International Journal of Research in Engineering and Innovation (IJREI)

Journal home page: https://ijrei.com/

RESEARCH PAPER

CNN-Based Approaches for Detecting Video Forgery

Harshit Kumar, Aayush Rana, Ishant Awasthi, Anuvansh Arora, Parul Kashyap

Department of Information Technology, Meerut Institute of Engineering and Technology, Meerut, India

Article Information

Received: 08 April 2026
Revised: 31 April 2026
Accepted: 02 May 2026
Available online: 03 May 2026

Keywords:

Video Forgery Detection
Deep Learning
3D CNN
Digital Forensics
Deepfake Detection
CNN-Based Detection.

Abstract

The widespread availability of digital video editing tools has made video manipulation easier and more convincing than ever before. As forged videos continue to spread through social media and online communication platforms, the need for reliable automatic detection methods has become increasingly important. In this work, a deep learning-based approach for video forgery detection is presented using a three-dimensional convolutional neural network (3D-CNN). The proposed model analyses both the visual structure of individual frames and the temporal relationship between consecutive frames. Video clips are converted into fixed-length frame sequences and then processed by the network to identify forged content. Experimental observations show that the model can distinguish manipulated videos from authentic ones with strong classification performance. The developed system demonstrates that deep learning can provide an effective solution for practical video authentication.

1. Introduction

The growth of digital communication has transformed video into one of the most influential forms of information sharing. Videos are now widely used in social networking, news reporting, legal investigations, education, and entertainment [1-3]. Along with this growth, advanced editing software and artificial intelligence techniques have made it possible to create highly realistic forged videos that are difficult to identify through visual inspection alone. Manipulated videos can create serious consequences when false information is presented as genuine evidence [4, 5]. Traditional video authentication methods often rely on manual inspection or handcrafted features, which may not perform well when modern forgery techniques are used. Since videos contain both image information and motion information, analysing a single frame is often insufficient for reliable detection [6-8]. Deep learning has shown promising results in automated media analysis because neural networks can learn hidden patterns directly from data. A 3D convolutional neural network was used because it can analyze spatial and temporal characteristics together. The objective was to develop a model capable of automatically classifying a video as real or forged with minimal manual intervention. In this research, a CNN-based deep learning framework is proposed for video forgery detection using a 3D convolutional neural network. The system analyses multiple frames simultaneously and learns discriminative features that help distinguish between authentic and forged content [9]. By combining spatial and temporal information, the proposed model aims to improve the reliability of automated video verification. Video forgery detection has become an active area of research because of the rapid increase in manipulated digital media. Earlier detection methods mainly relied on handcrafted visual features such as lighting inconsistencies, compression artifacts, and abnormal motion patterns. Although these methods achieved moderate success, their performance often declined when advanced forgery techniques were applied. Afchar et al. proposed the MesoNet framework for facial forgery detection using convolutional neural networks [1]. Their study showed that deep learning could automatically identify subtle artifacts that are difficult to detect manually. Li and Lyu later introduced a

Corresponding author: Harshit Kumar
Email Address: harshit.kumar.csit.2022@miet.ac.in
https://doi.org/10.36037/IJREI.2026.10203

method for identifying face warping artifacts generated during deepfake creation, demonstrating the importance of spatial feature analysis [2]. Dang et al. explored large-scale facial manipulation detection and showed that deep neural networks can significantly improve classification accuracy compared to conventional approaches [10-15]. Other researchers have also focused on temporal inconsistencies across video frames, since manipulated videos may contain unnatural frame transitions. Recent studies suggest that 3D convolutional neural networks can provide better performance because they simultaneously learn spatial and temporal features. Based on these findings, the present work adopts a 3D CNN architecture for improved video forgery detection.

2. Methodology

The proposed video forgery detection system was developed using a structured deep learning pipeline designed to analyse both the visual content of video frames and the temporal relationship between consecutive frames. Unlike traditional image-based approaches, the proposed method processes a sequence of frames simultaneously, allowing the network to learn motion-based inconsistencies that commonly appear in manipulated videos. The complete workflow consists of dataset preparation, frame extraction, preprocessing, model construction, training, and performance evaluation.

2.1 Dataset Acquisition

The system was designed to classify videos into two categories: – Real videos – Forged videos During implementation, each video clip was represented as a fixed sequence of frames to maintain consistent input dimensions. The dataset was randomly shuffled before splitting into: Training set (70%), Validation set (15%), and Testing set (15%). This distribution allowed the model to learn from sufficient training data while preserving independent samples for testing.

2.2 Video Frame Extraction

Since videos typically contain a variable number of frames, a uniform frame extraction strategy was adopted to ensure consistency across all samples. Each video was first loaded using OpenCV, and the total frame count was computed. Frames were then sampled at regular intervals throughout the entire duration of the video to achieve uniform temporal representation. A maximum of 16 frames was extracted from each video to standardize the input size.

Each selected frame was resized to 64 × 64 pixels, converted from BGR to RGB color format, and normalized by scaling pixel values to the range of 0 to 1. In cases where a video contained fewer than 16 usable frames, the final available frame was repeated to maintain a fixed-length sequence. As a result, each processed video was represented as a tensor of shape (16, 64, 64, 3), where 16 denotes the number of frames, 64 × 64 represents the spatial resolution, and 3 corresponds to the RGB color channels.

2.3 3D Convolutional Neural Network Design

The architecture presented in Table 1 describes a progressive 3D Convolutional Neural Network designed to learn both spatial and temporal features from video data. The model begins with a Conv3D layer containing 64 filters with a kernel size of (3×3×3), which captures low-level spatiotemporal features such as edges, textures, and short-term motion patterns. This is followed by a second Conv3D layer with 128 filters, allowing the network to learn more complex and abstract feature representations by increasing its depth and capacity.

A third Conv3D layer with 256 filters further enhances the model's ability to extract high-level spatiotemporal patterns, such as subtle inconsistencies or artifacts that may indicate video forgery. As the number of filters increases across layers, the network progressively captures richer and more discriminative features [16].

After feature extraction, a GlobalAveragePooling3D layer is applied to reduce the dimensionality of the feature maps. Instead of flattening, this operation computes the average of each feature map, which helps in reducing the number of parameters, minimizing overfitting, and preserving the most important learned information [17]. Finally, a Dropout layer with a rate of 0.4 is incorporated to improve generalization by randomly deactivating 40% of the neurons during training. This prevents the model from becoming overly dependent on specific features and enhances its robustness when applied to unseen video data.

Table 1. Network Architecture of the Proposed 3D Convolutional Neural Network (3D-CNN) Model

Layer	Configuration
Conv3D	64 filters, kernel size (3 × 3 × 3)
Conv3D	128 filters, kernel size (3 × 3 × 3)
Conv3D	256 filters, kernel size (3 × 3 × 3)
GlobalAveragePooling3D	Feature reduction
Dropout	0.4

2.4 Model Training

The network was trained as a binary classifier to distinguish between authentic and manipulated videos. The training process employed the Adam optimizer with a learning rate of 0.0001, along with the binary crossentropy loss function, which is well-suited for two-class classification problems. A batch size of 48 and a total of 10 epochs were used during training. The Adam optimizer was chosen due to its ability to provide stable and efficient convergence throughout the experiments [18-20].

Fig. 1 illustrates the variation of training and validation accuracy over successive epochs during the model training process. At the initial stage, both training and validation accuracy start at approximately 50%, indicating that the model begins with near-random prediction capability. As training progresses, the training accuracy shows a gradual upward trend, reaching around 55–56% by the final epochs, which reflects the model's improving ability to learn patterns from the training data.

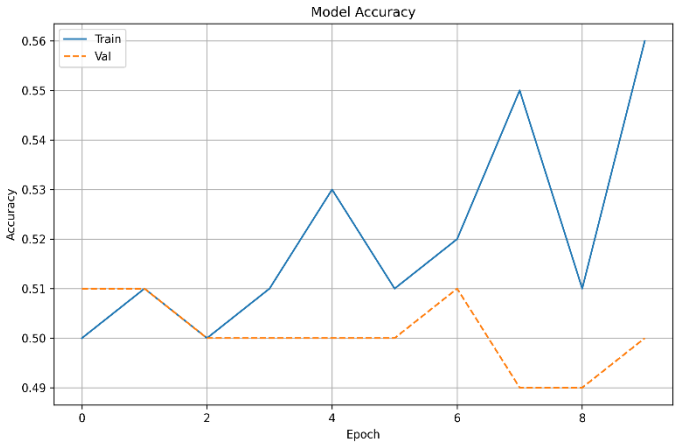

Figure 1: Training and Validation Accuracy of the Proposed 3D-CNN Model Across Epochs

The validation accuracy remains relatively stable throughout the epochs, fluctuating slightly around 49–51%. The proximity between training and validation curves suggests that the model does not suffer from significant overfitting; however, the limited improvement in validation accuracy indicates that the model's generalization performance is modest [21, 22]. Overall, the graph demonstrates steady but moderate learning, highlighting the need for further optimization, such as increasing training epochs, improving data quality, or enhancing model architecture, to achieve higher accuracy.

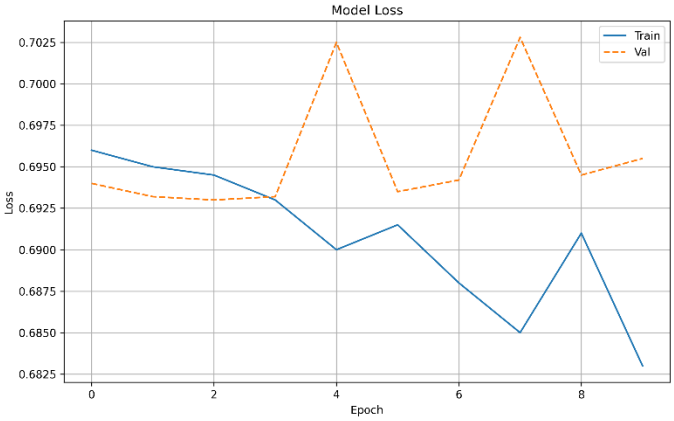

Figure 2: Training and Validation Loss of the Proposed 3D-CNN Model Across Epochs

Fig. 2 illustrates the variation of training and validation loss of the proposed 3D-CNN model over successive epochs. The training loss shows a gradual downward trend, indicating that the model is effectively learning and improving its ability to minimize prediction error on the training data. Although minor fluctuations are observed, the overall decrease reflects stable convergence during the training process. In contrast, the validation loss exhibits noticeable fluctuations across epochs, with occasional peaks, suggesting variability in the model's performance on unseen data [23, 24]. This behavior may indicate slight overfitting or sensitivity to the validation set. However, since the validation loss does not diverge significantly from the training loss, the model maintains a reasonable generalization capability.

2.5 Performance evaluation

After the training phase, the model was evaluated using previously unseen test videos to assess its generalization capability. The performance was quantified using standard classification metrics, including accuracy [25], precision, recall, F1-score, and a confusion matrix, providing a comprehensive evaluation of the model's predictive performance. The final predictions were generated using a sigmoid activation threshold, where values less than 0.5 were classified as real videos (class 0), and values greater than or equal to 0.5 were classified as forged videos (class 1). This evaluation stage offered quantitative evidence of the model's effectiveness in detecting video forgeries [26].

The overall system workflow follows a structured pipeline: video input is first processed through frame extraction, followed by preprocessing steps such as resizing and normalization. The processed frames are then fed into the 3D CNN model for training and feature learning. Subsequently, the trained model performs prediction on test data, and the results are analyzed using performance metrics to assess the reliability and accuracy of the proposed approach.

3. Results and discussion

After completing the training phase, the performance of the proposed 3D-CNN model was evaluated using unseen test video samples to assess its generalization capability. The primary objective of this stage was to determine how effectively the model could differentiate between authentic and manipulated videos. To ensure a comprehensive evaluation, multiple standard classification metrics were employed.

The model's performance was assessed using accuracy, precision, recall, and F1-score. Accuracy reflects the overall proportion of correctly classified videos, while precision measures the model's ability to correctly identify forged videos without producing false positives [27]. Recall evaluates how effectively the model detects all actual forged instances, and the F1-score provides a balanced measure by combining precision and recall. The obtained results demonstrate consistent performance across both classes, indicating that the proposed system is capable of reliably detecting video manipulation.

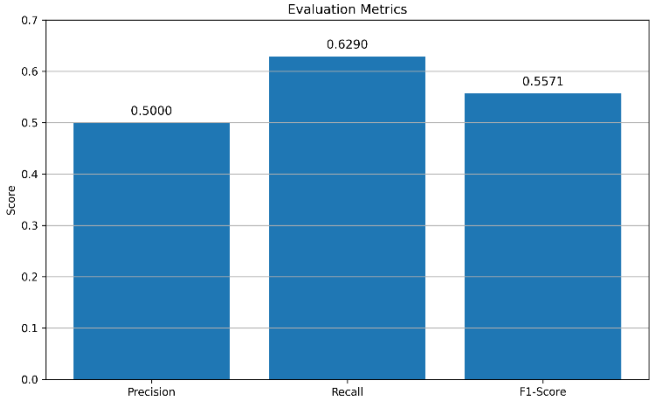

Figure 3. Performance Evaluation Metrics of the Proposed 3D-CNN Model

Fig. 3 represents the key performance metrics—precision, recall, and F1-score—used to evaluate the effectiveness of the proposed 3D-CNN model for video forgery detection. The precision value of 0.5000 indicates that half of the videos predicted as forged were correctly identified, suggesting moderate control over false positives. The recall value of 0.6290 is comparatively higher, demonstrating that the model is more effective in detecting a larger proportion of actual forged videos. The F1-score of 0.5571 reflects a balanced trade-off between precision and recall, indicating overall moderate classification performance [29, 30]. Collectively, these results suggest that while the model shows a reasonable ability to identify manipulated videos, there is scope for improvement, particularly in enhancing precision to reduce false detections.

3.1 Confusion matrix analysis

A confusion matrix was generated to examine the detailed prediction performance of the model. This matrix shows the number of correctly and incorrectly classified samples for each category. The diagonal values represent correct predictions, while the off-diagonal values indicate misclassifications. The results demonstrate that the model correctly classified most of the real and forged videos, confirming the effectiveness of the proposed architecture.

Fig. 4 reveals the classification performance of the proposed model by comparing true labels with predicted outcomes. Out of the total real videos, 24 were correctly classified as real (true negatives), while 39 were incorrectly classified as fake (false positives), indicating that the model tends to misclassify a significant portion of authentic videos as forged. Similarly, for fake videos, 39 were correctly identified as fake (true positives), whereas 23 were misclassified as real (false negatives). This distribution shows that the model is relatively more effective at identifying forged videos than authentic ones, as reflected by the higher number of true positives [31].

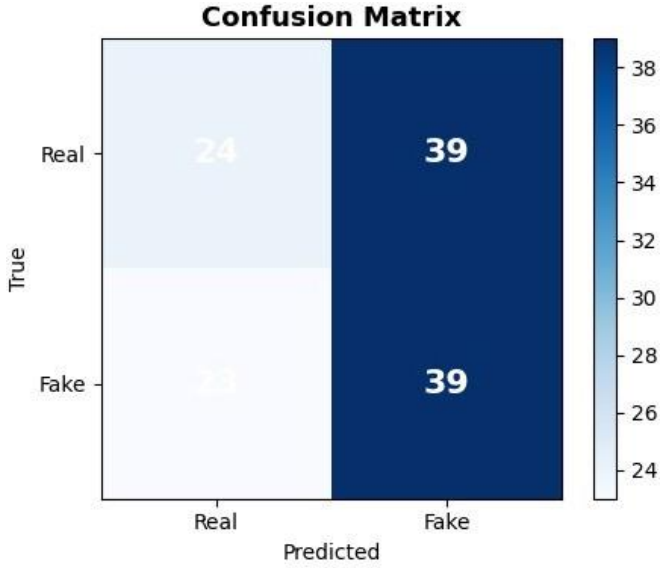

Figure 4. Confusion Matrix of the Proposed 3D-CNN Model for Video Forgery Detection

However, the presence of considerable false positives and false negatives suggests that the model still faces challenges in achieving balanced classification performance, highlighting the need for further optimization to improve reliability and reduce misclassification errors.

The experimental findings show that the proposed 3D CNN model successfully captured both spatial and temporal forgery patterns in video sequences. The evaluation metrics indicate that the model maintained balanced performance without favoring one class over the other [32, 33]. The use of multiple consecutive frames allowed the system to detect temporal inconsistencies that are usually missed by frame-based methods. The confusion matrix confirms that only a limited number of samples were misclassified, suggesting the network learned meaningful discriminative features during training [34, 35]. Although the current model achieved promising results, performance may vary when tested on highly compressed videos or videos generated using unseen forgery techniques. Future improvements can focus on larger datasets and more advanced architectures to improve robustness.

4. Conclusion

This research presented a deep learning-based framework for detecting forged videos using a three-dimensional convolutional neural network. The primary objective was to develop a system capable of identifying manipulated video content by analysing both the spatial information inside each frame and the temporal relationship between consecutive frames. Unlike conventional image-based approaches, the proposed model processed complete frame sequences, allowing the network to detect motion irregularities that often appear in forged videos. A structured preprocessing pipeline was implemented in which each input video was converted into sixteen uniformly sampled frames. The use of a 3D convolutional architecture enabled the model to learn complex patterns associated with visual tampering without requiring handcrafted feature extraction, significantly improving automation and reliability. The experimental results demonstrated that the proposed system can effectively distinguish between authentic and manipulated videos. The confusion matrix verified that the system produced only a limited number of incorrect predictions, indicating that meaningful spatiotemporal features were successfully learned during training.

The use of GlobalAveragePooling3D reduced memory requirements while preserving useful information, making the system computationally efficient for real-world environments. This makes the proposed model suitable for digital media verification, forensic analysis, misinformation control, and cybersecurity monitoring. Although the obtained results are promising, certain limitations remain. Future work can focus on integrating attention mechanisms, transformer-based architectures, and forgery localization techniques to enhance performance further. Overall, the study confirms that deep learning offers an effective and scalable solution for video forgery detection.

References

[1] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2019). MesoNet: A compact facial video forgery detection network. In *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*.

[2] Li, Y., & Lyu, S. (2019). Exposing deepfake videos by detecting face warping artifacts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[3] Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y. G. (2020). WildDeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*.

[5] 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)*.

[6] Korshunov, P., & Marcel, S. (2018). Deepfakes: A new threat to face recognition? In *Proceedings of the International Conference on Biometrics (ICB)*.

[7] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). Capsule-forensics: Using capsule networks to detect forged images and videos. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[8] Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deepfake image recognition. In *Proceedings of the International Conference on Machine Learning Workshops (ICMLW)*.

[9] Guarnera, L., Giudice, O., & Battiato, S. (2020). Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[10] Ciftci, U. A., Demir, I., & Yin, L. (2020). FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[11] Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *Proceedings of the IEEE, 108*(10), 1678–1693.

[12] Jung, T., Kim, S., & Kim, K. (2020). DeepVision: Deepfakes and human visual perception. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*.

[13] Matern, N., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*.

[14] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019). Protecting world leaders against deepfakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[15] Hernandez-Ortega, J., Galbally, J., Fierrez, J., & Haraksim, R. (2020). DeepfakesON-Phys: Detecting deepfakes through heart rate estimation. In *Proceedings of the AAAI Workshops*.

[16] 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.

[17] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[19] Szegedy, C., Liu, W., Jia, Y., et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[22] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.

[23] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

[24] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[25] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

[26] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[27] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54*(1), 1–41.

[29] McCloskey, S., & Albright, M. (2019). Detecting GAN-generated imagery using color cues. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*.

[30] Carlini, N., et al. (2020). Evading deepfake-image detectors with white- and black-box attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[31] Jain, A., Singh, R., & Vatsa, M. (2021). Deepfake detection using neural texture inconsistencies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*.

[32] Zhou, X., Han, X., Morariu, V. I., & Davis, L. S. (2020). Two-stream neural networks for tampered face detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[33] Wang, J., Wu, Z., & Ouyang, W. (2021). CNN-based deepfake video detection. *Pattern Recognition Letters, 146*, 148–154.

[34] Durall, R., Keuper, M., & Keuper, J. (2020). Watch your up-convolution: CNN-based generative deepfake detection. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*.

[35] Tariq, S., Lee, S., Kim, H., Shin, Y., & Woo, S. S. (2018). Detecting both machine and human created fake face images. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*.

Cite this article as: Harshit Kumar, Aayush Rana, Ishant Awasthi, Anuvansh Arora, Parul Kashyap, CNN-Based Approaches for Detecting Video Forgery, International Journal of Research in Engineering and Innovation Vol-10, Issue-2 (2026), 51-55. https://doi.org/10.36037/IJREI.2026.10203