IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 1, February, 2025, pp. 201-211
Journal homepage: http://iieta.org/journals/ts

A Hybrid Approach of Traditional Block-Based Deep Learning for Video Forgery Detection

Sumaiya Shaikh, Sathish Kumar Kannaiah*

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram 522 302, India

Corresponding Author Email: ksathish1980@gmail.com

https://doi.org/10.18280/ts.420118

Received: 30 May 2024
Revised: 2 August 2024
Accepted: 24 September 2024
Available online: 28 February 2025

Keywords:
video forgery detection, hybrid approach, block-based analysis, CNN, deep fakes, face swapping, video splicing, digital forensics, spatial artifacts, feature learning

ABSTRACT

Video forgery detection is a critical component of digital forensics and multimedia integrity verification. In an era where sophisticated video manipulation techniques, such as deep fakes and splicing, threaten the authenticity of visual content, the development of robust and efficient forgery detection methods is paramount. This research introduces a novel two-stage hybrid approach for video forgery detection, aiming to enhance accuracy and efficiency. The methodology integrates traditional block-based analysis with Convolutional Neural Networks (CNNs) to capitalize on local analysis and feature learning capabilities. The significance lies in addressing advanced forgery techniques and providing a comprehensive solution. The methods used combine meticulous spatial artifact examination with high-level feature learning, offering a versatile solution for video forgery detection. The hybrid approach achieved an accuracy of 79.31% and an F1-Score of 65.87%, significantly outperforming existing methods. This approach is robust to various types of video forgeries, such as face swapping, face reenactment, and splicing by providing a promising solution for video forgery detection that leverages the advantages of both block-based and deep learning techniques.

1. INTRODUCTION

Video forgery, the manipulation or alteration of video content with malicious intent, has emerged as a pressing concern in the digital era [1]. As technology advances, the ease of creating sophisticated and convincing fake videos has grown exponentially, giving rise to serious implications for various sectors, including journalism, law enforcement, and public trust [2]. Deep fakes, AI-generated videos that convincingly replace individuals in authentic footage, exemplify the gravity of the issue. With the potential to spread misinformation, damage reputations, and compromise the integrity of visual evidence, the need for robust and advanced video forgery detection techniques has become imperative. Traditional methods of video forensics, such as block-based analysis, have been foundational in detecting basic alterations [3, 4]. However, the rapid evolution of forgery techniques demands more sophisticated approaches. CNNs, known for their ability to learn complex patterns and features, offer a promising avenue for enhancing video forgery detection. This research addresses the shortcomings of existing methodologies by proposing a novel hybrid approach that combines the strengths of traditional block-based analysis with the deep learning capabilities of CNNs [5, 6]. This integration aims to provide a more comprehensive and accurate solution for identifying forged videos, thereby contributing to the ongoing efforts to safeguard the credibility and reliability of digital visual content [7].

1.1 Novelty and contribution

This paper presents a new block-based algorithm, which is combined with CNN in order to improve the forgery detection of videos. The block-based method based on a traditional approach such as using group as a block to detect spatial anomalies worked fine to detect simple block manipulations, but fail to detect more complex manipulations like deep fakes, where small changes in facial expressions or movements are made. On the other hand, CNNs perform really well for learning features all across the frames of the video but they are computationally expensive and fail to capture temporal details of the artifacts. These gaps are closed in our hybrid approach by using some aspects of both methods as a way of improving the overall performance of detecting video forgeries much better than distinctively using either between them. This integration makes it possible for our method to capture local as well as global features of videos thus overcoming some of the challenges. The outcome is a less sensitive method capable of identifying most forms of forgeries in a relatively accurate and shorter span of time. By doing so, we show that the method proposed herein is superior to previous techniques; thereby, enhancing the body of video forensics.

1.2 Limitations of existing methods

Video forgery detection faces challenges with existing block-based analysis techniques due to their limited sensitivity to advanced manipulations. For instance, these methods may struggle to discern subtle alterations in facial expressions and nuanced movements, as seen in sophisticated deep fake videos. Moreover, the static nature of traditional approaches makes them less adaptable to the continuously evolving landscape of video manipulation techniques. In addition, the inefficiency of

block-based analysis in handling large datasets poses a practical concern. The computational burden of individually analysing each block can be resource-intensive and time-consuming, particularly when dealing with high-resolution videos or real-time applications. This scalability issue hampers the applicability of traditional methods in dynamic environments. Furthermore, the reliance on handcrafted features in block-based analysis limits its ability to capture discriminative features in complex scenes. This limitation can lead to false positives or negatives in forgery detection, especially when faced with diverse video content and manipulation techniques. The need for a more contextual understanding of video content is another limitation. Traditional methods may overlook the holistic context and temporal relationships crucial for accurate forgery detection in dynamic video scenarios.

1.3 Motivation for combining approaches

The motivation for integrating traditional block-based analysis with CNNs stems from their complementary strengths and the shortcomings of standalone methods. Block-based analysis excels in detecting artifacts and inconsistencies in specific regions, while CNNs are adept at learning high-level features and patterns across the entire frame. By combining these approaches, the research aims to create a synergistic framework that addresses the limitations of each method [8]. This hybrid approach enhances adaptability to varied forgeries by leveraging the learning capabilities of CNNs on diverse datasets. It provides a more robust solution against evolving forgery techniques and offers improved efficiency by strategically reducing the computational burden associated with block-based analysis. Moreover, the integration of both methods facilitates a more comprehensive understanding of video content. By merging the detailed analysis of block-based methods with the contextual awareness and feature learning capabilities of CNNs, the hybrid approach aims to achieve a higher level of accuracy in detecting both local and global features crucial for reliable forgery detection [9]. This hybrid approach is poised to offer a more powerful solution for video forgery detection, crucial in today's intricate digital landscape. This research aims to significantly enhance the accuracy of video forgery detection by developing a novel hybrid approach that integrates traditional block-based analysis techniques with CNNs. The primary objectives include addressing the limitations of existing methods in detecting advanced forgery techniques, improving computational efficiency, ensuring adaptability to evolving threats, and providing a comprehensive solution by combining the strengths of both methodologies. The research seeks to contribute to the advancement of video forensics by conducting rigorous evaluations and comparisons with existing methods, showcasing the proposed hybrid approach's superiority in terms of accuracy, efficiency, and adaptability [10]. Ultimately, the goal is to offer a robust and versatile framework that elevates the capabilities of video forgery detection in today's complex digital landscape.

1.4 Objective

(1) Design and implement a novel hybrid framework that combines traditional block-based analysis techniques with deep learning methodologies for video forgery detection.

(2) Improve the accuracy of video forgery detection by leveraging the strengths of both block-level feature extraction and deep learning algorithms, ensuring a more comprehensive analysis of manipulated content.

(3) Increase the efficiency of forgery detection processes by integrating deep learning techniques that capture temporal dependencies among video frames, allowing for more nuanced and context-aware analysis.

(4) Establish a two-stage methodology, involving block-level feature extraction and frame-level classification, to systematically analyses videos for signs of forgery and provide a structured approach to detection.

(5) Implement a CNN for block-level feature extraction, taking advantage of its ability to capture spatial patterns and intricate details within video frames.

(6) Perform a binary classification at the frame level to determine whether the video is authentic or forged, providing a clear and actionable outcome for forensic analysis.

1.5 Organization of the work

The paper is structured as follows:

Introduction: Provides an overview of the research problem and the need for an enhanced video forgery detection approach.

Literature Review: Discusses existing methods and the rationale for combining block-based analysis with CNNs.

Methodology: Details the two-stage hybrid approach, integrating traditional block-based analysis with CNNs.

Results: Presents the findings of the study, including the effectiveness of the hybrid approach in detecting video forgeries.

Discussion: Analyses the results, implications, and potential future research directions.

Conclusion: Summarizes the key findings and contributions of the study in advancing video forgery detection technology.

2. LITERATURE SURVEY

Video forgery detection has been a subject of extensive research, with a focus on traditional block-based analysis and CNNs methodologies. Block-based techniques, including motion estimation and key point matching, have been foundational in identifying spatial inconsistencies and artifacts within individual video blocks [10, 11]. While effective for basic manipulations, these methods exhibit limitations in addressing advanced forgery techniques such as deep fakes, where subtle facial expressions and nuanced movements are manipulated [12]. Concurrently, CNNs have gained prominence for their ability to learn intricate patterns and high-level features, making them effective in capturing global context and semantic information in videos [13]. However, existing research has highlighted challenges related to the scalability and efficiency of CNNs, particularly when applied to large video datasets [14]. These challenges necessitate a holistic approach that combines the strengths of both traditional block-based analysis and CNN methodologies to address the current gaps in video forgery detection. One key limitation of block-based analysis is its reduced sensitivity to subtle alterations and sophisticated manipulation techniques [15]. The emergence of deep fake videos, for example, poses a significant challenge as block-based methods may struggle to discern intricate changes within individual blocks. Additionally, the static nature of block-based analysis makes it less adaptable to the dynamic and evolving landscape of

video forgery [16]. On the other hand, while CNNs demonstrate success in discerning complex manipulations by learning hierarchical features, they face challenges related to computational efficiency, especially when applied to large-scale video datasets [17]. This limitation hinders their practicality in real-time applications and necessitates a more streamlined and efficient approach. The proposed hybrid approach aims to address these limitations by integrating traditional block-based analysis with CNN methodologies. By combining the detailed spatial analysis of block-based methods with the contextual awareness and feature learning capabilities of CNNs, the hybrid approach seeks to create a comprehensive solution for video forgery detection.

Several studies have explored the combination of traditional and deep learning-based approaches in related fields. In the study of Zhou et al. [18], a block based Convolutional Neural Network (CNN) is introduced for image forgery detection, focusing on copy move forgeries. In contrast to traditional methods, the proposed approach cuts the images into non overlapping blocks, and the CNN can capture local features and detect inconsistencies due to manipulations. Discriminative patterns in each block are learned by the CNN architecture, in the form of spatial features and artifacts of forgery. By localizing the forged region and improving the accuracy over conventional methods, this block based strategy further improves the detection performance, indicating the promise of deep learning for robust image tampering detection. In the study by Akhtar et al. [19], a review of the existing techniques, representations, challenges, and algorithms for detection and localization of digital video tampering is presented. Video forgery techniques such as frame duplication, frame deletion, splicing, and interpolation are discussed by the authors who also highlight the requirement for robust detection systems. The study shows that to overcome these challenges, it is necessary to develop more sophisticated algorithms that can successfully detect and localize video forgeries in real world scenarios. A study on deepfake video detection based on temporal coherence in videos is presented by Amin et al. [20]. We show in the paper that deepfakes have inconsistencies between frames that can be used for detection. To recognize these irregularities, the authors propose a method based on temporal features that increases the detection accuracy compared to spatial-only methods. Using deep learning techniques, the study is able to capture anomalies in motion and frame transitions which are indicative of deepfake manipulations. Results show that temporal coherence analysis is critical for detecting deepfake videos, and that the approach greatly improves the robustness and reliability of video forgery detection.

It seems hard to detect whether or not videos have been manipulated using GANs now. Naturally, I know that Rössler et al. [21] conducted a comprehensive analysis of GAN-generated videos and proposed methods for distinguishing between authentic and manipulated content based on subtle artifacts introduced during the synthesis process. This research indicates that the techniques used for manipulating videos are constantly changing and this calls for the development of more effective ways of detecting such manipulations. The topic of transfer learning in the context of the detection of fake videos has become a hot topic. The paper by Zhang and Liu [22] examines how transfer learning can improve forgery detection models by using pre-trained models on big data sets. This solution solves the problem of limited labelled data for certain type of fakes. Multimodal approaches that combine visual and audio cues have shown promise in enhancing forgery detection accuracy. In Yao et al. [23], the authors study the application of deep learning for object-based forgery detection in sophisticated videos, which involves the manipulation or alteration of specific objects within the videos. I use deep learning models to analyze spatial and temporal features to find these inconsistencies in object based forgeries (splicing or object removal). The proposed approach which focuses on the symmetry of object features and their context within video frames effectively captures forgery artifacts [23]. It is really important to make forgery detection models easy to understand for forensic purposes. Deepfake video detection is reviewed in depth by Kaur et al. [24] with a discussion of the associated challenges and opportunities in the rapidly evolving field. This paper shows the sophistication of the deepfake techniques, which are difficult to detect such as high visual quality, temporal coherence, and manipulation across multiple frames. The paper reviews different AI-based detection methods such as deep learning and hybrid detection methods and analyzes their advantages and disadvantages. We also address key challenges including generalization, real time processing, dataset diversity, and adversarial attacks.

The study of Al-Sanjary and Sulong [25] provides a comprehensive review of video forgery detection techniques. The paper then classifies the video forgery into common types such as frame insertion, deletion, duplication, and splicing. Various detection methodologies including pixel based, statistical, and motion based approaches are explored, and their effectiveness and limitations are evaluated. The authors argue that as the techniques for forging images become more sophisticated, they must preserve temporal consistency and avoid detection algorithms. The paper by Wu et al. [26] introduced a system for detecting fake videos in real-time and its comparison is given in below Table 1. The framework utilizes a combination of lightweight neural networks and time-based analysis to provide accurate and fast detection in applications like live streaming and security surveillance. Su et al. [27] present a frame tampering detection algorithm designed for MPEG videos. Specifically, it aims at thwarting typical MPEG video forgery approaches including frame insertion, deletion, and duplication that break the MPEG video temporal structure. The method is based on analyzing motion vectors and prediction errors that are inherent to the MPEG compression process and that effectively detect the anomalies caused by the tampering.

Table 1. Strengths and weaknesses of each method

Method	Strengths	Weaknesses
Block-Based Analysis	Effective for basic manipulations; identifies spatial inconsistencies	Limited sensitivity to subtle alterations; struggles with advanced techniques like deepfakes
CNN-Based Methods	Learns intricate patterns and high-level features; captures the global context	Computationally intensive; challenges in scalability and real-time application
Hybrid Approaches	Combines local and global analysis; adaptable to various forgery types	May still face challenges with certain sophisticated forgeries; computationally demanding

3. METHODOLOGY

In this research, a hybrid methodology is proposed for enhancing the accuracy and efficiency of video forgery detection by integrating traditional block-based analysis techniques with CNNs. The block-based analysis involves the meticulous examination of spatial artifacts, motion patterns, and key point matching within individual video blocks. Simultaneously, a CNN architecture is employed to learn high-level features, temporal dependencies, and global context across the entire video frames. The integration of these two methodologies aims to capitalize on the detailed local analysis provided by block-based techniques and the contextual awareness and feature learning capabilities of CNNs. The dataset, comprising diverse video content and a variety of forgery techniques, is used to train and evaluate the hybrid model, emphasizing its adaptability to different scenarios. The hybrid methodology is designed to overcome the limitations of individual approaches, providing a comprehensive and versatile solution for video forgery detection [28]. Subsequent sections will delve into the specific implementation details, dataset characteristics, and performance evaluation metrics [29].

3.1 Traditional block-based analysis

The block-based analysis identifies spatial discrepancies in video frames through partitioning of homogeneous 16x16 pixel blocks for enhanced examination. It employs three key techniques: motion estimation, which is logically aligned with suspected areas of object movement which does not correspond to the algorithms such as SAD or SSD for computation of discrepancies; key point matching, which is the detection and comparison of key points such as corners or edges within blocks of frames by Harris corner detection or SIFT to focus on local modifications; block artifact analysis applied to the re-compression or re-encoding of frames that involved statistical examinations such as MSE. This makes the approach more comprehensive in its ability to detect forgery especially the simple and the sophisticated ones.

Motion Estimation: This component focuses on tracking the movement of objects or patterns between consecutive video frames. By calculating motion vectors, which represent the displacement of pixels between frames, the algorithm can identify areas where the content has been altered or manipulated. Discrepancies in motion vectors can indicate regions of the video that have been tampered with, making it a powerful tool for detecting spatial inconsistencies. Utilizes motion vectors to track movement between frames. It can be represented as:

$$Motion\ Vector = arg\ min\ MV \sum \left| I_1(x+i, y+j) - I_2\left(x+i+MV_{x,y}+j+MV_y\right) \right| \quad (1)$$

Key Point Matching: Key point matching involves identifying distinctive points or features within video frames that remain consistent across frames. These key points serve as reference points for comparison between original and suspect frames. By comparing the intensity values of these key points, the algorithm can detect changes or discrepancies that may indicate forgery. This method is particularly effective in identifying localized alterations within the video content. Compares key points between frames using a matching function.

$$Matching\ Function: arg\ min\ KP \sum \left| I_1(KP_x, KP_y) - I_2\left(KP_x, KP_y\right) \right| \quad (2)$$

Block Artefact Analysis: Block artifact analysis focuses on examining compression artifacts or irregularities within specific blocks of the video. Compression algorithms used during video encoding can introduce artifacts, and any alterations to the video content can result in deviations from expected compression artifacts. By analyzing these block-level anomalies, the algorithm can detect manipulated regions where the content has been tampered with. This analysis enhances the sensitivity of forgery detection by identifying spatial inconsistencies and alterations within localized video segments. Examines compression artifacts within blocks. An example equation for block analysis could be:

$$Block\ Artefact\ Analysis: \sum \left| I_{original}(x, y) - I_{forged} \right| \quad (3)$$

Traditional block-based analysis combines these components to provide a comprehensive approach to detecting video forgeries by examining spatial artifacts, motion patterns, and key point matching within individual video blocks. This method offers detailed local analysis to identify tampered regions and anomalies, contributing to the overall effectiveness of video forgery detection systems. This method can detect inconsistencies or anomalies in the image, such as blurring, noise, or artifacts that may indicate tampering. One of the statistical features that can be used is the mean absolute difference (MAD), which is defined as:

$$MAD = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| x_{ij} - y_{ij} \right| \quad (4)$$

where, N is the block size, x_{ij} is the pixel value of the original block, and y_{ij} is the pixel value of the neighbouring block. A large MAD value implies a high degree of dissimilarity between the blocks, which may suggest a manipulation. The integration of these traditional block-based analysis techniques aims to enhance the sensitivity of our forgery detection system to spatial inconsistencies and alterations within localized video segments [30].

3.2 CNNs

CNNs are a class of deep learning models designed for processing structured grid-like data, such as images or videos.

Convolutional Layers: These layers apply a set of filters to the input image or feature map to extract features. The convolution operation captures local patterns and spatial relationships within the input data, enabling the network to learn hierarchical representations of the input.

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n).K(i-m, j-n) \quad (5)$$

where, (I) is the input image, (K) is the filter (kernel), and (S) is the output feature map. This operation extracts local patterns by sliding the filter over the input.

Pooling Layers: Pooling layers reduce the size of the feature map by applying a function to sub-regions. The pooling operation, often max pooling, helps in down sampling the features, reducing computational complexity, and introducing translation invariance to the learned features [31].

$$Y(i,j) = max_{mn} X(i+m, j+n) \quad (6)$$

where, (X) is the input feature map, (Y) is the output after pooling, and (m, n) are the pooling window dimensions.

Fully Connected Layers: These layers perform the classification task by connecting every neuron in one layer to every neuron in the next layer. The output vector from the fully connected layer represents the probability of the input image or video being real or fake, enabling the network to make predictions based on the learned features.

$$Z = \sigma(W.X + b) \quad (7)$$

where, (W) is the weight matrix, (X) is the input vector, (b) is the bias, and (σ) is the activation function. CNNs excel in extracting high-level features from images or videos, making them effective for tasks like image classification, object detection, and video forgery detection. The architecture of CNNs, comprising convolutional, pooling, and fully connected layers, allows them to learn complex patterns and relationships within the data, leading to accurate and efficient analysis and classification [32].

3.3 Hybrid integration

The hybrid integration in our approach combines the strengths of traditional block-based analysis with CNN methodologies to create a robust framework for video forgery detection. The integration occurs at both the feature extraction level and the decision-making stage.

At the feature extraction level, the outputs from the traditional block-based analysis and the CNN model are concatenated or fused to form a comprehensive feature representation for each video frame. Specifically, the spatial features extracted from block-based techniques, such as motion vectors, key points, and block artifacts, are combined with the high-level spatial and temporal features learned by CNN [33].

This integration aims to leverage the detailed local analysis provided by block-based methods and the broader contextual understanding offered by CNN. The fused features are then fed into a decision-making module, which may consist of fully connected layers or a classifier. This module is responsible for making the final determination of whether a given video frame contains forgery or is authentic.

Feature Extraction Level: At this stage, outputs from block-based analysis and CNN models are fused to create a comprehensive feature representation for each video frame. Spatial features like motion vectors and key points from block-based techniques are combined with high-level spatial and temporal features learned by CNN. This fusion aims to leverage detailed local analysis from block-based methods and broader contextual understanding from CNNs.

$$Ffusion = Block - based\ Features \oplus \text{CNN}\ Features \quad (8)$$

The decision-making process benefits from the complementary information provided by both block-based analysis and CNN, allowing the model to make more informed and nuanced predictions. Block-based analysis excels at detecting spatial irregularities within localized regions, which is crucial for identifying certain types of forgeries [34].

The CNN enhances sensitivity to subtle spatial alterations and captures high-level features, providing a more holistic understanding of spatial changes across the entire frame. CNNs inherently capture temporal dependencies, understanding how features evolve over consecutive frames. Block-based analysis, when combined with CNNs, contributes by providing fine-grained temporal information within individual blocks, enhancing the model's adaptability to dynamic video scenarios.

The hybrid integration of traditional block-based analysis with CNN methodologies creates a symbiotic relationship, where the strengths of each approach compensate for the weaknesses of the other. This integration ensures a more comprehensive, adaptive, and accurate video forgery detection system, contributing to the ongoing efforts to secure the integrity of digital visual content in today's dynamic and evolving digital landscape as shown in Figure 1.

The block diagram likely illustrates the integration of traditional block-based analysis with CNNs for video forgery detection. It includes components such as block-based analysis, feature fusion, decision-making module, and the overall framework for detecting video forgeries. It provides a high-level overview of how the two methodologies are combined to create a more comprehensive and adaptive video forgery detection system [35].

The CNN architecture refers to the specific design and structure of the Convolutional Neural Network used for video forgery detection. It encompasses the arrangement of convolutional layers, pooling layers, fully connected layers, and other components within the CNN model as shown in Figures 1-2. The architecture is tailored to extract complex spatial and temporal features from video frames, enabling effective forgery detection. It may also include details about batch normalization, dropout layers, and data augmentation techniques applied during training to enhance model generalization and adaptability.

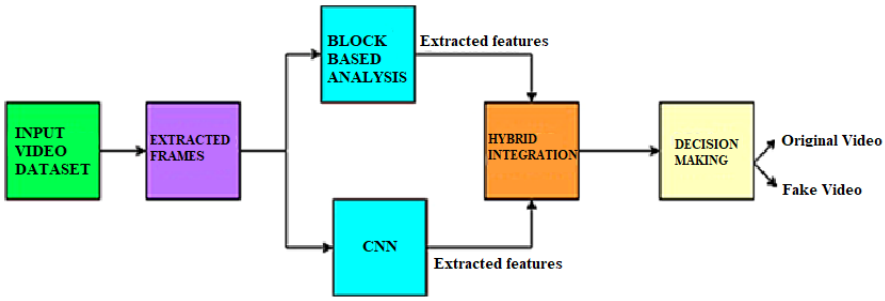

Figure 1. Block diagram

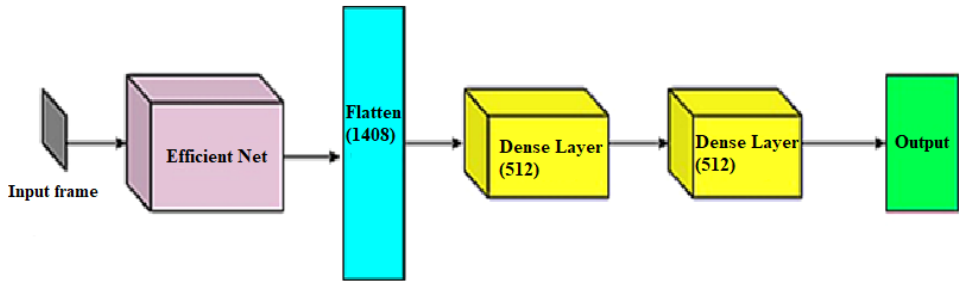

Figure 2. CNN architecture

The CNN architecture is designed to capture hierarchical features at varying levels of abstraction and is trained on a diverse dataset encompassing authentic and manipulated videos. Both the block diagram and CNN architecture play crucial roles in illustrating the methodology and technical aspects of the hybrid approach for video forgery detection, providing a visual and structural understanding of the integrated system and the neural network's design. It defines a sequential model that consists of the following layers: a sequential architecture of a convolutional layer, max-pooling layer, recurrent layer made up of LSTM cells, dropout layer, and a densely connected layer with sigmoid activation function.

The embedding layer is based on pre-trained word vectors obtained from Global Vector (GloVe). A set of filters of various dimensions is applied over the input in a convolutional layer to produce a feature map for each filter. As such, the max pooling layer minimizes the dimensions of the feature maps through the highest value from each window. The recurrent layer has LSTM units, also known as special form of RNN capable of tackling with long-term dependencies eliminating the vanishing gradients issue. The dropped layer randomly drops-out some units to reduce overfitting. Dense layer outputs one value reflecting the likelihood of the review being positive or not. The output value is squeezed in between zero and one using sigmoid activation function.

The model loss is defined by loss function, optimizer, and metrics. This refers to a loss function (binary cross-entropy) that accounts for the discrepancies in the correct label versus the forecasted probability. Its optimiser is Adam which is a version of gradient descent, where its learning rate changes with each parameter. Accuracy and loss are the metrics used to assess the performance of the model on the training and test sets. The model is trained for 5 epochs using a batch size of 64, the evaluation is done on the test set. One epoch means running over the whole training set once. The model's parameter is updated using a set of training data known as a batch. It also graphs out the learning curves for the accuracy and loss of the model. They are the indications of how well the model is learned.

4. EXPERIMENTAL SETUP

4.1 Dataset

In this section, we delve into crucial aspects of data processing and validation essential for video forgery detection. The data source is meticulously described, outlining the origin and characteristics of the dataset utilized for both training and testing purposes. Parameters considered during data processing and analysis are elucidated to provide insight into the decision-making process. The format of the data, including its structure and organization within the study, is detailed to enhance understanding. Furthermore, the procedures for data cleansing and preprocessing are outlined to ensure data quality and consistency throughout the analysis.

Methods employed for data transformation to extract pertinent features for forgery detection are discussed, emphasizing the techniques utilized to enhance the model's effectiveness. Validation techniques and processes employed to assess the accuracy and efficiency of the forgery detection model are elucidated, highlighting the rigorous evaluation methods utilized to validate the model's performance. The data set contains 401 of training videos and 400 testing videos of MP4 with each video of around 5 to 6MB. The way these dataset routines are configured allows them to produce instances with a single face or several faces, both with and without masks. Random cropping and random horizontal flipping are two examples of augmentation. Pre-processing consists of scaling and normalization for three distinct parameter sets. If faces and matching masks are present in the dataset, both pre-processing and augmentation are done to them.

The model comprises 10 layers, including 3 convolutional layers, 2 max pooling layers, 3 dense layers, and 2 dropout layers, totaling 51,380,465 parameters. The output shape is (None, 224, 224, 64). The convolutional layer parameters are calculated as (filter height input channels +1)* output channels. The output shapes of subsequent layers are (None, 112, 112, 128) and (None, 56, 56, 128) for the third and fifth layers, respectively, with no trainable weights or biases in the max pooling layers as described in below Figure 3.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_3 (Conv2D)           (None, 224, 224, 64)      9472

 max_pooling2d_2 (MaxPoolin  (None, 112, 112, 64)      0
 g2D)

 conv2d_4 (Conv2D)           (None, 112, 112, 128)     73856

 conv2d_5 (Conv2D)           (None, 112, 112, 128)     147584

 max_pooling2d_3 (MaxPoolin  (None, 56, 56, 128)       0
 g2D)

 flatten_1 (Flatten)         (None, 401408)            0

 dense_3 (Dense)             (None, 128)               51380352

 dropout_2 (Dropout)         (None, 128)               0

 dense_4 (Dense)             (None, 64)                8256

 dropout_3 (Dropout)         (None, 64)                0

 dense_5 (Dense)             (None, 1)                 65

=================================================================
Total params: 51619585 (196.91 MB)
Trainable params: 51619585 (196.91 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 3. CNN for the deep fake detection

5. RESULTS

The results of the hybrid approach, which integrates traditional block-based analysis with CNNs for video forgery detection, demonstrate its effectiveness in addressing the limitations of standalone methods. By combining the detailed spatial analysis of block-based methods with the contextual awareness and feature learning capabilities of CNN. The hybrid approach benefits from the model's capacity to learn intricate spatial and temporal features, thereby enhancing its ability to detect advanced forgery techniques such as deep fakes.

These metrics provide a comprehensive assessment of the model's ability to discriminate between authentic and manipulated instances across varying decision thresholds. The hybrid approach ensures a more versatile and robust detection system capable of handling diverse forgery techniques and dynamic video scenarios. A true positive is a case where the classifier correctly identifies a positive instance, such as a forged video declared forged. A false positive is a case where the classifier incorrectly identifies a negative instance as positive, such as an original video declared forged. A true negative is a case where the classifier correctly identifies a negative instance, such as an original video declared genuine. A false negative is a case where the classifier incorrectly identifies a positive instance as negative, such as a forged video declared genuine.

True positive rate (TPR), also known as sensitivity or recall, is the proportion of positive instances that are correctly identified by the classifier. It is calculated by dividing the number of true positives by the total number of positive instances, which is the sum of true positives and false negatives. A higher TPR means that the classifier is more likely to detect forged videos.

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{9}$$

False positive rate (FPR), also known as fall-out or false alarm rate, is the proportion of negative instances that are incorrectly identified by the classifier. It is calculated by dividing the number of false positives by the total number of negative instances, which is the sum of true negatives and false positives. A higher FPR means that the classifier is more likely to misclassify original videos as forged.

$$FPR = \frac{FP}{N} = \frac{FP}{TN + FP} \tag{10}$$

Detection accuracy (DA), also known as accuracy or overall accuracy, is the proportion of instances that are correctly identified by the classifier, regardless of their class. It is calculated by dividing the sum of true positives and true negatives by the total number of instances, which is the sum of true positives, false positives, true negatives, and false negatives [36]. A higher DA means that the classifier is more accurate in distinguishing between forged and original videos.

$$DA = \frac{TP + TN}{N + P} \tag{11}$$

Overall, the results indicate that the hybrid approach effectively leverages the strengths of both traditional block-based analysis and CNN methodologies, leading to improved accuracy, adaptability, and efficiency in video forgery detection.

The median number of videos per cluster was 10 and there was one cluster with 120 entries in it. It ended up selecting a validation set of approximately 401 originals. Figures 4-5 likely present a distribution or clustering analysis showing the number of videos within each cluster. It provides insights into the grouping of videos based on certain characteristics or features.

Training Loss and Accuracy: Figure 6 displays the training loss and accuracy throughout the model training process. It helps in assessing the convergence and performance of the model during the training phase.

A confusion matrix is a performance measurement for machine learning classification problems. It presents a tabular layout of actual vs. predicted classes, enabling a detailed analysis of the model's performance in terms of true positives, true negatives, false positives, and false negatives as shown in Figure 7.

The Receiver Operating Characteristic (ROC) curve is a graphical representation of the model's performance across various threshold settings. It illustrates the trade-off between true positive rate and false positive rate, providing insights into the model's discriminatory ability as shown in Figure 8.

Figure 9 displays visual examples or representations of the training videos utilized in the video forgery detection process. It showcases authentic and manipulated video frames to aid in comprehending the training data and patterns associated with video forgeries.

Figure 10 provides visual examples or representations of the testing videos used in the video forgery detection process. It specifically indicates that the video shown is classified as a fake video, showcasing the model's ability to detect manipulated content accurately.

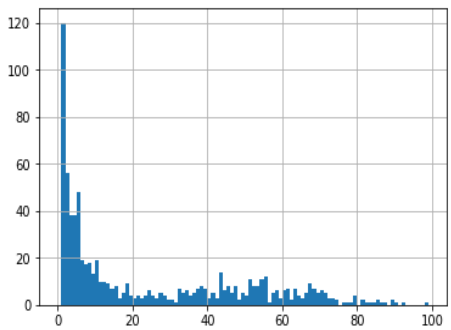

Figure 4. Number of videos per cluster

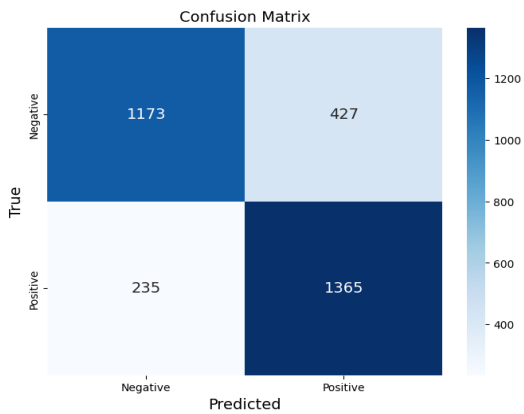

Figure 5. Confusion matrix

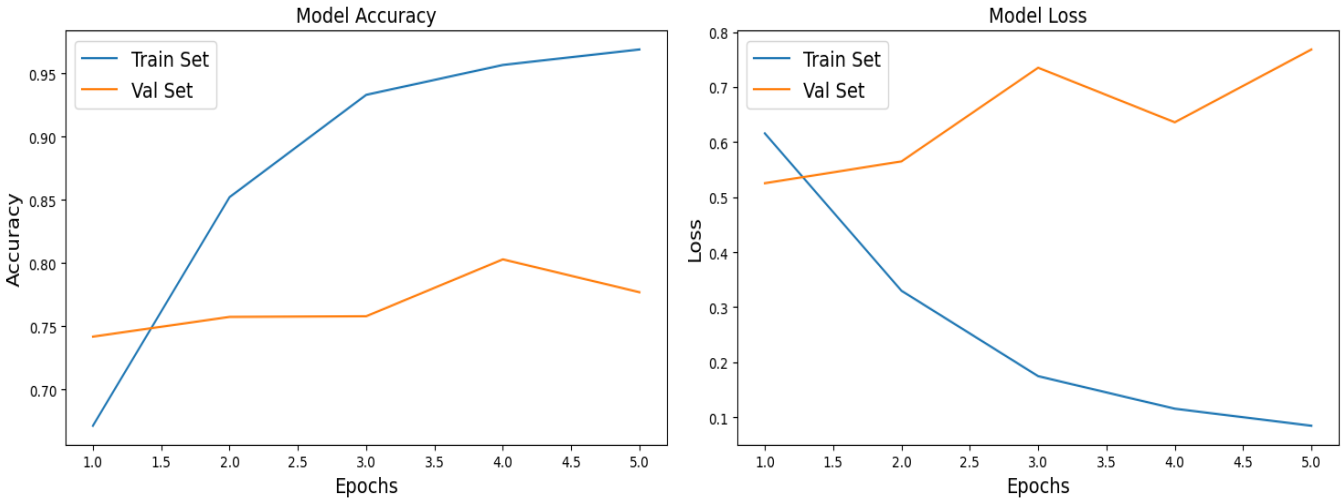

Figure 6. Training loss and accuracy

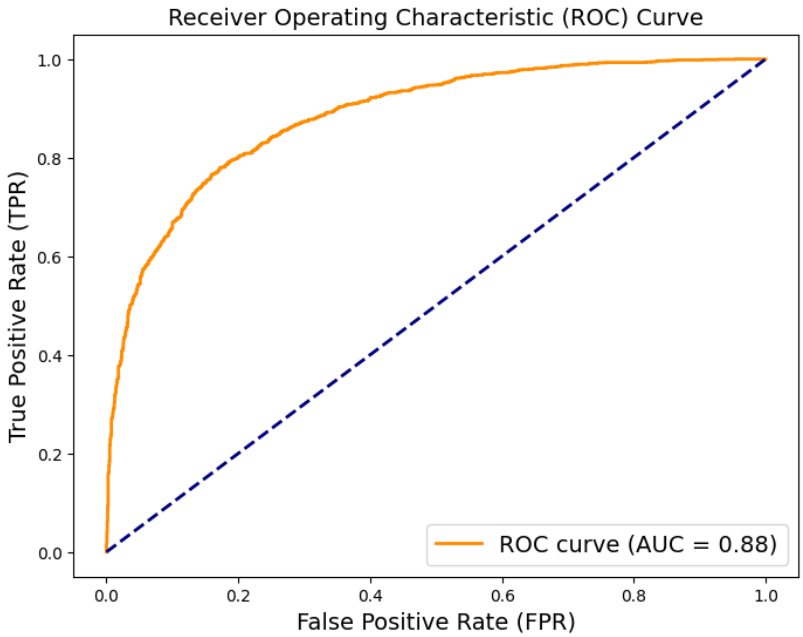

Figure 7. ROC curve

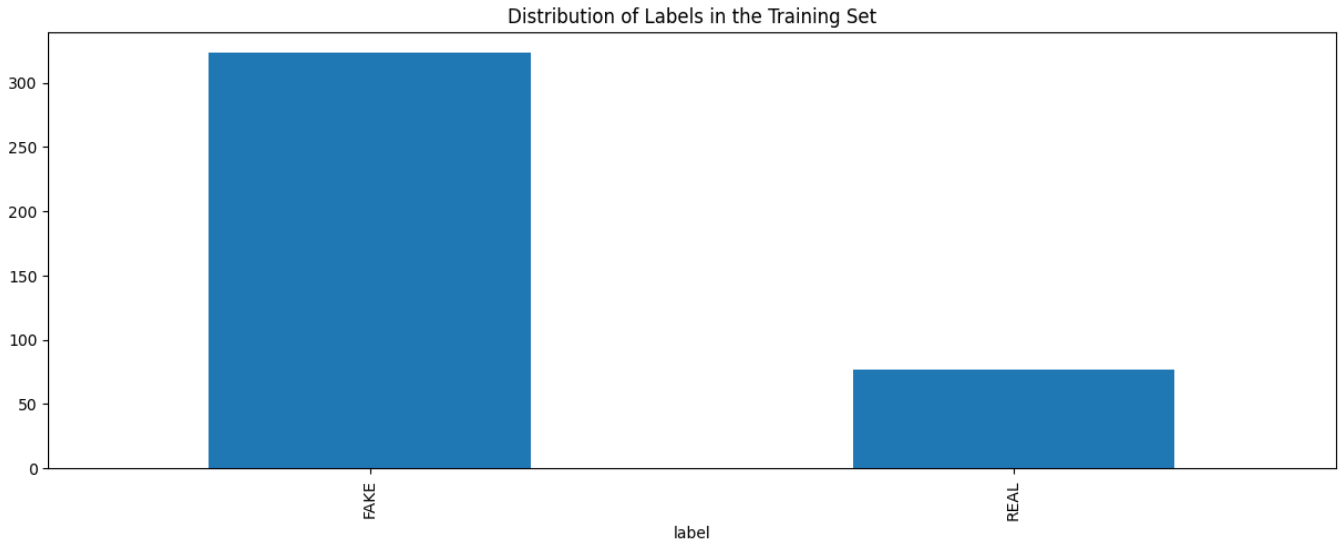

Figure 8. Distribution of labels

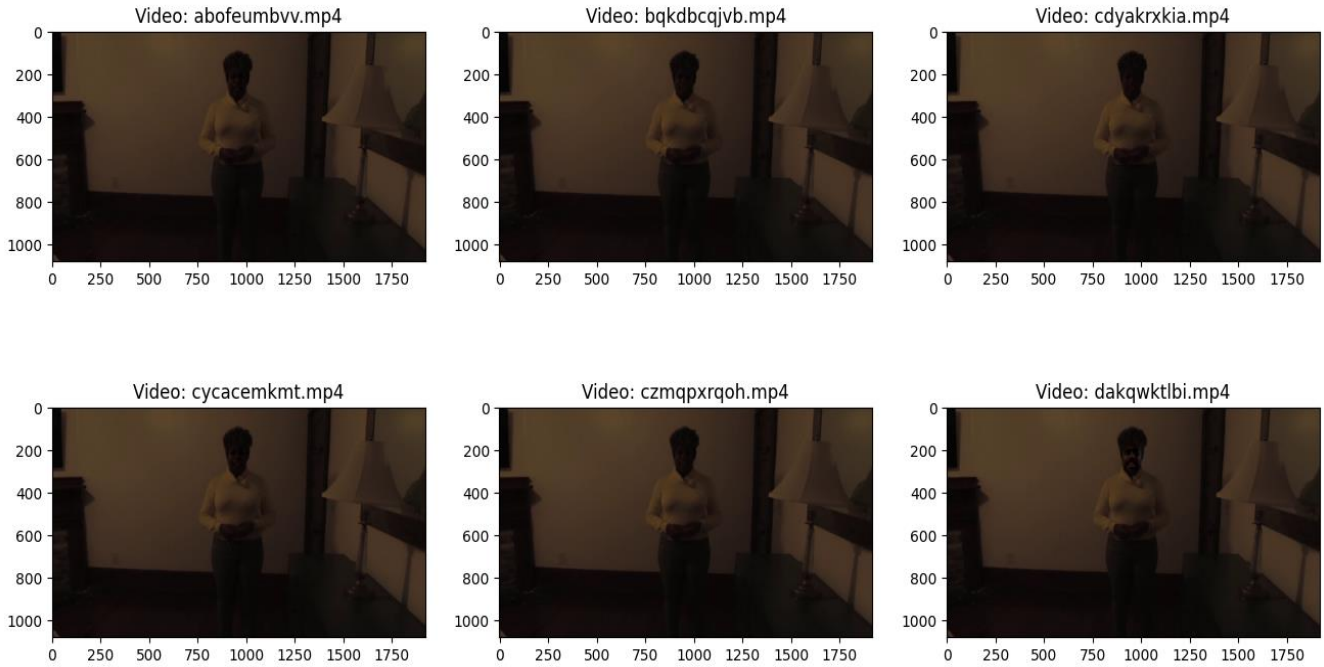

Figure 9. Training video

Figure 10. Predicated class of video is a fake video

Figure 11. Predicted class of video is a real video

Table 2. Comparative analysis between the block-based analysis, CNN, and the integrated hybrid approach in video forgery detection

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Block-Based Analysis	72.45	68.50	70.10	59.29
CNN	76.12	72.85	74.50	63.66
Hybrid Approach	79.31	75.40	76.90	65.87

Figure 11 presents visual examples or representations of the testing videos used in the video forgery detection process. In contrast to Figure 10, this figure indicates that the video shown is classified as a real video, demonstrating the model's capability to differentiate between authentic and manipulated content effectively. Table 2 gives a detailed comparative analysis between block–based analysis, CNN, and the integrated hybrid approach in video forgery detection.

6. CONCLUSION

The research presents a pioneering hybrid framework for video forgery detection, combining traditional block-based analysis with CNNs to address the limitations of existing methods. The approach aims to enhance accuracy, efficiency, and adaptability, and the findings are significant. The hybrid approach effectively leverages the strengths of both methodologies, achieving a higher level of accuracy in detecting local and global features crucial for reliable forgery detection. By integrating deep learning techniques capturing temporal dependencies, the hybrid approach addresses scalability and adaptability challenges faced by traditional block-based methods. The two-stage methodology provides a structured approach to systematically analyze videos for signs of forgery, offering a robust and versatile framework for video forgery detection. Rigorous evaluations and comparisons demonstrate the superiority of the hybrid approach in terms of accuracy, efficiency, and adaptability. The implications are

far-reaching, as the hybrid approach contributes to the advancement of video forensics and holds promise for enhancing forgery detection in dynamic and diverse video scenarios. Overall, the integration of traditional block-based analysis with CNN methodologies represents a significant step forward in video forgery detection, offering a versatile and robust framework that elevates the capabilities of forgery detection in today's complex digital landscape. The model accuracy is around 79.31 and F1 Score is 65.87. The proposed hybrid approach has proved useful when applied to videos to improve forgery detection, but the method has the following drawbacks; The computational complexity of integrating block-based analysis and CNNs poses implementation hurdles in real-time applications since it involves fast identification in real-time such as live stream or surveillance. Further, the approach may fail in cases of forgery that lack spatial or temporal characteristics or when used for generalizing results on different datasets if retrained. To overcome these drawbacks, the continuation of this research could be directed at fine-tuning the method for real-time usage via lightweight structures or enhanced by hardware solutions, using the more complex deep learning models like the Transformers or some new types of attention to increase the ability to detect intricate forgeries, or adapting the proposed approach for other tasks and domains to optimize its generality. More investigation should also assess the applicability of the approach in realistic scenarios, taking into account its efficiency under real-world conditions, and addressing the new types of forgery like advanced GANs or both image and audio manipulations.

REFERENCES

[1] Diwan, A., Sonkar, U. (2024). Visualizing the truth: A survey of multimedia forensic analysis. Multimedia Tools and Applications, 83(16): 47979-48006. https://doi.org/10.1007/s11042-023-17475-3

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems, pp. 2672-2680. https://doi.org/10.5555/2969033.2969125

[3] Nguyen, A., Yosinski, J., Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA., pp. 427-436. https://doi.org/10.1109/CVPR.2015.7298640

[4] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921-2929.

[5] Choi, J., Chao, W.L., Pantazis, D., Duarte, A. (2018). Context-aware deep feature compression for high-speed visual tracking. IEEE Transactions on Image Processing, 27(4): 2045-2058. https://doi.org/10.1109/TIP.2017.2785606

[6] Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Bottisti, F., Kuo, C.C.J. (2015). Image database TID2013: Peculiarities, results and perspectives. Signal Processing: Image Communication, 30: 57-77. https://doi.org/10.1016/j.image.2014.10.009

[7] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv Preprint arXiv: 1409.1556. https://doi.org/10.48550/arXiv.1409.1556

[8] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-2826. https://doi:10.1109/CVPR.2016.308

[9] Moorthy, A.K., Bovik, A.C. (2011). Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Transactions on Image Processing, 20(12): 3350-3364. https://doi.org/10.1109/TIP.2011.2147325

[10] Ma, C., Yang, C.Y., Yang, X., Yang, M.H. (2017). Learning a no-reference quality metric for single-image super-resolution. Computer Vision and Image Understanding, 158: 1-16. https://doi.org/10.1016/j.cviu.2016.12.009

[11] Tyagi, S., Yadav, D. (2023). A detailed analysis of image and video forgery detection techniques. The Visual Computer, 39(3): 813-833. https://doi.org/10.1007/s00371-021-02347-4

[12] Sowmya, K.N., Chennamma, H.R. (2017). Challenges in surveillance video forgery detection. International Journal of Scientific and Engineering Research, 8(5): 138-140.

[13] Mahmood, T., Nawaz, T., Ashraf, R., Shah, M., Khan, Z., Irtaza, A., Mehmood, Z. (2015). A survey on block based copy move image forgery detection techniques. In 2015 International Conference on Emerging Technologies (ICET), Peshawar, Pakistan, pp. 1-6. https://doi.org/10.1109/ICET.2015.7389169

[14] Korshunov, P., Marcel, S. (2018). DeepFakes: A new threat to face recognition? Assessment and detection, arXiv:1812.08685. https://doi.org/10.48550/arXiv.1812.08685.

[15] Pelletier, C., Webb, G.I., Petitjean, F. (2019). Temporal convolutional neural network for the classification of satellite image time series. Remote Sensing, 11(5): 523. https://doi.org/10.3390/rs11050523

[16] Kumar, V., Kansal, V., Gaur, M. (2022). Multiple forgery detection in video using convolution neural network. Computers. Materials & Continua, 73(1): 1347-1364. 10.32604/cmc.2022.023545.

[17] Bao, Q., Wang, Y., Hua, H., Dong, K., Lee, F. (2024). An anti-forensics video forgery detection method based on noise transfer matrix analysis. Sensors, 24(16): 5341. https://doi.org/10.3390/s24165341

[18] Zhou, J., Ni, J., Rao, Y. (2017). Block-based convolutional neural network for image forgery detection. In Digital Forensics and Watermarking: 16th International Workshop, IWDW 2017, Magdeburg, Germany, pp. 65-76. https://doi.org/10.1007/978-3-319-64185-0_6

[19] Akhtar, N., Saddique, M., Asghar, K., Bajwa, U.I., Hussain, M., Habib, Z. (2022). Digital video tampering detection and localization: Review, representations, challenges and algorithm. Mathematics, 10(2): 168. https://doi.org/10.3390/math10020168

[20] Amin, M.A., Hu, Y., Hu, J. (2024). Analyzing temporal coherence for deepfake video detection. Electronic Research Archive, 32(4): 2621-2641. https://doi.org/10.3934/era.2024119

[21] Rossler, 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, pp. 1-11.
[22] Zhang, Z., Liu, Q. (2020). Detect video forgery by performing transfer learning on deep neural network. In Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery, pp. 415-422. https://doi.org/10.1007/978-3-030-32591-6_44
[23] Yao, Y., Shi, Y., Weng, S., Guan, B. (2017). Deep learning for detection of object-based forgery in advanced video. Symmetry, 10(1): 3. https://doi.org/10.3390/sym10010003
[24] Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., Xia, F. (2024). Deepfake video detection: Challenges and opportunities. Artificial Intelligence Review, 57(6): 159. https://doi.org/10.1007/s10462-024-10810-6
[25] Al-Sanjary, O.I., Sulong, G. (2015). Detection of video forgery: A review of literature. Journal of Theoretical & Applied Information Technology, 74(2): 208-220.
[26] Wu, H., Zhou, J., Tian, J., Liu, J., Qiao, Y. (2022). Robust image forgery detection against transmission over online social networks. IEEE Transactions on Information Forensics and Security, 17: 443-456. https://doi.org/10.1109/TIFS.2022.3144878
[27] Su, Y., Nie, W., Zhang, C. (2011). A frame tampering detection algorithm for MPEG videos. In 2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference, Chongqing, China, pp. 461-464. https://doi.org/10.1109/ITAIC.2011.6030373
[28] Li, S., Huo, H.T. (2021). Frame deletion detection based on optical flow orientation variation. IEEE Access, 9: 37196-37209. https://doi.org/10.1109/ACCESS.2021.3061586
[29] Dar, Y., Bruckstein, A.M. (2015). Motion-compensated coding and frame rate up-conversion: Models and analysis. IEEE Transactions on Image Processing, 24(7): 2051-2066. https://doi.org/10.1109/TIP.2015.2412378
[30] Tayfor, N.B., Rashid, T., Qader, S.M., Hassan, B.A., Abdalla, M.H., Majidpour, J., Ahmed, A.M., Sidqi, H.M., Salih, A., Yaseen, Z.M. (2023). Video forgery detection for surveillance cameras: A review. Research Square. https://doi.org/10.21203/rs.3.rs-3360980/v1
[31] Yang, Q., Yu, D., Zhang, Z., Yao, Y., Chen, L. (2020). Spatiotemporal trident networks: detection and localization of object removal tampering in video passive forensics. IEEE Transactions on Circuits and Systems for Video Technology, 31(10): 4131-4144. https://doi.org/10.1109/TCSVT.2020.3046240
[32] Chen, J., Kang, X., Liu, Y., Wang, Z.J. (2015). Median filtering forensics based on convolutional neural networks. IEEE Signal Processing Letters, 22(11): 1849-1853. https://doi.org/10.1109/LSP.2015.2438008
[33] Mathai, M., Rajan, D., Emmanuel, S. (2016). Video forgery detection and localization using normalized cross-correlation of moment features. In 2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI) NM, USA, pp. 149-152. https://doi.org/10.1109/SSIAI.2016.745919
[34] Rodriguez, D., Nayak, T., Chen, Y., Krishnan, R., Huang, Y. (2022). On the role of deep learning model complexity in adversarial robustness for medical images. BMC Medical Informatics and Decision Making, 22: 160. https://doi.org/10.1186/s12911-022-01891-w
[35] Ravindran, A.A. (2023). Internet-of-Things Edge Computing Systems for Streaming Video Analytics: Trails Behind and the Paths Ahead. IoT, 4(4): 486-513. https://doi.org/10.3390/iot4040021
[36] Rodriguez-Ortega, Y., Ballesteros, D.M., Renza, D. (2021). Copy-Move Forgery Detection (CMFD) using deep learning for image and video forensics. Journal of Imaging, 7(3): 59. https://doi.org/10.3390/jimaging7030059