International Journal of Applied Engineering Research ISSN 0973-4562 Volume 21, Number 2 (2026) pp. 148-154

FSBI++: Improving Deepfake Detection with a Hybrid CNN-ViT Model

Neethu James
Postgraduate Student (M.Tech)
Department of Computer Science and Engineering
St. Joseph's College of Engineering and Technology ,Palai
Kerala, India
neethujames87@gmail.com
*Corresponding author

Vimal Babu P
Assistant Professor
Department of Computer Science and Engineering
St. Joseph's College of Engineering and Technology ,Palai
Kerala, India
vimalbabu@sjcetpalai.ac.in

***Abstract*—Deepfakes are an evolving and persistent threat, creating a growing need for reliable detection methods. In response, this paper introduces Frequency Enhanced Self-Blended Images++ (FSBI++), a deepfake detection framework that integrates multi-frequency analysis with Self-Blended Images (SBI) augmentation and a hybrid convolutional neural network (CNN)–vision transformer (ViT) architecture. The proposed model utilizes a parallel dual branch feature extraction strategy, consisting of an EfficientNet-B5 CNN and a ViT, whose representations are fused and processed through a multi-layer perceptron (MLP) classifier. The framework captures both spatial and frequency domain features, enabling improved detection capability. Experimental results on the FaceForensics++ dataset achieve an accuracy of 96.58% and an AUC of 98.71%. The findings suggest that combining spatial, frequency, and hybrid learning paradigms strengthens adaptability and generalization in deepfake detection.**

***Index Terms*—Frequency Enhanced Self-Blended Images++ (FSBI++), Self-Blended Images (SBI), Multi-Frequency Feature Generator (MFG), Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT), Convolutional Neural Network (CNN), Vision Transformer (ViT), Multi-Layer Perceptron (MLP), Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Multi-task Cascaded Convolutional Networks (MTCNN), Area Under the Curve (AUC), Receiver Operating Characteristic (ROC), Area Under the Precision Recall Curve (PR-AUC).**

I. INTRODUCTION

Deepfakes have emerged as a significant threat to individual privacy, public security, and democratic systems. They can generate fake images and videos that are imperceptible from authentic ones. This technique involves altering a target individual's facial region by replacing it with another person's identity. It is achieved by the integration of synthesized facial components into the original image. Accordingly, systems capable of automatically evaluating the authenticity of digital visual media are increasingly important. With rapid progress in computer vision and deep learning, deepfake techniques have grown increasingly sophisticated. Hence, demand for more resilient and effective detection methods has intensified. An ideal deepfake detection system can distinguish highly realistic manipulated or synthetic content from authentic content. Deep learning models have proven successful in automatically extracting discriminative features for deepfake detection, avoiding reliance on hand-crafted features [1][2].

Data-driven methods rely on neural networks to automatically extract features from labeled datasets. These models are then used to categorize inputs as either authentic or manipulated. However, they typically require substantial amounts of annotated data for training. Moreover, their ability to handle unseen manipulation techniques is often limited, which reduces their effectiveness in practical applications.

These challenges motivate the proposed FSBI++, a deepfake detection framework that combines a hybrid CNN–ViT architecture with multi-frequency feature analysis designed to increase generalization and detection accuracy.

The main innovations of this work are as follows:

- **FSBI++ framework:** A new deepfake detection framework developed to deliver high accuracy.
- **Self-Blended Images (SBI) augmentation:** Improves generalization by simulating realistic unseen manipulations
- **Multi-frequency feature generator:** incorporates Discrete Wavelet Transform (DWT) and Discrete Cosine Transform (DCT) to capture subtle artifacts across frequency bands.
- **Hybrid CNN-ViT model:** Learns both fine grained local features and global contextual dependencies for comprehensive representation.

II. RELATED WORKS

Recent deepfake detection approaches focus on advancing generalization by combining spatial, frequency, and contextual representations. The FSBI++ framework improves upon the FSBI model described by Hasanaath et al. [3]. FSBI integrates Self-Blended Images (SBI) augmentation with frequency-aware learning (DWT) using a CNN backbone. SBI, presented by Shiohara et al., creates pseudo forgeries through image blending, allowing models to learn more manipulation artifacts [4].

Frequency domain analysis has proven useful in revealing subtle inconsistencies introduced during deepfake generation

and compression. Younus et al. showed that frequency based features can achieve high detection performance with lower computational cost. DWT separates an image into different frequency sub-bands, making hidden artifacts more detectable beyond the spatial domain [5]. Moreover, Pontorno et al. observed that the Discrete Cosine Transform (DCT) captures frequency domain patterns that can act as distinctive signatures of manipulated content [6]. More recently, data-driven models based on modern neural network architectures, including CNN and ViT architectures, have shown success in deepfake detection tasks.

A. Deepfake detection with CNN

Convolutional Neural Networks (CNNs) are widely used for their ability to learn hierarchical spatial features from data, often outperforming traditional methods [7]. In this line of work, Ilhan et al. employed NASNetLarge for feature extraction and classification based on facial regions, although the method relies only on spatial features [8]. Taking a different approach, Nguyen et al. introduced Capsule Networks that retain spatial relationships using capsule layers and dynamic routing, facilitating geometry-aware detection [9]. Earlier, Afchar et al. proposed MesoNet, which focuses on mesoscopic properties of an image [10].

To incorporate temporal information, Tipper et al. and Al-Dulaimi et al. explored CNN–LSTM frameworks, where CNN extracts spatial features while LSTM captures temporal inconsistencies across frames [11][12]. Although these approaches introduce temporal modeling, most CNN based methods still place greater emphasis on spatial information. This can limit their ability to capture global dependencies and subtle frequency artifacts.

B. Deepfake detection with CNN and ViT

Vision Transformers (ViT) use self-attention to model global relationships and complement CNN based local feature extraction. Hybrid CNN–ViT architectures can capture both spatial detail and contextual information [13].

Several works explore this idea in different ways. For instance, Wodajo et al. proposed a Convolutional Vision Transformer (CViT), where CNN features are passed to a transformer to identify global inconsistencies [14]. Wang et al. followed a related direction by designing a sequential hybrid model that augments ViT with convolutional pooling and a re-attention mechanism [15] [13]. Other studies have looked at parallel designs instead [13]. Coccomini et al., for example, introduced EfficientViT and Convolutional Cross-ViT, both of which combine convolutional feature extractors with transformers [16]. Zhao et al. extended this further by developing a spatiotemporal framework. It incorporated temporal modeling alongside CNN based spatial features and ViT based attention to capture dynamic inconsistencies [17].

Despite these advances, many approaches still tend to focus on spatial, frequency, or contextual information in isolation. The proposed FSBI++ framework addresses this gap by bringing these aspects together, integrating SBI augmentation, DWT and DCT based multi frequency feature extraction, and a parallel hybrid CNN–ViT architecture. By doing so, FSBI++ is able to learn complementary spatial, frequency, and global contextual features, leading to effective deepfake detection.

III. FSBI++ FRAMEWORK

The proposed FSBI++ deepfake detection framework improves upon the FSBI framework [3] by integrating three modules: Self-Blended Images (SBI), a Multi-Frequency Feature Generator (MFG), and a hybrid CNN–ViT model. The framework is built upon a parallel feature extraction strategy that brings together spatial and frequency domain representations to successfully detect deepfake manipulations. The overall pipeline of the FSBI++ framework, including data preprocessing, feature extraction, and classification stages, is presented in Fig. 1.

The deepfake detection pipeline begins with video preprocessing. Frames are extracted from video sequences to construct the dataset. Frames are extracted at fixed intervals (every ten frames) to reduce redundancy while preserving temporal diversity. Each frame is resized to a fixed resolution of 224 × 224 pixels. This ensures compatibility with the hybrid model and standardizes input dimensions. To avoid dataset imbalance caused by longer videos, a maximum of thirty frames per video is retained.

Face detection on the extracted frames is performed using Multi-Task Cascaded Convolutional Networks (MTCNN). MTCNN identifies facial bounding boxes along with key landmarks, including the eyes, nose, and mouth. If MTCNN fails to detect a face, an OpenCV Haar Cascade detector is used as a fallback mechanism. Detected faces are cropped with a margin around the bounding box and aligned using eye landmarks before being resized to 224 × 224 resolution. This process normalizes pose variations and reduces background bias in the dataset [18]. The FSBI++ framework consists of the following modules.

A. SELF-BLENDED IMAGES (SBI)

A Self-Blended Image is generated by blending an original image with its transformed version using a mask. Self-Blended Images are used as a self supervised augmentation strategy to improve model generalization [4]. So, in this case, synthetic forgeries are generated by blending an original face image with its transformed version. A transformed version of the original face image is generated through a series of stochastic transformations, including horizontal flipping, affine geometric warping, brightness and contrast adjustment, Gaussian blurring, color perturbations in the hue and saturation channels, and the introduction of JPEG compression artifacts. These transformations mimic distortions commonly observed in real manipulated image.

To create the synthetic sample, a smooth elliptical mask is generated that defines the blending region between the original image and its transformed version.

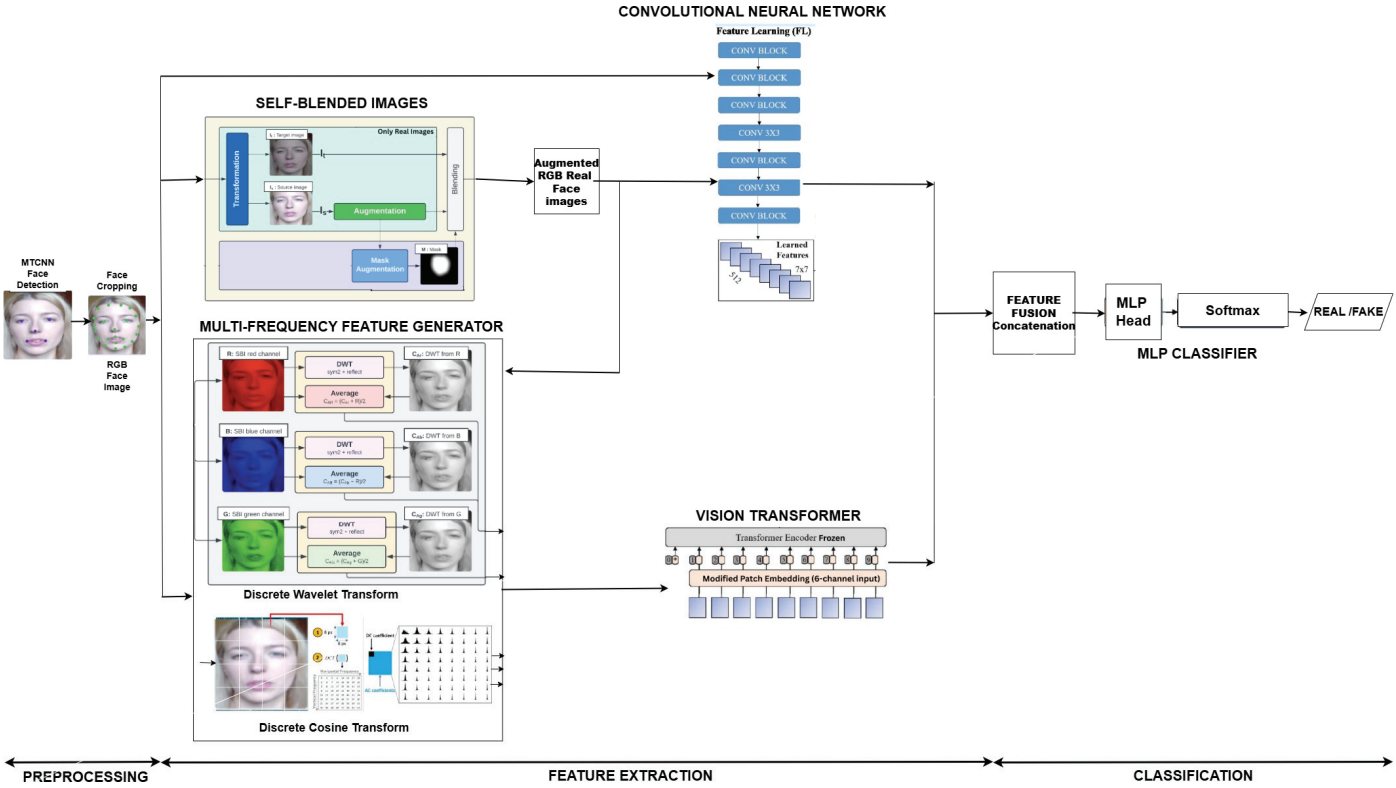

Fig. 1: FSBI++ Framework Architecture

Mathematically it is represented as:

$$I_{SBI} = I_s \odot M + I_t \odot (1 - M) \quad (1)$$

where, I_{SBI} is the generated self-blended image, I_t is the original face image, I_s is the transformed version of the image, M: is the elliptical blending mask [3].

Self-Blended Images (SBI) simulate common spatial artifacts observed in deepfakes, including blending boundary artifacts, facial landmark misalignment, and color or illumination inconsistencies, to help the model learn manipulation patterns during training. SBI augmented real face images are used as input for both the CNN and the Multi-Frequency Feature Generator (MFG) incorporating DWT and DCT.

B. MULTI-FREQUENCY FEATURE GENERATOR(MFG).

Deepfake generation and compression often introduce subtle artifacts in the frequency domain. To identify these inconsistencies, a Multi-Frequency Feature Generator (MFG) is constructed using both DWT and DCT.

DWT separates the cropped input face image into multiple frequency components (sub-bands). The low frequency component retains the overall structure, while the LH and HL components capture directional and texture related variations. The HH component is discarded. DWT, based on the Haar wavelet, provides a simple way to separate low and high frequency information [5].

Simultaneously, DCT is performed by dividing the image into non-overlapping 8 × 8 blocks and converting each block into frequency coefficients. Low and mid frequency components represent overall appearance and texture, whereas high frequency components highlight fine grained details, including noise and compression artifacts [6].

Finally, the outputs from DWT and DCT, each of size 224 × 224 × 3, are concatenated to form a 224 × 224 × 6 multi-frequency feature map.

$$MFG = \text{Concat}(DWT_{LL}, DWT_{LH}, DWT_{HL}, DCT_{Low}, DCT_{Mid}, DCT_{High}) \quad (2)$$

This combined feature map highlights both local and global spectral inconsistencies and is provided as input to the Vision Transformer for learning frequency domain relationships.

C. HYBRID CNN-ViT MODEL

Deepfake manipulations are detected using a hybrid model that combines a convolutional neural network, a vision transformer, and a Multi-Layer Perceptron (MLP) classifier. The hybrid architecture incorporates spatial and frequency domain features. This module performs parallel feature extraction by processing two types of inputs. These include spatial domain representations from the SBI module and frequency domain representations from the MFG module.

The spatial branch of the parallel feature extraction strategy employs EfficientNet-B5 as the convolutional

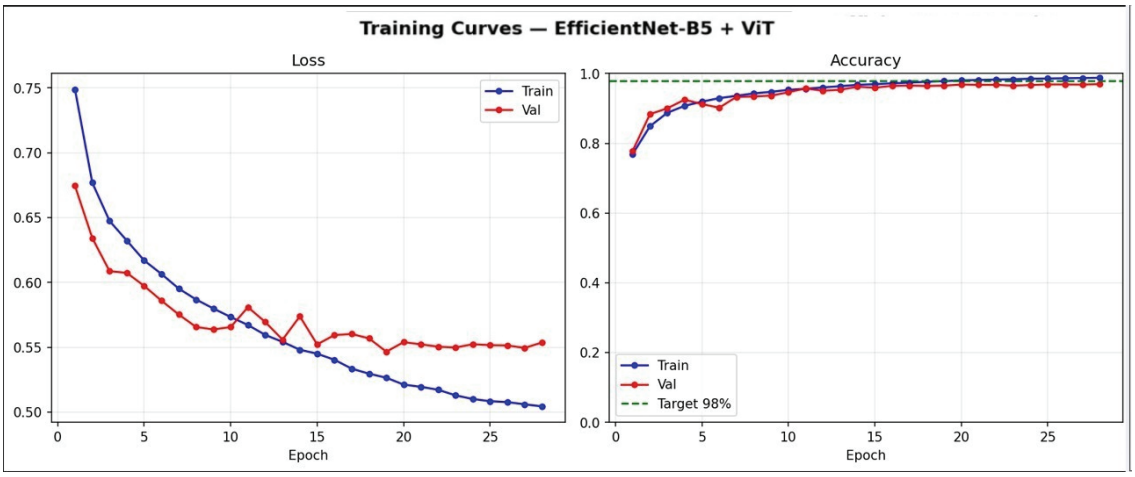

Fig. 2: Training and validation curves of the FSBI++ framework show loss reduction and accuracy improvement, indicating stable convergence.

backbone to process both fake and SBI augmented images. It captures high level semantic features from RGB inputs. The model utilizes ImageNet pre-trained weights for initialization and is then fully fine-tuned. This branch outputs a 2048 dimensional feature vector and is trained using a learning rate of 1e-5.

Parallel to this, the frequency domain ViT branch employs a Vision Transformer (ViT) to process six channel MFG inputs. The patch embedding layer is modified from its standard three channel configuration to accept six input channels. Transformer encoder layers remain frozen to reduce overfitting. This branch yields a 768 dimensional feature vector, trained with a learning rate of 1e-4.

The features from both branches are fused via concatenation, resulting in a 2816 dimensional representation. This representation is then fed into a four layer MLP classifier (2816 → 1024 → 512 → 128 → 2) with Batch Normalization, GELU activation, and dropout for improved generalization. The final Softmax layer outputs probabilities for real and fake classes.

The CNN based spatial branch extracts high level semantic features, including facial geometry, skin texture, lighting inconsistencies, blending artifacts, and identity-related patterns. In contrast, the ViT based frequency branch captures global frequency relationships, artifact distribution, and long range spectral (frequency) dependencies.

IV. EXPERIMENTAL RESULTS

A. TRAINING

The image dataset is constructed from facial regions extracted from the FaceForensics++ C23 dataset (FF++). FF++ is a widely used benchmark containing real and manipulated videos generated using multiple techniques and compression levels [19]. All 7000 video samples from the FF++ dataset are utilized in FSBI++. The extracted faces in the image dataset are augmented using Self-Blended Images (SBI) and further processed through a Multi-Frequency Generator (MFG) based on DWT and DCT. The dataset is partitioned into training, validation, and test sets using an 80:10:10 ratio to support model training.

The hybrid CNN–ViT model is trained using transfer learning, fine-tuning, and advanced optimization strategies to achieve a target accuracy of 98% while ensuring convergence. Transfer learning is applied to both the CNN and ViT components. Different learning rates are assigned across model components, with lower rates for the CNN backbone and higher rates for the ViT patch embedding and classification head.

The model is trained on input images of size 224 × 224 × 3. Training is conducted for 30 epochs with a batch size of 16. The best performance is observed at epoch 27. Optimization is performed using the AdamW optimizer. A cosine annealing learning rate scheduler is used to progressively reduce the learning rate to 1e-6. Carefully selected hyperparameters are employed during training, including learning rate selection, weight decay, optimizer configuration, learning rate scheduling, gradient clipping, and mixed-precision training.

Class imbalance is addressed using a weighted cross-entropy loss, with higher emphasis on real samples (7.75) than fake samples (1.00). Regularization techniques such as label smoothing (0.1), gradient clipping (1.0), batch normalization, dropout, and mixed-precision training are further applied to improve generalization and training stability.

Model checkpointing is implemented to retain the best performing model based on validation performance. Early stopping is triggered if the validation AUC does not exhibit improvement over seven consecutive epochs.

Figure 2 illustrates the training dynamics of the proposed hybrid model. Both training and validation losses exhibit a consistent downward trend, decreasing from approximately

0.75 to 0.50 and 0.67 to 0.55, respectively. Minor fluctuations in validation loss are observed, likely due to regularization techniques such as batch normalization and dropout.

Similarly, training and validation accuracies improve steadily, reaching approximately 99% and 98%, respectively. Thus, the model's learning dynamics are reflected in the loss and accuracy curves, which demonstrate consistent convergence.

B. TESTING

The proposed EfficientNet-B5 and Vision Transformer hybrid model demonstrated strong performance on the test dataset. It achieved an accuracy of 96.58% along with an AUC of 98.71%. As presented in Fig. 3, the model is further assessed using standard performance metrics, including precision, recall, and F1-score. The results are 98.82%, 97.31%, and 98.06%, respectively.

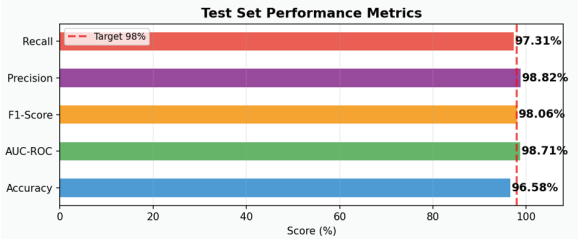

Fig. 3: Evaluation metrics of the FSBI++ model, including precision, recall, and F1-score

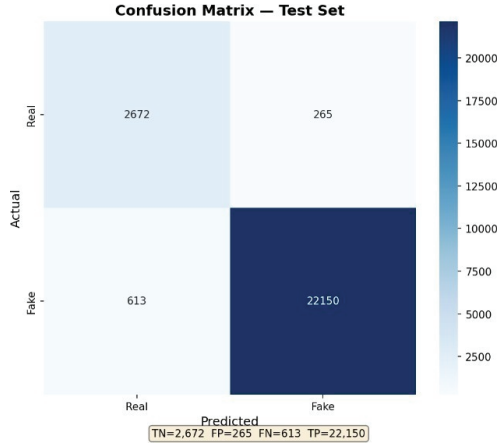

Fig. 4: Confusion matrix illustrating accurate classification of real and fake samples by the FSBI++ model.

The confusion matrix in Fig. 4 provides additional insight into the model's classification behavior. Out of 2937 real frames extracted from videos, 2672 are correctly identified, while 22150 out of 22763 fake frames are accurately classified. This highlights the model's strong ability to detect fake samples, especially within the majority class.

Class	Precision	Recall	F1-Score
Real	0.8134	0.9098	0.8589
Fake	0.9882	0.9731	0.9806
Macro Avg	0.9008	0.9415	0.9198

TABLE 1: Per-class performance metrics of the model.

However, as shown in Table 1, the precision for the real class is relatively lower (0.8134). This is mainly due to the imbalance in the dataset, where fake samples account for nearly 88% of the total data. Because of this, the model tends to favor the majority class during training. This causes the decision boundary to lean more toward fake predictions.

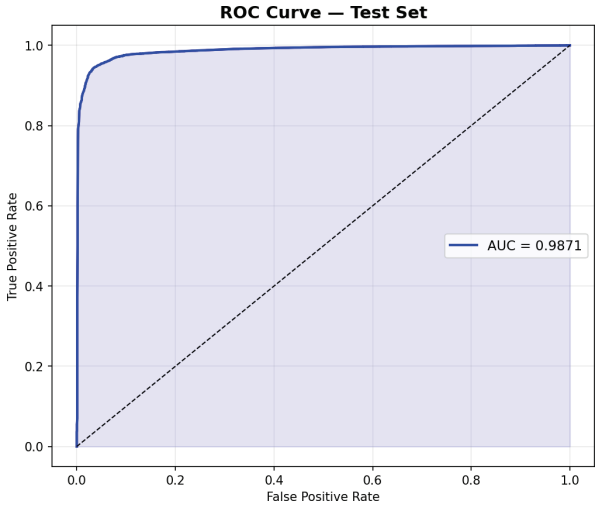

Fig. 5: ROC curve indicating class separability.

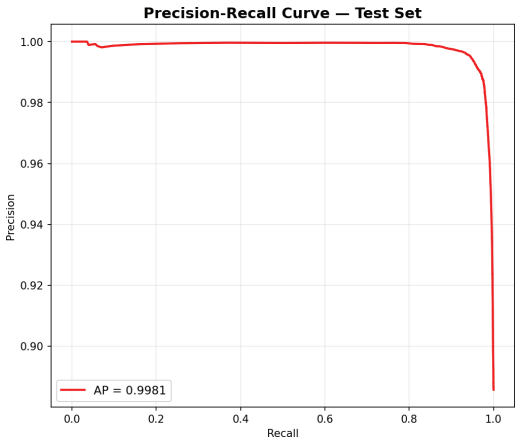

Fig. 6: PR curve showing model behavior under class imbalance.

Fig. 5 shows a high ROC-AUC of 0.9871, suggesting strong class separability. Fig. 6 also supports this with a PR-AUC of 0.9981, which reflects discriminative ability in imbalanced class settings. Consistently high precision across varying recall levels indicates that false positives remain low

even as recall increases. These results suggest that the model constraints are more closely related to threshold selection than to feature representation.

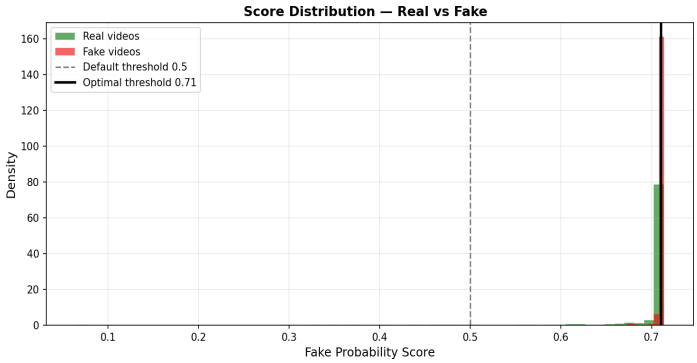

Fig. 7: Distribution of prediction scores for real and fake samples, illustrating the impact of threshold adjustment.

To handle this issue, threshold calibration is applied using the validation dataset. As shown in Fig. 7, the default threshold of 0.5 causes class overlap. When the threshold is adjusted to 0.71, class separation and decision balance become better. This increases real class precision while maintaining strong fake detection performance. Although there is a slight reduction in F1-score due to the precision–recall trade-off, threshold calibration improves overall class balance.

The model outputs predictions as confidence scores that reflect the probability that an input image is manipulated. Scores near 0 indicate real images, while scores near 1 indicate fake images. Using the calibrated threshold of 0.71, scores above it are classified as fake and those below as real.

For cross dataset evaluation, the model is tested on the CelebA dataset and achieves an accuracy of 65%. This result shows that generalizing across different datasets is still challenging. Even with this performance drop, the model behaves in a fairly stable manner, suggesting a reasonable level of robustness [20]. Further improvements are needed to achieve better generalization for deepfake detection under actual conditions.

V. CONCLUSION

The continuous evolution of deepfake generation methods and their widespread accessibility require techniques for distinguishing manipulated content from authentic media. In this work, SBI is utilized to simulate localized manipulation artifacts, while MFG (DWT and DCT) reveals these artifacts in the frequency domain. The hybrid CNN–ViT architecture enables the joint learning of local inconsistencies and global contextual dependencies, increasing deepfake detection performance.

The proposed FSBI++ framework uses both spatial and frequency domain representations along with hybrid learning strategies. Experimental results demonstrate that the approach achieves improved class balance compared to conventional methods. Threshold calibration further strengthens efficacy in diverse scenarios. Overall, FSBI++ offers a constructive and scalable solution for deepfake detection, with potential extensions to cross domain generalization and real time deployment.

VI. FUTURE WORKS

Even though FSBI++ performs well, there is still room for improvement in cross dataset evaluation. Models tend to struggle when faced with unseen datasets because manipulation methods, compression levels, and data distributions continuously vary.

For future work, three aspects need to be explored. The first involves the use of domain adaptation techniques, particularly adversarial training. These methods help learn domain-invariant representations. This raises generalization across datasets. The second focuses on the use of dynamic thresholding instead of a fixed decision threshold of 0.71. This enables predictions to adapt based on input characteristics such as quality and compression, thus increasing reliability in real world scenarios. The third aspect focuses on temporal modeling techniques to capture inconsistencies across frames. Architectures such as Recurrent Neural Networks (RNNs) or 3D Convolutional Neural Networks (3D CNNs) can further refine detection performance. These approaches can augment both the efficiency and the practical use of the proposed framework.

VII. DECLARATIONS

A. DATA AVAILABILITY STATEMENT

The FaceForensics++ and CelebA datasets used in this work are publicly available.FaceForensics++ is available at https://github.com/ondyari/FaceForensics , and CelebA is available at https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html. The data preparation pipeline, including frame extraction (preprocessing), SBI-based data augmentation, and multi-frequency feature extraction using DWT and DCT, is described in the paper. Additional processed data and implementation details can be made available by the authors upon reasonable request.

B. ACKNOWLEDGEMENTS : AI USAGE DISCLOSURE

During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.3) for language editing and grammar correction. The AI tool was not used for generating scientific content, results, or interpretations. All outputs were carefully reviewed, validated, and edited by the authors, who take full responsibility for the final content.

C. CONFLICT OF INTEREST STATEMENT

The authors report there are no competing interests to declare.

D. AUTHORS CONTRIBUTION STATEMENT

Neethu James was involved in the conception and design of the study, selection of the methodology, drafting of the manuscript,preparation of Figures 1–7 and Table 1, and revising it critically for important intellectual content. Vimal Babu P contributed to the analysis and interpretation of the data, supervised the work, and approved the final version of the manuscript. Both authors reviewed the manuscript, approved the final version, and agree to be accountable for all aspects of the work.

E. ETHICAL STATEMENT

This study utilizes publicly available FaceForensics++ dataset. No human subjects were directly involved, and no personal or sensitive data were collected by the authors. The research complies with standard ethical guidelines for the use of publicly available datasets in computer vision and deep learning. The proposed method is intended solely for research and security purposes, particularly for detecting manipulated media and mitigating the misuse of deepfake technologies.

F. FUNDING

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

REFERENCES

[1] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep learning for deepfakes creation and detection: A survey," *IEEE Access*, vol. 7, pp. 107625–107642, 2019.

[2] M. S. Rana and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 8, pp. 188707–188724, 2020.

[3] A. A. Hasanaath, "Fsbi: Deepfakes detection with frequency enhanced self-blended images," in *2024 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pp. 1–6, IEEE, 2024.

[4] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18720–18729, 2022.

[5] M. A. Younus and T. M. Hasan, "Effective and fast DeepFake detection method based on Haar wavelet transform," in *2020 International Conference on Computer Science and Software Engineering (CSASE)*, pp. 186–190, 2020.

[6] O. Giudice, L. Guarnera, and S. Battiato, "On the exploitation of DCT-traces in the generative-AI domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1672–1681, 2021.

[7] A. Richards and S. L. Mary, "Deep fake face detection using convolutional neural networks," in *2021 International Conference on Innovative Computing, Information and Control (ICICIC)*, pp. 1–5, 2021.

[8] I. Ilhan, "An improved DeepFake detection approach with NASNetLarge CNN," *Signal, Image and Video Processing*, pp. 1–9, 2023.

[9] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2307–2311, 2019.

[10] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a compact facial video forgery detection network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7, 2018.

[11] S. Tipper, "An investigation into the utilisation of CNN with LSTM for video deepfake detection," *arXiv preprint arXiv:2305.12345*, 2023.

[12] O. A. H. H. Al-Dulaimi, "A hybrid CNN-LSTM approach for precision deepfake image detection based on transfer learning," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 5, 2023.

[13] Z. Wang, Z. Guo, and J. Huang, "A timely survey on vision transformer for deepfake detection," *arXiv preprint arXiv:2108.07112*, 2021.

[14] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv preprint arXiv:2102.11126*, 2021.

[15] T. Wang *et al.*, "Deep convolutional pooling transformer for deepfake detection," *Information Sciences*, vol. 642, p. 119154, 2023.

[16] D. Coccomini, N. Messina, G. Gennaro, and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection," in *Proceedings of the 15th International Conference on Distributed Smart Cameras*, pp. 1–8, 2021.

[17] X. Zhao, Y. Wang, and S. Lyu, "Exploring complementarity of global and local spatiotemporal information for fake face video detection," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2543–2547, 2021.

[18] N. Zhang, J. Luo, and W. Gao, "Research on face detection technology based on mtcnn," in *2020 International Conference on Computer Science and Software Engineering (CSASE)*, pp. 154–158, 2020.

[19] A. Ro¨ssler, D. Cozzolino, V. Luisa, C. Riess, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11, 2019.

[20] A. V. Nadimpalli and A. Rattani, "On improving cross-dataset generalization of deepfake detectors," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 262–271, 2023.