International Journal of Advanced Technology in Engineering
Volume. No. 14, Issue No. 05, May 2026
(ISSN: 2348-7550)

Hybrid CNN–ViT with Frequency Domain Enhancement for GI Bleeding Classification

[1]Dr. M. Aparna,[2]Lavanya Chowdam,[3]Lakshmi Sai Suvarnika Janapati, [4]Sudhamayi Ganji, [5]Swathi Gumma, [6]Rajeev Chowdary Gutta

[1]Associate Professor,[2] UG Student, [3]UG Student,[4]UG Student,[5]UG Student, [6]UG Student Department of Computer Science and Engineering (AI & ML),
Tirumala Engineering College, Narasaraopet, India
[1]mudiyalaaparna.89@gmail.com,[2]lavanyakrishna7013@gmail.com,[3]janapatisuvarnika@gmail.com
[4]sudhamayiganji@gmail.com,[5]ganganagaswathi@gmail.com,[6]stu.rajeevg@gmail.com

► ***Abstract***— In conventional architectures, passing raw convolutional neural network (CNN) feature maps directly to vision transformers (ViT) often results in suboptimal performance. This is mainly due to redundant activations and noise that can obscure relevant patterns, particularly in the case of small, localized anomalies. Additionally, directly transmitting these feature maps to the ViT limits the model's ability to effectively capture long-range dependencies. A Fourier transform to the CNN-generated feature maps is proposed to address these limitations, converting them to the frequency domain. This transformation reduces noise and amplifies periodic patterns, thereby providing ViT with cleaner, frequency-enhanced inputs that allow for more effective self-attention across the entire image. The proposed method CNN-Fourier-ViT was evaluated on a wireless capsule endoscopy (WCE) dataset comprising 2,618 images focused on detecting bleeding versus non-bleeding conditions in gastrointestinal imaging. Experimental results demonstrated that our model outperformed traditional CNNs and CNN-ViT hybrids, achieving an accuracy of 96% with a loss of 0.11. This approach illustrates the advantages of combining local feature extraction, frequency-based enhancement, and global attention for precise bleeding detection in WCE images, underscoring its potential in resource-limited clinical settings.

Index Terms— CNN, Vision Transformer, Fourier Transform, WCE, Bleeding Detection, Frequency Enhancement, Medical Imaging, Deep Learning.

I. Introduction

In recent years, deep learning techniques, particularly Convolutional Neural Networks (CNNs), have shown promising results in medical image analysis due to their ability to extract local spatial features. However, CNNs often struggle to capture global contextual information and long-range dependencies, which are essential for identifying subtle and small-scale anomalies in medical images. On the other hand, Vision Transformers (ViTs) have demonstrated superior capability in modeling global relationships through self-attention mechanisms, but their performance heavily depends on the quality of input features.

A major limitation in existing CNN–ViT hybrid models is the direct use of CNN feature maps as input to the transformer, which may contain redundant activations and noise. This can reduce the model's ability to focus on relevant patterns, especially in cases involving small bleeding regions. To overcome this limitation, this work proposes a novel hybrid architecture, CNN-Fourier-ViT, which incorporates frequency-domain transformation using the Fourier Transform to enhance feature representation.

The proposed approach converts CNN-extracted feature maps into the frequency domain, enabling noise reduction and amplification of important patterns. This transformation provides cleaner and more informative inputs to the Vision Transformer, allowing it to better capture global dependencies and improve detection accuracy. The model is evaluated on a Wireless Capsule Endoscopy dataset consisting of 2,618 images for binary classification of bleeding and non-

Experimental results demonstrate that the proposed CNN-Fourier-ViT model outperforms traditional CNN and CNN–ViT architectures, achieving an accuracy of 96% with a loss of 0.11. These findings highlight the effectiveness of integrating spatial feature extraction, frequency-domain enhancement, and global attention mechanisms for improved medical image analysis. The proposed system offers a reliable and efficient solution for automated bleeding detection, particularly in resource-constrained clinical environments.

II. METHODLOGY:

- The proposed CNN-Fourier-ViT model combines spatial feature extraction, frequency-domain enhancement, and global attention for accurate bleeding detection in WCE images. The overall workflow consists of the following steps:
- **1. Data Collection and Preprocessing** The dataset consists of 2,618 Wireless Capsule Endoscopy images categorized into bleeding and non-bleeding classes. Images are resized, normalized, and augmented (such as rotation and flipping) to improve model generalization and reduce overfitting.

- **2. Feature Extraction using CNN**
A Convolutional Neural Network (CNN) is used to extract low-level and high-level spatial features from input images. CNN captures local patterns such as edges, textures, and small bleeding regions.
- **3. Fourier Transform for Feature Enhancement**
The feature maps obtained from the CNN are transformed into the frequency domain using the Fourier Transform. This step helps in reducing noise and highlighting important patterns, especially periodic and subtle features that may not be clearly visible in the spatial domain.
- **4. Vision Transformer (ViT) for Global Learning**
The enhanced feature maps are passed to a Vision Transformer (ViT). The self-attention mechanism in ViT captures long-range dependencies and global relationships across the optimized using appropriate loss functions. Performance is evaluated using metrics such as accuracy and loss. The proposed model achieves an accuracy of 96% with a loss of 0.11, outperforming traditional CNN and CNN–ViT models.

III. DESIGN:

The proposed system follows a hybrid deep learning architecture called **CNN-Fourier-ViT**, designed to combine local feature extraction, frequency enhancement, and global attention for accurate bleeding detection.

1. Input Layer
The system takes Wireless Capsule Endoscopy images as input. All images are resized and normalized to maintain consistency across the dataset.

2. CNN Module (Local Feature Extraction)
A Convolutional Neural Network processes the input images to extract spatial features such as edges, textures, and small bleeding regions. This module focuses on capturing local patterns effectively.

3. Fourier Transform Module (Feature Enhancement)
The feature maps generated by the CNN are transformed into the frequency domain using the Fourier Transform.
This module:
- Reduces noise and redundant information
- Enhances important frequency patterns
- Improves feature quality before passing to the transformer

4. Vision Transformer (ViT) Module (Global Learning)
The enhanced feature maps are fed into the Vision Transformer. Using the self-attention mechanism, ViT captures long-range dependencies and global relationships across the image.

entire image, improving the detection of bleeding regions.

- **5. Classification Layer**

The output from the ViT is fed into fully connected layers followed by a softmax classifier to categorize images into bleeding and non-bleeding classes.

- **6. Model Training and Evaluation**

The model is trained using labeled data and

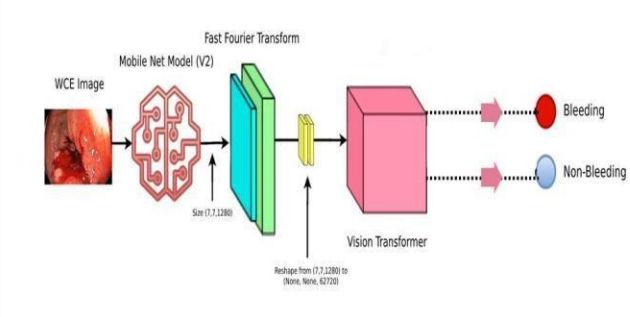


FIG 1: SYSTEM ARCHITECHTURE

Classification Layer

The output from the ViT is passed through fully connected layers and a softmax function to classify the image into:

- Bleeding
- Non-bleeding

5. Output Layer

The final output provides the predicted class label along with confidence scores, enabling accurate and reliable detection.

D. System Architecture

IV. IMPLEMENTATION

The implementation of the proposed CNN-Fourier-ViT model is carried out in the following steps:

1. Dataset Preparation

A dataset of 2,618 Wireless Capsule Endoscopy images is collected and labeled into bleeding and non-bleeding classes. The dataset is divided into training and testing sets.

2. Data Preprocessing

All images are resized to a fixed dimension and normalized. Data augmentation techniques such as rotation, flipping, and scaling are applied

4. Fourier Transform Integration The feature maps obtained from the CNN are transformed into the frequency domain using the Fourier Transform. This step enhances important

improve model robustness and prevent overfitting.

3. CNN Model Implementation A Convolutional Neural Network is implemented to extract spatial features from the input images. The network consists of convolutional layers, activation functions (ReLU), and pooling layers to capture important local patterns. patterns and reduces noise before passing features to the next stage.

5. Vision Transformer Implementation The frequency-enhanced features are fed into the Vision Transformer. The transformer uses self-attention mechanisms to capture global dependencies and relationships within the image.

6. Model Training

The model is trained using labeled data with an appropriate loss function (such as cross-entropy loss) and an optimizer (such as Adam). Training is performed over multiple epochs until the model converges.

7. Model Evaluation

The performance of the model is evaluated using metrics such as accuracy and loss. The proposed model achieves an accuracy of 96% and a loss of 0.11.

8. Prediction and Output

The trained model is used to classify new WCE images into bleeding or non-bleeding categories, providing reliable predictions for medical analysis.

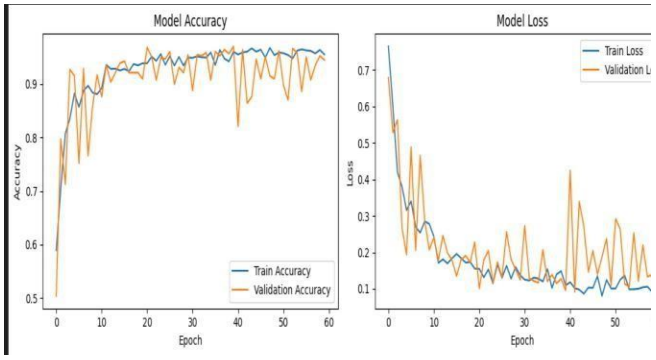
V. Experimental Results and Analysis

The proposed CNN-Fourier-ViT model is evaluated on a dataset of 2,618 Wireless Capsule Endoscopy images for binary classification of bleeding and non-bleeding conditions.

1. Performance Metrics

To assess the effectiveness of the model, standard evaluation metrics are used:

- Accuracy
- Loss



5. Results

The proposed model achieves:

- **Accuracy:** 96%
- **Loss:** 0.11

These results indicate high reliability in detecting bleeding regions from WCE images.

6. Comparative Analysis

The performance of the proposed model is compared with existing approaches:

- **Traditional CNN:**

Captures local features but fails to model global dependencies → lower accuracy

- **CNN–ViT Hybrid:**

Improves global understanding but suffers from noisy feature inputs

- **Proposed CNN-Fourier-ViT:**
  - Reduces noise using Fourier Transform
  - Enhances important patterns
  - Improves attention learning in ViT

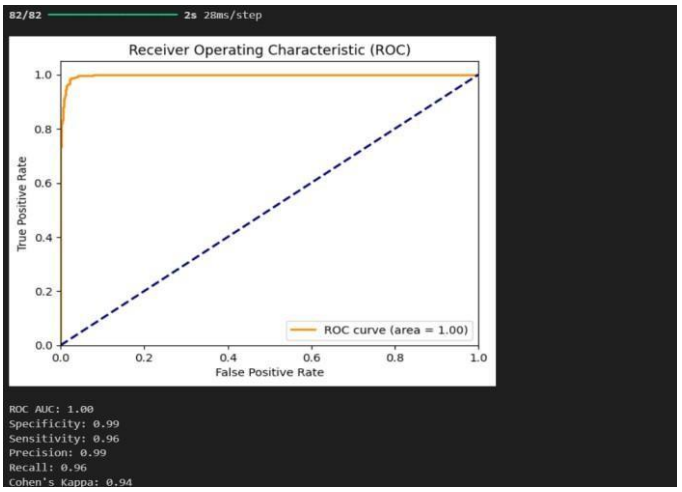
→ Achieves **better accuracy and lower loss**

7. Analysis of Results

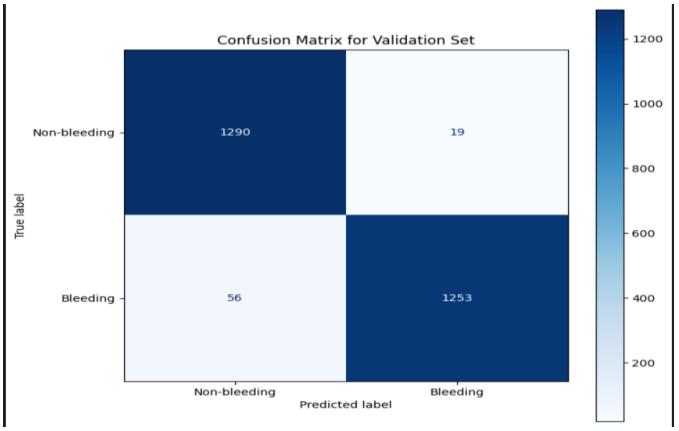
The improved performance of the proposed model is mainly due to:

- **Noise Reduction:** Fourier Transform filters irrelevant information
- **Feature Enhancement:** Important frequency components are highlighted
- **Better Global Learning:** ViT captures long-range dependencies effectively
- **Hybrid Strength:** Combines CNN (local) + Fourier (enhancement) + ViT (global)

Receiver Operating Characteristic



Confusion Matrix



V. Conclusion

The proposed system, WCE Bleeding Classification, successfully integrates deep learning techniques such as MobileNetV2, Fast Fourier Transform, and Vision Transformer to accurately classify gastrointestinal images into bleeding and non-bleeding categories. The use of MobileNetV2 enables efficient feature extraction, while the Fourier Transform enhances important patterns by reducing noise. The Vision Transformer further improves performance by capturing global relationships in the image using attention mechanisms. This combined approach results in high classification accuracy and better detection of small bleeding regions. The system is effective, computationally efficient, and suitable for real-time medical applications. It helps doctors in early diagnosis, reduces manual effort, and minimizes the chances of missing critical bleeding cases.

VI. References

1. Da˘gı, G¨oktu˘g E., G¨ok¸cay, E., and Tora, H. Fourier-Based ImageClassification Using CNN. Journal of Science, Technology and Engineering Research, 5(1), 92–101.
2. D. Pantelaios, P.-A. Theofilou, P. Tzouveli, and S. Kollias, “Hybrid CNN-ViT Models for Medical Image Classification,” in 2024 ISBI, 2024, pp. 1–4.
3. Dong, K., Zhou, C., Ruan, Y., and Li, Y. (2020). MobileNet-v2 model for image classification. In ITCA-2020, pages 476–480. IEEE.
4. Duan, H., Liu, Y., Yan, H., He, Q., He, Y., and Guan, T. Fourier ViT: A multi-scale ViT with FT for histopathological image classification. In CACRE-2022, IEEE.
5. Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
6. Iddan, G., Meron, G., Glukhovsky, A., and Swain, P. (2000). Wireless capsule endoscopy. Nature, 405(6785), 417.
7. Nair, V., Chatterjee, M., Tavakoli, N., Namin, A. S., and Snoeyink, C. (2020). Optimizing CNN using fast Fourier transformation for object recognition. In Proceedings of the 2020 ICMLA, pages 234–239. IEEE.
8. O’Shea, K. (2015). An introduction to convolutional neural networks.
9. Reddy, A. Sai Bharadwaj and Juliet, D. Sujitha. (2019). Transfer learning with ResNet-50 for malaria cell-image classification. In 2019 ICCSP, pages 0945–0949.
10. Rustam, F., Siddique, M. A., Siddiqui, H. U. R., Ullah, S., Mehmood, A., Ashraf, I., and Choi, G. S. (2021). Wireless capsule endoscopy bleeding images classification using CNN based model. IEEE Access, 9, 33675–33688.
11. Thota, Gokaramaiah and Nagaraju, K and Korra, Sathya Babu. (2025). SVDGrad-CAM: Singular Value Decomposition filtered Gradient Weighted Class Activation Map. In ICPR, pages 90–105. Springer.
12. Thota, Gokaramaiah and Nagaraju, K and Korra, Sathya Babu. (2025). Quantitative Assessment of Class Activation Maps: An Empirical Study on Musculoskeletal Disorders. In 9th CVIP. Springer.
13. U¸can, Murat, Kaya, Buket, and Kaya, Mehmet. (2022). Multi-class gastrointestinal images classification using EfficientNet-B0 CNN Model. In 2022 ICDABI, pages 1–5.
14. Wang, Cheng, Chen, Delei, Hao, Lin, Liu, Xuebo, Zeng, Yu, Chen, Jianwei, and Zhang, Guokai. (2019). Pulmonary image classification based on inception-v3 transfer learning model. IEEE Access, 7, 146533–146541. IEEE.
15. Yadav, Salini and Aparna, P. (2024). Performance Comparison of Transformers and Convolutional Neural Networks (CNNs) Based Architecture on Endoscopy Images. In 2024 IEEE CONECCT, pages 1–5. IEEE.