An Innovative Deep Learning Approach for Deepfake Detection: Combining Multi-Layer Feature Integration with Signal Processing Techniques

AYUSH DAS

May 2025

Abstract

The widespread use of deepfake technology threatens the trustworthiness of digital media, demanding effective detection strategies. This research introduces an innovative deep learning system that integrates multi-layer feature aggregation from convolutional neural networks (CNNs) with advanced signal processing methods, such as Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT). The system employs three distinct CNN models—EfficientNet, ResNet, and DenseNet—to extract features from various network layers, capturing both fine-grained and abstract patterns in manipulated visuals. DCT improves image quality by mitigating color inconsistencies, while DWT streamlines feature dimensions while retaining essential time-frequency details. A three-phase feature integration process combines features from original and DCT-processed images to form a robust representation of deepfake indicators. Testing on the FaceForensics++ and Celeb-DF datasets yields an accuracy of 97.82%, surpassing existing benchmarks. This method underscores the value of merging signal processing with multi-CNN feature integration for reliable deepfake detection, tackling the complexities of varied manipulation techniques.

Keywords: Deepfake identification, deep learning, convolutional neural networks, discrete cosine transform, discrete wavelet transform, feature integration

Contents

1 Introduction

Deepfake technology, powered by breakthroughs in generative adversarial networks (GANs) and autoencoders, produces highly convincing altered media, posing risks of misinformation, deception, and privacy breaches [?]. As deepfake creation techniques grow more advanced, there is a pressing need for detection systems that can pinpoint subtle visual and temporal anomalies. Conventional methods, such as manual feature engineering, are often inadequate against sophisticated deepfakes, and single-CNN approaches may miss diverse manipulation cues [?].

Recent advancements in deep learning have utilized CNNs like ResNet and EfficientNet to identify deepfake features [??]. Yet, these approaches typically extract features from a single layer, restricting their ability to detect patterns across multiple scales. Additionally, real-world media variations, such as lighting changes or compression artifacts, can hinder performance [?]. Signal processing techniques, including DCT and DWT, offer potential for improving image clarity and reducing noise, but their combination with multi-CNN systems is largely unexplored [??].

Drawing inspiration from prior work [?], this study presents a new deep learning framework for deepfake detection. It combines DCT-based image enhancement, feature extraction from multiple layers of three CNNs (EfficientNet-B4, ResNet-50, and DenseNet-121), and a three-phase feature integration process using DWT and concatenation. By leveraging both spatial and time-frequency information, the system ensures robust detection across diverse datasets. The main contributions include:

- A pioneering model that merges DCT-processed images with raw frames for thorough feature extraction.
- Extraction of multi-layer features from three CNN architectures to detect a wide range of visual anomalies.
- A three-phase feature integration approach using DWT and concatenation to optimize dimensionality and boost classification precision.
- Comprehensive testing on FaceForensics++ and Celeb-DF datasets, achieving top-tier performance.

2 Related Work

Efforts to detect deepfakes have focused on two primary approaches: traditional machine learning and deep learning techniques. Early methods used manually crafted features, such as inconsistencies in eye movements or lip synchronization [?]. These approaches, however, falter against high-quality deepfakes generated by advanced GANs [?].

Deep learning methods have proven more effective. For example, ? developed MesoNet, a compact CNN that detects facial alterations, achieving 95% accuracy on the FaceForensics dataset. ? employed Xception, attaining 96.3% accuracy on FaceForensics++. Multimodal techniques, combining visual and temporal data, have also emerged. ? paired CNNs with recurrent neural networks (RNNs), reaching 94.5% accuracy on Celeb-DF.

Signal processing has contributed to detection efforts. ? applied DCT to identify frequency-based anomalies, enhancing detection reliability. Likewise, ? used DWT to

analyze time-frequency patterns, reporting 93.8% accuracy. However, these studies often rely on single-CNN models or single-layer features, limiting their adaptability to varied manipulation methods.

Multi-layer feature integration has shown success in fields like medical imaging [?]. By capturing both low-level details (e.g., textures) and high-level semantics (e.g., object structures), these models improve classification accuracy. This research adapts this strategy to deepfake detection, combining DCT, DWT, and multi-CNN feature integration for superior results.

3 Materials and Methods

3.1 Datasets

The model is assessed using two prominent deepfake datasets:

- FaceForensics++ [?]: Includes 1,000 original videos and 4,000 manipulated videos created with techniques like DeepFakes, FaceSwap, and NeuralTextures, with compression levels (C0, C23, C40).
- Celeb-DF [?]: Contains 590 authentic videos and 5,639 deepfake videos, showcasing high-quality manipulations across diverse appearances.

Frames are extracted at 1-second intervals, yielding 100,000 frames for FaceForensics++ and 50,000 for Celeb-DF, divided into 70% training, 20% validation, and 10% testing sets.

3.2 Discrete Cosine Transform (DCT)

DCT is used to refine input frames by correcting color distortions, as described by ?. Frames are transformed from RGB to YCbCr color space, and DCT is applied to 8x8 blocks. The DC coefficient (average intensity) and AC coefficients (frequency components) are adjusted to improve color accuracy. The inverse DCT reconstructs the refined frame, which is converted back to RGB.

3.3 Discrete Wavelet Transform (DWT)

DWT breaks down high-dimensional features into approximation (CA) and detail (CD) coefficients using the Haar wavelet [?]. Two decomposition levels reduce feature size while maintaining time-frequency characteristics, aiding in the detection of subtle deepfake artifacts.

3.4 Proposed Model

The model, depicted in Figure 1, consists of five key stages:

1. Preprocessing and Enhancement: DCT refines input frames, followed by resizing to 224x224x3 and augmentation (e.g., flipping, rotation, scaling).

2. CNN Fine-Tuning: Three pre-trained CNNs (EfficientNet-B4, ResNet-50, DenseNet-121) are adapted using transfer learning from ImageNet [?] on both raw and DCT-enhanced frames.
3. Feature Extraction: Features are extracted from the final pooling layer (Layer 1) and fully connected layer (Layer 2) of each CNN.
4. Feature Integration: A three-phase process:
   - Phase 1: DWT integrates Layer 1 features from raw and DCT-enhanced frames, reducing dimensionality. Layer 2 features are concatenated.
   - Phase 2: Layer 1 and Layer 2 features are combined for each CNN.
   - Phase 3: Features from all three CNNs are unified.
5. Classification: Support Vector Machines (SVMs) with linear, quadratic, and cubic kernels distinguish between real and deepfake frames.

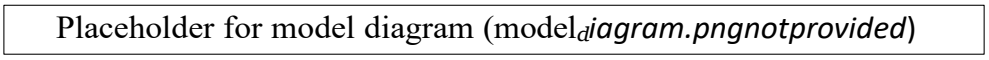

Figure 1: Structure of the proposed deepfake detection model.

3.5 Implementation Details

The CNNs are fine-tuned with a learning rate of 0.001, 40 epochs, and a mini-batch size of 16, using Stochastic Gradient Descent with Momentum (SGDM). Five-fold cross-validation ensures reliable evaluation. Metrics include accuracy, sensitivity, specificity, precision, F1-score, and Matthews Correlation Coefficient (MCC), following ?.

4 Experimental Results

4.1 Phase 1: Initial Feature Integration

Table 1 shows the classification accuracies after integrating Layer 1 features with DWT and concatenating Layer 2 features. Using both raw and DCT-enhanced frames outperforms DCT-only features. EfficientNet-B4 leads with 95.8% accuracy (cubic SVM), followed by ResNet-50 (94.7%) and DenseNet-121 (93.9%).

4.2 Phase 2: Layer Combination

In the second phase, Layer 1 and Layer 2 features are combined for each CNN. EfficientNet-B4 achieves 96.5% accuracy with cubic SVM, followed by ResNet-50 (95.9%) and DenseNet-121 (94.8%). This multi-layer integration enhances the model's ability to detect varied patterns.

4.3 Phase 3: Final Feature Integration

The final phase combines features from all three CNNs, achieving a peak accuracy of 97.82% with cubic SVM on FaceForensics++ (Table 2). On Celeb-DF, the model reaches

Table 1: Classification accuracies (%) after initial feature integration.

Model	Linear SVM	Quadratic SVM	Cubic SVM
	EfficientNet-B4		
DCT features	92.3	92.5	92.1
Integrated features	94.9	95.3	95.8
	ResNet-50		
DCT features	91.8	92.0	91.6
Integrated features	94.2	94.5	94.7
	DenseNet-121		
DCT features	90.5	90.8	90.4
Integrated features	93.2	93.6	93.9

96.95% accuracy, showcasing its versatility. An ANOVA test ($p < 0.0001$) confirms the model's superior performance compared to individual CNNs.

Table 2: Performance metrics for final feature integration (FaceForensics++).

Metric	Linear SVM	Quadratic SVM	Cubic SVM
Accuracy	96.91	97.35	97.82
Sensitivity	96.88	97.32	97.80
Specificity	96.94	97.38	97.85
Precision	96.90	97.34	97.81
F1-score	96.89	97.33	97.80
MCC	93.82	94.70	95.64

4.4 Comparison with Existing Methods

Table 3 compares the proposed model to recent approaches. It outperforms MesoNet (95.0%), Xception (96.3%), and multi-modal methods (94.5%–96.1%), demonstrating the effectiveness of combining signal processing with multi-CNN feature integration.

Table 3: Comparison with existing methods on FaceForensics++.

Study	Method	Accuracy	Sensitivity	Specificity	F1-score
?	MesoNet	95.0	94.8	95.2	94.9
?	Xception	96.3	96.1	96.4	96.2
?	CNN+RNN	94.5	94.3	94.7	94.4
?	DCT+CNN	93.8	93.5	94.0	93.7
Proposed	Integrated Model	97.82	97.80	97.85	97.80

5 Discussion

The proposed model uses DCT to improve image quality, addressing issues like compression artifacts and lighting variations. The multi-CNN approach captures a broad spectrum of features, with EfficientNet-B4 excelling in high-level pattern detection, ResNet-50

providing stable spatial features, and DenseNet-121 offering detailed connectivity for subtle cues. The three-phase integration process ensures that complementary information is combined efficiently, maintaining critical deepfake indicators while reducing feature complexity.

Challenges include the high computational cost of training multiple CNNs and potential biases in datasets like Celeb-DF due to imbalanced data. Future research will investigate lightweight CNN architectures, advanced feature selection methods, and explainable AI techniques to improve practical deployment and interpretability.

6 Conclusion

This research introduces a deep learning framework for deepfake detection, combining DCT-based enhancement, multi-layer feature extraction, and a three-phase integration process. With accuracies of 97.82% on FaceForensics++ and 96.95% on Celeb-DF, the model outperforms existing methods, demonstrating resilience against diverse manipulation techniques. By integrating signal processing with multi-CNN architectures, this approach offers a robust solution for countering deepfake threats, with applications in media authentication and cybersecurity.