Iraqi Journal of Computers, Communications, Control and Systems Engineering
Journal homepage: https://ijccce.uotechnology.edu.iq

Video Deep Fake Detection Based on Spatiotemporal Analysis

Muna Ghazi Abdulsahib
College of Computer Science, University of Technology, Baghdad, Iraq.
MUNA.G.Abdulsahib@uotechnology.edu.iq

DOI: https://doi.org/10.33103/uot.ijccce.25.2.3

HIGHLIGHTS

- Hybrid deep-fake videos detection.
- VGG16 deep-learning.
- DFDC dataset.

ARTICLE HISTORY

Received: 11/ March /2025

Revised: 18/ May /2025

Accepted: 28/ June /2025

Available online: 30/ October /2025

Keywords:
Deepfake, Spatial feature, Temporal feature and VGG model.

ABSTRACT

Deep fake videos are a threat to information integrity by using advanced deep learning techniques to manipulate visual content and make it hard to distinguish from real videos. This includes videos where a person's image is changed or replaced using deep learning techniques. In this paper, we propose a hybrid approach for deep fake video detection that combines spatial and temporal features. We use a pre-trained VGG16 to extract deep spatial features from individual frames, and then global average pooling at both spatial and temporal level to convert variable length video sequences into fixed dimensional representations. A fully connected classifier with dropout regularization is then used to classify the video as real or fake. We evaluated our method on the DFDC dataset and got 95% accuracy and 0.81 AUC. Our framework not only uses transfer learning to reduce the need for large amount of training data but also computational efficiency, so it's suitable for large scale deep fake detection.

I. INTRODUCTION

A fake video is a video that has been digitally created or altered to show content that is not real. This includes videos where a person's image is changed or replaced using deep learning algorithms, often to look like someone else or create completely made-up situations. Videos that have been edited or manipulated by splicing, filtering or other video editing techniques to change the original content's context or meaning. Videos made entirely of computer graphics that, if done well, could pass as real video. Because deepfakes can spread misinformation, sway public opinion and invade personal privacy, they are a big problem. The research community is facing a special challenge with deepfakes because

of how fast the technology is advancing. Deepfakes are different from regular video forgeries in that they have minute artifacts and temporal irregularities that are often invisible to the human eye. This requires creating advanced detection algorithms that can detect even the slightest changes in spatiotemporal patterns. Even with great progress, current detection methods often struggle to generalize to other types of manipulations and are still vulnerable to attacks [1]-[4].

Deep fake video detection is a new field that uses deep learning and AI to detect videos that have been modified. Since these videos, created by complex algorithms like GANs, can mimic real people and events so well it's hard to tell real from fake information. The methods used by researchers in this field can range from traditional feature based to end-to-end deep learning models. Modern methods use deep neural networks to automatically learn from large datasets while traditional methods rely on hand crafted features such as anomalies in motion patterns, inconsistent face landmarks and abnormal lighting. Since videos have an extra temporal component, to detect fakes you need to look at both the consistency of individual frames over time and their individuality. Hybrid methods that combine temporal analysis with spatial feature extraction have been explored recently. Both temporal irregularities between frames and visual signals in individual frames have been designed to be captured by these methods [5]-[8].

Spatial features refer to visual patterns or characteristics extracted from individual frames in a video or image dataset that capture information about the structure, texture, edges, colors and local/global relationships between pixels in a 2D spatial domain. Regardless of the temporal (time based) information, these are important for analyzing the content of individual frames. Spatial features are important for computer vision and deepfake detection because they give an accurate representation of what an image looks like at a particular time. These were previously obtained through handcrafted methods such as edge detectors, local binary patterns, histograms of oriented gradients and scale invariant feature transform (SIFT). Deep network models have become the go to method for spatial feature extraction in recent years because they can learn complex structure from data. Small inconsistencies in illumination, color distribution or texture anomalies are examples of small artifacts that can be detected by looking at spatial features. Along with temporal analysis these give deepfake detection techniques overall better performance by giving frame level data and how that data changes over time [9]-[13].

Temporal features capture the change of visual information over time, versus spatial features which capture the static quality of individual frames, like texture, color, borders. When working with sequential data, temporal features are key especially when dealing with video and dynamic scenes. They give a more detailed and richer interpretation of images by giving insight into motion dynamics, temporal coherence and the flow of events. Temporal features are often extracted using frame-to-frame variation techniques. These have led to big improvements in temporal analysis tasks as they combine local motion dynamics with global sequence level context. So, understanding temporal features is important for understanding motion and dynamics in movies and building robust and reliable video analysis algorithms [14]-[18].

The visual geometry group at Oxford University created the deep model VGG16. The model has 16 layers, convolutional and fully connected, hence the 16 in the name. In the "pretrained VGG16 model" we mean this architecture has been trained on a large image dataset, often ImageNet, which has millions of images and 1000 classes. The pretrained VGG16 model performs well on many image analysis tasks because of its good feature extraction and transfer learning capabilities. VGG16 is known for its simple and consistent architecture, easy to understand and apply. It uses small 3x3 conv filters throughout the network. You can apply it to many computer vision tasks because it learns a hierarchy of visual cues, from simple edges in early layers to more complex patterns in deeper layers. You can use the rich feature representations ImageNet has learned for you by using the pretrained weights instead of training the model from scratch on your own dataset. Tasks with little training data benefit a lot from

this. It's a base for many transfer learning applications in computer vision, like deep fake detection, and is used a lot for object detection and image classification [19]-[23].

This paper contribution a new hybrid approach that combines a pre-trained deep model with temporal aggregation. Specifically, it uses a pre-trained VGG16 to extract spatial features from individual video frames and then aggregate them over time. Together these features allow the system to capture the temporal dynamics and fine-grained visual information to tell real videos from deepfakes. Also, it converts a variable length video sequence into a fixed dimensional feature vector by doing global average pooling at spatial and temporal level. Besides making learning and classification more effective, this simplified representation also makes the system scalable and suitable for large dataset like DFDC. Moreover, the method uses a powerful pipeline for frame extraction and preprocessing which includes face detection, scaling and normalization. By focusing on the most important regions (like faces) this pipeline improves the features extraction and strengthen the detection against video artifacts.

Because deepfakes use deep learning to alter identities, lip movements and facial emotions in videos, they are a serious threat to digital media. Current methods use complex temporal modeling or hybrid architectures but often require a lot of computational resources and big datasets. Many methods struggle to find a balance between accuracy and efficiency especially when dealing with variable length video sequences. To overcome these challenges we propose a simple yet effective hybrid system that combines temporal aggregation to detect inconsistencies in deepfakes with transfer learning for spatial feature extraction. We want to make it deployable on large datasets like DFDC and achieve robust detection with low computational overhead.

II. RELATED WORK

In 2023 O. S. A. Aboosh et al. [24] proposed hybrid deep learning models for detecting fake videos. This is part of the wider effort to tackle video counterfeiting. Hybrid deep learning models that combine recurrent neural networks (RNN) and convolutional neural networks (CNN) are used to detect fake videos. Video classification was done by training simpleRNN and Gated Recurrent Unit (GRU) models using facial features extracted from the frames using inceptionV3 model. The proposed models were evaluated using deepfake detection challenge (DFDC) dataset. SimpleRNN and GRU both have 98.5% and 98.9% detection accuracy respectively. The models' AUC is 0.979 and 0.986. Fusing the advantages of multiple architectures can give great accuracy and robustness but they can be computationally expensive and require huge resources for real-time processing and training.

In 2023 A. Jellali et al. [25] proposed an approach that combines CNN with Haar cascade classifiers. First facial regions are detected and extracted from video frames using Haar cascades. A CNN then processes these facial photos to classify them as real or fake. Using DFDC dataset they were able to make 91 out of 100 videos correct. The accuracy of the binary classification into true or fake faces was almost 99%. In summary the approach classified facial photos well. But future work should consider overfitting.

In 2023 F. Mira [26] proposed a method to detect deepfakes which uses long short-term memory and convolutional neural networks to distinguish between real and fake video frames. it uses YOLO face detector to detect facial video frames, processes them using CNN and then classifies them using XGBoost. The suggested methods like YOLO-CNN-XGBoost and CNN with LSTM are new ways to improve detection accuracy. The suggested methods may not be reproducible and applicable as there is no experimental validation and information about the dataset.

In 2024 N. Nibras et al. [27] proposed to use algorithms that will give the most accurate and fastest results to detect fake videos. To extract facial features from video frames, use InceptionV3 model. Two types of recurrent neural network—Simple Recurrent Neural Networks (SimpleRNN) and Gated

Recurrent Units (GRU)—are then trained with these features. The goal is to represent the temporal and spatial irregularities in fake videos. the method will be to upload a video to the web application then it will be passed through several machine learning approaches including feature extraction, classification and filtering. After the process is done the output will show if the video is real or not. For sequence length 10, 20 and 40 our model gave accuracy of 83.456%, 87.349% and 88.965% respectively. For sequence length 40 the highest accuracy was achieved. The model can capture irregularities in both space and time domain by combining RNNs for temporal analysis with InceptionV3 for spatial feature extraction. But combining deep learning models will result to higher computational complexity. There are concerns with the model's generalizability because its performance on unknown data or different type of fake videos is not covered.

In 2024 K. Vyshegorodtsev et al. [28] tackled the growing problem of video deepfakes. The model was designed to work on videos with multiple faces which is common in real world scenarios like online meetings. To improve detection accuracy, they proposed a novel approach that uses geometric-fakeness features (GFF). Temporal discrepancies between frames are captured by these features. After processing these temporal patterns, a deep learning model is trained to make a final deepfake prediction. This works well for videos with multiple faces like those from internet video conferencing. Many experiments were performed on well-known datasets like FaceForensics++, DFDC, Celeb-DF, and WildDeepFake. According to the results the proposed method outperformed the state-of-the-art methods. On the DFDC dataset it got 0.98 accuracy and 0.91 f-score. The focus on temporal anomalies in GFF allows to detect small deepfakes that are ignored by methods that only look at individual frames. However, there is additional computational cost to extract GFF and process temporal patterns which can be a problem for real-time applications.

In 2024 M. Liao et al. [29] proposed a two-stage feature extraction framework to improve the feature representation over single-branch methods. The framework consisted of a frequency domain feature extraction branch and a spatial domain feature extraction branch based on EfficientNet. The collected features are then combined with the Transformer's encoder structure and cross-attention mechanism to simulate feature correlations across global regions. The detection accuracies for Deepfake, Face2Face and NeuralTextures forged images are 83.5%, 70.25%, and 78.5%, respectively. uses Transformer architecture and advanced feature extraction methods to get high accuracy on various forged image formats while facing the problem of amount of data and computing.

In 2025 E. Tchaptchet et al. [30] presented a way to detect deepfakes by analyzing iris of the eyes. You can visualize the biological features of the iris by applying a gradient map to the iris. They suggest that you can differentiate between real and modified content by looking at specific iris features such as size, shape and reflection patterns. Using the Flickr-Faces-HQ (FFHQ) dataset and StyleGAN2 images they showed that this algorithm achieved a great detection accuracy of 0.979 and sensitivity of 0.921. Compared to methods that analyze the whole face, focusing on specific iris features may reduce the complexity. But focusing only on the iris may miss other face irregularities present in deepfakes.

III. METHODOLOGY

This section introduces the proposed method with an algorithm and block diagram for deep fake video detection based on spatiotemporal analysis combining spatial and temporal features.

The proposed method detects deep fake videos by combining spatial and temporal analysis using a pre-trained VGG16 model on the DFDC dataset (DeepFake Detection Challenge). First, frame extraction and preprocessing. A certain number of frames (e.g. 16 frames per video) are sampled for each video. Then normalization and downsize each frame to a lower resolution 64×64, then face detection to focus on the region of interest.

Then extract spatial features from each video frame using pre-trained VGG16 model. Deep spatial information of each frame including textures, edges and other visual patterns are extracted by this model. Equation (1) represents extracting spatial features for each frame I_t.Where R is the set of real numbers. H, W, and C represent, respectively, tensor shape height, width and depth.

$$F_t = fvgg16(I_t) \in R^{H \times W \times C} \tag{1}$$

After that, global average pooling on every frame as shown in Equation (2). Then global average pooling on temporal features along the time dimension. So, the whole video is represented by a single fixed dimension vector. Aggregate the frame level features as shown in Equation (3), Where N represents the number of frames.

$$S_t = GAP(F_t) \in R^C \tag{2}$$

$$v = {}^{1}/_{N} \sum_{t=1}^{N} S_t \in R^C \tag{3}$$

Finally, a fully connected classifier with dropout regularization receives the aggregated video level information and produces a binary prediction of real or fake video. Binary crossentropy loss is used to train the model and class weighting is used to address class imbalance. Accuracy and AUC is used to evaluate the performance. Equation (4) represents classification, whereas σ is sigmoid function, b is bias and W are learnable weights.

$$y = \sigma\,(W^t\, v + b) \tag{4}$$

To produce a powerful video level representation this combines temporal aggregation with the spatial feature extraction capabilities of a pre-trained VGG16 model. This representation is then used by the classifier to decide if the input video is real or a deepfake. When frame extraction, spatial analysis and temporal aggregation is combined a full system for deepfake video detection is produced. The advantages of their method, memory efficiency with smaller frame sizes, variable frame duration with pooling and transfer learning with VGG16. *Fig. 1* illustrates the block diagram of the proposed method.

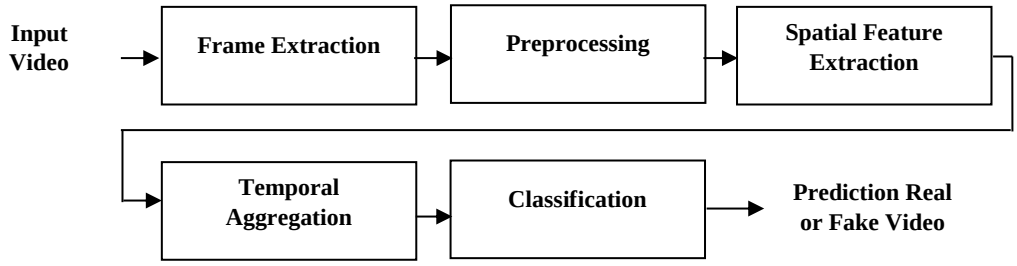

FIG.1. BLOCK DIAGRAM OF THE PROPOSED METHOD.

Algorithm below is the proposed method for deep fake video detection based on spatiotemporal analysis.

Algorithm 1. Proposed deepfake video detection based on spatiotemporal analysis methods.

Input: Video file.
Output: Prediction labels (0) real, (1) fake.
Process:
Step 1: Load video and DFDC dataset.
Step 2: For each video.
Step 3: Extract frames.
Step 4: For each frame:
Step 4.1: Preprocessing by apply resize, normalize and face detection.
Step 4.2: Extract spatial features by VGG16 model.
Step 4.3: Use global average pooling to get a fixed-length feature vector for every frame.
Step 4.4: Sequence feature vector of frame-level.
Step 5: Aggregate the sequence over the temporal dimension utilizing global average pooling.
Step 6: Represent video level by single vector of fixed dimensions.
Step 7: Classification using sigmoid layers for binary classification after fully connected layers.
Step 8: Apply prediction 1 if prediction > 0.5 else 0.
Step 9: Return prediction (0) real, (1) fake.
Step 10: End.

IV. RESULTS and DISCUSSION

This section shows the experimental results of the proposed method. The accuracy and AUC measures are used to evaluate the proposed deepfake detection method.

AUC stands for Area Under the Receiver Operating Characteristic (ROC) Curve. These metric measures how well binary classification models can separate the two classes at all thresholds. An AUC of 1.0 is perfect discrimination, 0.5 is no discriminative power.

The DFDC dataset (DeepFake Detection Challenge) is a large video dataset created for the deepfake detection problem. It has a large and hard dataset to train models to detect manipulated videos. It's a benchmark to measure how well algorithms can tell real vs fake videos. There are over 3,000 identities and almost 100,000 videos. Videos are high quality (1080p) and 30fps. Changing the faces of the actors is the main purpose of deepfakes.

The proposed method detects deep fake videos by combining spatial and temporal analysis using a pre-trained VGG16 model on the DFDC dataset that gives above 95% accuracy and AUC 0.81. *Fig. 2* shows the AUC-ROC curve of the proposed method.

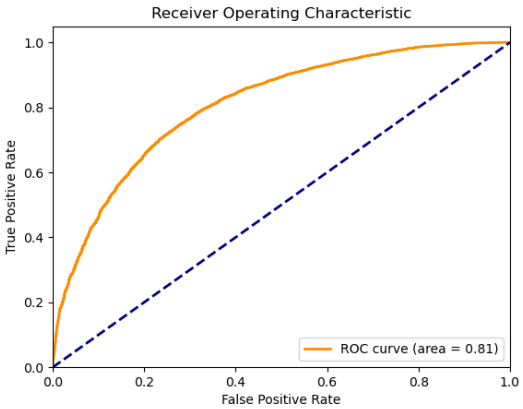

FIG. 2. AUC-ROC CURVE OF THE PROPOSED METHOD.

Table I illustrate the comparison of the different video deep fake detection methods to the proposed method, focusing on the methodology, datasets and accuracy of each method.

TABLE I. COMPARISON OF VARIOUS VIDEO DEEP FAKE DETECTION METHODS ON THE METHODOLOGY, DATASETS AND ACCURACY

Reference	Methodology	Dataset	Accuracy
[24]	Used hybrid deep learning models that combine RNNand CNN to detect fake videos.	DFDC dataset.	SimpleRNN and GRU both have 98.5% and 98.9% detection accuracy respectively.
[25]	combines CNN with Haar cascade classifiers.	DFDC dataset.	The accuracy of the binary classification into true or fake faces was almost 99%.
[26]	uses LSTM and CNN to distinguish between real and fake video frames.	———	———
[27]	Used both space and time domain by combining RNNs for temporal analysis with InceptionV3 for spatial feature extraction.	———	For sequence length 10, 20 and 40 our model gave accuracy of 83.456%, 87.349% and 88.965% respectively.
[28]	Used geometric-fakeness features. After processing these temporal patterns, a deep learning model is trained to make a final deepfake prediction.	FaceForensics++, DFDC, Celeb-DF, and WildDeepFake datasets.	On the DFDC dataset it got 0.98 accuracy.
[29]	Used Transformer architecture and advanced feature extraction methods that consisted of a frequency and a spatial domain.	———	The detection accuracies for Deepfake, Face2Face and NeuralTextures forged images are 83.5%, 70.25%, and 78.5%, respectively.
[30]	Used biological features of the iris by applying a gradient map to the iris to detect deepfakes.	FFHQ dataset and StyleGAN2 images.	Achieved a great detection accuracy of 0.979.
The proposed method	Used hybrid approach that combines a pre-trained deep model with temporal aggregation.	DFDC dataset.	Achieved an accuracy of 95% and 0.81 AUC.

As indicated in the table above, which illustrates of various video deep fake detection methods, they are all diverse in terms of methodology, datasets and accuracy. Each has different approaches, showing the many ways to do deepfakes. To capture the spatial and sequential properties of deepfakes, methods use Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks. For example, [24] uses a CNN-RNN hybrid deep learning model (SimpleRNN and GRU) and gets 98.5% and 98.9% respectively. Similar to this, [26] processes video frames with CNN-LSTM and highlights the importance of temporal consistency in deepfakes. Others have new feature extraction methods. [25] gets 99% with CNN and Haar cascade classifier. To increase robustness, [27] combines InceptionV3 with RNN and tests both spatial and temporal domains. Accuracy increases with sequence length. [29] tries transformer-based architectures with spatial and frequency domain feature extraction and gets good results for various types of manipulated images. [28] uses geometric-fakeness features from temporal patterns. A very innovative

one [30] uses gradient map to analyze iris features to focus on biological traits. With 97.9% detection accuracy, this shows how well human biometric features can be used.

Many methods use the DFDC dataset, a standard for deepfake detection, so we have a fair comparison. By testing on multiple datasets: FaceForensics++, DFDC, Celeb-DF, WildDeepFake, [28] expands the range and generalization ability. [30] uses FFHQ and StyleGAN2 images, so it's more focused on generating face images rather than deepfake videos.

Above results show that hybrid methods that combine CNNs and RNNs (or LSTMs) usually perform well, often above 98%. CNNs with Haar cascade classifiers are used in [25] which has the highest accuracy of 99%. Geometric-fakeness approach [28] and biological iris-based detection [30] also have impressive results, so using individual feature extraction with deep learning can greatly improve deepfake detection. This comparison shows the effectiveness of hybrid deep learning methods, especially those that combine temporal and spatial information.

V. CONCLUSIONS

In this paper we presented a hybrid deep fake detection method that combines spatial and temporal analysis by using a pre-trained VGG16 model with global average pooling. By extracting deep spatial features from individual frames and aggregating them temporally our method captures the fine grain visual details and the dynamic inconsistencies of deep fakes. Results on the DFDC dataset show our method achieves high accuracy and robustness. This work shows the potential of using transfer learning and temporal aggregation to tackle the evolving deep fake challenge. Future work will focus on improving temporal modeling with advanced attention mechanisms and exploring additional multimodal cues to generalize across different deep fake methods.

REFERENCES

[1] S. K. Pandey, L. Kumar, G. Kumar, A. Kumar, K. U. Singh and T. Singh, "An Overview of Video Tampering Detection Techniques: State-of-the-Art and Future Directions," 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2023, pp. 171-175, doi: 10.1109/CISES58720.2023.10183511.

[2] M. T. Jafar, M. Ababneh, M. Al-Zoube and A. Elhassan, "Forensics and Analysis of Deepfake Videos," 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 053-058, doi: 10.1109/ICICS49469.2020.239493.

[3] U. Patil and P. M. Chouragade, "Deepfake Video Authentication Based on Blockchain," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2021, pp. 1110-1113, doi: 10.1109/ICESC51422.2021.9532725.

[4] P. Joshi and N. V, "Deep Fake Image Detection using Xception Architecture," 2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST), Jamshedpur, India, 2024, pp. 533-537, doi: 10.1109/ICRTCST61793.2024.10578398.

[5] R. Chauhan, R. Popli and I. Kansal, "A Comprehensive Review on Fake Images/Videos Detection Techniques," 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2022, pp. 1-6, doi: 10.1109/ICRITO56286.2022.9964871.

[6] A. Singh, R. Bharne, R. Kadu, P. B. Dasarwar and G. Buddhawar, "Impact of Deep Learning Techniques on Deep Fake Image Identification for Digital Investigation," 2024 International Conference on Modeling, Simulation & Intelligent Computing (MoSICom), Dubai, United Arab Emirates, 2024, pp. 325-329, doi: 10.1109/MoSICom63082.2024.10881036.

[7] M. Carter, M. Tsikerdekis and S. Zeadally, "Approaches for Fake Content Detection: Strengths and Weaknesses to Adversarial Attacks," in IEEE Internet Computing, vol. 25, no. 2, pp. 73-83, 1 March-April 2021, doi: 10.1109/MIC.2020.3032323.

[8] D. H. Choi, H. J. Lee, S. Lee, J. U. Kim and Y. M. Ro, "Fake Video Detection With Certainty-Based Attention Network," 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 2020, pp. 823-827, doi: 10.1109/ICIP40778.2020.9190655.

[9] M. Kavitha and R. Gayathri, "Joint Spectral-Spatial Feature Using Deep 3-D CNN for Hyperspectral Images," 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC), Chennai, India, 2022, pp. 281-285, doi: 10.1109/ICESIC53714.2022.9783563.

[10] L. Qiu, W. Qin, H. Yang and Y. Chen, "HRNet: Local-Spatial Feature Fusion Network for Texture Recognition," 2024 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 2024, pp. 535-540, doi: 10.1109/IPEC61310.2024.00097.

[11] P. Liu and X. Wang, "Spatial Features Acquisition for Airplane Target Detection in Hyperspectral Remote Sensing Using Virtual RGB Images," in IEEE Geoscience and Remote Sensing Letters, vol. 21, pp. 1-5, 2024, Art no. 5509405, doi: 10.1109/LGRS.2024.3451691.

[12] W. Wang, Y. Yuan and D. Ma, "Adaptive Spectral and Spatial Feature Extraction Framework for Hyperspectral Classification," 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 2021, pp. 3629-3632, doi: 10.1109/IGARSS47720.2021.9554853.

[13] F. Zhou and B. Zhao, "Local Feature Descriptor Construction Technique Based on Point Pair Features and Spatial Features and Its Quantitative Evaluation," 2024 4th International Conference on Neural Networks, Information and Communication Engineering (NNICE), Guangzhou, China, 2024, pp. 1802-1807, doi: 10.1109/NNICE61279.2024.10498236.

[14] H. Li, W. Yang and Q. Liao, "Temporal Feature Enhancing Network for Human Pose Estimation in Videos," 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 579-583, doi: 10.1109/ICIP.2019.8803783.

[15] Y. Li, X. Wang, Y. He, Y. Wang, Y. Wang and S. Wang, "Deep Spatial-Temporal Feature Extraction and Lightweight Feature Fusion for Tool Condition Monitoring," in IEEE Transactions on Industrial Electronics, vol. 69, no. 7, pp. 7349-7359, July 2022, doi: 10.1109/TIE.2021.3102443.

[16] O. Ye, T. Liu, Y. Fu, J. Deng, J. Feng and Y. Zhang, "The Video Captioning Method Based on The Spatial- Temporal Information and Attention Mechanism," 2021 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Xi'an, China, 2021, pp. 1-5, doi: 10.1109/ICSPCC52875.2021.9564796.

[17] J. -T. Lee and S. Yun, "Multi-Scale Temporal Feature Fusion for Few-Shot Action Recognition," 2023 IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 2023, pp. 1785-1789, doi: 10.1109/ICIP49359.2023.10223132.

[18] Z. Dong, "Fast Action Recognition Based on Local and Nonlocal Temporal Feature," 2021 IEEE 4th International Conference on Information Systems and Computer Aided Education (ICISCAE), Dalian, China, 2021, pp. 196-200, doi: 10.1109/ICISCAE52414.2021.9590691.

[19] J. Tao, Y. Gu, J. Sun, Y. Bie and H. Wang, "Research on vgg16 convolutional neural network feature classification algorithm based on Transfer Learning," 2021 2nd China International SAR Symposium (CISS), Shanghai, China, 2021, pp. 1-3, doi: 10.23919/CISS51089.2021.9652277.

[20] R. H. K, R. L. S, M. K. M, B. K, F. P. B. A and T. T, "Study on VGG16 Transfer learning Model for Goat/Sheep Image Classification," 2023 International Conference on Recent Advances in Science and Engineering Technology (ICRASET), B G NAGARA, India, 2023, pp. 1-8, doi: 10.1109/ICRASET59632.2023.10420202.

[21] G. Singh, K. Guleria and S. Sharma, "A Transfer Learning-based Pre-trained VGG16 Model for Skin Disease Classification," 2023 IEEE 3rd Mysore Sub Section International Conference (MysuruCon), HASSAN, India, 2023, pp. 1-6, doi: 10.1109/MysuruCon59703.2023.10396942.

[22] A. Kaur, V. Kukreja, M. Kumar, A. Choudhary and R. Sharma, "A Fine-tuned Deep Learning-based VGG16 Model for Cotton Leaf Disease Classification," 2024 5th International Conference for Emerging Technology (INCET), Belgaum, India, 2024, pp. 1-4, doi: 10.1109/INCET61516.2024.10593164.

[23] D. Kusumawati, A. A. Ilham, A. Achmad and I. Nurtanio, "Vgg-16 And Vgg-19 Architecture Models In Lie Detection Using Image Processing," 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 2022, pp. 340-345, doi: 10.1109/ICITISEE57756.2022.10057748.

[24] O. S. A. Aboosh, A. N. Hassan and D. K. Sheet, "Fake Video Detection Model Using Hybrid Deep Learning Techniques," 2023 6th International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, pp. 499-504,2023. doi: 10.1109/ICOIACT59844.2023.10455952.

[25] A. Jellali, I. Ben Fredj and K. Ouni, "An Approach of Fake Videos Detection Based on Haar Cascades and Convolutional Neural Network," 2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET), Hammamet, Tunisia, pp. 01-06,2023. doi: 10.1109/IC_ASET58101.2023.10150604.

[26] F. Mira, "Deep Learning Technique for Recognition of Deep Fake Videos," 2023 IEEE IAS Global Conference on Emerging Technologies (GlobConET), London, United Kingdom, pp. 1-4,2023. doi: 10.1109/GlobConET56651.2023.10150143.

[27] N. Nibras, S. Fahim, S. Sakib, S. U. Rashid and A. Rahman, "An Efficient Algorithm for Fake Video Detection," 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), Dhaka, Bangladesh, pp. 1337-1343,2024. doi: 10.1109/ICEEICT62016.2024.10534423.

[28] K. Vyshegorodtsev, D. Kudiyarov, A. Balashov, and A. Kuzmin "Deepfake detection in videos with multiple faces using geometric-fakeness features", arXiv, 2024. doi:10.48550/arxiv.2410.07888.

[29] M. Liao and M. Chen," A new deepfake detection method by vision transformers", In International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2024), vol. 13403, pp. 953-957. SPIE, 2024. Doi: 10.1117/12.3051840.

[30] E. Tchaptchet, E. Fute Tagne, J. Acosta, D. B. Rawat and C. Kamhoua, "Deepfakes Detection by Iris Analysis," in IEEE Access, vol. 13, pp. 8977-8987, 2025, doi: 10.1109/ACCESS.2025.3527868.