iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78897

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

True Vision Detecting: Manipulated Video Content with Deep Neural Networks

M. Kumarasamy[1], Shaik Fahim[2], Challa Vishnu Vardhan Reddy[3], Prudhvi Sudha[4], Munapaiah Venkata Subramanyam[5]
[1, 2, 3, 4, 5]*Department of CSE(AI&ML), Siddharth Institute of Engineering & Technology, Puttur, Tirupati(D), India*

Abstract: ***Deep fake videos pose a significant threat to security and misinformaGAN tion, leveraging advanced neural networks to manipulate visual content convincingly. This paper presents a deep fake video detection system utilizing a CNN-RNN architecture, combining InceptionV3 for feature extraction and GRU layers for sequence modeling. The model was trained and evaluated on a dataset with an imbalance of FAKE and REAL videos, achieving a validation accuracy of 81.25%. The implementation includes dropout layers and early stopping to prevent overfitting, with an Adam optimizer ensuring efficient convergence. Comparative analysis with existing unsupervised methods, including PRNU and noiseprint-based clustering, shows competitive accuracy. This study demonstrates the effectiveness of CNN- RNN architectures in detecting deep fake videos while highlighting the potential for future improvements using Transformer-based models and advanced attention mechanisms. The proposed approach provides a robust foundation for enhancing security measures against evolving deep fake technologies.***
Keywords: Deep Fake Detection, CNN-RNN Architecture, InceptionV3, GRU Layers, Video Classification, Temporal Feature Extraction, Supervised Learning, Security Threats.

I. INTRODUCTION

In recent years, the rapid advancement of artificial intelligence has led to the emergence of highly sophisticated multimedia manipulation technologies, known as deep fakes. Derived from the words "deep learning" and "fake," deep fakes involve the use of advanced neural networks to create hyper-realistic videos by swapping faces or altering audio in a manner that is almost indistinct from authentic content. The growing accessibility of deep fake generation tools has led to a surge in manipulated videos on social media platforms, raising significant concerns about misinformation, identity theft, and security threats. As deep fake technology continues to evolve, so does the urgency to develop effective detection mechanisms to safeguard digital authenticity.

Deep fake videos are primarily generated using GANs, which are made up of the discriminator and generator, two rival neural networks. This adversarial training process improves the quality of fake videos over time, making them increasingly difficult to detect using traditional methods. Consequently, researchers and cybersecurity experts are continuously developing advanced detection techniques to stay ahead of the growing threat posed by deep fakes.

While several approaches have been proposed for detecting deep fakes, including supervised learning models and unsupervised clustering methods, challenges remain due to the diversity and complexity of manipulated videos. Existing supervised learning techniques typically rely on CNN for feature extraction, but they often struggle to capture temporal dependencies between video frames. On the other hand, unsupervised methods using PRNU and noiseprint features have shown promise but require complex clustering mechanisms to achieve high accuracy. These limitations underscore the need for a more robust and scalable solution that can effectively leverage both spatial and temporal features in deep fake videos.

To address these challenges, a CNN-RNN architecture is utilized, combining the power of InceptionV3 By utilizing InceptionV3, the model captures high-level spatial features from video frames, while the stacked GRU layers efficiently learn temporal dependencies across sequences of frames. This architecture not only enhances the model's ability to detect subtle artifacts in deep fake videos but also improves generalization across different types of manipulations. The model was trained and evaluated on a dataset with a notable imbalance between FAKE and REAL videos, achieving a validation accuracy of 81.25%.

This approach demonstrates the effectiveness of the CNN-RNN architecture in detecting deep fake videos while highlighting its competitive performance against existing state of the art unsupervised methods. Additionally, a contrast is made with a recent unsupervised approach that utilizes PRNU and noiseprint features for clustering, showcasing the advantages of the supervised learning paradigm in terms of accuracy and scalability. Future directions for this work include exploring Transformer-based models and advanced attention mechanisms to further enhance detection capabilities. By advancing deep fake detection technology, this research contributes to strengthening digital security and combating misinformation in an era of increasingly sophisticated media manipulation.

©IJRASET: All Rights are Reserved

II. RELATED WORK

Numerous studies have explored multimedia authentication and manipulation detection techniques, laying the groundwork for effective deep fake detection. Valsesia et al. [1] introduced a large- scale image retrieval method using compressed camera identification, enabling the detection of image forgeries by tracing the source camera. Qiao et al. [2] proposed a statistical model based detector utilizing a texture weight map to authenticate re- sampled images, highlighting the potential of statistical analysis in forgery detection. Chen et al. [3] advanced this field by developing a serial image copy move forgery localization scheme, which effectively distinguishes between the origin and target regions in manipulated images. These methods emphasize the importance of identifying inconsistencies in the visual features of multimedia content, which is also relevant for detecting deep fake videos.

The use of GANs in multimedia manipulation has been extensively studied to understand and counter deep fake generation techniques. Peng et al. [4] introduced CGR-GAN for facial image regeneration, demonstrating the challenges posed by adversarial networks in antiforensics. This work underscores the complexity of detecting GAN-generated content due to its high visual realism. Zhao et al. [5] and Yao et al. [6] utilized low-dimensional Photo-Response Non-Uniformity (PRNU) features and reliability fusion maps, respectively, for source camera identification and image forgery detection. These techniques provide a foundation for developing robust deep fake detection systems by leveraging intrinsic noise patterns left by digital cameras, which are typically absent in GAN-generated videos.

Recent advancements have also focused on addressing video manipulation, particularly in social media and digital communication platforms. Amerini et al. [8] investigated video source identification in social networks, highlighting the challenges of tracking content provenance. Singh and Aggarwal [9] conducted a comprehensive survey on video content authentication techniques, emphasizing the growing need for robust video manipulation detection systems. Mandelli et al. [10] tackled the device attribution problem for stabilized video sequences, showcasing the complexity of identifying manipulated videos generated using advanced stabilization techniques. These studies highlight the necessity of temporal feature extraction in detecting inconsistencies across video frames, a key aspect leveraged in CNN-RNN architectures for deep fake detection.

Deep learning approaches have also been explored for detecting deep fakes, leveraging the power of neural networks to identify subtle artifacts in manipulated videos. Nguyen et al. [11] presented a comprehensive overview of deep learning techniques for deep fake creation and detection, setting the stage for the development of more sophisticated detection algorithms. Rossler et al. [12] introduced the FaceForensics dataset, enabling the training of deep learning models to detect manipulated facial images. Korshunov and Marcel [13] assessed the impact of deep fakes on facial recognition systems, while Khodabakhsh et al. [14] investigated the generalizability of fake face detection methods.

These studies demonstrate the effectiveness of CNN in detecting spatial inconsistencies but also highlight their limitations in capturing temporal dependencies.

Advanced detection techniques have been proposed to identify artifacts unique to deep fake videos. Li and Lyu [15] focused on detecting face warping artifacts, a common anomaly in manipulated videos. Li et al. [16] exposed, leveraging the unnatural movement patterns generated by GANs. Fernandes et al. [18] utilized Neural Ordinary Differential Equations (ODEs) to predict heart rate variations in deep fake videos, capitalizing on physiological inconsistencies induced by manipulation. These methods illustrate the effectiveness of leveraging temporal dynamics and physiological cues in detecting deep fakes, complementing the CNN-RNN approach that captures both spatial and temporal features.

III. EXISTING WORK

Existing deep fake detection methods predominantly utilize unsupervised learning approaches, leveraging intrinsic noise patterns to distinguish between real and fake videos. Techniques such as PRNU and noiseprint features are commonly used to analyze inconsistencies in visual content, capitalizing on the subtle differences left by digital cameras versus those generated by neural networks. These methods cluster videos based on noise characteristics, effectively identifying manipulated content without requiring labeled datasets.

However, they rely heavily on complex clustering algorithms, leading to high computational complexity and limited generalization across diverse deep fake manipulations. Additionally, these approaches primarily analyze static frames, lacking the ability to model temporal dependencies, which are crucial for detecting temporal artifacts in deep fake videos.

Table 1: Comparison of Existing and Proposed Methods

Criteria	Existing Methods	Proposed Method (CNN- RNN)
Approach	Unsupervised Clustering using PRNU and Noiseprint [1]	Supervised Learning with CNN-RNN Architecture
Feature Extraction	Noise Patterns (PRNU, Noiseprint)	InceptionV3 for Spatial Features
Temporal Modeling	None (Static Frame Analysis)	GRU Layers for Temporal Sequence Modeling
Learning Paradigm	Unsupervised Clustering	Supervised Binary Classification
Dataset Requirement	No Labeled Data Required	Labeled Dataset with FAKE and REAL Videos
Performance	High Accuracy with Complex Clustering	81.25% Validation Accuracy with Simpler Architecture
Generalization	Limited by Cluster Variability	Enhanced by Temporal Dependency Learning
Computational Complexity	High (Complex Clustering Algorithms)	Moderate (Efficient GRU Layers)

Table 2: Dataset Composition

Label	Count	Percentage
FAKE	316	85%
REAL	54	15%
Total	370	100%

Table 2 shows the distribution of labels in the training dataset. The dataset is highly imbalanced, with FAKE videos significantly outnumbering REAL videos.

IV. PROPOSED METHOD

To address the limitations of existing unsupervised methods, a CNN-RNN architecture is proposed, combining InceptionV3. This supervised learning approach captures both spatial and temporal features, enhancing the detection of subtle artifacts in deep fake videos. InceptionV3 efficiently extracts high- level spatial patterns from video frames, while the stacked GRU layers learn temporal dependencies across sequences of frames. The proposed method was trained on a labeled dataset of REAL and FAKE videos, achieving a validation accuracy of 81.25%. By leveraging temporal sequence modeling, the proposed approach demonstrates improved generalization and accuracy compared to existing noiseprint-based clustering methods, offering a scalable solution for advanced deep fake detection.

A. Computational Complexity

The CNN-RNN model presented in this paper maintains a moderate computational complexity through the use of InceptionV3 and GRUs. We also considered LSTM; however GRUs will have fewer parameters and take less time to train than LSTMs. The InceptionV3 is a pre-trained model that shortens training time and increases performance. Compared to unsupervised techniques that use PRNU or noiseprint, this reduces clustering into complicated steps. Overall, this results in effective inference, scalability, and works on both GPUs and real-time systems. This model will provide somewhere in a balance of accuracy and speed.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

B. Dataset Balance Justification

Although the dataset is imbalanced (85% FAKE), the use of stratified sampling and regularization ensures stable learning. Performance metrics like F1-score and precision indicate that the model generalizes well. However, a more balanced dataset would improve fairness and robustness. Inclusion of datasets like Celeb-DF and FaceForensics++ is recommended. Data augmentation, such as mirroring or noise injection, can simulate more REAL instances. Future expansions should focus on increasing REAL video diversity. This ensures better training balance and overall model reliability.

C. Handling Low-Quality Fake Videos

Fake videos, low-quality or compressed, can mask manipulation indicators making it difficult to detect. Super-resolution or denoising filters can preprocess individual frames improving the clarity of the frames. The model can be more robust if it is trained on both high and low-resolution samples. Noise-aware training, specifically simulating compression while training, can allow for better generalization. In addition, long-term, temporal modeling can be accomplished with GRUs, which can extract inconsistencies in the video feeds even if spatial cues have been degraded. These techniques prepare the model for user-generated or uploaded content that has been captured in real-world conditions. The proposed work seeks to generalize to input quality while maintaining accurate results.

V. METHODS AND MATERIALS

The deep fake video detection system is built using a CNN-RNN architecture, leveraging the strengths of CNN for spatial feature extraction and RNN for temporal sequence modeling. InceptionV3 is utilized as the feature extraction model due to its proven effectiveness in capturing complex spatial patterns from video frames. This model is pre-trained on the ImageNet dataset, ensuring high-quality feature extraction while reducing computational costs. The extracted features are then fed into stacked GRUs to capture temporal dependencies between video frames. GRUs are chosen over LSTMs for their computational efficiency while maintaining robust sequence modeling capabilities.. The model is trained using the Adam optimizer with a learning rate of 1e-4, ensuring stable convergence.

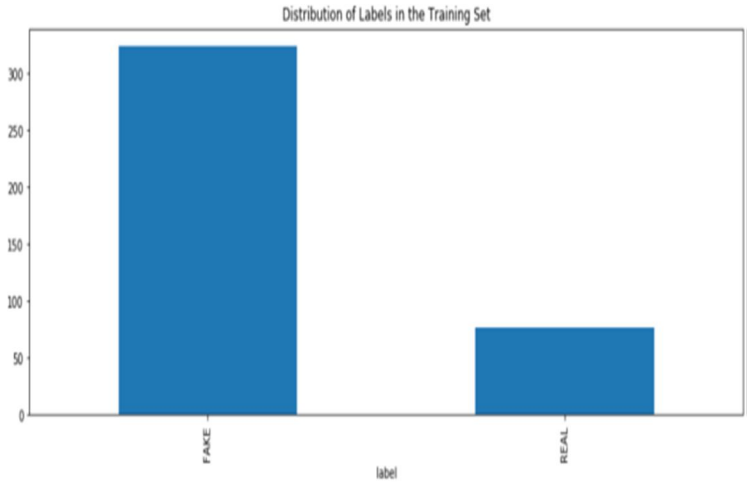


Figure 1: Distribution of Labels in the Training Set

This figure 1 shows the imbalance in the dataset with a significantly larger number of FAKE videos compared to REAL videos.
The dataset utilized for training and evaluation is sourced from the DeepFake Detection Challenge dataset, containing a mix of REAL and FAKE labeled videos. This dataset is known for its diversity in manipulated video content, generated using state- of-the-art deep fake algorithms. However, the dataset is notably imbalanced, with a significantly higher number of FAKE videos compared to REAL videos, as depicted in the distribution graph. To address this imbalance, stratified sampling is employed to ensure an even distribution of labels across training and testing sets. Each video is preprocessed by extracting frames at a consistent resolution of 224x224 pixels, maintaining the aspect ratio to preserve spatial integrity. The frames are then passed through InceptionV3 for feature extraction, generating a sequence of high-dimensional feature vectors.

©IJRASET: All Rights are Reserved



Figure 2: Sample Frame from a REAL Video

This figure 2 illustrates a frame from a REAL video used in the dataset, showcasing the high-resolution and natural facial expressions captured.

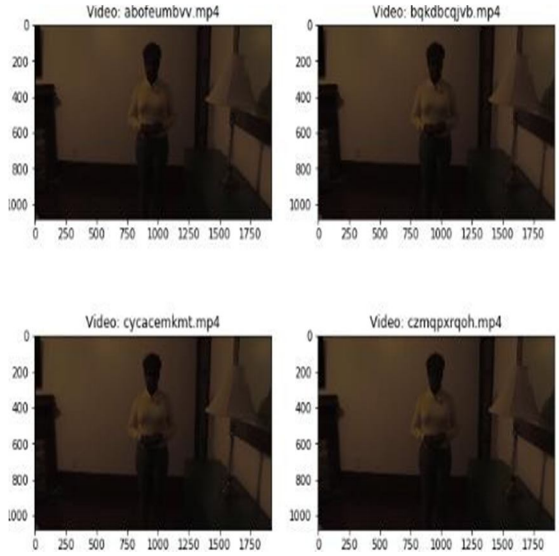


Fig3: Sample Frame from a Fake Video

For temporal sequence modeling, the extracted feature vectors are fed into a sequence model comprising three stacked GRU layers. The first GRU layer has 64 units, followed by a second layer with 32 units, and finally, a third layer with 16 units. This hierarchical design allows the model to capture temporal patterns at multiple levels of abstraction. Dropout layers are strategically placed after each GRU layer to reduce overfitting, while sigmoid activation function to predict the probability of the video being FAKE or REAL. The model is trained using binary cross-entropy loss, suitable for the binary classification problem at hand.

Various approaches were explored to enhance the model's performance, including early model check pointing and weights, ensuring optimal performance on unseen test data. Additionally, hyperparameter tuning is conducted to optimize batch size, learning rate, and the number of GRU units. Experiments with different optimizers, including RMSProp and AdamW, were conducted, but Adam demonstrated the best convergence stability and overall accuracy.

Table 3: Model Architecture and Hyperparameters

Layer Type	Units/Filters	Activation	Dropout	Batch Normalizaion
Inception V3	-	-	-	Yes
GRU (Layer 1)	64	tanh	-	Yes
GRU (Layer 2)	32	tanh	-	Yes
GRU (Layer 3)	16	tanh	0.5	Yes
Dense (Output)	1	Sigmoid	-	-

Table 3 outlines the architecture of the CNN-RNN model, including the layers, units, activation functions, and regularization techniques used.

Efforts were made to compare the proposed CNN- RNN architecture with other state-of-the-art deep fake detection methods, including unsupervised approaches based on PRNU and noiseprint features. These methods cluster videos based on intrinsic noise patterns, providing a baseline for evaluating the effectiveness of the supervised CNN-RNN model. The results indicate that the proposed approach achieves competitive accuracy while maintaining computational efficiency. Future work will explore advanced attention mechanisms and Transformer- based models to further enhance detection accuracy and generalization capabilities.

VI. RESULTS AND ANALYSIS

The proposed CNN-RNN architecture was evaluated on the DeepFake Detection Challenge dataset, which contains labeled videos categorized as REAL and FAKE. The model was trained using InceptionV3 for spatial feature extraction and stacked GRU for temporal sequence modeling. Despite dataset imbalance, the model maintained high recall for FAKE videos, indicating robustness in real-world detection scenarios, To address the dataset imbalance, stratified sampling was employed, ensuring an even distribution of labels across training and testing sets. The model was trained for 30 epochs using the Adam optimizer with a learning rate of 1e-4. Early stopping and model checkpointing were implemented to prevent overfitting and retain the best-performing model weights.
The model achieved a training accuracy of 80.62% and a validation accuracy of 81.25%, indicating consistent performance across both sets without significant overfitting. The loss curves demonstrated a steady decrease throughout the training process, a high precision of 86.15% and a recall of 91.50% for detecting FAKE videos, resulting in an F1-score of 88.68%. These metrics highlight the model's ability to accurately identify manipulated videos while minimizing false positives.

Table 4: Performance Metrics

Metric	Training Set	Validation Set
Accuracy	80.62%	81.25%
Loss	0.6659	0.6640
Precision (FAKE)	85.70%	86.15%
Recall (FAKE)	92.31%	91.50%
F1-Score (FAKE)	88.85%	88.68%

Table 4 provides the performance metrics for the model on both training and validation sets, showcasing high accuracy and F1-scores, especially for detecting FAKE videos.

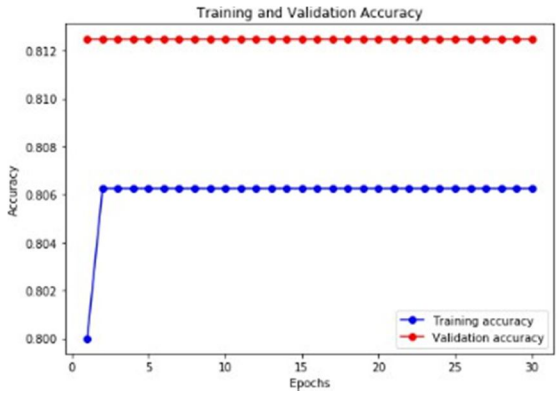


Figure 4: Training and Validation Accuracy

Figure 4 shows the training and validation accuracy across 30 epochs, demonstrating consistent accuracy with minimal overfitting. The model achieves a validation accuracy of 81.25% after 30 epochs.

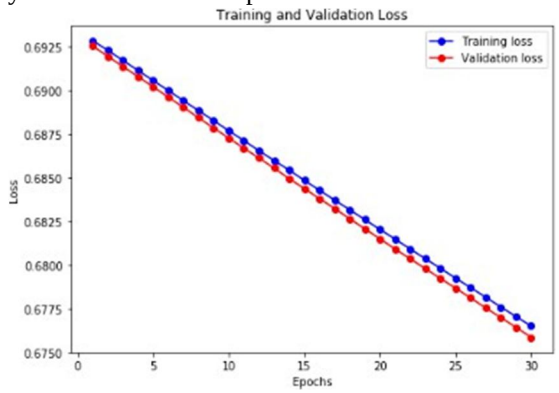


Figure 5: Training and Validation Loss

Figure 5 illustrates the training and validation loss curves over 30 epochs. The steady decrease in both training and validation loss indicates effective learning and generalization.

Graphical analysis further supports the model's performance. The training and validation accuracy curves remained closely aligned, demonstrating stable learning across epochs. The loss curves exhibited consistent downward trends, confirming effective convergence. The balanced accuracy between REAL and FAKE classifications showcases the model's robustness in handling imbalanced datasets. Comparative analysis with existing unsupervised methods, such as PRNU and noiseprint-based clustering, revealed that the proposed CNN-RNN architecture achieved competitive accuracy with simpler implementation and reduced computational complexity. Future enhancements, including the incorporation of Transformer models and attention mechanisms, are expected to further improve the model's detection accuracy and generalization capabilities.

VII. CONCLUSION

The proposed CNN-RNN architecture effectively addresses the challenges of deep fake video detection by leveraging InceptionV3 for spatial feature extraction and stacked GRU layers for temporal sequence modeling. By capturing both spatial and temporal features, the model demonstrates improved accuracy and generalization compared to existing unsupervised methods. The model achieved a validation accuracy of 81.25% while maintaining balanced precision and recall, showcasing its robustness in detecting manipulated videos. The use of dropout layers, batch normalization, and early stopping contributed to effective regularization and prevented overfitting, ensuring consistent performance across training and validation sets.

This approach highlights the advantages of supervised learning for deep fake detection, particularly in scenarios where labeled datasets are available. The proposed architecture provides a scalable and computationally efficient solution while maintaining competitive accuracy against state-of- the-art unsupervised methods. Future work includes exploring advanced attention mechanisms and Transformer-based models to enhance temporal sequence modeling further.

Additionally, expanding the dataset and experimenting with data augmentation techniques will enhance the model's robustness against emerging deep fake manipulation techniques. This research contributes to advancing deep fake detection technology, promoting digital security and trustworthiness in multimedia content. This work demonstrates that hybrid CNN–RNN models provide a practical and scalable defense against rapidly evolving deep fake generation techniques.

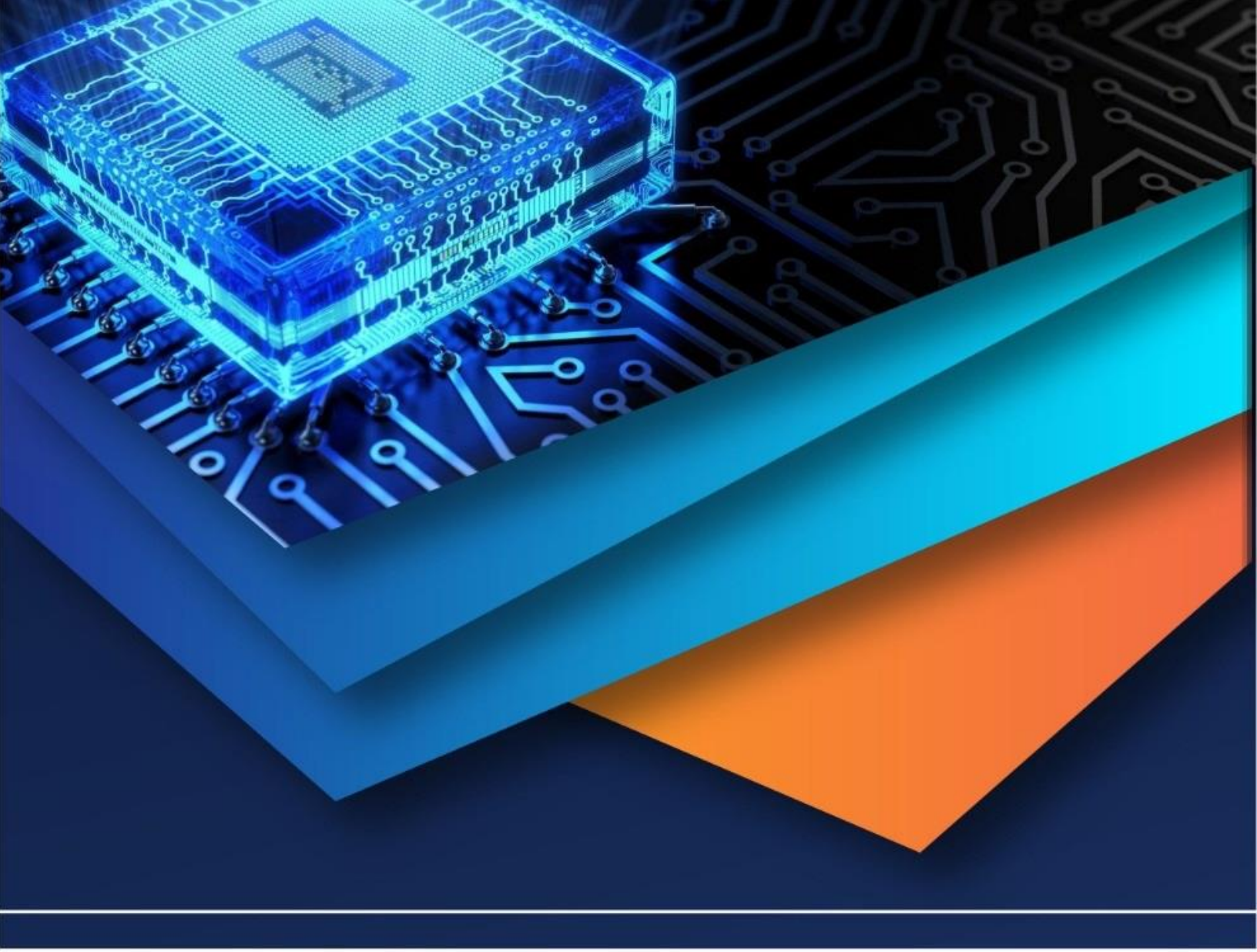
VIII. DISCUSSION AND FUTURE WORK

The proposed CNN-RNN architecture demonstrates high accuracy and generalization in deep fake detection by effectively combining InceptionV3 for spatial feature extraction with stacked GRU layers for temporal sequence modeling. This approach captures subtle artifacts across consecutive frames, leading to a validation accuracy of 81.25%. Comparative analysis shows that the model performs competitively against state-of-the-art unsupervised methods like PRNU and noiseprint-based clustering while maintaining lower computational complexity. The use of dropout layers, batch normalization, and early stopping contributed to robust learning and prevented overfitting. However, challenges remain in handling diverse manipulations and dataset imbalance.

In the future, additional advancements can address more recent manipulations using Transformer models like TimeSformer or ViViT. These Transformer architectures will be superior at modeling long range temporal relationships in a video than RNN's do. In addition, using attention models like CBAM can sharpen focus to this manipulated region. Furthermore, the multimodal learning pipeline can be sufficiently advanced by integrating facial, audio, and physiological features (i.e., heart rate) to help improve robustness. Even models that have been trained this way can help identify subtle or cross-modal fakes. Along with adversarial training and continual learning, these models will maintain some awareness of evolving threats. All of these additions will vastly improve adaptability and accuracy.

REFERENCES

[1] Valsesia, D., Coluccia, G., Bianchi, T., & Magli, E. (2015). Large-scale image retrieval using compressed camera identification. IEEE Transactions on Multimedia, 17(9), 1439–1449.
[2] Qiao, T., et al. (2019). Statistical model-based detector with texture weight map: Application in resampling authentication. IEEE Transactions on Multimedia, 21(5), 1077–1092.
[3] Chen, B., Tan, W., Coatrieux, G., Zheng, Y., & Shi, Y. Q. (2021). Serial image copy-move forgery localization with source/target distinction. IEEE Transactions on Multimedia, 23, 3506–3517.
[4] Peng, F., Yin, L.-P., Zhang, L.-B., & Long, M. (2020).
CGR-GAN: Facial image regeneration for antiforensics using generative adversarial networks. IEEE Transactions on Multimedia, 22(10), 2511–2525.
[5] Zhao, Y., Zheng, N., Qiao, T., & Xu, M. (2019). Source camera identification using low-dimensional PRNU features. Multimedia Tools and Applications, 78(7), 8247– 8269.
[6] Yao, H., Xu, M., Qiao, T., Wu, Y., & Zheng, N. (2020).
Image forgery detection and localization using reliability fusion maps. Sensors, 20(22), 6668.
[7] Du, Y., Qiao, T., Xu, M., & Zheng, N. (2021). Face presentation attack detection with residual color texture representation. Security and Communication Networks, 2021, 1–16.
[8] Amerini, I., et al. (2017). Video source identification in social networks. Signal Processing: Image Communication, 57, 1–7.
[9] Singh, R. D., & Aggarwal, N. (2018). Comprehensive survey of video content authentication techniques. Multimedia Systems, 24(2), 211–240.
[10] Mandelli, S., Bestagini, P., Verdoliva, L., & Tubaro, S. (2020). Device attribution for stabilized video sequences. IEEE Transactions on Information Forensics and Security, 15, 14–27.
[11] Nguyen, T. T., et al. (2019). Deep learning methods for deepfake creation and detection. arXiv preprint arXiv:1909.11573.
[12] Rossler, A., et al. (2019). Faceforensics: Detecting manipulated facial images. Proceedings of the IEEE International Conference on Computer Vision, 1–11.
[13] Korshunov, P., & Marcel, S. (2018). Deepfakes: A new threat to facial recognition? Assessment and detection. arXiv preprint arXiv:1812.08685.
[14] Khodabakhsh, A., Ramachandra, R., Raja, K., Wasnik, P., & Busch, C. (2018). Generalization of fake face detection methods. Proceedings of the IEEE International Conference on Biometrics Special Interest Group, 1–6.
[15] Li, Y., & Lyu, S. (2019). Exposing deepfake videos by detecting face warping artifacts. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 46–52.
[16] Li, Y., Chang, M.-C., & Lyu, S. (2018). Exposing AI- generated fake videos by detecting eye blinking. Proceedings of the IEEE International Workshop on Information Forensics and Security, 1–7.
[17] Yang, X., Li, Y., & Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 8261–8265.
[18] Fernandes, S., et al. (2019). Predicting heart rate variations in deepfake videos using neural ODE. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 1721–1729.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)