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.damani58@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 cells responsible for the skin's color, and is among the most terrible forms of skin cancer. It is characterized by uncontrolled growth of transformed cells, which can mostly be attributed to DNA damage due to over-exposure to ultraviolet (UV) radiation. Melanoma, which represent 1% all skin cancers, is the highest contributor to skin cancer death resulting to It's being the second most common cancer among women under 30. The American Cancer Society put the global figure of melanoma cases at about 100,000 in 2023 in the United States alone and estimated that about Eight Thousand will lead to death if not mitigated in time [1]. This tends to inundate the work of American physicians, and such therapies have little effect leading to a very low survival rate thus early diagnosis is critical for management of these cases exposure to UV in outdoor activities. Diagnosis is however challenging due to the difficulties in distinguishing between similar appearing benign and malignant skin lesions and this makes it one of the most difficult types of skin cancers to classify accurately. Most of the Common diagnostic methodologies include dermoscopic examination of the ABCDE features of skin lesions i.e. asymmetry, border, color, diameter and evolution. However, curtains up for intermediate dermatologists, the identification of mellanoma from benign lesions is made complex due to mild dissimilitude in the texture, color and structure.

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

With deep learning, primarily through Convolutional Neural Networks (CNNs), there have been major strides made towards the detection of medical images such as skin lesions. Image-focused tasks such as photos are easily converged by CNN models, the key component of which are the filters dedicated to image details. Nevertheless, the issue with CNNs stems from the fact that they are designed only to capture small regions of an image and therefore, large and more complex patterns that could serve as indicators of melanoma may be left out and overlooked. This has sparked researchers to shift their focus to Vision Transformers (ViTs) which have a fundamentally different way of analyzing images. Contrary to CNNs, ViTs perceive images as collections of patches and solicits attention mechanisms capable of capturing wider, deeper relationships which suits them to address the subtler patterns that are often associated with melanoma. Focusing on only spatial patterns such as shape and arrangement has its disadvantages where it might miss other significant aspects in melanoma detection. For images, methods such as the Fourier Transform and Wavelet Transform are useful which view images in frequencies that depict both mini and macro perspectives that are pertinent to the problem at hand. The Fourier Transform decomposes the actual image to reveal its overall frequency structure, which can be useful in removing pits and obtrusions as background noise. On the other hand, the Wavelet transform can be advantageously used to focus on the details and processes characteristics of smaller and larger physical dimensions that depict edges and textures present in the image. The combination of these methods provides

In this study we present SpectraViT, which is a new hybrid model that takes advantage of the Vision Transformer backbone, incorporating Fourier and Wavelet transforms into the model. By incorporating Fourier transform for global features and wavelet transforms for local features, it enhances classification accuracy.

II. Background work

Melanoma, though only 5 percent of skin cancers, causes over 75 percent of related deaths. Early detection raises the five-year survival rate to 98 percent. Dermatoplasty aids

in diagnosis but requires expertise and may yield varying interpretations.

For melanoma detection, CNNs outperform Vision Transformers (ViTs), especially with smaller datasets. CNNs handle image transformations like rotation and translation efficiently, capturing critical features without needing large datasets. ViTs, in contrast, often require extensive data and are more complex to train for medical imaging, making CNNs the preferred choice for effective and resource-efficient melanoma detection.

Fourier Vision Transformers (Fourier-ViTs)All Fourier-ViTs use Fourier Transform methods to move spatial data, i.e., images, into the frequency domain in a novel way to achieve image analysis. This allows them to effectively capture high and low-frequency details to aid in recognizing complex patterns in images. Fourier-ViTs reduce the risk of overfitting, excluding in the case of smaller datasets. Hence, they are very useful in medical imaging, where we need very accurate and precise detection.

Wavelet Vision Transformers are a new spin on vision transformers that use Wavelet Transforms to decompose images into multiple scales and directions. This technique enables both the spatial and frequency information capturing which is an essential part for extracting a localized feature. This is what makes Wavelet-ViTs excel in high-resolution based fine detail attention using computer vision tasks, for example in medical imaging. They are good at capturing complex patterns but maintain the local structure of images. Combining the principles of wavelet decomposition and transformer architecture, these models enhance feature extraction while minimizing increases in computational requirements, achieving both performance and efficiency in the natural language and tasks relevant spaces.

Image processing techniques such as Fourier and Wavelet Transforms make the analysis of complicated patterns much easier by highlighting particular features. Fourier Transforms shows promise in locating patterns and textures and thus is an integrated production in this laboratory to efficiently and accurately perform Vision Transformers. Wavelet Transforms, on the other hand, performs a multi-resolution analysis, decomposing images into multiple scales and orientations, representing spatial and frequency information more effectively. This two-in-one capability gives Wavelet Transforms an edge in local feature detection as well as in identifying high-level and intricate patterns —as seen often in medical imaging Together, these techniques create a formidable framework for enhanced image analysis and pattern recognition.

The aim is to present our approach which combines the best of Fourier and Wavelet Vision Transformers. Our goal is to have the best of both worlds by extracting only the dominant frequency components from Fourier whilst retaining a huge amount of localized features with this data analytical technique that are suited for image analysis, specifically for medical diagnostics. This hybrid approach minimizes the effects of noise and overfitting whilst allowing for the representation of complex patterns while staying computationally efficient. Thus such combinational integration could provide enhanced detection/inference to improve the diagnostic accuracy, making excellent potential tool from medical imaging to clinical applications.

III. Dataset

The Melanoma Cancer Dataset consists of two parts - train dataset and test dataset. The train dataset consists of 9600 images, used to train the model and other consists of 1000 images to test the model's performance. Therefore, total dataset size is around 10000 images.It contains balanced number of both benign and malignant variety of skin lesions images which can be used for the effective training and evaluation of the model.

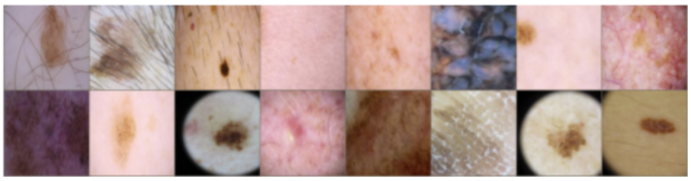

Fig. 1. Overview of the Melanoma Skin Cancer Dataset.

IV. Proposed System: SpectraViT

A. Overview

The proposed system, termed SpectraViT, aims to enhance the classification accuracy of melanoma skin cancer images by combining the strengths of Fourier Transform (FT) and Wavelet Transform (WT). This hybrid approach leverages both frequency and spatial analysis, providing a robust framework for feature extraction and classification in deep learning models.

B. Fourier Transform

The **Fourier Transform** is an essential technique in image processing that transforms images from the spatial domain into the frequency domain. It decomposes an image into its sine and cosine components (complex exponentials) of various magnitudes, frequencies, and phases. Each point in the frequency domain represents a specific frequency in the original image, facilitating various applications such as image enhancement, filtering, and compression. The two-dimensional Fourier Transform of a discrete spatial function $f(m, n)$ is defined as:

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

where ω_1 and ω_2 are frequency variables in radians per sample. The frequency-domain representation $F(\omega_1, \omega_2)$ provides crucial information about the image's frequency components, which can be utilized to highlight important features relevant for melanoma detection.

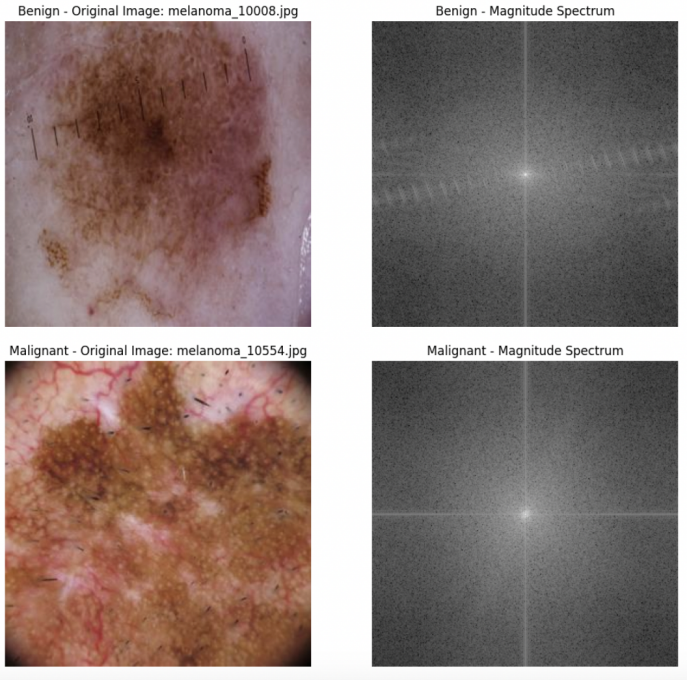

Fig. 2. Fourier transform applied on lesions to capture frequency component

C. Wavelet Transform

The **Wavelet Transform** (WT) analyzes images at multiple resolutions using wavelets, or "small waves," allowing for simultaneous assessment of spatial and frequency information in a multiresolution framework.

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

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

where:

- $W_f(a,b)$ represents the wavelet coefficient at scale a and translation b,
- a is the scaling parameter,
- b is the translation parameter,
- $\psi(t)$ is the mother wavelet, and
- ψ^* is the complex conjugate of ψ.

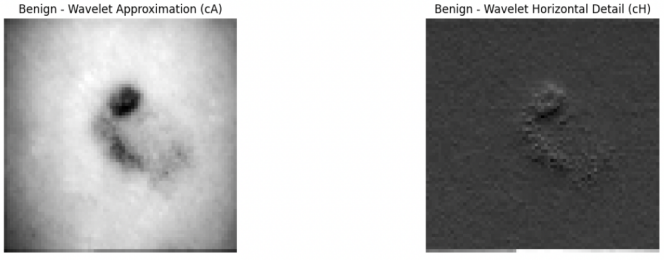

Fig. 3. Haar wavelet transform applied on lesions to capture spatial features

V. Implementation

In this this section, our proposed model **SpectraViT** that fuses Fourier and wavelet transformations with the Vision Transformer (ViT) architecture has been implemented for skin cancer classifications.The model architecture, data preparation, training procedure, and evaluation metrics of the model will be discussed further.

A. Model Architecture

The architecture of **SpectraViT** is designed to explore 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 1**, the model comprises the following main components:

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

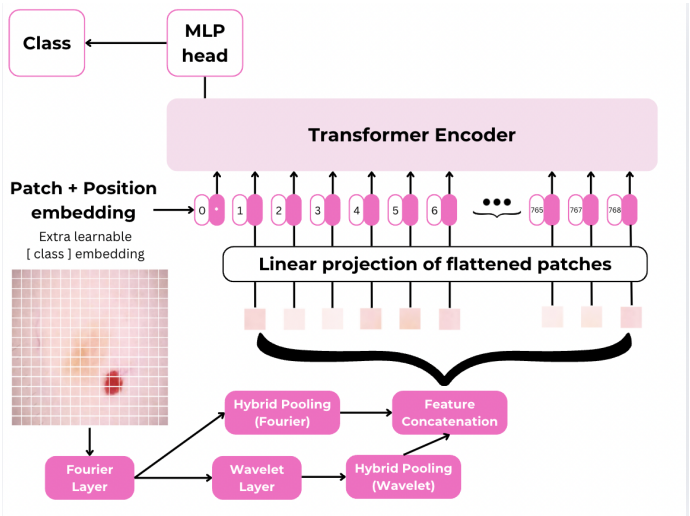

Fig. 4. Architecture diagram of model

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.

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.

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]

B. Data Preparation

For the training and validation of the SpectraViT model, we used the **Melanoma Skin Cancer Dataset**, consisting of a diverse collection of images. 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.

1) Evaluation Metrics: To assess model performance, we computed training loss, validation loss, and validation accuracy after each epoch. These metrics provide insight into the model's ability to generalize and perform classification accurately.

VI. RESULTS

In this section, we present the results obtained from our proposed SpectraViT model for melanoma classification, along with analysis. The performance metrics and visualizations highlight the effectiveness of our hybrid approach.

The classification report in Figure 5 summarizes the precision, recall, and F1-scores for each class. The SpectraViT model demonstrates high performance, accurately distinguishing between malignant and benign cases.

Figure 6 illustrates the predictions made by the SpectraViT model on the test dataset, showing its ability to correctly classify melanoma cases.

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

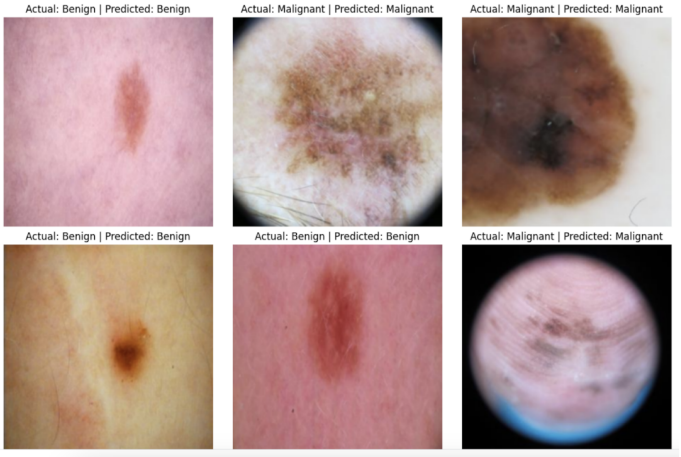

Fig. 6. Model Predictions on Test Dataset

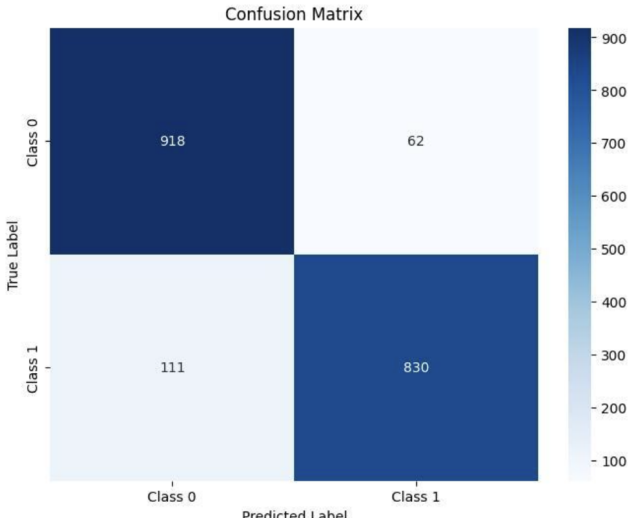

Fig. 7. Confusion Matrix

The AUC of the Precision-Recall curve (Figure 9) is 0.95, indicating that the model minimizes false positives while effectively identifying melanoma cases.

The ROC curve in Figure 10 shows an AUC of 0.96, demonstrating that the SpectraViT model is highly effective at distinguishing between malignant and benign lesions.

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.

Figure 11 provides a comparison of the SpectraViT model's performance with other baseline models, highlighting its superior classification capabilities for melanoma.

VII. 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.

VIII. 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

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.

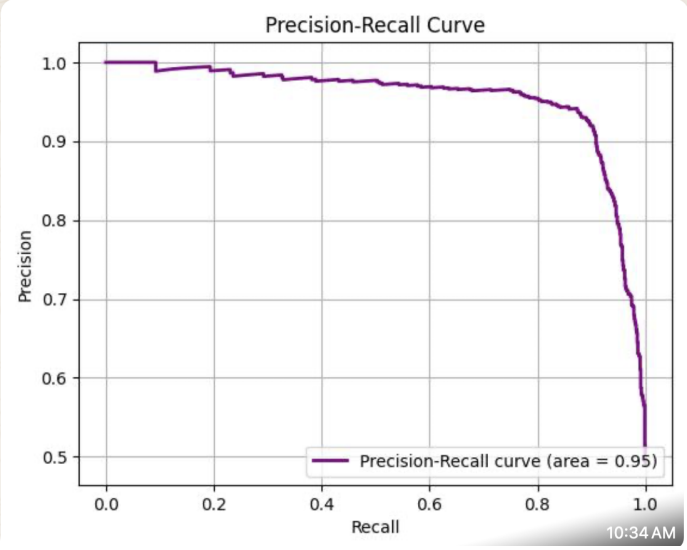

Fig. 8. Precision-Recall Curve

```
Testing: 100%|██████████| 32/32 [00:17<00:00,  1.82batch/s]
Testing completed. Accuracy: 91.50%
```

Fig. 9. Precision-Recall Curve

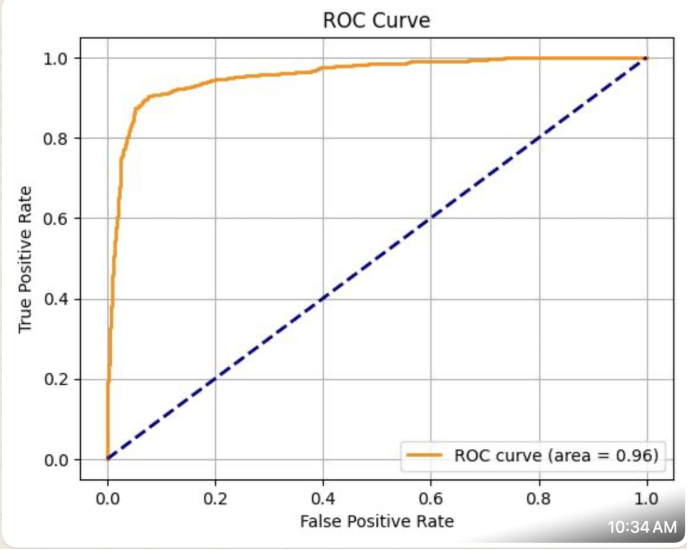

Fig. 10. ROC Curve

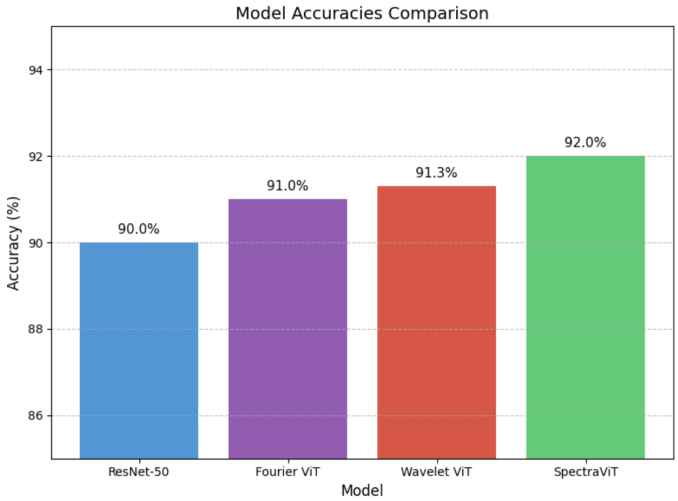

Fig. 11. Model Performance Comparison