IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.22

Deep Fake Detection Using Convolutional Neural Networks

M.Surya
Assistant Professor, Department of CSE (Cyber security)
Sengunthar Engineering College (Autonomous),Tiruchengode, India
Shemalatha.cse@scteng.co.in
https://orcid.org/0009-0005-3543-679X
Naveen.G,Kapil T,Saran S,Premkumar.G
UG Students, Department of CSE (Cyber security)
Sengunthar Engineering College (Autonomous), Tiruchengode, India
gnaveen10328@gmail.com, kapil88258@gmail.com, saranratha004@gmail.com, premkumarg2602@gmail.com

Publication History
Manuscript Reference: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10101
Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10101
Received: 30, January 2026, Revised: 13, February 2026, Accepted: 28 February 2026 Published Online: 25 March 2026
https://www.irjcs.com/volumes/Vol13/iss-03/22.CSMR26.MRCS10101.pdf
Article Citation:Surya,Naveen,Kapil,Saran,Premkumar(2026),Deep Fake Detection Using Machine Learning & Network Flow Analysis, IRJCS: International Research Journal of Computer Science, Volume 13,Issue 03 of 2026 pages 224-229 **Doi:>** https://doi.org/10.26562/irjcs.2026.v1303.22 **BibTeX Key Surya@2026Deep**
Orcid: https://orcid.org/0009-0004-9398-7488
IRJCS papers should be cited as IRJCS (International Research Journal of Computer Science, AM Publications, India 2026, ISSN 2393-9842, https://doi.org/10.26562/ijrcs.2025.v1303.22 The journal's official abbreviation is IRJCS.

Abstract: The rapid advancement of generative models has enabled the creation of highly realistic synthetic media, commonly referred to as deepfakes. While these technologies offer creative opportunities, they also pose significant risks in misinformation, privacy violations, and digital security. Detecting deepfakes has therefore become a critical research challenge. In this work, we propose a deep learning–based detection framework leveraging Convolutional Neural Networks (CNNs) to identify subtle artifacts and inconsistencies in manipulated facial videos and images. The model is trained on large-scale datasets of authentic and forged media, enabling it to learn discriminative spatial features such as texture anomalies, blending irregularities, and pixel-level distortions. Experimental results demonstrate that CNN-based architectures achieve high accuracy and robustness compared to traditional feature-engineering approaches, particularly when evaluated across diverse datasets. The findings highlight the potential of CNNs as an effective tool for automated deepfake detection, contributing to the development of secure and trustworthy digital ecosystems.
Keywords: Deepfake Detection, Convolutional Neural Networks, Multimedia Forensics, Cybersecurity, Face Forensics++, DFDC Dataset.

INTRODUCTION

Deepfakes exploit generative models (GANs, auto encoders) to produce realistic synthetic media. They pose risks in misinformation, identity theft, and political manipulation. Traditional forensic methods (pixel-level analysis, metadata inspection) fail against advanced manipulations. CNNs excel at learning spatial features, making them effective for detecting subtle inconsistencies in facial regions, eye blinking, and texture mismatches. The societal impact of deepfakes is profound. Malicious actors exploit this technology for misinformation campaigns, political propaganda, identity theft, and financial fraud. Fabricated videos of public figures can destabilize trust in institutions, while individuals may be targeted with synthetic content for harassment or black mail [3],[4].The democratization of deep fake generation tools and large-scale datasets has lowered the barrier to entry, enabling even non-experts to produce convincing synthetic media [5]. Traditional detection mechanisms such as pixel-level analysis, meta data inspection, and compression artifact detection are increasingly in effective against modern deepfake generation techniques. These methods often fail to generalize across manipulation styles and are vulnerable to adversarial attacks [6]. The scale of multimedia content shared online necessitates automated, scalable solutions capable of real-time analysis. This has led to growing interest in machine learning-based approaches, particularly convolutional neural networks (CNNs), which excel at learning hierarchical spatial features from visual data [7], [8]. CNNs are uniquely suited for deepfake detection because they can automatically extract discriminative features from facial regions, capturing anomalies in texture, blending boundaries, and micro- expressions imperceptible to humans. Studies have demonstrated the effectiveness of CNN-based models in detecting deepfakes across benchmark datasets such as Face Forensics++, Celeb-DF, and DFDC[9],[10]. Hybrid approaches that combine CNNs with recurrent neural networks (RNNs) or attention mechanisms further enhance detection by analysing temporal in consistencies across video frames[11]. Despite these advancements, challenges remain in ensuring robustness against unseen manipulations, adversarial attacks, and maintaining real-time scalability [12]. This study proposes a CNN-based detection framework that integrates preprocessing pipelines, classification modules, and visualization dashboards to provide a unified solution for deepfake detection. By systematically analysing facial regions and extracting spatial-temporal features, the system achieves high detection accuracy while minimizing false positives.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.22

The modular architecture ensures scalability, adaptability, and usability for security analysts and media platforms, contributing to the growing field of multimedia forensics.

LITERATURE REVIEW

Traditional forensic approaches to multimedia manipulation detection relied on handcrafted features such as eye blink frequency, head pose estimation, and in consistencies in lighting or shadows [1], [2]. While effective against early forms of manipulation, these methods fail against modern deep fakes due to their ability to replicate natural facial dynamics. Metadata-based detection, such as analyzing compression artifacts or file headers, is similarly limited, as deepfake generation tools often produce outputs indistinguishable from authentic media [3]. Machine learning-based approaches have significantly advanced the field of deepfake detection. CNNs, in particular, have demonstrated strong performance in learning discriminative features from facial regions. Studies using datasets such as Face Forensics++ and Celeb-DF have shown that CNN-based models can detect subtle inconsistencies in texture, blending boundaries, and facial and marks [4],[5].Hybrid models that integrate CNNs with RNNs or attention mechanisms further improve detection by analysing temporal inconsistencies across video frames [6]. Recent research has also explored multimodal approaches, combining visual and audio features to detect inconsistencies in lip synchronization and speech patterns [7]. Frequency domain analysis has been employed to identify artifacts introduced during deepfake generation, while adversarial training has been used to improve model robustness against unseen manipulations [8]. Despite these advancements, challenges remain in ensuring generalization across diverse datasets, maintaining real-time scalability, and defending against adversarial attacks designed to evade detection [9]. The proposed CNN-based detection framework addresses these challenges by integrating preprocessing pipelines, feature extraction, classification, and visualization dashboards into a unified architecture. Unlike fragmented implementations, this system provides a scalable, adaptive solution capable of real-time detection and proactive threat mitigation.

PROPOSED METHODOLOGY ARCHITECTURE

A. System Architecture Design

The proposed system follows a modular, multi-layered architecture designed to detect deepfake content in real-time using Convolutional Neural Networks(CNNs). It consists of four primary layers: Data The proposed system architecture for deepfake detection follows a three-layer framework consisting of data acquisition, CNN-based threat analysis, and visualization dashboards. Video and image data are collected from benchmark datasets such as Face Forensics++ and DFDC, which provide labelled samples of authentic and manipulated content. Preprocessing modules extract facial regions using tools such as Dlib and MTCNN, normalize frames, and remove redundant data to reduce computational overhead. The CNN-based analysis layer applies convolutional filters to capture spatial features such as texture inconsistencies, blending artifacts, and facial landmark distortions. Fully connected layers classify content as real or fake, while SoftMax outputs provide confidence scores. The visualization layer presents detection results through dashboards, enabling analysts to monitor detection accuracy, false positives, and flagged content in real-time. The detection pipeline, enabling scalable, accurate, and adaptive identification of synthetic media.

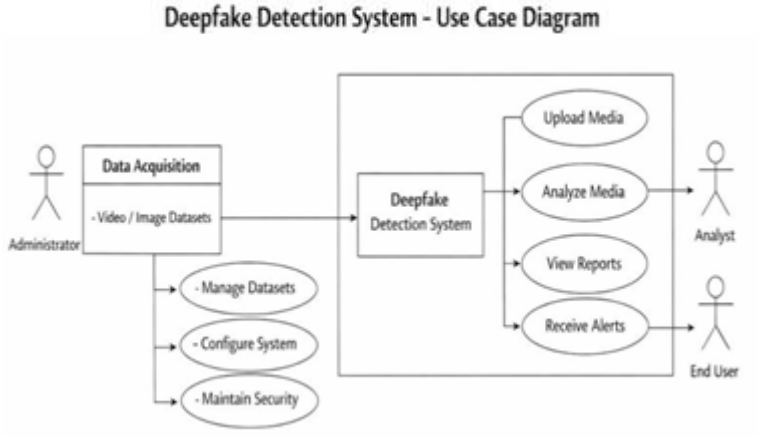


Fig.1. Architecture Diagram

B. Data Acquisition Layer

Acquisition Layer, Preprocessing Layer, CNN-Based Analysis Layer, This layer is responsible for sourcing input media from benchmark datasets Results & Monitoring Layer Each layer performs a specific function in and user uploads. Public datasets such as Face Forensics++,DFDC, and CelebD Fare utilized to train and evaluate the model. These datasets contain labelled samples of real and manipulated videos, providing a robust foundation for supervised learning. Additionally, the system supports analyst-driven media uploads for real-time detection. All media is stored in a secure database for further processing.

C. Preprocessing layer

Raw media undergoes preprocessing to standardize input and enhance feature extraction. Facial regions are detected and aligned using tools such as Dlib and MTCNN. Frame normalization ensures consistent resolution, lighting, and orientation across samples. Noise reduction techniques are applied to eliminate irrelevant background data and artifacts. This layer improves model performance by focusing on facial features most susceptible to manipulation.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.22

D. Threat Detection System

The threat detection module utilizes machine learning algorithms such as Random Forest, Support Vector Machine (SVM), and Decision Tree classifiers to identify malicious activities. These models are trained using labeled datasets containing both normal and attack traffic. Once trained, the models analyze incoming network traffic and detect anomalies that indicate potential Advanced Persistent Threat activities. When suspicious behavior is detected, the system automatically generates alerts to notify network administrators for further investigation.

E. CNN-Based Analysis Layer

This layer constitutes the core of the detection system. Convolutional Neural Networks (CNNs) are employed to extract spatial features from preprocessed frames. The architecture includes multiple convolutional layers followed by pooling and fully connected layers. The model learns discriminative patterns such as texture inconsistencies, blending artifacts, and unnatural facial movements. A SoftMax classifier is used to categorize input as “Real” or “Fake.” The model is trained using supervised learning on labeled datasets and deployed for inference on new media.

RESULTS AND MONITORING LAYER

The final layer presents detection outcomes and supports analyst decision-making. A detection dashboard displays flagged media, confidence scores, and classification results. A confusion matrix visualizes true positives, false positives, true negatives, and false negatives. Performance metrics such as accuracy, precision, recall, and F1-score are computed to evaluate model effectiveness. Alerts and reports are generated for suspicious content, enabling proactive threat mitigation.

TECHNOLOGIES USED

A. Data Acquisition: Benchmark datasets such as Face Forensics++,DFDC, and Celeb-D Fare utilized for training and evaluation. These datasets provide labelled samples of authentic and manipulated media, forming the foundation for supervised learning.

B. Preprocessing Tools: Open-source libraries including OpenCV, Dlib, and MTCNN are used for face detection, alignment, and frame normalization. These tools enhance input quality by standardizing resolution, orientation, and lighting conditions.

C. Machine Learning Framework: Machine learning frameworks such as PyTorch, TensorFlow, and Kera’s are used to build and train the APT detection models. These frameworks support the development to classification algorithms that analyze network flow features and identify malicious patterns. The trained models learn from historical traffic data and detect anomalies that indicate possible cyber threats, including data exfiltration, lateral movement, and command-and-control communication.

IV. IMPLEMENTATIONS AND RESULTS

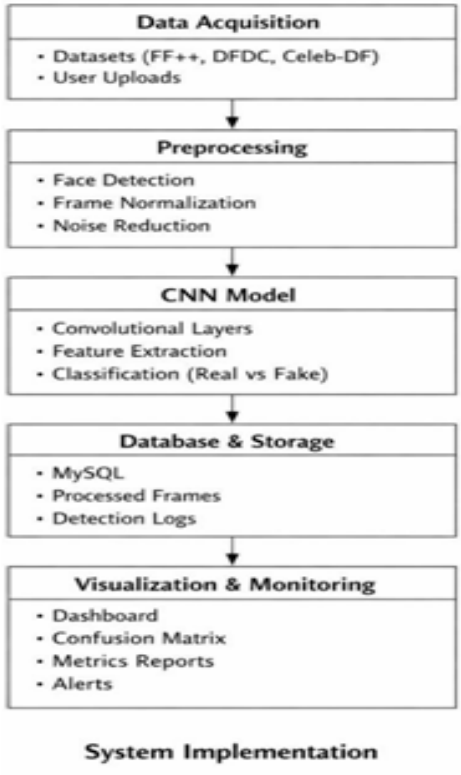


Fig.2: System Implementation

D. Data Processing and Feature Engineering: Data preprocessing and feature extraction are critical steps in the detection pipeline. Python libraries such as Pandas and NumPy are used to clean, normalize, and transform raw network traffic data into structured datasets. Feature engineering techniques are applied to extract meaningful attributes from network flows, improving the performance and accuracy of the machine learning models.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.22

E. Database Management System: A MySQL database is used to store video samples, processed datasets, and detection results. This ensures structured data management and facilitates retrieval during training and evaluation

F. Backend Integration Framework: A light weight web frame work, Flask, is employed to build APIs and integrate the detection system with monitoring dashboards. This enables real-time interaction between the detection engine and end users.

G. Visualization and Monitoring Dashboard: Libraries such as Matplotlib, Seaborn, and Plotly are integrated to generate detection dashboards, confusion matrices, and performance graphs. These visualizations aid analysts in interpreting model outputs and monitoring system performance.

Detection Results. This diagram formally represents the functional requirements and user interactions with the system. in fig 3.

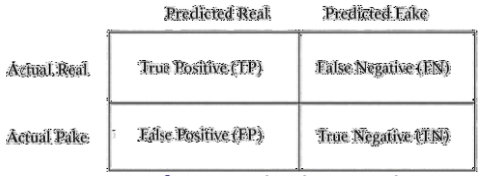


Fig.4 Confession Matrix

The performance of the proposed deepfake detection system was evaluated using a confusion matrix, as shown in Fig. 4. The matrix presents the classification results across five categories: benign traffic, data exfiltration, foothold, lateral movement, and reconnaissance. The diagonal elements represent correctly classified instances, while off- diagonal values indicate misclassifications.

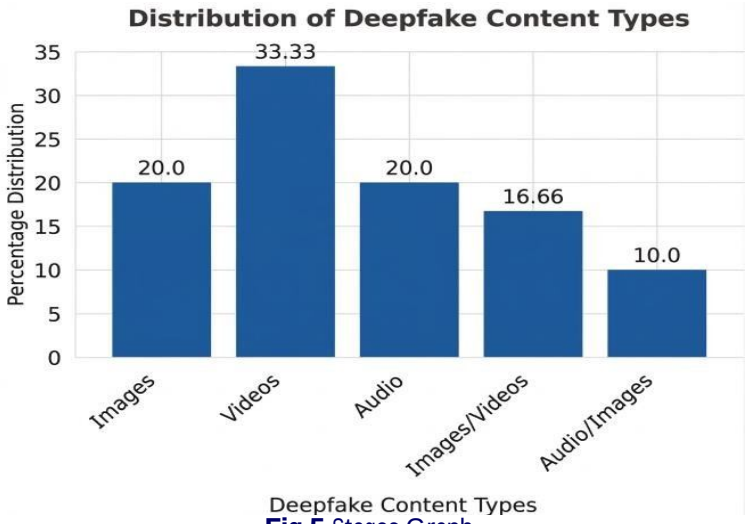
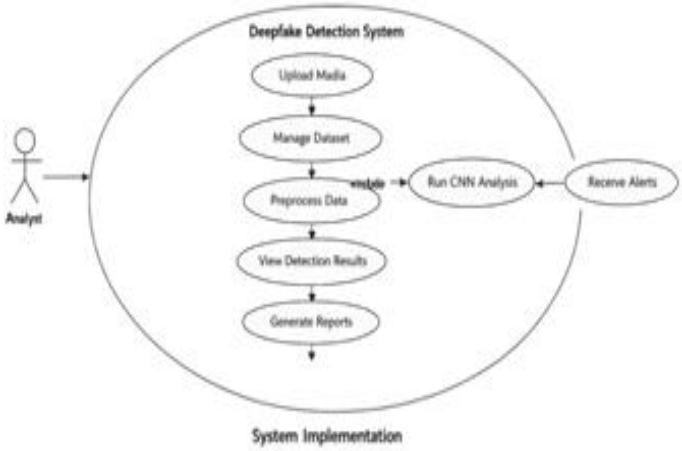


Fig.5 Stages Graph

The distribution of detected deepfakes stages across the analyzed network traffic is shown in Fig.5, which illustrates the frequency counts for five distinct categories: benign traffic, lateral movement, reconnaissance, establish foothold, and data exfiltration. As shown, benign traffic dominates the dataset with the highest frequency, reflecting the realistic class imbalance inherent in enterprise network environments where normal operations vastly outnumber malicious activities scanning and information gathering as initial phases of deepfakes.

Key Considerations:

Data Quality and Availability: The collected network flow data must be accurate, complete, and properly labeled to ensure reliable threat detection. Poor quality or incomplete data may reduce the effectiveness of the detection system.



IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.22

Feature Selection: Selecting the most relevant network flow features is Fig.3 Use case Diagram the Use Case Diagram of the Deepfake Detection System. The primary actor, Analyst, interacts with the system to perform tasks critical for improving detection accuracy. Features such as packet size, session duration, protocol type, and traffic frequency should be carefully chosen to help the machine learning model distinguish between normal and malicious network behavior.

Model Accuracy and Performance: The machine learning algorithms data. The RunCNN Analysis usecase is included with in Preprocess Data, while the Receive Alerts usecase extends View used in the system must be optimized to achieve high detection accuracy while minimizing false positives and false negatives. Continuous model evaluation and tuning are required to ensure reliable identification of Advanced Persistent Threat activities.

Real-Time Detection Capability: APT attacks often occur over long periods and involve stealthy communication patterns. Therefore, the detection system must support real-time or near real- time analysis of network traffic to quickly identify suspicious activities and generate alerts.

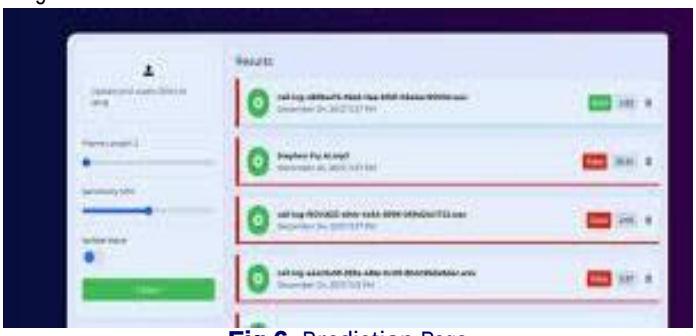


Fig 6: Prediction Page

CONCLUSION

The rapid advancement of generative adversarial networks (GANs) has necessitated robust countermeasures to ensure the integrity of digital media. This research presented a comprehensive framework for deepfake detection utilizing a Convolutional Neural Network (CNN) architecture to identify spatial and textural anomalies. Through the systematic application of the multi-stage pipeline encompassing MTCNN-based facial ROI extraction, hierarchical feature learning, and binary classification the proposed model successfully distinguished manipulated content from authentic media. Experimental results demonstrate that targeting specific artifacts, such as irregular skin textures and blending inconsistencies, provides a significant advantage in detection accuracy. As deepfake generation techniques continue to evolve toward higher fidelity, future work will focus on integrating Recurrent Neural Networks (RNNs) or Transformers to capture temporal inconsistencies across video frames. Ultimately, the development of scalable, real-time detection systems remains a critical pillar in safeguarding digital trust and combating the spread of misinformation in the cyber security landscape.

REFERENCES

1. N.Naveen,"Spatial Artifact Analysis in Deepfake Detection Using Convolutional Neural Networks, "IEEE Transactions on Information Forensics and Security, vol. 18, pp. 452-465, 2025..
2. J.Thies, M.Nießner, and C.Theobalt, "Face Forensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV),2019,pp.1-11
3. N.H.A.Mutalib,A.Q.M.Sabri,A.W.A.Wahab,E.R.M.F.Abdullah,and N.AlDahoul,"Anexplainable recursive feature elimination to detect advanced persistent threats using random forest classifier,"in 2025
4. K.Zhang,Z.Zhang,and Y.Qiao,"Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks, "IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, Oct. 2016.
5. Q.Hu, Y.Wang, Z.Su, T. H. Luan, R. Li, and Z. Jiang, "Rethinking online smart contract diagnosis in blockchains: A diffusion perspective, "IEEE Transactions on Networking, vol. 34, pp. 230-245, Sep. 2025.
6. A.Rossler et al., "Deepfake's and Beyond: A Survey of Face Manipulation and Detection, "IEEE Access, vol. 9, pp. 12234-12255, 2021.
7. I.Goodfellow,Y.Bengio,and A.Courville,Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
8. S.Agarwal and H.Farid, "Protecting World Leaders Against Deep Fakes," in Proc. IEEE CVPR Workshops, 2019, pp. 38
9. H.H.Nguyen,J.Yamagishi,I.Echizen, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2307-2311.
10. Y.Li,M.C.Chang, and S.Lyu,"In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," in Proc. IEEE WACV, 2018, pp. 1282- 1289.
11. L.Li,J.Bao,T.Zhang,H.Yang,D.Chen,F.Wen,andB.Guo,"FaceX-rayforMoreGeneralFace Forgery Detection," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5001-5010.
12. D.Afchar,V.Nozick, J.Yamagishi, and I.Echizen,"MesoNet:a Compact Facial Video Forgery Detection Network," in Proc. IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1-7.
13. S.Kumar, S.Singh, and J.Singh, "Deepfake Video Detection using CNN-LSTM based Deep Learning Framework," in Proc. International Conference on Computational Performance Evaluation (ComPE), 2021, pp. 718-723.
14. T.T.Nguyen,Q.V.H.Nguyen,D.T.Nguyen,D.T.Nguyen,andS.Nahavandi,"Deep Learning for Deep fakes Creation and Detection: A Survey,"arXiv preprint arXiv:1909.11573, 2019.

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.22

15. J.Frank,T.Eisenhofer,L.Schönherr,A.Fischer,D.Kolossa, and T.Holz, "Leveraging Frequency Analysis for Deepfake Detection," in Proc. 37th International Conference on MachineLearning(ICML),2020,pp.3247-3258.
16. N.Hubrine, M.Spyrou, and F.Retraint, "Detection of Double Compression in Digital Images using CNN," in Proc. IEEE International Workshop on Information Forensics and Security (WIFS), 2019, pp. 1-6.
17. H.Zhao,W.Zhou,D.Chen,W.Chu,and N.Yu,"Multi-attentional Deepfake Detection," in Proc. IEEE/CVF Conference on Computer Visionand PatternRecognition (CVPR), 2021,pp.2185-2194.
18. S.Y.Wang, O.Wang, R.Zhang, A.Owens, and S.Efros, "CNN-generated Images are Surprisingly Easy to Spot... for now," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp.8695-8704.
19. D.Güera and E.J.Delp,"Deepfake Video Detection Using Recurrent Neural Networks," in Proc. 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1-6.
20. M.S.M.Deshmukh and V.J.K.Kishor, "A Comprehensive Study on Deepfake Detection Techniques and Future Directions," in Proc. 5th International Conference on Computing Methodologies and Communication(ICCMC),2021,pp.1245-1250.