IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2674∼2684
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2674-2684

Explainable deep learning framework for advanced deepfake video manipulation detection

Shahrin Islam[1], Bibhas Roy Chowdhury Piyas[1], Fatama Jannat Tisha[1], Abu Saleh Musa Miah[2], Sadia Rahman[3], Shazzad Hossen[4], Md Abdus Samad Kamal[5]

[1]Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh
[3]Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chittagong, Bangladesh
[4]Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh
[5]Cluster of Electronics and Mechanical Engineering, School of Science and Technology, Gunma University, Kiryu, Japan

Article Info

Article history:

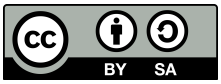
Received Dec 10, 2025
Revised Mar 5, 2026
Accepted Apr 22, 2026

Keywords:

Computer vision
Deepfake detection
Explainability
Gradient-weighted class activation mapping
Shapley additive explanations
Transfer learning
Video manipulation

ABSTRACT

The growing sophistication of deepfake technologies has emerged as a critical threat to the credibility of digital media by generating highly realistic yet fabricated visual content. This erodes public trust, elevates security vulnerabilities, and challenges information integrity across online platforms. Despite notable advancements, existing research still suffers from limited data diversity, insufficient model explainability, and inadequate model evaluation. To overcome this limitation, a framework for detecting deepfake video manipulation by using a transfer learning approach was introduced. Each extracted frame was processed by a convolutional neural network (CNN)-based model to obtain frame-level predictions, which were subsequently aggregated to produce the final video-level prediction using a predefined threshold. The publicly available, widely adopted FaceForensics++ dataset was used, which contains high-quality videos generated using advanced manipulation techniques. Various CNN architectures, including Xception, Densenet121, InceptionResNetV2, ResNet50, and EfficientNetB3, were explored along with rigorous hyperparameter tuning. Among these, the Xception architecture outperformed others by achieving a test accuracy of 94.5%. Gradient-weighted class activation mapping (Grad-CAM), generalized gradient-based visual explanations (Grad-CAM++), and Shapley additive explanations (SHAP) were employed to enhance model explainability by visualizing the key regions that influence deepfake detection. The research offers an effective approach to address deepfake threats and safeguard information integrity in contemporary industry 4.0.



Corresponding Author:

Md Abdus Samad Kamal
Cluster of Electronics and Mechanical Engineering, School of Science and Technology, Gunma University
Kiryu, Gunma 376-8515, Japan
Email: maskamal@gunma-u.ac.jp

1. INTRODUCTION

Deepfakes are hyper-realistic videos created by manipulating or synthesizing facial expressions, voices, or even entire actions utilizing modern deep learning approaches, particularly generative adversarial network (GAN) architectures. Such manipulated clips appear so convincing that they are virtually

indistinguishable from reality. Despite their technological sophistication, deepfakes pose serious risks due to their potential for misuse [1], [2]. Beyond spreading misinformation, deepfakes have been weaponized for political sabotage, identity theft, defamation, and cyberstalking [3], [4]. Given the severity of these risks, it is essential to develop robust and reliable detection systems to safeguard digital ecosystems and restore public trust in online content [5].

Despite advancements in deepfake detection, existing studies still struggled with limited data diversity, inadequate model evaluation, and poor interpretability [6], [7]. They often relied on small datasets that lacked high-quality videos and advanced manipulation techniques, which restricted their ability to detect sophisticated deepfakes. Additionally, these approaches provided minimal benchmarking and lacked explainability of model behavior. This work focuses on addressing the following gaps:

- Most existing studies relied on small and non-representative datasets that lacked high-quality videos and advanced manipulation techniques, which limited the adaptability to generalize and identify sophisticated deepfakes.
- Model explainability is often overlooked in existing deepfake video manipulation research, which makes it challenging to understand and interpret the reasoning behind predictions.
- Most prior studies offered minimal benchmarking and inadequate error analysis, which restricted the clarity about the model's behavior and the root causes of misclassification.

In this study, a deepfake manipulation recognition framework was introduced using transfer learning in which frame-level predictions are generated using convolutional neural network (CNN)-based architectures and subsequently aggregated for video classification. Comprehensive hyperparameter tuning was performed, followed by detailed error and explainability analysis to evaluate model performance, identify sources of misclassification, and enhance model interpretability. The core contributions of this study are listed as follows:

- Presented a deepfake video manipulation recognition framework trained on the extensive FaceForensics++ dataset, which includes high-quality videos and advanced manipulation techniques. The framework incorporates advanced preprocessing and architectural enhancements to mitigate dataset constraints and strengthen model generalization.
- Used explainable artificial intelligence (XAI) methods such as gradient-weighted class activation mapping (Grad-CAM), Shapley additive explanations (SHAP), and generalized gradient-based visual explanations (Grad-CAM++) to emphasize the areas of the visual content that were most influential in detecting video manipulation to improve model interpretability.
- Benchmarked the proposed model against several pretrained CNN architectures and prior studies to ensure rigorous comparison. Moreover, we performed detailed hyperparameter optimization and error analysis to identify key sources of misclassification.

Deepfake detection has emerged as an important area of research due to the rising complexity and societal influence of synthetic media. Recently, researchers have proposed a wide range of deepfake detection frameworks, primarily using deep learning techniques. Heidari *et al*. [8] conducted a comprehensive review of these techniques, highlighting CNN-based and region-based convolutional neural network (RCNN)-based models with transfer learning and adversarial training are the most common, achieving over 90% accuracy on benchmarks. Raza *et al*. [9] carried out a study focused on deepfake image identification using deep learning methods. In their research, they utilized several models, including NASNet, Xception, MobileNet, VGG16, and their proposed deep feature pooling (DFP) method. Among these, the DFP approach achieved the highest accuracy of 94%. Similarly, Chang *et al*. [10] introduced noise-level analysis visual geometry group (NA-VGG), a refined VGG16 model that incorporates steganalysis rich model (SRM) filtering and image augmentation to identify deepfake faces. Area under the curve (AUC) score of 85.7% on Celeb-DF dataset, outperforming the baseline VGG16 as well as other approaches. To enhance generalization, Hsu *et al*. [11] introduced common fake feature network (CFFN) with a pairwise learning strategy for deepfake detection, achieving 0.930 precision and 0.936 recall on self-attention generative adversarial network (SA-GAN). Coccomini *et al*. [12] compared vision transformer (ViT)-Base and EfficientNetV2-M for deepfake detection on ForgeryNet. While EfficientNetV2-M performed better on known forgery (up to 81.1% accuracy), ViT-Base showed stronger generalization with up to 77.5% accuracy and lower variance on unseen forgery. In related effort, Ghita *et al*. [13] used ViT-based model for deepfake detection, trained on 40,000 Kaggle images. It achieved 89.91% accuracy, showing strong performance and fast convergence, comparable to existing methods. Joshi and Nivethitha [14] used transfer learning based Xception model for deepfake image and

video detection, achieving 93.01% accuracy and demonstrating strong effectiveness. Younus and Hasan [15] proposed an approach for identifying deepfakes that utilizes the Haar wavelet transform to identify blur mismatches between the face and background. Tested on the UADFV dataset, it achieved 90.5% accuracy. This research provides useful insights into deepfake detection; however, most of them depend solely on image-level analysis or fail to address low-resolution, real-world video content. To overcome these limitations, a transfer-learning-based deepfake detection model that aggregates frame-level outputs to generate reliable video-level predictions using a widely accepted benchmark dataset is presented. Table 1 provides a comprehensive overview of prior research in related domains.

Table 1. Summary of related studies

Ref	Context	Proposed approach	Limitation	Data description
[16]	Deepfake video detection	Spatiotemporal model with 3DCNN, 3DResNet, TCN	Limited generalization	FF++, DFDC, Celeb-DF
[17]	Identification of deepfake videos	EfficientNet-B0 TL + EfficientNet-GRU with particle swarm optimization (PSO) optimization	High computational complexity and lack of explainability	Celeb-DFv2, DFDC subset, YT-Faces
[18]	General forged image and video detection	Capsule-Forensics: VGG-19 feature extractor + Capsule Network	Lacks robustness against adversarial/mixed manipulations.	Meso-data(frame-level)
[19]	Identity-aware deepfake video identification	Biometric feature matching using 3DMM + adversarial 3DMM Generative Network	Identity-dependency with high 3DMM computation	VoxCeleb2, DFD, FF++, DFDC-preview, Celeb-DF
[20]	Cross-dataset deepfake video recognition	StyleGRU with style-attention for temporal artifact detection	High preprocessing cost with StyleGAN latent dependency	Trained on FF++; tested on CDF, DFD, FaceShifter, DeeperForensics

2. METHOD

2.1. Dataset description

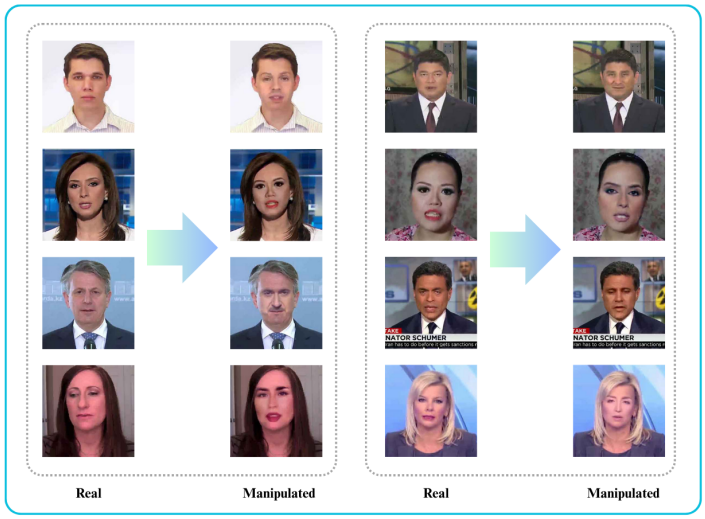
This study employs the FaceForensics++ dataset, which includes 2,000 video samples with 1,000 real and 1,000 manipulated. It is a broadly recognized standard dataset used for deepfake identification. The corpus consists of 1,000 real video clips sourced from YouTube, each featuring front-facing subjects with diverse facial expressions and body movements. These original videos were manipulated using different DeepFake techniques to generate the corresponding fake videos. Figure 1 illustrates the data visualization of the FaceForensics++ dataset. Figure 1(a) presents representative sample manipulated frames alongside their corresponding real ones from the dataset and Figure 1(b) illustrates the dataset distribution.

2.2. Data preprocessing

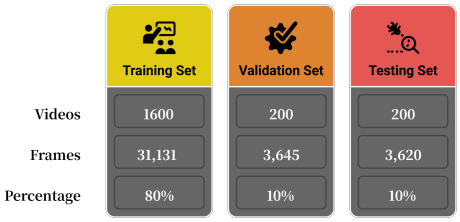
By extracting one frame per second from all videos, 38,396 frames were obtained, with each labeled as 1 for real and 0 for manipulated. These labeled frames were then used to train, validate, and test our deepfake detection model. To maintain uniformity and reduce computational complexity, each extracted frame was scaled to a resolution of 128×128×64 pixels. Additionally, the training data were augmented using horizontal flips, color jittering, and rotations. In contrast, the validation set was only resized and normalized to maintain consistency during model evaluation.

2.3. Overview of the proposed approach

The proposed framework initiates the process by extracting frames from the input video. One frame per second is extracted from each video to obtain a representative sample set. These frames are then passed through a preprocessing pipeline that includes resizing and normalization to make them suitable for model input. The preprocessed frames are subsequently fed into a CNN-based architecture comprising alternating convolution and pooling layers. The rectified linear unit (ReLU) activation function follows each convolutional layer and is chosen for its proven effectiveness in deep learning, helping prevent the vanishing gradient problem. The feature representations generated by the convolutional layers are then reshaped and passed through one fully connected layer, which produces a frame-level classification of either real or manipulated. To obtain a video-level prediction, we perform aggregation by averaging the prediction scores across all frames of a video. A threshold of 0.5 is applied to the average prediction score (PA): if $PA > 0.5$, the video is classified as real; otherwise, it is classified as manipulated. Figure 2 depicts the overall framework of our designed methodology.



(a)



(b)

Figure 1. Dataset visualization: (a) sample deepfake manipulation in FaceForensics++ dataset and (b) dataset distribution analysis

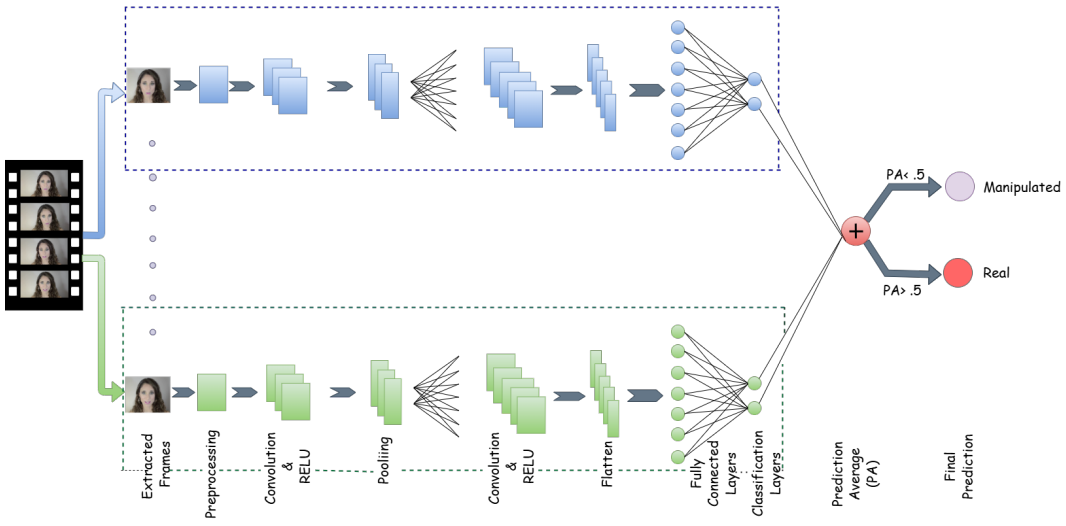


Figure 2. Workflow illustration of the designed methodology

2.4. Model training

The dataset was split into 80% for training, 10% for validation, and 10% for testing. The training data are used to learn the model's weights, while the validation split is used to tune hyperparameters and reduce overfitting. The testing subset, which contains unseen samples, is used to assess the model's performance. The

model is trained for 50 epochs, and the checkpoint with the highest validation score is saved and later used for final evaluation on the test data. Through extensive experimentation, the best hyperparameter configuration is determined, which is outlined in Table 2.

Table 2. Summary of hyperparameter configurations for the models

Hyperparameter	Hyperparameter space	Selected hyperparameter
Learning rate	[1e-5, 1e-4, 3e-4, 1e-3]	3e-4
Optimizer	[adam, Nadam, adamW]	adamW
batch size	[8, 16, 32, 64]	32
Epochs	[5, 10, 25, 20, 50]	50
Dropout rate	[0.1, 0.15, 0.25, 0.35, 0.5, 0.55]	0.1

3. RESULTS AND DISCUSSION

3.1. Comparison of performance

Various CNN architectures, including Xception, Densenet121, InceptionResNetV2, ResNet50, EfficientNetB3, have been explored. Table 3 presents the evaluation results of these models using multiple performance metrics. Among them, the Xception model outperforms all others by achieving a test accuracy of 94.5%. Conversely, the inception model has performed the worst on this deepfake detection task.

Table 3. Comparison of model performance

Method	Real			Manipulated			Accuracy
	Precision	Recall	F1-score	Precision	Recall	F1-score	
Xception	0.94	0.95	0.95	0.95	0.94	0.94	0.95
Densenet121	0.93	0.94	0.94	0.94	0.93	0.93	0.94
Inception_ResNet_V2	0.95	0.91	0.93	0.91	0.95	0.93	0.93
ResNet50	0.90	0.94	0.91	0.90	0.89	0.91	0.92
EfficientNet_B3	0.88	0.87	0.87	0.87	0.88	0.88	0.88

3.2. Analysis of model performance and computational efficiency

This section compares the performance and computational efficiency of the explored models. As illustrated in Table 4, each architecture offers a different balance between performance and complexity. The experimental results show that Xception performs best, achieving the highest accuracy with the lowest error rate and a moderate number of parameters. With strong accuracy and faster inference, Xception stands out as the most practical and deployment-ready model for deepfake detection.

Table 4. Analysis of model performance and computational efficiency

Model	Training time (Hr)	Inference time (ms)	Parameter count (million)	Model size (Gb)	Error rate (%)	GPU memory usage (Gb)
Xception	12.3	34.8	22.9	0.89	5.0	5.6
Densenet121	14.1	39.5	8.0	0.35	6.0	3.8
Inception_ResNet_V2	17.4	53.2	55.8	1.10	7.5	6.3
ResNet50	15.6	44.1	25.6	1.00	8.0	7.1
EfficientNet_B3	16.8	58.7	12.0	0.55	12.0	4.2

3.3. Result of stratified K-Fold cross validation

To evaluate the robustness of the model, a stratified 5-fold validation approach was applied. The mean and standard deviation of the performance metrics presented in Table 5 demonstrate the model's stability and generalizability, confirming its effectiveness for real-world deployment. The consistent performance across all folds indicates that the model is not biased toward any particular subset of the data.

3.4. Result analysis

The strong performance of the Xception architecture in Figure 3 is due to its depthwise separable convolutions (yellow 3×3 blocks), which reduce computational cost while preserving key spatial features. Strided convolutional layers (red 3×3/2 blocks) downsample feature maps, which allows the model to capture complex patterns and subtle manipulations. Finally, the GAP and Softmax layers (purple and green blocks) enable robust classification and improved generalization.

Table 5. Performance metrics across folds

Fold	Precision	Recall	F1-score	Accuracy
1st fold	0.941	0.952	0.946	0.948
2nd fold	0.938	0.947	0.942	0.945
3rd fold	0.945	0.953	0.949	0.951
4th fold	0.936	0.944	0.940	0.943
5th fold	0.948	0.956	0.952	0.954
Mean	0.942	0.950	0.946	0.948
Standard deviation	0.0043	0.0040	0.0045	0.0042

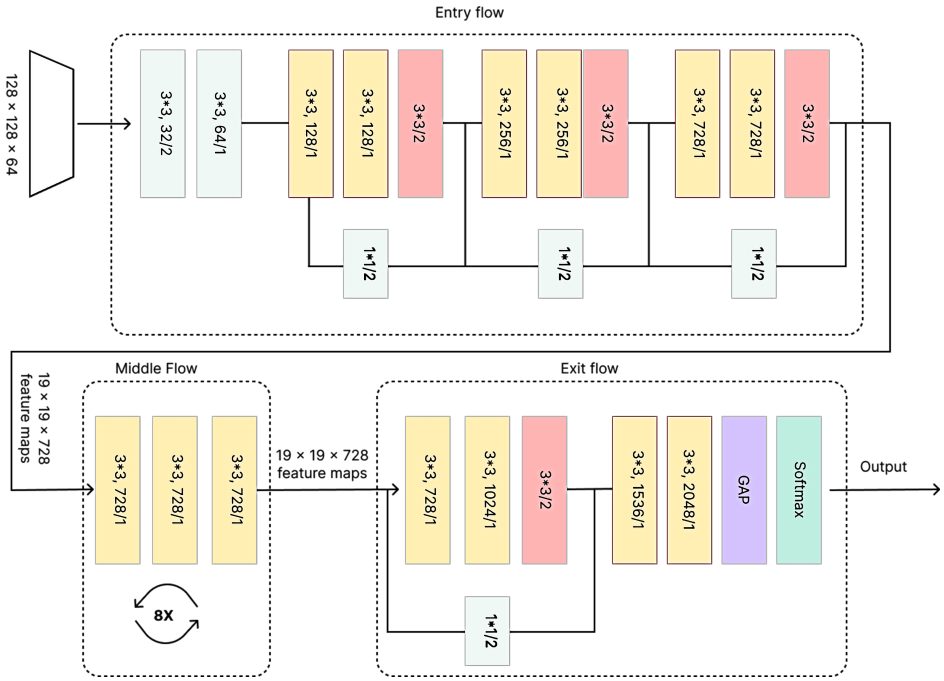


Figure 3. Xception model architecture

Figure 4 illustrates the outcomes of the proposed model, providing a comprehensive overview of its training behavior and classification performance across multiple evaluation metrics. The training and validation trends in Figure 4(a) show that the Xception model maintains consistently high accuracy and steadily decreasing loss over the training epochs. The close correspondence between the training and validation curves suggests that the model is learning effectively and generalizing well, with no evidence of overfitting. The ROC curve in Figure 4(b) shows excellent classification performance with an AUC of 0.99, which indicates nearly perfect discrimination between real and manipulated samples.

3.5. Error analysis

The confusion matrix in Figure 4(c) shows 11 misclassifications: 6 manipulated videos misidentified as real and 5 genuine videos misidentified as manipulated. The model has achieved 94 true positives for manipulated videos and 95 true negatives for real videos. There are 6 false negatives and 5 false positives, which indicates a slightly higher rate of false negatives. Error analysis reveals that misclassifications are primarily due to subtle differences between real and manipulated frames, as well as indistinguishable facial movements and expressions in these instances.

3.6. Explainability

This section illustrates the explainability within the proposed model and verifies its predictions using class activation mapping (CAM), Grad-CAM, Grad-CAM++, and SHAP. XAI [21], [22] helps reveal how the model makes decisions, enhances transparency, trust, and practical effectiveness in addressing deepfake threats. Figure 5 illustrates the explainability of the proposed model, providing visual and feature-level insights that support a deeper understanding of the model's behavior and decision-making process.

3.6.1. Grad-CAM

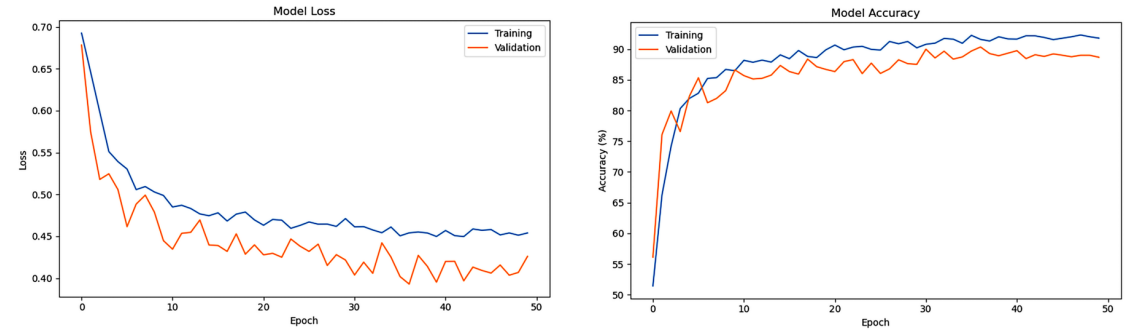
Grad-CAM emphasizes the regions of the frame that mostly influence the model's decisions, helping reveal which features it uses to distinguish real from manipulated videos [23]. Figure 5(a) shows explainability maps generated using CAM, Grad-CAM, and Grad-CAM++ for both real and manipulated frames. CAM highlights the broader facial area but lacks precision, whereas Grad-CAM refines the focus to class-relevant regions, and Grad-CAM++ provides the sharpest, most localized activations, offering a clearer understanding of how the model arrives at its predictions. The heatmaps show significant regions, including the eyes, mouth, and overall facial features, which are most significant in decision-making for the model. The outcome suggests that the proposed model detects manipulation-induced inconsistencies by providing a concrete visual representation of subtle cues. Overall, visualizations show that model focuses on critical facial areas to distinguish deepfakes.

3.6.2. SHAP

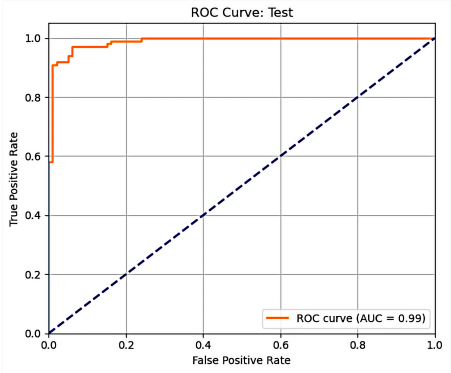
The SHAP visualizations in Figure 5(b) show that the model focuses on crucial facial areas, including eyes and mouth, in both real and manipulated frames. For real videos, these regions have high SHAP values, indicating their importance in predicting. In manipulated videos, while these same regions are still highlighted, the SHAP values vary, indicating subtle inconsistencies or artifacts introduced during manipulation. The heatmaps suggest that the model detects these discrepancies by concentrating on the most important and vulnerable parts of the face. Overall, the SHAP visualizations provide a clear picture of how the model makes its decisions by highlighting the facial region that most influences its predictions.

3.7. Comparison with related works

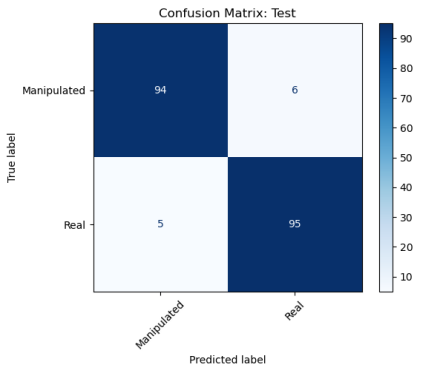
Table 6 presents a performance assessment contrasting our model with prior techniques, highlighting the superiority and improvements of our approach over current methodologies. The comparative results demonstrate that the proposed model consistently achieves higher accuracy and robustness across multiple evaluation metrics. These improvements indicate the effectiveness of the proposed design in addressing existing limitations and advancing the state of the art in deepfake detection.



(a)

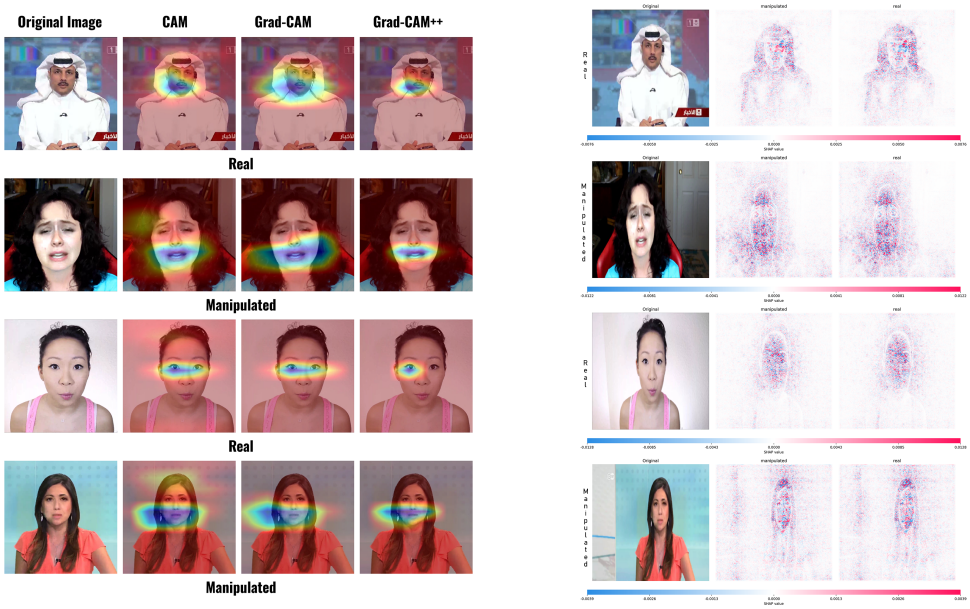


(b)



(c)

Figure 4. Outcomes of the proposed model: (a) training curve, (b) ROC curve, and (c) confusion matrix



(a) (b)

Figure 5. Explainability of the proposed model: (a) CAM, Grad-CAM and Grad-CAM++ and (b) SHAP

Table 6. Comparison with existing work

Paper	Methodology	Dataset	Accuracy (%)
Cozzolino *et al*. [24]	Noiseprint + siamese network	FaceForensics++	92.14
Wu *et al*. [25]	SSTNet	FaceForensics++ (c40)	90.11
Masi *et al*. [26]	Two-branch recurrent network	FaceForensics++ (c40)	91.1
Li *et al*. [27]	Patch&Pair CNN	FaceForensics++ (c40)	93.1
Our method	Proposed fine-tuned Xception model	FaceForensics++	94.50

4. CONCLUSION

Deepfake detection remains a challenging task as generative tools continue to produce highly realistic and nearly indistinguishable media. In this paper, these challenges are addressed by proposing a deep learning framework to detect advanced video manipulation. The method combines frame-level predictions to produce a video-level decision using a defined threshold. To overcome limitations in earlier studies, including limited data diversity and a lack of explainability, the high-quality FaceForensics++ dataset was used, which includes advanced manipulation techniques to improve model generalization. Multiple CNN architectures, including Xception, DenseNet121, InceptionResNetV2, ResNet50, and EfficientNet B3, was explored with rigorous hyperparameter tuning. Among these, the Xception model demonstrated superior performance, attaining a test accuracy of 94.5%. The model was benchmarked against prior works and conducted error analysis to identify sources of misclassification. Additionally, XAI methods was employed, including Grad-CAM, Grad-CAM++, and SHAP, to highlight the most influential areas in deepfake detection. The strong performance of Xception can be credited to its use of depthwise separable convolutions, which help reduce computational load while preserving essential spatial features. Error analysis revealed that misclassifications stemmed from subtle differences between real and manipulated frames, as well as indistinguishable facial movements. As deepfake generation tools continue to evolve, future work should explore more sophisticated detection techniques and evaluate performance on real-world datasets that include diverse forms of advanced manipulation.

ACKNOWLEDGMENTS

We gratefully acknowledge the contributions and support of all coauthors who facilitated this research.

FUNDING INFORMATION

This research is supported by Daffodil International University, Bangladesh.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Shahrin Islam	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Bibhas Roy Chowdhury Piyas	✓	✓	✓	✓		✓		✓	✓	✓	✓			
Fatama Jannat Tisha	✓		✓	✓		✓			✓		✓		✓	
Abu Saleh Musa Miah	✓			✓		✓			✓		✓	✓	✓	
Sadia Rahman			✓		✓		✓			✓	✓			
Shazzad Hossen	✓		✓	✓		✓			✓		✓		✓	
Md Abdus Samad Kamal				✓	✓		✓			✓		✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

The authors report no financial or personal conflicts regarding this study.

DATA AVAILABILITY

The data associated with this work are publicly available in the GitHub repository at https://github.com/Shazzad-Hossen/deepfake-detection-ml.

REFERENCES

[1] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: learning to detect manipulated facial images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 1–11, doi: 10.1109/ICCV.2019.00009.

[2] E. Kim and S. Cho, "Exposing fake faces through deep neural networks combining content and trace feature extractors," *IEEE Access*, vol. 9, pp. 123493–123503, 2021, doi: 10.1109/ACCESS.2021.3110859.

[3] N. Basit, F. Khalid, Q. U. Ain, and M. Andleeb, "Faceswap finder: a fusion-based deepfake detection technique," in *2025 6th International Conference on Advancements in Computational Sciences (ICACS)*, Feb. 2025, pp. 1–6, doi: 10.1109/ICACS64902.2025.10937811.

[4] A. A.-M. Alrawahneh, S. N. A. S. Abdullah, S. N. H. S. Abdullah, N. H. Kamarudin, and S. K. Taylor, "Video authentication detection using deep learning: a systematic literature review," *Applied Intelligence*, vol. 55, no. 4, Feb. 2025, doi: 10.1007/s10489-024-05997-8.

[5] L. X. Ying, A. H. M. Aman, M. S. Jalil, T. M. Omar, Z. S. Attarbashi, and M. A. Abuzaraida, "Malaysia cyber fraud prevention application: features and functions," *Asia-Pacific Journal of Information Technology and Multimedia*, vol. 12, no. 02, pp. 312–327, Dec. 2023, doi: 10.17576/apjitm-2023-1202-10.

[6] K. R. Sheth and V. S. Vora, "A comparative study on image forgery-facial retouching," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 851–859, Apr. 2023, doi: 10.11591/eei.v12i2.4481.

[7] K. Duhan and A. Kajal, "A comparative analysis of deep learning based approaches for deepfake identification," *Procedia Computer Science*, vol. 259, pp. 482–493, 2025, doi: 10.1016/j.procs.2025.03.350.

[8] A. Heidari, N. J. Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: a systematic and comprehensive review," *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 2, Mar. 2024, doi: 10.1002/widm.1520.

[9] A. Raza, K. Munir, and M. Almutairi, "A novel deep learning approach for deepfake image detection," *Applied Sciences*, vol. 12, no. 19, Sep. 2022, doi: 10.3390/app12199820.

[10] X. Chang, J. Wu, T. Yang, and G. Feng, "Deepfake face image detection based on improved VGG convolutional neural network," in *2020 39th Chinese Control Conference (CCC)*, Jul. 2020, pp. 7252–7256, doi: 10.23919/CCC50068.2020.9189596.

[11] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, "Deep fake image detection based on pairwise learning," *Applied Sciences*, vol. 10, no. 1, Jan. 2020, doi: 10.3390/app10010370.

[12] D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, and G. Amato, "Cross-forgery analysis of vision transformers and CNNs for deepfake image detection," in *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, Jun. 2022, pp. 52–58, doi: 10.1145/3512732.3533582.
[13] B. Ghita, I. Kuzminykh, A. Usama, T. Bakhshi, and J. Marchang, "Deepfake image detection using vision transformer models," in *2024 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*, Jun. 2024, pp. 332–335, doi: 10.1109/BlackSeaCom61746.2024.10646310.
[14] P. Joshi and V. Nivethitha, "Deep fake image detection using xception architecture," in *2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)*, Apr. 2024, pp. 533–537, doi: 10.1109/ICRTCST61793.2024.10578398.
[15] 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)*, Apr. 2020, pp. 186–190, doi: 10.1109/CSASE48920.2020.9142077.
[16] P. Agrawal, D. Pathak, V. Madaan, P. K. Verma, and W. O. Choo, "Spatiotemporal deep learning for real-time video-based deepfake detection using 3DCNN, 3DResNet, TCN, and VAE," in *Scientific Reports*, 2026, doi: 10.1038/s41598-026-49090-1.
[17] L. Cunha, L. Zhang, B. Sowan, C. P. Lim, and Y. Kong, "Video deepfake detection using particle swarm optimization improved deep neural networks," *Neural Computing and Applications*, vol. 36, no. 15, pp. 8417–8453, May 2024, doi: 10.1007/s00521-024-09536-x.
[18] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: using capsule networks to detect forged images and videos," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 2307–2311. doi: 10.1109/ICASSP.2019.8682602.
[19] D. Cozzolino, A. Rossler, J. Thies, M. Niesner, and L. Verdoliva, "ID-Reveal: identity-aware deepfake video detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 15088–15097, doi: 10.1109/ICCV48922.2021.01483.
[20] J. Choi, T. Kim, Y. Jeong, S. Baek, and J. Choi, "Exploiting style latent flows for generalizing deepfake video detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 1133–1143, doi: 10.1109/CVPR52733.2024.00114.
[21] N. Mansoor and A. I. Iliev, "Explainable AI for deepfake detection," *Applied Sciences*, vol. 15, no. 2, Jan. 2025, doi: 10.3390/app15020725.
[22] W. H. Abir *et al*., "Detecting deepfake images using deep learning techniques and explainable AI methods," *Intelligent Automation and Soft Computing*, vol. 35, no. 2, pp. 2151–2169, 2023, doi: 10.32604/iasc.2023.029653.
[23] S. Venkateswarulu and A. Srinagesh, "DeepExplain: enhancing deepfake detection through transparent and explainable AI model," *Informatica*, vol. 48, no. 8, May 2024, doi: 10.31449/inf.v48i8.5792.
[24] D. Cozzolino, G. Poggi, and L. Verdoliva, "Extracting camera-based fingerprints for video forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019, pp. 130–137.
[25] X. Wu, Z. Xie, Y. Gao, and Y. Xiao, "SSTNet: detecting manipulated faces through spatial, steganalysis and temporal features," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, *pp. 2952–2956*, doi: 10.1109/ICASSP40776.2020.9053969.
[26] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Computer Vision – ECCV 2020*, Cham, Switzerland: Springer International Publishing, 2020, pp. 667–684, doi: 10.1007/978-3-030-58571-6_39.
[27] X. Li, K. Yu, S. Ji, Y. Wang, C. Wu, and H. Xue, "Fighting against deepfake: patch&pair convolutional neural networks (PPCNN)," in *Companion Proceedings of the Web Conference 2020*, Apr. 2020, pp. 88–89, doi: 10.1145/3366424.3382711.

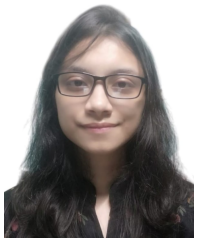
BIOGRAPHIES OF AUTHORS

Shahrin Islam is a lecturer in the Department of Software Engineering at Daffodil International University, Bangladesh. She received B.Sc. and M.Sc. in Computer Science and Engineering from Daffodil International University, Bangladesh. Her research interests include artificial intelligence, machine learning, deep learning, and optimization. She has published 5 journal papers with renowned publishers such as Elsevier and IEEE. She is open to international collaborations focused on addressing real-world and societal challenges through the use of advanced technologies and artificial intelligence. She can be contacted at email: shahrinislam.swe@diu.edu.bd.

Bibhas Roy Chowdhury Piyas is a lecturer in the Department of Software Engineering at Daffodil International University, Bangladesh. He received his B.Sc. in Computer Science and Engineering from Chittagong University of Engineering and Technology, Bangladesh, and is currently pursuing an M.Sc. in Computer Science and Engineering from Jahangirnagar University, Bangladesh. His research interests include computer vision, natural language processing, and deep learning. He has published 5 journal papers, 2 conference papers, and 1 book chapter with renowned publishers such as Elsevier, Springer Nature, and IEEE. He received the best paper award at ICCCT-2024. He welcomes international collaborations to address real-world and societal challenges using cutting-edge technologies and artificial intelligence. He can be contacted at email: piyas.swe@diu.edu.bd.

Fatama Jannat Tisha is a lecturer in the Department of Software Engineering at Daffodil International University, Bangladesh. She received her B.Sc. and M.Sc. in Computer Science and Engineering from Jahangirnagar University, Bangladesh. Her research interests span natural language processing, deep learning, explainable artificial intelligence (XAI). She received National Science and Technology (NST) Fellowship 2023-24 in the M.S. program by the Ministry of Science and Technology (MoST) and published 3 journal paper with the renowned publisher Elsevier. She actively welcomes international and interdisciplinary collaborations to translate research into impactful and sustainable solutions. She can be contacted at email: tisha.swe@diu.edu.bd.

Abu Saleh Musa Miah is a postdoctoral researcher at the University of Aizu, Japan, funded by the Public University Corporation Fukushima. He obtained his Ph.D. in Computer Vision from the University of Aizu under the Japanese Government (MEXT) Scholarship. During his doctoral studies, he developed vision-based sign language recognition systems using graph-based and deep neural networks. He also worked as an AI engineer at Eyes Japan Co., Ltd., where he contributed to the development of a vision-based Parkinson's disease detection system. Additionally, he served as a research assistant at the Center for Language Research, University of Aizu, applying machine learning and NLP techniques to detect machine-generated scientific texts. His research focuses on computer vision, deep learning, and multimodal AI applications, with particular emphasis on sign language recognition, human activity recognition, and AI for healthcare. Earlier in his career, he conducted research on EEG-based motor imagery classification. His name has been included in the world's top 2% Elsevier researchers' list. He has authored 102 peer-reviewed publications, including 66 journal articles, 13 book chapters, and 23 conference papers, in prestigious venues such as scientific reports (Nature), IEEE Access, IEEE Open Journal of the computer society, pattern analysis and applications, multimedia tools and applications, multimedia systems, and MDPI journals. He has also presented his work at international conferences organized by IEEE, Springer, and ACM. He is member of IEEE. He can be contacted at email: abusalehcse.ru@gmail.com.



Sadia Rahman received her B.Sc. in Computer Science and Engineering from Chittagong University of Engineering and Technology, Bangladesh and is currently serving as assistant software engineer at IDLC, Bangladesh. Her research interests include computer vision, natural language processing, and deep learning. She has published 3 journal papers and 1 conference paper with renowned publishers such as Elsevier and Springer Nature. She received the best paper award at ICCCT-2024. She is interested in collaborating internationally to solve practical and societal issues by utilizing state-of-the-art technologies and AI innovations. She can be contacted at email: sadia1704064@gmail.com.

Shazzad Hossen is a software engineer. He received B.Sc. in Computer Science and Engineering from Daffodil International University, Bangladesh. His research interests include artificial intelligence, machine learning, deep learning, and blockchain development. He has published 1 journal paper and 3 conference papers with renowned publishers such as Springer Nature, and IEEE. He is open to international collaborations focused on addressing real-world and societal challenges through the use of advanced technologies and artificial intelligence. He can be contacted at email: shazzad15-2420@diu.edu.bd.

Md Abdus Samad Kamal is an associate professor at the Graduate School of Science and Technology, Gunma University, Japan. He received Ph.D. degree from Kyushu University, Japan, in 2006. Earlier, he was a researcher at The University of Tokyo, Japan, a visiting researcher with Toyota Central R & D Labs., Inc., Japan, and a senior lecturer at the School of Engineering, Monash University, Malaysia campus. His research interests include intelligent transportation systems, cooperative control of connected and automated vehicles, and the applications of model predictive control. He is a member of the Society of Instrument and Control Engineers (SICE), Japan Society of Automotive Engineers (JSAE), and a senior member of the Institute of Electrical and Electronic Engineers (IEEE). He can be contacted at email: maskamal@gunma-u.ac.jp.