

Hybrid Spatial–Temporal & Frequency Deepfake Detector

MS Meenu, Mohd Faizanul Haque

Dept. of AI & DS, Dr. Akhilesh Das Gupta Institute of Professional Studies

Dept. of AI & DS, Dr. Akhilesh Das Gupta Institute of Professional Studies

Abstract

Recent developments in deep learning have enabled the creation of highly realistic synthetic videos, often referred to as deepfakes. These videos can manipulate a person's facial features, expressions, or even their voice in a manner that appears convincing to human viewers. Because of this, they pose significant challenges in areas such as journalism, cybersecurity, and digital forensics, where visual media is generally treated as trustworthy evidence.

This project presents a hybrid approach for detecting deepfakes by combining spatial, temporal, and frequency based analysis within a unified framework. Rather than depending on a single type of feature, the system draws on multiple aspects of video data. Visual details at the frame level are captured using EfficientNet, temporal dependencies across frames are learned through a Transformer encoder, and frequency-related patterns are analyzed using the Fast Fourier Transform (FFT). These feature representations are then merged to produce the final prediction.

The model was trained and evaluated on a Kaggle dataset consisting of 731 videos, with nearly equal numbers of real and manipulated samples. On a validation subset of 147 videos, the system reached an accuracy of approximately 92%, while maintaining balanced precision, recall, and F1 scores. When compared with simpler baseline models, the results indicate that integrating multiple feature types improves overall performance as well as robustness.

Keywords

Deepfake detection; Hybrid spatial temporal model, Frequency domain analysis, FFT-based spectral features, Transformer-based temporal modeling, EfficientNet encoding, Video forensics, Deep learning.

Abbreviations

- AI – Artificial Intelligence
- CNN – Convolutional Neural Network

- FFT – Fast Fourier Transform
- LSTM – Long Short-Term Memory
- UI – User Interface
- GANs – generative adversarial networks

1. Introduction

Deepfakes refer to digitally altered or fully synthetic media created using deep learning techniques, most commonly generative adversarial networks (GANs) and similar models. These systems can modify a person's facial appearance, expressions, or voice while maintaining a high level of visual and audio realism.

The increasing availability of deepfake generation tools has significantly reduced the technical expertise required to produce convincing fake videos. This growing accessibility has raised concerns related to identity verification, trust in online content, and the overall reliability of information within digital environments.

Such highly realistic manipulated videos can be exploited for various harmful purposes, including spreading misinformation, influencing political narratives, committing financial fraud, damaging reputations, and enabling targeted harassment. As a result, developing dependable deepfake detection methods has become an important challenge in both research and practical applications.

Earlier techniques in media forensics largely depended on handcrafted features or easily noticeable inconsistencies, such as irregular blending, color differences, or compression artifacts. However, modern deepfake generation methods produce content with minimal visible defects, making manual inspection insufficient. Detecting these forgeries now requires automated, data driven approaches capable of identifying subtle patterns across spatial, temporal, and frequency domains that are not easily perceived by humans.

Current detection systems increasingly rely on deep neural networks to examine fine-grained pixel details, motion consistency between frames, alignment between

audio and video signals, and frequency based traces introduced during synthesis or re-encoding processes.

In response to these challenges, this project develops a hybrid deepfake detection system designed specifically for video-based facial manipulation. The approach processes a video as a sequence of frames, extracts spatial features using EfficientNet, captures temporal dependencies through a Transformer-based attention mechanism, and incorporates frequency-domain information derived from Fast Fourier Transform (FFT). These combined features are then used to classify the content as real or fake.

In addition to evaluating performance metrics, the study also compares the proposed model with simpler baseline approaches, highlighting the advantages of integrating multiple feature representations for improved accuracy and robustness.

1.1 Challenges

Deepfake detection involves a number of technical and practical challenges that make it difficult to build systems suitable for real-world use. These difficulties largely arise from the increasing realism of synthetic media, limitations in generalizing across different datasets, and the need to maintain efficiency during deployment.

1. **High visual realism:** Recent deepfake generation models, trained on large-scale face and video datasets, can produce high resolution outputs with convincing lighting conditions, detailed skin textures, and natural looking facial expressions. Because of this, there are very few obvious visual signs that can be easily identified by humans or traditional forensic approaches.
2. **Subtle and multi-level artifacts:** The remaining imperfections are often very minor, appearing as slight inconsistencies in motion, small variations in texture distribution, or compression related irregularities. Detecting such artifacts requires models that can capture detailed patterns across multiple levels and feature domains.
3. **Generalization and dataset shift:** Many deepfake detection models perform well on the data they are trained on but show reduced accuracy when applied to unseen datasets. Differences in generation techniques, compression settings, or real world social media content often expose a lack of generalization, suggesting that models may rely too heavily on specific learned patterns.
4. **Rapid evolution of generation techniques:** The field of deepfake generation evolves rapidly, often in

response to advances in detection. As newer models are designed to reduce detectable artifacts, detection systems must also adapt continuously, creating a dynamic and competitive cycle.

5. **Complexity of video data:** Detecting deepfakes in videos is inherently more complex than in images, as it requires analyzing temporal information such as facial movements, head dynamics, and frame-to-frame consistency. This additional dimension increases both modeling difficulty and computational requirements.

6. **Computational constraints:** Models designed to process high resolution videos and extended frame sequences often require significant GPU memory and processing time. This creates challenges when deploying such systems in real-time scenarios or on devices with limited computational resources, such as mobile or edge platforms.

To address these challenges, the proposed hybrid model integrates multiple complementary feature representations within a unified architecture. The design aims to capture diverse characteristics of manipulated content while remaining feasible to train on moderately resourced hardware through optimized preprocessing and batching strategies.

1.2 Need for Hybrid Spatial–Temporal & Frequency Deepfake Detector

Detection models that rely on a single modality whether spatial appearance or temporal dynamics often struggle in real-world scenarios. In many cases, deepfakes are designed in such a way that artifacts are minimized in one domain while remaining detectable in another, which reduces the effectiveness of single-focus approaches.

A hybrid strategy that combines spatial, temporal, and frequency-based analysis offers greater robustness and improves the chances of identifying complex manipulations. By examining multiple aspects of video data simultaneously, the system becomes less dependent on any one type of signal.

Frame-level spatial analysis allows the model to detect inconsistencies in facial geometry, texture details, blending regions, and lighting conditions. These features can reveal localized alterations, even when the overall motion in the video appears smooth and natural.

Temporal modeling across sequences helps uncover irregularities in motion, such as unnatural facial dynamics, identity shifts between frames, or mismatched lip synchronization. These issues are commonly

associated with face-swapping and reenactment techniques.

Frequency-domain analysis, typically performed using Fast Fourier Transform (FFT), highlights patterns that are not easily visible in the spatial domain. This includes periodic artifacts, spectral imbalances, and unusual high-frequency components introduced during generation or compression processes.

By integrating spatial, temporal, and frequency based features, the proposed model is designed to:

- Improve generalization across different deepfake generation techniques and varying compression conditions.
- Enhance sensitivity to subtle inconsistencies without relying too heavily on a single type of artifact.
- Support more reliable deployment in practical settings, where media content may come from diverse sources and undergo multiple transformations.

1.3 Objectives

The primary objectives of this project are outlined below:

- To develop a deepfake detection framework that integrates spatial (frame level), temporal (sequence level), and frequency domain features for classifying videos as real or fake.
- To implement a multi-branch architecture combining EfficientNet based feature extraction, Transformer-based temporal modeling, and FFT-driven spectral analysis using modern deep learning frameworks.
- To build an end to end processing pipeline that includes data loading, frame extraction, preprocessing, model training, and evaluation using a separate validation set.
- To evaluate the performance of the proposed model using standard classification metrics such as accuracy, precision, recall, F1 score, and confusion matrix, with particular attention to maintaining balanced results across both classes.
- To carry out a comparative analysis against baseline approaches, including CNN only, CNN with Transformer, and CNN with FFT models trained on the same dataset, in order to measure the effectiveness of the hybrid architecture.

1.4 Scope of Work

The scope of this work is defined across the following aspects:

- **Media type:** The system is designed specifically for video based deepfakes that involve manipulation of facial identity or expressions. Audio-only deepfakes and

non-facial forms of manipulation are not considered within this study.

- **Dataset:** The experiments use a Kaggle deepfake dataset consisting of 731 RGB facial videos (363 real and 368 fake), arranged into class-specific directories to support supervised binary classification.

- **Modeling focus:** The work emphasizes architectural design, multi-branch feature integration, and performance evaluation. It does not cover areas such as dataset watermarking, content provenance systems, or blockchain based verification methods.

- **Deployment:** A basic prediction pipeline is developed to demonstrate practical feasibility. However, aspects such as large scale deployment, real time streaming integration, and extensive adversarial testing are left for future work.

- **Explainability:** Although interpretability is acknowledged as an important direction for future improvements, advanced explainable AI methods, such as SHAP or CAM based visualizations, are not fully implemented in the current version of the system.

2. Methodology

Deepfake detection has become an important research area as AI-generated videos continue to grow in realism and accessibility. Identifying manipulated facial content is challenging because modern deepfake techniques can imitate natural expressions, movements, and visual details with high accuracy. To address this issue, this study explores a hybrid deepfake detection framework that combines spatial, temporal, and frequency-domain analysis within a unified architecture. The proposed system uses EfficientNet for visual feature extraction, Transformer layers for temporal sequence modeling, and FFT-based spectral analysis to capture hidden artifacts. By integrating multiple feature perspectives, the model aims to improve detection reliability and achieve stronger performance compared to conventional single-branch approaches.

2.1 Dataset and Pre-processing

The dataset used in this project is the “Deep Fake Detection (DFD) Entire Original Dataset” available on Kaggle. It contains a well organized collection of both real and AI-generated facial videos, making it suitable for training and evaluating deepfake detection models. For the purposes of this study, a subset of 731 RGB videos was selected, consisting of 363 real and 368 fake samples. The videos are arranged into separate class-

specific folders, enabling straightforward binary classification. They capture human faces under a variety of conditions, providing diverse input data for the model. The preprocessing pipeline includes several key steps:

- **Video loading and frame extraction:** A fixed number of frames (e.g., 8 frames) are uniformly sampled from each video along the temporal dimension to capture representative motion patterns.
- **Frame resizing and normalization:** Extracted frames are resized to a resolution of 128×128 and normalized to ensure compatibility with both the EfficientNet based spatial branch and the FFT-based frequency branch.
- **Class labeling:** The directory structure of real and fake videos is used to assign ground-truth labels, which supports supervised learning and helps maintain balanced mini-batches during training.
- **Dataset split:** A separate validation set of 147 videos is held out from training to evaluate the model's performance on unseen data and reduce the risk of overfitting.

These preprocessing steps ensure that the video data is consistently prepared for input across the spatial, temporal, and frequency components of the proposed model.

2.2 Model Architecture

The central architecture is a Multi-Branch Spatio-Temporal Transformer with EfficientNet encoding and FFT based spectral feature fusion.

The design integrates three complementary streams: a spatial encoder, a temporal Transformer, and a frequency domain branch, followed by a shared classification head.

2.2.1 Spatial Encoder (Frame Level)

- **Input:** Individual RGB frames of size $128 \times 128 \times 3$.
- **Backbone:** A pre-trained EfficientNet-B0 model serves as the spatial feature extractor, leveraging transfer learning from large-scale image datasets to capture rich visual patterns.
- **Frozen layers:** Most convolutional layers are frozen to reduce training time and prevent overfitting on the relatively small video dataset.
- **Global average pooling:** The encoder output is passed through global average pooling to obtain a compact latent vector per frame that encodes facial appearance, texture, and local structure. This branch establishes a strong visual representation for subsequent temporal modeling.

2.2.2 Temporal Transformer (Sequence Level)

- **Input sequence:** For each video, a sequence of frame level feature vectors (e.g., 8 time steps) generated by the EfficientNet encoder.
- **Positional embeddings:** Learnable positional encodings are added to preserve temporal order information for the Transformer layers.
- **Multi-head self-attention:** Transformer blocks with multiple attention heads model dependencies across frames, capturing both short term and long-range temporal relationships.
- **Residual connections and layer normalization:** These components stabilize training and ensure effective gradient flow through the sequence.
- **Temporal aggregation:** Global average pooling across time converts the sequence output into a single temporal feature vector summarizing motion and dynamics.

This branch enables the detector to recognize unnatural motion, temporal inconsistencies, and identity instability in deepfake videos.

2.2.3 Frequency-Domain (FFT) Branch

- **Input:** The same $128 \times 128 \times 3$ frames used for spatial encoding.
- **2D FFT:** A two-dimensional Fast Fourier Transform is applied to each frame to convert from spatial space to the frequency domain.
- **Magnitude spectrum:** The magnitude of the complex FFT output is computed, revealing periodic structures and spectral characteristics.
- **Log scaling:** A logarithmic transform is applied to compress dynamic range and highlight subtle spectral variations.
- **Dense projection:** The flattened spectral representation is passed through a dense layer to produce a compact frequency feature vector per frame, followed by pooling to generate a video-level spectral feature. Frequency analysis helps reveal generator-specific patterns, compression anomalies, and high-frequency artifacts that may be challenging to model purely in the spatial domain.

2.2.4 Feature Fusion and Classification Head

- **Concatenation:** The pooled temporal feature vector from the Transformer branch is concatenated with the pooled frequency feature vector from the FFT branch.

- Dense layer with ReLU: A fully connected layer with ReLU activation projects the fused features into a lower-dimensional space, encouraging joint learning across modalities.
- Dropout: A dropout layer with probability 0.5 acts as regularization, reducing co-adaptation and overfitting.
- Output layer: A final dense neuron with sigmoid activation outputs a probability score for the fake class, enabling binary classification via a threshold (e.g., 0.5). This multi branch fusion strategy yields a unified representation that harnesses spatial, temporal, and frequency cues for robust deepfake detection.

2.3 Baseline Models

To contextualize the performance of the proposed model, three baseline architectures were implemented and trained on the same dataset:

- CNN Only: A convolutional network that processes each frame independently and aggregates predictions at the video level, focusing solely on spatial information.
 - CNN + Transformer: A model where CNN-based frame features are passed to a Transformer for temporal modeling, combining spatial and temporal cues but without frequency domain analysis.
 - CNN + FFT: A hybrid architecture that merges CNN based spatial features with FFT based spectral features, omitting explicit temporal modeling.
- These baselines reflect common deepfake detection approaches that rely on subsets of the feature domains considered in the proposed multi-branch system.

2.4 Training Configuration

Training was performed as a supervised binary classification task with the following configuration:

- Optimizer: Adam optimizer with a learning rate of 0.0001.
 - Loss function: Binary cross-entropy, appropriate for two-class problems.
 - Epochs: 10 training epochs, monitoring validation accuracy at each epoch.
 - Batch size: 2 video sequences per batch, constrained by GPU memory and the complexity of the hybrid architecture.
 - Validation protocol: A dedicated validation set of 147 videos was used to track out-of-sample performance and inform model selection.
- Regularization strategies such as dropout, partial freezing of the EfficientNet backbone, and conservative

batch sizing contributed to stable convergence and reduced overfitting.

2.5 Evaluation Metrics

Standard binary classification metrics were used to evaluate the model:

- Accuracy: Proportion of correctly classified videos over the total number of validation samples.
 - Precision: Fraction of predicted fake videos that are truly fake, indicating resistance to false positives.
 - Recall (Sensitivity): Fraction of actual fake videos correctly identified, reflecting the model's ability to detect attacks.
 - F1-score: Harmonic mean of precision and recall, balancing Type I and Type II errors.
 - Confusion matrix: A 2×2 table summarizing true positives, false positives, true negatives, and false negatives for the real/fake classes.
- These metrics are widely used in deepfake detection research and in broader biometric security evaluations.

3. Results and Discussion

3.1 Quantitative Results for the Proposed Model

On the 147-sample validation set, the Multi-Branch Spatio Temporal Transformer with EfficientNet encoding and FFT-based spectral feature fusion achieved the following performance:

- Validation accuracy: 0.92
- Precision (class 0 – real): 0.91
- Recall (class 0 – real): 0.93
- F1-score (class 0 – real): 0.92
- Precision (class 1 – fake): 0.93
- Recall (class 1 – fake): 0.91
- F1-score (class 1 – fake): 0.92

The confusion matrix is:

- True real classified as real: 68
- True real classified as fake: 5
- True fake classified as real: 7
- True fake classified as fake: 67

These results indicate that the model correctly classifies the vast majority of both real and fake videos, with relatively few misclassifications and no strong bias toward one class. Balanced F1 scores demonstrate that the system maintains a similar trade off between precision and recall across both real and fake classes, which is important in forensic contexts where both false positives and false negatives are costly.

3.2 Comparative Study with Baseline Models

The best validation accuracy of each model trained on the same dataset is summarized as follows:

Table 1
Model Accuracy Comparison

Model	Best Validation Accuracy
CNN Only	0.8916
CNN + Transformer	0.8844
CNN + FFT	0.8707
Proposed model	0.92

This comparison highlights several observations:

- The CNN-only model benefits from spatial feature learning but lacks explicit temporal or spectral modeling, limiting its ability to capture motion and generator specific fingerprints.
- The CNN+Transformer model introduces temporal attention but still omits frequency information, which may contain important cues about synthesis artifacts.
- The CNN+FFT model leverages spatial and frequency-domain features but does not exploit temporal dynamics, reducing its sensitivity to motion-based inconsistencies.
- The proposed hybrid model, which unifies spatial, temporal, and frequency features, achieves the highest validation accuracy, suggesting that multi-cue analysis yields more discriminative representations for deepfake detection.

These findings are consistent with trends in the literature that advocate for hybrid or multimodal deepfake detectors combining multiple feature domains.

3.3 Discussion

The results indicate that a well structured hybrid architecture can achieve clear performance improvements over simpler baseline models while maintaining a balanced distribution of errors. By jointly modeling spatial features and temporal relationships, the system is able to detect inconsistencies that only become apparent across multiple frames, such as identity instability, irregular lighting, or unnatural facial movements.

The inclusion of an FFT-based frequency branch further enhances performance by capturing spectral patterns associated with generative processes and compression artifacts. This observation is consistent with recent

findings that highlight the value of incorporating frequency domain information into deepfake detection systems.

From a practical standpoint, the achieved validation accuracy of approximately 92%, along with balanced precision and recall, suggests that the model can serve as a useful component within larger verification frameworks, where automated systems assist in identifying potentially manipulated content for further human inspection.

At the same time, the relatively limited dataset size and the possibility of dataset-specific biases indicate the need for additional evaluation across diverse datasets before considering deployment in real-world scenarios. Furthermore, the computational demands of the architecture, particularly in terms of GPU memory, may restrict its immediate applicability on edge devices. This underscores the importance of optimization and model compression strategies, as discussed in the future work section.

4. Conclusion and Future Work

4.1 Conclusion

This work introduces a hybrid deepfake detection system that combines spatial, temporal, and frequency based analysis within a unified framework. The model is implemented as a multi-branch spatio temporal architecture, integrating EfficientNet for visual feature extraction with Transformer-based temporal modeling and FFT-driven spectral feature fusion.

The system treats video input as a sequence of RGB frames, from which it extracts spatial representations, captures temporal dependencies, and incorporates frequency-domain information into a single classification pipeline.

Experiments conducted on a Kaggle deepfake dataset containing 731 videos show that the proposed model achieves a validation accuracy of approximately 92% on a held out set of 147 samples.

The model also maintains balanced performance, with F1-scores of 0.92 for both real and fake classes, and a confusion matrix reflecting relatively low misclassification rates.

A comparison with baseline models including CNN-only, CNN combined with Transformer, and CNN combined with FFT demonstrates that the hybrid approach provides measurable improvements in performance. This supports the idea that integrating

spatial, temporal, and frequency features leads to more reliable deepfake detection.

Overall, the project highlights the feasibility of developing an end to end deepfake detection system using modern deep learning techniques, while also providing empirical support for the effectiveness of multi-branch architectures in this domain.

4.2 Future Work

Spatio-temporal attention and localization:

Incorporating attention visualization methods or class activation mapping can help highlight suspicious regions or frames, improving interpretability for forensic analysis.

- **Model ensembles:** The proposed approach can be combined with alternative architectures such as other EfficientNet variants, ResNet-based encoders, or lightweight Vision Transformers using ensemble techniques like averaging or voting to improve both accuracy and robustness.

- **Lightweight deployment:** Converting the model into optimized formats such as TensorFlow Lite or ONNX, along with applying pruning, quantization, or knowledge distillation, can make deployment feasible on mobile and edge devices.

- **Uncertainty-aware predictions:** Techniques such as Monte Carlo dropout or Bayesian inspired methods can be used to estimate prediction confidence, allowing uncertain cases to be flagged for manual inspection.

- **Multimodal expansion:** The system can be extended to include audio features, lip-sync consistency checks, and motion based cues, enabling a more comprehensive multimodal detection framework that analyzes both visual and audio signals.

- **Cross-dataset and compression-robust evaluation:** Testing the model across multiple public deepfake datasets such as FaceForensics++, Celeb-DF, and DFDC under different compression settings can help evaluate generalization and reduce reliance on dataset specific patterns.

- **Explainable AI:** Integrating interpretability techniques like SHAP or LIME can provide clearer explanations of model decisions, improving trust and supporting applications in areas such as journalism and digital forensics.

- **Adversarial robustness:** Evaluating the system against adversarially generated deepfakes and applying strategies such as adversarial training, data augmentation, or feature-level regularization can strengthen resilience.

- **Real-time and streaming pipelines:** Adapting the model for near real-time processing would allow integration with streaming platforms, social media systems, or broadcast monitoring tools.

Pursuing these directions can bring the system closer to practical deployment while improving its ability to adapt to the rapidly evolving landscape of deepfake generation techniques.

References

- [1] Kietzmann, T., McClure, J., & Goodfellow, I. (2022). Deepfakes: Trick or treat? Artificial intelligence and the future of deception. *Annual Review of Vision Science*, 8, 1–25.
- [2] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). DeepFakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131–148.
- [3] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [4] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.
- [5] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2020). The Deepfake Detection Challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*.
- [6] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- [7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [8] Zhuang, H., Wang, P., & Zhang, X. (2022). Deepfake detection using frequency-based features and CNNs. *Pattern Recognition Letters*, 156, 54–60.
- [9] Tiwarekar, S. (2026). Deep Fake Detection (DFD) Entire Original Dataset. *Kaggle Dataset*. Available at: <https://www.kaggle.com>