

A LIGHTWEIGHT EXPLAINABLE DEEFAKE DETECTION FRAMEWORK USING SELF-BLENDED AUGMENTATION AND MULTI-DOMAIN FEATURE FUSION

Ms. Y. Tresa¹, Dr. K. Santhosh Kumar²

¹Research Scholar, PG and Research Department of Computer Science, Providence College (Autonomous), Coonoor. Assistant Professor, Sri Ramakrishna College of Arts & Science for Women, Coimbatore. Email: tresalija@gmail.com

²Research Guide, PG and Research Department of Computer Science, Providence College (Autonomous), Coonoor. Email: bbksan@gmail.com

ABSTRACT

Deepfake detectors trained only on manipulation-specific artefacts often lose accuracy when tested on unseen datasets and compression settings. This paper proposes a lightweight and explainable framework for facial video deepfake detection that combines self-blended augmentation, spatial artefact learning, frequency-domain analysis, and physiological consistency cues. The proposed design uses self-blended images to synthesise hard negative samples from pristine faces, encouraging the detector to learn generic forgery boundaries rather than memorising dataset-specific manipulation traces. A compact convolutional backbone extracts spatial features, while frequency residual maps and remote photoplethysmography-inspired temporal descriptors provide complementary evidence for manipulated content. A gated multi-domain fusion block integrates the feature streams and an explainability module produces Grad-CAM heatmaps and domain-level contribution scores. The framework is designed for resource-constrained deployment while retaining interpretability for forensic analysts and platform moderators. A full experimental protocol is proposed using FaceForensics++, Celeb-DF, DFDC, and cross-dataset validation. The expected contribution is a practical architecture that improves generalisation, reduces computational overhead, and provides transparent evidence for its predictions.

KEYWORDS: Deepfake detection, explainable artificial intelligence, self-blended images, feature fusion, remote photoplethysmography, frequency analysis, lightweight CNN.

INTRODUCTION

The emergence of realistic face manipulation tools has created a major challenge for multimedia authentication, digital forensics, journalism, and public safety. Deepfake videos can modify identity, facial expression, lip movement, or visual attributes while preserving a natural-looking facial appearance. As generative models improve, purely visual detectors become vulnerable because manipulation artefacts become weaker, more localised, and more dependent on the generation pipeline.

FaceForensics++ established a large benchmark for manipulated facial imagery using DeepFakes, Face2Face, FaceSwap, and NeuralTextures, enabling systematic evaluation of detector robustness under compression.^[2] However, the same literature also shows that detectors may overfit to dataset-specific visual traces. This motivates frameworks that learn manipulation-agnostic evidence while remaining computationally efficient enough for practical screening.

The research problem addressed in this paper is therefore: how can a deepfake detector remain lightweight, interpretable, and robust under cross-dataset conditions? The proposed answer is a multi-domain detection framework that fuses three complementary cues: spatial artefacts, frequency inconsistencies, and physiological-temporal signals. Spatial cues capture boundary blending and texture discontinuities; frequency cues reveal abnormal spectral patterns introduced by synthesis and compression; physiological cues exploit the difficulty of preserving subtle skin-colour pulse signals in manipulated videos.

The main contributions of this paper are fourfold. First, it proposes a lightweight detection pipeline that can be implemented without depending on computationally expensive transformer-only architectures. Secondly, it introduces self-blended augmentation as a practical method for learning manipulation-agnostic facial artefacts from clean data. Thirdly, it combines spatial, spectral, and physiological cues through a gated fusion strategy so that the detector can adapt to different video quality conditions. Finally, it incorporates explainability at both region and feature-domain levels, which is essential for responsible use in digital forensic and moderation settings.

In the Indian context, the need for such frameworks is particularly relevant because social media platforms, messaging applications, online news channels, and public communication systems are widely used for rapid information sharing. A lightweight and explainable detector can support fact-checking teams, law-enforcement laboratories, academic researchers, and platform safety teams without requiring high-end server infrastructure for every screening task.

II. RELATED WORK

A. CNN-Based Deepfake Detection: Early deepfake detectors relied on convolutional networks trained on aligned facial crops. MesoNet demonstrated that a compact network can identify facial video forgeries by focusing on mesoscopic image properties, reporting strong results for Deepfake and Face2Face manipulations.^[3] Xception-style classifiers later became a common baseline for FaceForensics++ because depthwise separable convolutions capture fine-grained manipulation details with manageable complexity.^[7]

B. Self-Blended Images: Self-blended images are generated by blending pseudo source and target faces from a single pristine image, thereby reproducing common forgery artefacts such as blending boundaries and statistical inconsistencies. Unlike two-face blending, self-blending creates difficult synthetic fakes that are visually subtle. This encourages the model to learn generic forensic representations and improves transfer to unknown manipulations.^[1]

C. Physiological and Frequency-Based Detection: Remote photoplethysmography-based methods use subtle skin-colour variation to estimate pulse-related signals, and they are valuable because deepfake generation often fails to preserve physiological consistency.^[4] Frequency-domain methods, meanwhile, analyse artefacts that may be difficult to observe in RGB space. Combining these signals can improve robustness because each domain responds differently to compression, lighting, and synthesis errors.

D. Explainability: Grad-CAM produces visual explanations by using gradients flowing into the final convolutional layer to highlight image regions that most influence a target prediction.^[5] For deepfake detection, explainability is especially important because a binary real/fake score alone is insufficient for forensic review. Heatmaps and feature-domain contribution scores help analysts determine whether a prediction is based on meaningful facial evidence rather than background bias.

E. Research Gap: Existing detectors normally emphasise either visual artefacts, temporal inconsistency, or model accuracy. However, fewer approaches jointly address lightweight deployment, cross-dataset robustness, and explanation quality. A detector that performs well but cannot justify its decision may not be suitable for forensic reporting. Similarly, a detector that is accurate only on one benchmark may not be useful when the manipulation method changes. This paper therefore positions explainable multi-domain fusion as a balanced approach between accuracy, efficiency, and practical interpretability.

III. Proposed Framework

The proposed framework contains five stages: face preprocessing, self-blended augmentation, multi-domain feature extraction, gated feature fusion, and explainable decision reporting. During preprocessing, each video is decomposed into frames, faces are detected and aligned, and short facial clips are formed. The use of clips rather than isolated frames allows the framework to include temporal and physiological cues without requiring a heavy video transformer.

Self-blended augmentation is applied only during training. For each pristine face, a pseudo source version is produced using small transformations such as color jitter, sharpening, resizing, affine translation, or local warping. A soft facial mask is generated from landmarks and convex hull regions, then the transformed source is blended into the target. The resulting synthetic fake retains identity and pose but introduces subtle inconsistencies around facial components. This strategy reduces dependence on collecting every possible manipulation type and strengthens cross-manipulation generalization.

The spatial branch uses a lightweight backbone such as MobileNetV3, EfficientNet-B0, or a compact Xception variant.^[7,8] Its objective is to extract local evidence from boundaries, skin texture, and facial component inconsistencies. The frequency branch receives residual maps produced by high-pass filtering or discrete cosine transform statistics. The physiological branch estimates temporal colour traces from cheek and forehead regions and converts them into compact rPPG descriptors such as spectral peak stability, inter-region phase difference, and pulse-consistency variance.

To fuse these signals, the framework projects each branch output into a shared embedding space. A gated attention unit assigns adaptive weights to the spatial, frequency, and physiological embeddings. The final fused vector is passed to a binary classifier. The same gate values are retained as domain-level explanations, indicating whether the final decision relied mainly on visual artefacts, spectral inconsistencies, or physiological irregularities.

Table I: Proposed Framework Components

Module	Input	Output	Purpose
Self-blended augmentation	Pristine facial frames	Synthetic subtle fakes	Improve generalisation
Spatial branch	Aligned face crops	Texture and boundary embedding	Detect visible artefacts
Frequency branch	Residual/spectral maps	Frequency embedding	Detect spectral anomalies
Physiological branch	Face clips	rPPG-inspired descriptors	Measure temporal consistency
Fusion gate	All embeddings	Weighted representation	Adaptive evidence integration
Explanation unit	Classifier gradients and gates	Heatmap and contribution report	Support forensic interpretation

A. Architectural Design

The framework is deliberately modular. Each feature branch can be trained jointly with the complete system or evaluated independently during ablation. The spatial branch focuses on frame-level appearance, the frequency branch focuses on transform-domain abnormality, and the physiological branch focuses on temporal skin-colour consistency. This separation makes the model easier to analyse because the contribution of each domain can be inspected rather than hidden inside a single opaque feature vector.

Let F_s , F_f , and F_p denote the feature vectors obtained from the spatial, frequency, and physiological branches, respectively. Each branch output is projected using a small fully connected layer and batch normalisation. The fusion gate produces

three non-negative weights α_s , α_f , and α_p , where $\alpha_s + \alpha_f + \alpha_p = 1$. The fused representation is computed as $F = \alpha_s F_s + \alpha_f F_f + \alpha_p F_p$. This design keeps the fusion block simple, interpretable, and suitable for deployment on low-power hardware.

B. Self-Blended Augmentation Pipeline

The self-blended augmentation pipeline starts with a real facial image and creates an artificial manipulation without requiring a second identity. First, the facial region is identified using landmarks. Secondly, a source copy is transformed using colour adjustment, boundary smoothing, local scaling, and small affine changes. Thirdly, a soft blending mask is created around semantically important facial regions such as the cheeks, nose, eyes, and mouth. Finally, the transformed source is blended back into the target. The resulting sample appears realistic but contains weak artefacts similar to those found in many deepfake videos.

This augmentation is useful because it reduces the dependence on a fixed list of generation tools. A detector trained only on known fake-generation methods may learn the noise pattern of that dataset instead of a general concept of manipulation. In contrast, self-blended images repeatedly expose the detector to variations in mask shape, blending strength, colour shift, and boundary smoothness. This provides a broader training distribution and helps the model recognise subtle inconsistencies in unseen data.

C. Explainability and Decision Reporting

The explainability module produces two outputs. The first is a spatial heatmap generated from the final convolutional features, which highlights the facial regions that influenced the prediction. The second is a domain contribution report derived from the fusion gate. For example, if the frequency branch receives a high contribution score, the system indicates that spectral inconsistency played a major role. If the physiological branch dominates, the system points to temporal pulse inconsistency as the main evidence. These explanations are not treated as absolute proof, but as supporting evidence for human review.

IV. METHODOLOGY

A. Data Preparation: Videos are sampled at a fixed frame rate to reduce redundancy. Faces are detected, aligned, and cropped with a margin to preserve contextual boundaries. Clips of consecutive frames are grouped for temporal analysis. The training set includes real clips, dataset-provided manipulated clips, and self-blended clips generated from real facial frames. The validation and test sets exclude self-blended samples to measure performance on genuine manipulations.

B. Training Objective: The model is trained with a weighted binary cross-entropy loss and auxiliary consistency regularization. Let y be the ground-truth label and p be the predicted fake probability. The classification loss is computed as $BCE(y, p)$. A domain-consistency term penalizes unstable fusion weights across adjacent frames, and an explanation-sparsity term encourages Grad-CAM maps to focus on facial regions rather than irrelevant background. The total loss is $L = L_{cls} + \lambda_1 L_{gate} + \lambda_2 L_{exp}$, where λ_1 and λ_2 control the regularization strength.

C. Inference: At inference time, no synthetic augmentation is required. Each test video is processed into clips, and each clip receives a fake probability and explanation output. Video-level prediction is obtained using average probability pooling with confidence calibration. The explanation report includes a representative heatmap, branch contribution weights, and the most suspicious temporal segment. This makes the system usable as both an automated screening tool and a decision-support component.

D. Algorithmic Summary

The complete training procedure can be summarised as follows. Step 1: collect real and manipulated video clips from benchmark datasets. Step 2: extract aligned face crops and organise them into short temporal clips. Step 3: generate self-blended samples from real frames and add them to the training set. Step 4: compute spatial features from RGB crops, frequency features from residual maps, and physiological descriptors from temporal colour traces. Step 5: fuse domain features using the gated fusion block. Step 6: train the classifier using classification loss, gate consistency loss, and explanation regularization. Step 7: validate the model on unseen videos and tune only using the validation split.

During inference, the same preprocessing steps are used, but self-blended generation is disabled. Each clip is passed through the three feature branches, and the model outputs a clip-level probability. The final video-level decision is obtained by aggregating clip probabilities. The system also stores the top suspicious intervals so that an investigator can review the most relevant parts of the video instead of manually inspecting the full recording.

E. Implementation Details

The framework can be implemented using PyTorch or TensorFlow. Input frames may be resized to 224×224 pixels to maintain compatibility with lightweight CNN backbones. A clip length of 16 or 32 frames can be selected depending on the available GPU memory. Standard augmentations such as random cropping, horizontal flipping, Gaussian blur, JPEG compression, and colour jitter may be used carefully so that the detector does not overfit to clean laboratory conditions. Training may use the AdamW optimiser with cosine learning-rate scheduling and early stopping based on validation AUC. For deployment, model compression techniques such as pruning, quantisation-aware training, and knowledge distillation may be considered. Quantisation is particularly relevant when the detector is intended for edge devices or low-cost screening servers. However, compression should be applied only after verifying that explanation quality and cross-dataset performance are not degraded significantly. A small reduction in in-dataset accuracy may be acceptable if the model becomes faster and maintains reliable behaviour on unseen manipulations.

V. Experimental Design

The evaluation should use both in-dataset and cross-dataset protocols. FaceForensics++ can be used for controlled training because it contains multiple manipulation methods and compression levels. Celeb-DF and DFDC can be used to measure real-world transfer because they contain different identities, capture conditions, and synthesis pipelines. A robust detector

should not only achieve high accuracy on the training distribution but should also retain strong AUC and F1-score when the manipulation source changes.

Baselines should include MesoNet, XceptionNet, EfficientNet-B0, EfficientNet-B4, and a spatial-only version of the proposed model.^[3,6,7,8] Ablation studies should remove one feature branch at a time: no frequency branch, no physiological branch, no self-blended augmentation, and no explainability regularization. Metrics should include accuracy, precision, recall, F1-score, AUC, equal error rate, parameter count, floating-point operations, and inference latency per video clip.

Table II: Suggested Evaluation Protocol

Protocol	Training Data	Testing Data	Purpose
In-dataset	FaceForensics++ train split	FaceForensics++ test split	Measure controlled benchmark performance
Cross-dataset I	FaceForensics++	Celeb-DF	Evaluate identity and generation transfer
Cross-dataset II	FaceForensics++	DFDC	Evaluate real-world robustness
Ablation	Proposed variants	All test sets	Measure contribution of each module
Efficiency	Final model	Representative video clips	Report latency and model size

VI. Expected Results And Discussion

The central hypothesis is that self-blended augmentation improves cross-dataset generalisation by exposing the model to subtle and diverse blending artefacts during training. The frequency branch is expected to improve performance under visually realistic manipulations because generative synthesis often leaves spectral traces. The physiological branch is expected to contribute most strongly when face texture artefacts are weak but temporal skin-colour dynamics are inconsistent.

Explainability should be evaluated qualitatively and quantitatively. Qualitatively, heatmaps should concentrate on manipulated facial regions such as cheeks, mouth boundaries, eyes, jawline, or blending regions. Quantitatively, explanation faithfulness can be tested by masking the top-ranked heatmap regions and measuring the resulting change in prediction confidence. If the prediction changes strongly after removing highlighted regions, the explanation is more likely to reflect meaningful evidence.

A potential limitation is that physiological cues require stable face tracking and sufficient video quality. Low resolution, motion blur, heavy compression, makeup, lighting variation, and occlusion may weaken rPPG descriptors. The fusion gate helps mitigate this issue by reducing the weight of unreliable domains, but the system should still report uncertainty when the evidence is weak. Another limitation is that self-blended samples approximate but do not fully reproduce all future generation artifacts. For this reason, continual evaluation on newer manipulation methods is necessary.

A. Security and Ethical Considerations

Deepfake detection systems should be used as decision-support tools rather than final arbiters of truth. False positives may wrongly label authentic media as manipulated, while false negatives may allow harmful forged content to pass undetected. Hence, the framework should provide calibrated confidence scores and explanations so that human reviewers can assess the evidence. The system should also be evaluated across diverse lighting, camera quality, skin tone, compression, and demographic conditions to minimise unfair performance differences.

The ethical use of such a detector also requires careful data handling. Training and evaluation datasets should be obtained from authorised sources, and any personally identifiable video data should be handled according to applicable institutional and legal requirements. In academic work, the model should be reported with sufficient methodological transparency so that other researchers can reproduce the study and identify possible weaknesses.

B. FUTURE SCOPE

Future work may extend the proposed framework in several directions. First, larger and more recent datasets can be included to cover emerging generation models. Secondly, the physiological branch can be improved using more robust face tracking and signal-cleaning methods. Thirdly, the framework can be extended to audio-visual deepfake detection by adding speech-lip synchronisation and voice consistency features. Finally, privacy-preserving deployment can be explored so that videos are processed locally wherever possible.

VII. CONCLUSION

This paper Presented a lightweight explainable framework for deepfake detection using self-blended augmentation and multi-domain feature fusion. The proposed design integrates spatial, frequency, and physiological evidence through a gated fusion block and provides explanations through visual heatmaps and branch contribution scores. The framework addresses three key limitations in existing detectors: poor cross-dataset generalisation, lack of interpretability, and high computational cost. Future work should implement the complete system, evaluate it on multiple benchmark datasets, compare it against strong CNN and transformer baselines, and test its robustness under real-world compression and mobile-device constraints.

REFERENCES

1. K. Shiohara and T. Yamasaki, "Detecting Deepfakes with Self-Blended Images," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2022.

2. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE/CVF Int. Conf. Computer Vision, 2019.
3. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in Proc. IEEE Workshop on Information Forensics and Security, 2018.
4. K. S. Kammari et al., "A Comprehensive Review of Deepfake Detection Techniques Utilizing Remote Photoplethysmography," IEEE Access, 2025.
5. R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," International Journal of Computer Vision, 2020.
6. D. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and Vision Transformers for Video Deepfake Detection," 2021.
7. F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2017.
8. M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. Int. Conf. Machine Learning, 2019.