iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** I **Month of publication:** January 2026

DOI: https://doi.org/10.22214/ijraset.2026.76863

Transformer-Based Temporal Learning for Robust Frame Deletion Detection in Videos

Ch. V. V. S. R. Harshadeep[1], Dr. K. V. Ramana[2], K. Ravi Kiran[3]
[1]Student, [2]Professor, [3]Assistant Professor, Department of Computer Science and Engineering, JNTUK, Kakinada, India

***Abstract:** Video data plays an increasingly important role as digital evidence in areas such as surveillance systems, media verification, and forensic analysis. Nevertheless, the widespread availability of advanced video editing software has made it possible to perform complex temporal manipulations, including frame deletion, which can hide significant events and undermine the reliability of video evidence. Most existing frame deletion detection techniques are based on convolutional neural networks and handcrafted temporal descriptors. These approaches often exhibit limitations in modelling long-range temporal relationships and show reduced performance in low-motion scenes or under varying illumination conditions.To address these challenges, this paper proposes a Transformer-based temporal learning framework for reliable frame deletion detection in videos. The proposed method employs self-attention mechanisms to capture global temporal dependencies across video frames, allowing the identification of subtle temporal inconsistencies introduced by frame deletion. In contrast to conventional CNN-based techniques, the framework reduces dependence on frame differencing operations and manually defined statistical thresholds. By combining frame-level feature extraction with a temporal Transformer encoder, the proposed model enhances robustness across diverse motion patterns. This study demonstrates the potential of attention-driven temporal modelling in video forensics and establishes a scalable basis for future research in deep learning–based video manipulation detection.*

***Keywords:** Video Forensics, Frame Deletion Detection, Transformer-Based Models, Temporal Attention Mechanisms, Video Forgery Detection*

I. INTRODUCTION

The rapid advancement of digital video technologies has led to the widespread use of video content in applications such as surveillance systems, media verification, social platforms, and forensic investigations. Videos are often considered reliable sources of evidence because they capture events over time and provide contextual continuity. However, the growing accessibility of advanced video editing software has significantly increased the risk of video manipulation, raising serious concerns about the authenticity and reliability of video data used in critical decision-making processes.

Among various video tampering techniques, temporal manipulations pose a unique challenge for forensic analysis. Frame deletion, in particular, involves the removal of selected frames from a video sequence to conceal actions or alter the perceived timeline of events. Unlike spatial forgeries, frame deletion does not necessarily introduce visible artifacts within individual frames, making detection difficult through visual inspection. Traditional detection methods based on handcrafted temporal features, such as motion continuity or correlation analysis, often fail in low-motion scenes, static environments, or under changing illumination conditions. These approaches also rely heavily on manually defined thresholds, which limits their robustness and adaptability to diverse real-world video content.

To address these limitations, deep learning–based methods using convolutional neural networks have been widely explored for frame deletion detection. While CNN-based models can automatically learn spatio-temporal features, most existing approaches operate on fixed-length frame windows and are therefore limited in their ability to capture long-range temporal dependencies across entire video sequences. Recent progress in Transformer architectures has demonstrated their effectiveness in modelling global relationships through self-attention mechanisms. Motivated by this capability, this paper proposes a Transformer-based temporal learning framework for robust frame deletion detection. By capturing global temporal dependencies and reducing reliance on handcrafted temporal descriptors, the proposed approach aims to improve detection performance across diverse motion patterns and visual conditions, providing a scalable foundation for future video forensic research.

In addition to detection accuracy, practical video forensic systems must demonstrate robustness and generalization across diverse recording conditions, including variations in camera motion, scene dynamics, and environmental lighting. Many existing approaches are evaluated under constrained experimental settings and may not generalize well to real-world scenarios where video content is highly heterogeneous.

The ability to model temporal relationships consistently across different datasets and manipulation patterns is therefore critical for reliable deployment. By leveraging global temporal attention, Transformer-based models offer a flexible and data-driven mechanism to adapt to such variations, making them particularly suitable for scalable and general-purpose video forgery detection systems.

II. LITERATURE REVIEW

Frame deletion detection has been an active research area within video forensics, with early studies primarily relying on handcrafted temporal features and statistical analysis of inter-frame relationships. Initial approaches focused on detecting abrupt temporal discontinuities by analysing motion patterns, frame correlation, or frequency-domain characteristics. While these methods provided a foundation for identifying temporal tampering, their effectiveness was limited by sensitivity to noise, compression artifacts, and variations in motion intensity, particularly in low-motion or static scenes.

With the advancement of deep learning, Convolutional Neural Networks (CNNs) have been widely adopted for inter-frame video forgery detection. Hoogs et al. [1] introduced a C3D-based convolutional neural network for frame dropping detection, demonstrating that spatio-temporal features learned directly from video data outperform traditional handcrafted descriptors. Similarly, Voronin et al. [2] employed a 3D CNN architecture to detect deleted frames in videos, achieving improved detection accuracy by jointly modelling spatial and temporal information. Despite their effectiveness, these 3D CNN-based approaches operate on fixed-length frame clips, which restricts their ability to capture long-range temporal dependencies across entire video sequences.

To reduce computational complexity and improve interpretability, several hybrid methods have been proposed that combine deep feature extraction with statistical analysis. Kumar et al. [3] presented a framework for detecting multiple inter-frame forgeries using CNN-based feature extraction followed by Pearson Correlation Coefficient (PCC) analysis. While this approach is computationally efficient and interpretable, it relies heavily on manually defined thresholds, making it sensitive to illumination changes and prone to false positives. Gong et al. [4] proposed an improved residual feature–based method to enhance frame deletion cues using deep learning; however, the method still focuses on local temporal relationships and does not explicitly model global temporal context.

Motion-based approaches have also been explored to identify frame deletion, particularly in surveillance scenarios. Su et al. [5] utilized velocity field analysis to detect deleted frames by identifying inconsistencies in motion patterns. Although effective in motion-rich environments, such methods struggle in low-motion or static scenes, where temporal inconsistencies are subtle and difficult to distinguish from normal variations. These limitations highlight the dependency of motion-based techniques on scene dynamics.

More recent studies have addressed general inter-frame tampering detection, including deletion, insertion, and duplication. Xing et al. [6] and Tan et al. [7] proposed CNN-based and hybrid deep learning frameworks for inter-frame forgery detection, demonstrating improved performance across multiple manipulation types. Girish et al. [8] further explored deep spatio-temporal learning models to improve robustness across datasets. While these approaches show strong detection capability, they remain computationally intensive and rely on localized temporal modelling. Akhtar et al. [9] incorporated recurrent neural networks to enable frame-level localization of tampering; however, recurrent models often struggle with long video sequences due to vanishing gradients and limited temporal memory.

Recent advances in attention-based architectures have introduced new opportunities for temporal modelling in video analysis. Zhu et al. [10] explored the integration of CNNs with Vision Transformers for inter-frame forgery detection, demonstrating the potential of attention mechanisms to enhance temporal reasoning. However, the use of Transformers in video forensics remains limited, and existing works often employ hybrid architectures without fully exploiting temporal self-attention. Survey studies, such as the work by Ali et al. [11], emphasize the dominance of CNN-based methods in inter-frame forgery detection and explicitly identify the lack of Transformer-driven temporal modelling as a key research gap. Ceron et al. [12] further reinforced this observation through an extensive comparison of supervised and unsupervised CNN-based frame deletion detection methods, concluding that future research should explore Transformer-based architectures to improve robustness and generalization.

From the reviewed literature, it is evident that while CNN-based and hybrid approaches have achieved promising results, they are fundamentally constrained by localized temporal analysis, motion dependency, and heuristic thresholding. The limited exploration of Transformer-based temporal modelling presents a significant opportunity for advancing frame deletion detection. These observations motivate the development of an attention-driven temporal learning framework capable of capturing global temporal dependencies and detecting subtle frame deletion artifacts across diverse video conditions.

III. EXISTING MODELS

Existing approaches for frame deletion detection can be broadly categorized into traditional feature-based methods, deep learning–based convolutional models, and hybrid frameworks that combine learned features with statistical analysis. Each category has contributed to improving detection accuracy; however, significant limitations remain, particularly in modelling long-range temporal dependencies and ensuring robustness across diverse video conditions.

Traditional models rely on handcrafted temporal features such as motion continuity, optical flow variations, frame correlation coefficients, or frequency-domain characteristics to identify abrupt temporal inconsistencies. These methods are computationally efficient and interpretable, making them attractive for early video forensic applications. However, their effectiveness is highly dependent on scene dynamics and recording conditions. In low-motion or static scenes, temporal inconsistencies introduced by frame deletion may be subtle and difficult to distinguish from natural variations, leading to missed detections or false alarms. Additionally, traditional approaches often require manual parameter tuning, limiting their adaptability to different datasets and real-world scenarios.

Deep learning–based models, particularly those using Convolutional Neural Networks (CNNs), have been widely adopted to overcome the limitations of handcrafted methods. Three-dimensional CNNs extend conventional CNN architectures by jointly modelling spatial and temporal information, enabling automatic extraction of spatio-temporal features from video clips. These models have demonstrated strong performance in detecting frame deletion by learning discriminative patterns directly from data. Despite their success, CNN-based approaches typically operate on fixed-length temporal windows, restricting their ability to capture long-range temporal dependencies across entire video sequences. As a result, temporal inconsistencies caused by frame deletion may remain undetected when they occur outside the local receptive field of the model.

Hybrid models have been proposed to reduce computational complexity and improve interpretability by combining CNN-based feature extraction with statistical analysis. In such frameworks, deep features are extracted from consecutive frames and analysed using correlation measures or threshold-based decision rules to detect temporal anomalies. While these approaches offer efficiency and transparency, they rely heavily on manually defined thresholds, making them sensitive to illumination changes, compression artifacts, and noise. Furthermore, hybrid methods continue to focus on local temporal relationships and do not explicitly model global temporal context.

Motion-based and recurrent architectures have also been explored to enhance temporal modelling. Motion-centric approaches exploit velocity or optical flow information to identify inconsistencies caused by missing frames, but their performance degrades in low-motion environments. Recurrent neural networks, such as LSTMs or GRUs, have been used to model temporal sequences and enable frame-level localization of tampering. However, recurrent models often struggle with long video sequences due to limited memory capacity and vanishing gradient issues, reducing their effectiveness in capturing global temporal dependencies.

In summary, while existing models have achieved promising results in frame deletion detection, they are fundamentally constrained by localized temporal analysis, motion dependency, and heuristic decision mechanisms. These limitations highlight the need for a more flexible and globally aware temporal modelling approach. Attention-based architectures, particularly Transformers, offer a promising alternative by enabling direct modelling of long-range temporal relationships across video frames. This motivates the development of a Transformer-based temporal learning framework, which is presented in the next section.

IV. PROPOSED FRAMEWORK

The proposed framework is designed to overcome the limitations of existing frame deletion detection methods by incorporating Transformer-based temporal modelling to capture global dependencies across video sequences. Unlike conventional CNN-based and hybrid approaches that rely on localized temporal analysis or handcrafted statistical thresholds, the proposed method leverages self-attention mechanisms to model long-range temporal relationships and identify subtle temporal inconsistencies introduced by frame deletion.

The overall workflow of the proposed Transformer-based temporal framework is illustrated in **Fig. 1**. As shown in the figure, the framework follows a sequential pipeline consisting of frame extraction, preprocessing, feature extraction, temporal modelling, and final classification. Initially, the input video is processed to extract frames using uniform sampling. The extracted frames are then pre-processed through resizing and normalization to ensure consistency across different video resolutions and recording conditions.

Following preprocessing, frame-level spatial features are extracted using a lightweight CNN-based feature extraction module. This stage generates compact frame-level embeddings that capture essential visual characteristics while maintaining computational efficiency. To preserve the temporal order of frames, temporal positional encoding is applied to the extracted features, enabling the model to distinguish the relative positions of frames within the video sequence.

The position-aware feature representations are subsequently fed into a temporal Transformer encoder, as depicted in **Fig. 1**, where self-attention mechanisms are employed to model global temporal dependencies across the entire video. This allows the framework to capture subtle disruptions in temporal continuity caused by frame deletion, even in low-motion or visually consistent scenes. By attending to relationships between all frames, the Transformer encoder overcomes the limitations of fixed temporal windows commonly used in CNN-based approaches.

Finally, the output of the Transformer encoder is passed to a classification layer consisting of fully connected layers followed by a SoftMax function to determine whether the input video is original or contains frame deletion. The final detection output is generated based on globally attended temporal features rather than local frame differences. Overall, the proposed framework, illustrated in Fig. 1, provides a robust and scalable solution for frame deletion detection by combining CNN-based feature extraction with Transformer-driven temporal attention.

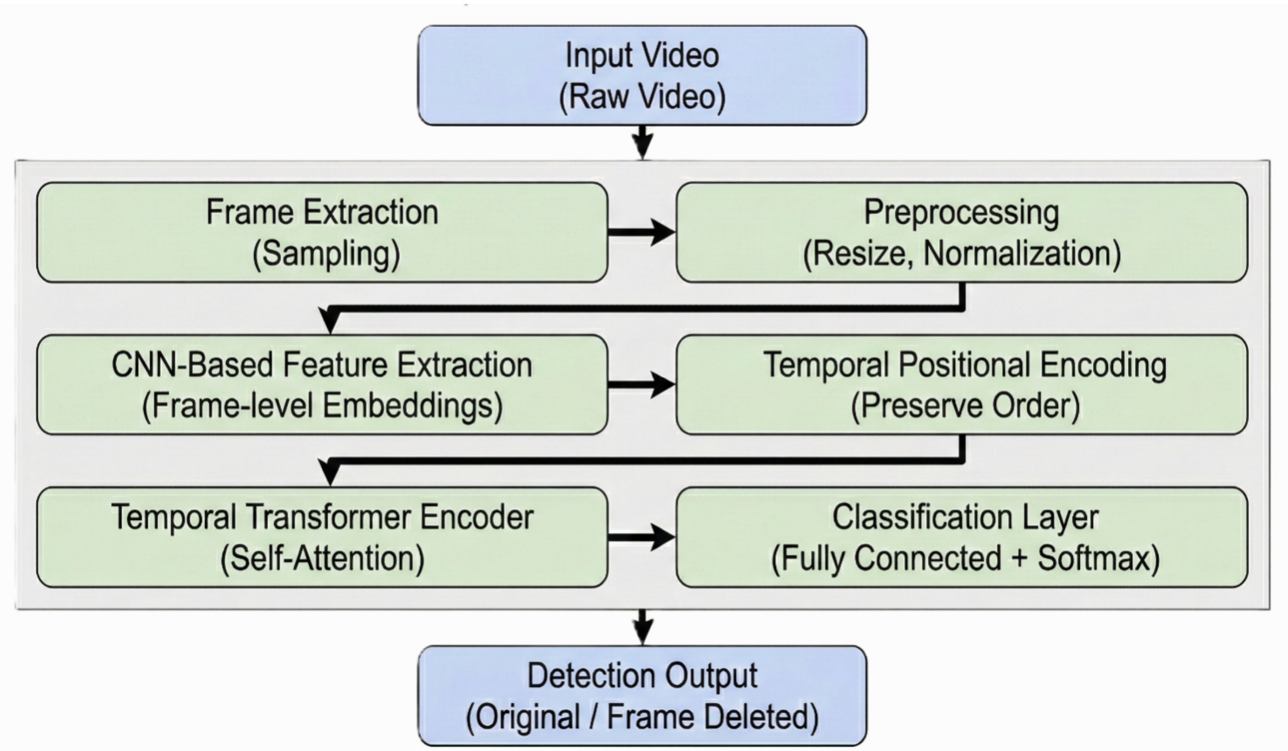


Fig. 1 Workflow of the proposed Transformer-based temporal framework for frame deletion detection

V. RESULTS AND DISCUSSION

The proposed Transformer-based temporal framework is designed to address key limitations observed in existing frame deletion detection methods, particularly those related to localized temporal modelling and motion dependency. Although the present work focuses on framework design rather than extensive empirical evaluation, the expected performance of the proposed approach can be analysed conceptually by comparing it with conventional CNN-based and hybrid methods discussed in the literature.

By leveraging self-attention mechanisms, the proposed framework enables global temporal reasoning across entire video sequences. Unlike CNN-based models that operate on fixed-length temporal windows, the Transformer encoder can model long-range temporal dependencies by attending to relationships between all frames. This capability is particularly beneficial for detecting subtle frame deletion artifacts that are distributed across time or occur in low-motion and visually consistent scenes, where traditional motion-based cues are weak or unreliable.

The integration of CNN-based frame-level feature extraction with temporal positional encoding further enhances the robustness of the framework. Frame-level embeddings capture essential spatial information, while positional encoding preserves temporal order, allowing the Transformer to distinguish between normal temporal variations and inconsistencies introduced by frame deletion. This design reduces reliance on explicit frame differencing operations and handcrafted temporal descriptors, which are known to be sensitive to illumination changes and compression artifacts.

Compared to hybrid approaches that rely on statistical thresholds or correlation measures, the proposed framework offers a fully learnable and data-driven decision mechanism. The classification stage operates on globally attended temporal representations rather than local frame differences, improving adaptability across diverse video content and recording conditions. Additionally, the modular structure of the framework allows flexibility in selecting backbone networks and Transformer configurations, making it suitable for deployment under varying computational constraints.

Overall, the proposed Transformer-based temporal framework demonstrates strong potential for improving the reliability and generalization of frame deletion detection systems. While experimental validation is required to quantify performance gains, the conceptual analysis indicates that attention-driven temporal modelling provides a more robust foundation than existing localized temporal approaches. These observations support the suitability of the proposed framework for future extensions involving large-scale evaluation and real-world video forensic applications.

VI. CONCLUSION

This paper presented a Transformer-based temporal learning framework for reliable frame deletion detection in videos. By addressing the limitations of existing CNN-based and hybrid approaches, the proposed framework leverages self-attention mechanisms to capture global temporal dependencies across video sequences. This enables the detection of subtle temporal inconsistencies introduced by frame deletion, particularly in challenging scenarios involving low motion, static backgrounds, or visually consistent content.

The proposed framework integrates CNN-based frame-level feature extraction with temporal positional encoding and a Transformer encoder to model long-range temporal relationships in a fully learnable and data-driven manner. By reducing dependence on handcrafted temporal descriptors, frame differencing operations, and heuristic thresholds, the framework improves robustness and adaptability across diverse video conditions. The modular design also allows flexibility in selecting network components based on computational requirements and application constraints.

Although the present work focuses on conceptual framework design rather than extensive experimental evaluation, the proposed approach establishes a strong foundation for future research in video forensics. Future work will involve quantitative performance evaluation on benchmark datasets, frame-level localization of deletion points, and optimization for real-time deployment. Overall, this study highlights the potential of attention-driven temporal modelling as a scalable and effective direction for deep learning–based video manipulation detection.

REFERENCES

[1] J. Hoogs, J. Wilkins, M. A. McCord, and K. A. Thompson, “A C3D-Based Convolutional Neural Network for Frame Dropping Detection in a Single Video Shot,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017, pp. 1–8.

[2] V. Voronin, A. Zelensky, and S. Fedotov, “Detection of Deleted Frames on Videos Using a 3D Convolutional Neural Network,” in Proc. SPIE—Applications of Digital Image Processing XLI, vol. 10752, 2018, pp. 1–9.

[3] A. Kumar, S. Kansal, and M. S. Gaur, “Multiple Forgery Detection in Video Using Convolution Neural Network,” Multimedia Tools and Applications, vol. 81, no. 4, pp. 1–21, 2022.

[4] C. Gong, X. Liu, and Z. Wang, “IReF: Improved Residual Feature for Video Frame Deletion Forensics,” in Proc. International Conference on Digital Image and Signal Processing (ICDIS), 2022, pp. 1–6.

[5] Y. Su, H. Zhang, and L. Chen, “Velocity Field-Based Surveillance Video Frame Deletion Detection,” in Advances in Computer Vision and Pattern Recognition, Springer, 2024, pp. 1–15.

[6] Y. Xing, J. Li, and Z. Wang, “Inter-Frame Video Tampering Detection Based on Deep Convolutional Neural Networks,” ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 18, no. 2, pp. 1–21, 2022.

[7] L. Tan, Y. Zhang, and X. Li, “A Hybrid Deep Learning Framework for Object-Based Inter-Frame Video Forgery Detection,” Signal Processing: Image Communication, vol. 98, pp. 116–129, 2022.

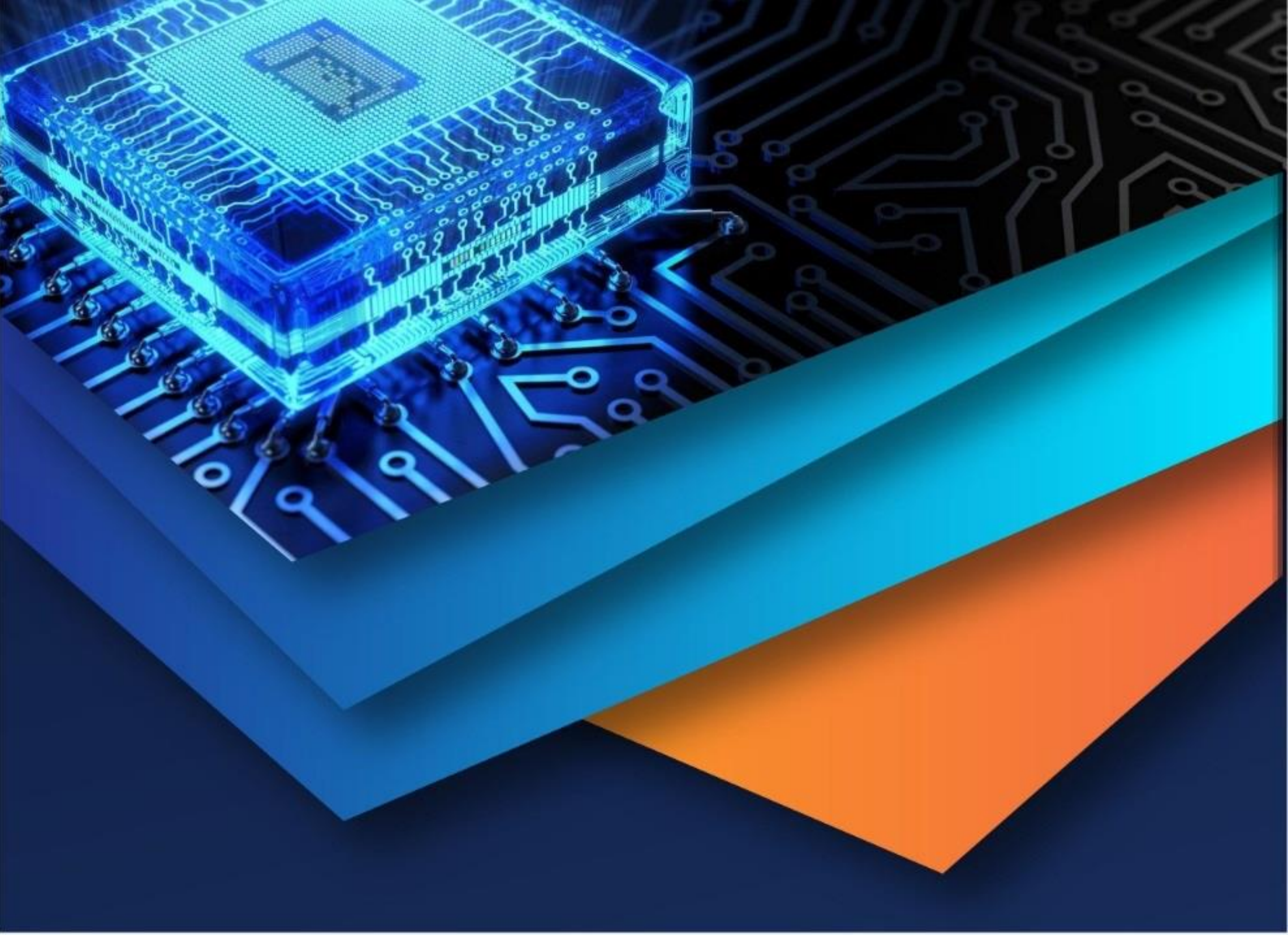
[8] R. Girish, P. R. Kumar, and S. R. Mahadeva Prasanna, “Deep Learning-Based Inter-Frame Video Forgery Detection,” International Journal of Multimedia Information Retrieval, vol. 12, no. 1, pp. 1–14, 2023.

[9] Z. Akhtar, A. K. Singh, and P. Gupta, “Deep Learning-Based Detection and Localization of Inter-Frame Video Tampering,” Mathematics, vol. 12, no. 3, pp. 1–19, 2024.

[10] Y. Zhu, Q. Wang, and H. Li, “Video Inter-Frame Forgery Detection Based on CNN and Vision Transformer,” in Proc. SPIE—Media Watermarking, Security, and Forensics, vol. 12528, 2025, pp. 1–10.

[11] M. Ali, R. Khalid, and S. Hussain, “Inter-Frame Forgery Video Detection: Datasets, Methods, and Challenges,” Electronics, vol. 14, no. 2, pp. 1–27, 2025.

[12] J. Ceron, L. Verdoliva, and P. Bestagini, “Detecting Frame Deletion in Videos Using Supervised and Unsupervised Learning,” IEEE Transactions on Information Forensics and Security, vol. 19, pp. 1–15, 2024.



10.22214/IJRASET

INDEX COPERNICUS

45.98

ISRA JIF

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR:
7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY