	[1, 6, 112, 112]
Wavelet Pooling	Pooling Wavelet Output	[1, 6, 28, 28]
Feature Concatenation	Combined Fourier and Wavelet Features	[1, 79968]
Projection Layer	Projected Features	[1, 768]
Reshape	Reshaped for ViT Input	[1, 1, 768]
Output Layer	Final Classification Output	[1, 2]

A. Evaluation Metrics

To assess model performance, we computed several metrics after each epoch, including training loss, validation loss, and validation accuracy. These metrics provide insight into the model's ability to generalize and perform classification accurately. In addition to these general metrics, more specific evaluation measures such as **Precision**, **Recall**, and **F1 Score** are crucial for understanding the model's performance, especially in imbalanced datasets.

a) Precision: Precision, also known as positive predictive value, measures the accuracy of positive predictions. It is the proportion of true positive results among all instances that were predicted as positive. The formula for Precision is:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Where: - TP is the number of true positives (correctly predicted positive instances), - FP is the number of false positives (incorrectly predicted positive instances).

Precision is particularly useful in scenarios where the cost of false positives is high (e.g., diagnosing a disease, where a false positive may lead to unnecessary treatments).

b) Recall: Recall, also known as sensitivity or true positive rate, measures the model's ability to identify all relevant instances. It is the proportion of true positives among all actual positive instances. The formula for Recall is:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where: - FN is the number of false negatives (instances that were missed by the model).

Recall is crucial when the cost of false negatives is high (e.g., detecting fraud or rare diseases), as it focuses on ensuring that as many positive instances as possible are captured by the model.

c) F1 Score: The F1 Score is the harmonic mean of Precision and Recall. It provides a balanced measure that takes both false positives and false negatives into account. The formula for F1 Score is:

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

The F1 Score is particularly useful when the class distribution is imbalanced, as it balances the trade-off between Precision and Recall. A higher F1 Score indicates a better balance between the two metrics.

d) Why These Metrics Matter: Precision, Recall, and F1 Score are particularly important for evaluating models in scenarios with imbalanced classes or when both false positives and false negatives have significant costs. Unlike accuracy, which can be misleading in such cases, these metrics provide a more nuanced understanding of model performance. By considering these metrics, we can ensure the model is not just predicting the majority class correctly but also handling the minority class effectively.

VII. Results

In this section, we present the results of our proposed SpectraViT model for melanoma classification together with an analysis. The performance metrics and some visualizations capture the effectiveness of our hybrid approach.

The classification report shown in Figure 5 summarizes the precision, recall, and F1-scores for each class. The SpectraViT model performs excellently in differentiating malignant from benign cases.

Figure 6 shows the predictions made by the SpectraViT model on the test dataset, showing its good performance in being able to classify melanoma.

The confusion matrix in Figure 7 shows that the model achieves an accuracy of 91.5%, correctly classifying the majority of the skin lesions in the dataset.

The area under the Precision-Recall curve (Figure 8) is 0.95, meaning that the model reduces false positives in cases of identified melanoma effectively.

The ROC curve in Figure 9 also shows that the SpectraViT model is an efficient one by yielding an AUC of 0.96 which makes it very clear that it is indeed possible to identify the malignant and benign lesions by this model.

Figure 10 provides a comparison of the SpectraViT model's performance with other baseline models, highlighting its superior classification capabilities for melanoma.

Classification Report:

	precision	recall	f1-score	support
0	0.89	0.94	0.91	980
1	0.93	0.88	0.91	941
accuracy			0.91	1921
macro avg	0.91	0.91	0.91	1921
weighted avg	0.91	0.91	0.91	1921

Fig. 5. Classification Report

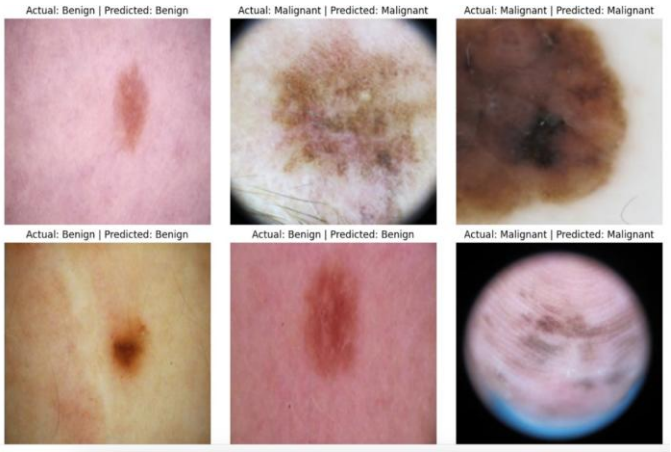

Fig. 6. Model Predictions on Test Dataset

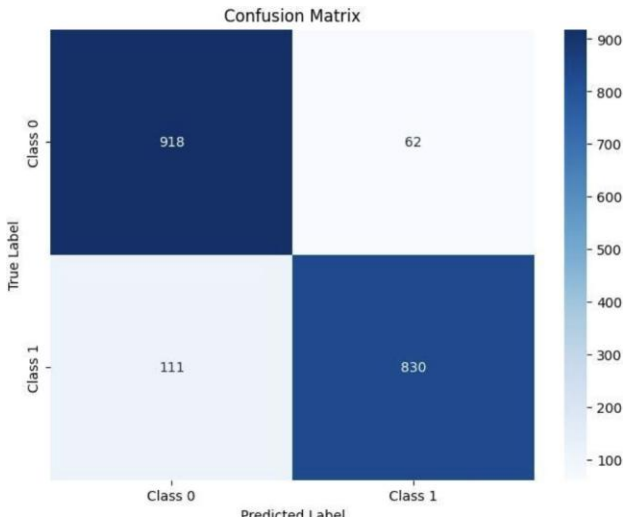

Fig. 7. Confusion Matrix

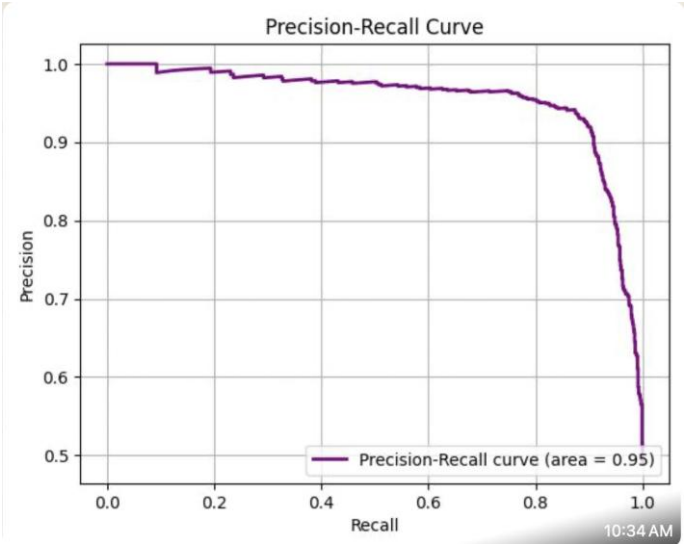

Fig. 8. Precision-Recall Curve

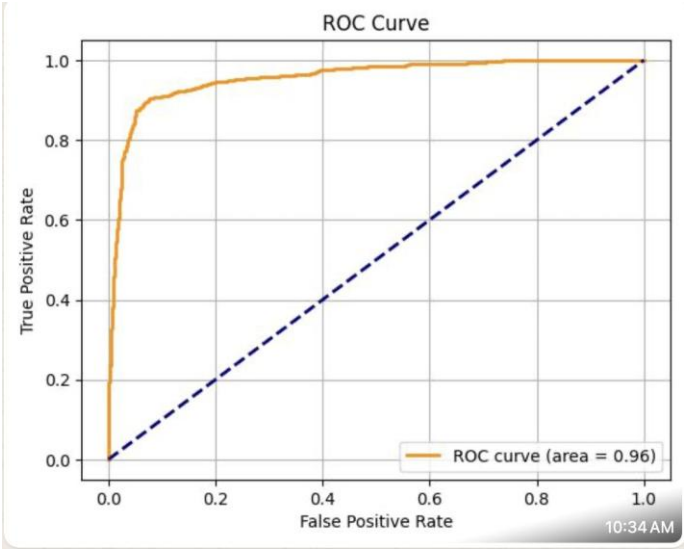

Fig. 9. ROC Curve

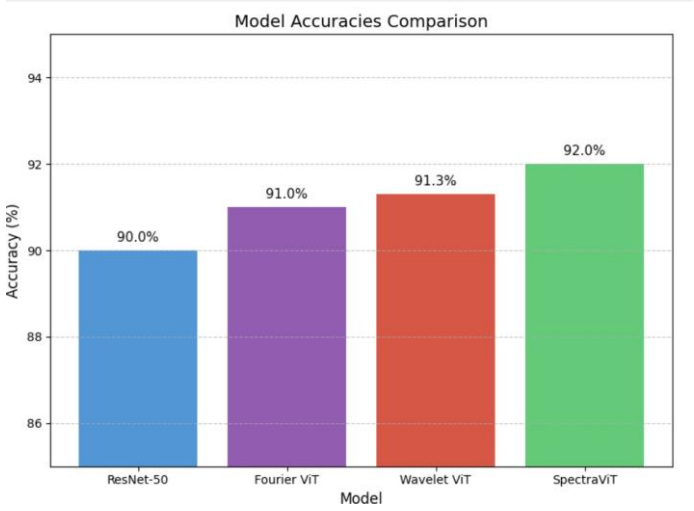

Fig. 10. Model Performance Comparison

VIII. Conclusion

In this work, we introduce SpectraViT: a Fourier and wavelet transform-derived melanoma classification model that improves upon the Vision Transformer architecture. By leveraging frequency and spatial domain information, SpectraViT provides skin lesion images with high frequency and global perception. This makes it easier to get more detailed information regarding intricate skin patterns and lesions, addressing cardinal problems posed by conventional CNNs and pure ViTs.

IX. Future Work

Future work for SpectraViT will focus on fine-tuning its hyperparameters, such as optimizing the learning rate schedule, increasing the number of epochs, and adjusting dropout rates to further enhance the model's generalization and prevent overfitting. Exploring different learning rate schedulers, like cosine annealing or cyclical learning rates, could improve convergence and stability. We also aim to experiment with other wavelet and Fourier configurations, testing various decomposition levels and kernel sizes to refine SpectraViT's sensitivity to both global and local melanoma patterns. Additionally, incorporating explainability tools like saliency maps or Grad-CAM could offer greater transparency and insight, which is critical for clinical integration. This refined approach aims to make SpectraViT not only more accurate but also a reliable tool for early melanoma detection in dermatology patient care.

References

[1] American Cancer Society, "Cancer Facts & Figures 2023," American Cancer Society, Atlanta, GA, USA, 2023.

[2] G. M. S. Himel, M. M. Islam, K. A. Al-Aff, S. I. Karim, and M. K. U. Sikder, "Skin Cancer Segmentation and Classification Using Vision Transformer for Automatic Analysis in Dermatoscopy-Based Noninvasive Digital System," 2023.

[3] L. Gamage, U. Isuranga, D. Meedeniya, S. De Silva, and P. Yogarajah, "Melanoma Skin Cancer Identification with Explainability Utilizing Mask Guided Technique," 2023.

[4] X. Shi, X. Dong, S. Ye, W. Li, and H. Li, "Wavelet Integrated Multiscale Feature Fusion Network for Imbalanced Skin Lesion Classification," Kunming University of Science and Technology, Yunnan University, The Third Affiliated Hospital of Kunming Medical University, 2023.

[5] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, "Deep Learning for Medical Image-Based Cancer Diagnosis," 2023.

[6] G. Cirrincione, S. Cannata, G. Cicceri, F. Prinzi, T. Currieri, M. Lovino, C. Militello, E. Pasero, and S. Vitabile, "Transformer-Based Approach to Melanoma Detection," 2023.

[7] A. Nekoozadeh, M. R. Ahmadzadeh, and Z. Mardani, "Multiscale Attention via Wavelet Neural Operators for Vision Transformers," Department of Electrical and Computer Engineering, Isfahan University of Technology, 2023.

[8] M. A. Arshed, S. Mumtaz, M. Ibrahim, S. Ahmed, M. Tahir, and M. Shafi, "Multi-Class Skin Cancer Classification Using Vision Transformer Networks and Convolutional Neural Network-Based Pre-Trained Models," 2023.

[9] G.-I. Kim and K. Chung, "ViT-Based Multi-Scale Classification Using Digital Signal Processing and Image Transformation," Kyonggi University, South Korea, 2023.

[10] S. I. Hussain and E. Toscano, "An Extensive Investigation into the Use of Machine Learning Tools and Deep Neural Networks for the Recognition of Skin Cancer: Challenges, Future Directions, and a Comprehensive Review," 2023.

[11] M. F. Aslan, "Comparison of Vision Transformers and Convolutional Neural Networks for Skin Disease Classification," Electrical and Electronics Engineering, Karamanoglu Mehmetbey University, Turkey, 2023.

[12] Y. Gulzar and S. A. Khan, "Skin Lesion Segmentation Based on Vision Transformers and Convolutional Neural Networks—A Comparative Study," 2023.

[13] H. Duan, Y. Liu, H. Yan, Q. He, Y. He, and T. Guan, "Fourier ViT: A Multi-scale Vision Transformer with Fourier Transform for Histopathological Image Classification," Tsinghua University, Shenzhen, China, 2023.

[14] T. C. Cahoon, M. A. Sutton, and J. C. Bezdek, "Breast Cancer Detection Using Image Processing Techniques," University of West Florida, Pensacola, FL, 2023.

[15] V. J. Pawar, K. D. Kharat, S. R. Pardeshi, and P. D. Pathak, "Lung Cancer Detection System Using Image Processing and Machine Learning Techniques," University College of Engineering, Osmania University, Hyderabad, India, 2023.