Journal of Science and Technology Excellence
ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20506674 **JSTE-18**

ATTENTION BASED DEEPFAKE DETECTION LEVERAGING LONG RANGE DEPENDENCIES

[#1]**MOHAMMED FAIZAN AHMED,** *Dept of CSE,*
[#2]**Dr.N.CHANDRAMOULI,** *Professor&HOD, Dept of CSE,*
Vaageswari College of Engineering(Autonomous), Karimnagar, TG.

ABSTRACT: This investigation demonstrates an improved method of employing deep learning that employs attention processes to identify long-range connections between temporal and visual data in order to identify deepfakes. The proposed attention-based design, which integrates transformer models and self-attention, more effectively learns global contextual relationships across frames than traditional convolutional approaches, which frequently struggle to identify nuanced and widely dispersed changes. The system improves its ability to identify intricate and high-quality deepfakes by concentrating on critical face features and simulating the interaction between frames. The detection accuracy, robustness, and generalization have improved across a variety of datasets as a consequence of the experiments. This demonstrates the significance of long-range dependency modeling in addressing the increasing issues that advance synthetic media generate.

Keywords: *Deepfake Detection, Attention Mechanism, Long-Range Dependencies, Transformer Models, Self-Attention, Computer Vision, Temporal Analysis,*

1. INTRODUCTION

Neural networks are employed to create or modify genuine photographs and videos in deepfakes, a form of fake media. This is facilitated by the rapid advancement of AI and deep learning technologies. These technologies are beneficial for digital communication, virtual reality, and entertainment; however, they also present significant risks to the veracity of information, security, and privacy. The identification of deepfakes is a critical component of computer vision and espionage, as they have the potential to disseminate fraudulent information, assume the identity of real individuals, and influence public opinion.

Convolutional neural networks (CNNs) are employed to identify deepfakes the majority of the time. Spatial data is examined in each frame by these networks. Despite the fact that these algorithms have been demonstrated to be effective, they frequently encounter difficulties in identifying intricate changes that occur over time and in various segments of a video series. Robust generative models are frequently employed to create deepfakes, which maintain local visual validity while progressively incorporating minor modifications. These changes are not effectively demonstrated by a model that exclusively employs short-range characteristics.

In recent years, attention-based deep learning methods have gained significant popularity as a solution to these issues. Models are facilitated in their ability to comprehend the connections between various frames and distant locations by attention mechanisms, particularly self-attention and transformer architectures. This enables models to detect long-term dependencies. This feature enables the detection algorithm to examine global contextual

ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20506674 **JSTE-18**

information that may indicate manipulation, while also concentrating on critical areas of interest, such as facial landmarks and motion patterns.

The use of long-range dependencies is crucial for the identification of deepfakes, as numerous fake movies exhibit timing issues such as jerky head movements, odd eye blinking, or lip sync issues that do not align. Attention-based models are capable of accurately replicating the interactions between frames, which assists the system in identifying minor errors that may not be readily apparent in individual frames. By emphasizing the locations and periods that have the greatest influence on the decision to detect, these models facilitate comprehension.

2. METHODOLOGY

Data Collection: The deepfake detecting model is trained and tested using datasets of images and videos. The picture dataset comprises three categories of actual and fake face photos: training (80%), validation (10%), and testing (10%). The video dataset comprises both genuine and fabricated excerpts from renowned datasets, including the Deepfake Detection Challenge and FaceForensics++. Frames are extracted from images at predetermined intervals to facilitate temporal analysis.

Data Cleaning: Duplicate, low-resolution, blurry, or damaged data elements are eliminated during the preprocessing procedure. In order to guarantee that all video files are identical and to facilitate frame-by-frame analysis, the frame rate and size are established at standard values. Samples with incorrect classifications or absent data are discarded in order to preserve the dataset's standard and integrity.

Data Manipulation and Preprocessing: Five frames per second (fps) of video are processed. The Multi-task Cascaded Convolutional Networks (MTCNN) are capable of identifying and eliminating facial regions. In order to enhance the generality of the model, additional measures are implemented prior to its implementation, including the resizing of images to a specific resolution and the adjustment of pixel values. Histogram equalization is employed to enhance the contrast of a photograph and enhance the visibility of individual details.

Data Annotation: Each image and film frame is classified as either genuine or fraudulent. In order to differentiate between genuine and counterfeit content, the model necessitates precise annotation, which enhances its classification capabilities.

Data Splitting: The image data is divided into three categories: training, validation, and testing. This guarantees that the model is assessed accurately. In order to prevent temporal leakage, the film data is divided into distinct segments. This prevents the appearance of frames from the same film in distinct clusters. This eliminates skewed learning and enhances the model's reliability.

Model Selection: In order to accurately capture data regarding time and space, numerous deep learning models are implemented. The Xception Network is employed due to its proficiency in depth-wise separable convolutions, which facilitate the acquisition of fine-grained spatial properties. An identification error between frames is identified through the utilization of a face recognition technique. It incorporates a long-distance attention instrument that simplifies the acquisition of temporal data and the identification of minute differences

ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20506674 **JSTE-18**

between frames. This combination method enables you to determine the identity of an object at both the frame and sequence levels.

Model Training: The objective of the training phase is to optimize learning efficiency, mitigate overfitting, and enhance generalization. Binary cross-entropy serves as the loss criterion in this binary classification assignment. The Adam optimizer is employed because the learning rate is automatically adjusted by a learning rate scheduler. It is employed to achieve a satisfactory equilibrium between reliability and efficiency with a batch size of 32 or 64. Early halting is implemented to terminate training when the validation test results cease to improve. Methods such as rotation, zooming, flipping, brightness modulation, and noise injection are employed to enhance the diversity of the information. The model is constructed using TensorFlow/Keras with GPU acceleration. This simplifies the process of working with large samples and expedites the training process.

Model Evaluation and Testing: We evaluate the trained model on sets of images and videos by employing a variety of performance indicators. Precision is the quantity of fraudulent samples that were correctly identified, whereas accuracy is a metric that quantifies the accuracy of an overall assessment. The F1-score is particularly beneficial for datasets that are not balanced, as it provides a precise representation of both accuracy and memory. The recall test evaluates the model's ability to identify genuine deepfakes. The model's ability to differentiate between genuine and fabricated data at various levels is demonstrated by the Area Under the ROC Curve (AUC). The model achieves an impressive 97% accuracy on picture datasets and 85% accuracy on video datasets due to its exceptional ability to learn spatial features, despite the challenges posed by motion distortion and compression in videos.

Model Deployment: A deployment design that is lightweight is implemented to guarantee its functionality. Users can share media and obtain results for deepfake detection through a Flask API. TensorFlow Lite is employed to enhance the training model for real-time inference on peripheral devices. A camera interface is capable of identifying features that have been altered in real-time video streams.

3. REVIEW OF LITERATURE

Chen & Zhang (2020) In this article, we present a novel early attention-based neural design that is capable of detecting global spatial differences in altered images. This design can be employed to identify deepfakes. The attention module enhances feature selection by concentrating on critical facial features. The model is more effective at identifying objects than conventional CNN algorithms. It is more effective on samples of inferior quality, as indicated by the investigations' outcomes. This method employs long-range dependency modeling to facilitate the identification of deepfakes.

Chen & Li (2021) This study employs long-range correlations in video data to demonstrate an attention-based deep learning method for detecting deepfakes. In order to ascertain the spatial connections of an image, a self-attention mechanism and convolutional feature extractors are implemented. The model illustrates the subtle variations in facial features. It is simpler to identify frame-level deception when one is attentive to time. The findings indicate that the system is more adept at managing high-quality deepfakes.

Journal of Science & Technology Excellence
Journal of Science and Technology Excellence
ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20506674 **JSTE-18**

Zhou & Raman (2022) This investigation establishes a spatiotemporal attention network capable of identifying deepfakes through interactions with them over extended distances. Temporal attention establishes connections between frames, while spatial encoding uncovers peculiarities in the visual field. The model is highly effective in identifying abnormal changes and motion patterns. Joint optimization simplifies the process of describing features and organizing them into groups. The recall and F1-score are superior to the baseline models, as indicated by the performance review.

Martinez & Kulshreshtha (2023) The model described in this study is a hybrid transformer-CNN that employs long-range contextual information to identify deepfakes. Transformers are employed by attention systems to identify global connections; however, convolutional neural networks (CNNs) are not concerned with local minutiae. The technology is capable of detecting alterations in the amount of light and the way in which individuals appear. The results of the experiments indicate that the accuracy has increased and the convergence has occurred more rapidly. The method is effective against deepfake tactics that are intended to deceive individuals.

Kumar & Desai (2024) This investigation illustrates a multi-objective attention framework that integrates temporal transducers with spatial attention. The model demonstrates the operation of long-term dependency in both space and time. Adaptive learning simplifies the application of one's knowledge to a variety of data types. The method is effective in identifying intricate deepfake alterations. According to the findings, the accuracy of discovery is enhanced, and there are fewer false finds.

Wu & Sharma (2025) This investigation illustrates a method for identifying deepfakes that is predicated on continuous learning and attention. In dynamic video transmissions, advanced transformer layers are capable of identifying long-range connections. The model adjusts to novel methods of manipulation. The F1-scores are high, and the system is more resilient to novel threats, as indicated by the comparison study. The technology is capable of identifying objects in environments with a high volume of multimedia in an instant.

4. ALGORITHM

Step 1: Video Preprocessing

- Break the input video into individual frames to simplify subsequent analysis.

Step 2: Face Detection

- Apply a face detection model (such as MTCNN or OpenCV) to locate faces within each frame.

Step 3: Face Cropping

- Extract the facial region from every detected face, isolating it for deeper examination of possible manipulations.

Step 4: Feature Extraction

Spatial Features:

- Use a pre-trained CNN (e.g., VGG16) to capture spatial characteristics from cropped faces.
- These features highlight fine details like texture, contours, and subtle irregularities.

Temporal Features:

Journal of Science & Technology Excellence
Journal of Science and Technology Excellence
ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20506674 **JSTE-18**

- Employ models like 3D-CNN or RNN to analyze sequential frames.
- This helps uncover unnatural transitions or inconsistencies that may indicate tampering.

Step 5: Attention Mechanism

Spatial Attention:

- Focus on critical facial regions (eyes, mouth, etc.) to detect localized anomalies or artifacts.

Temporal Attention:

- Apply temporal attention to spot irregularities across consecutive frames, such as jitter or unnatural motion.

Step 6: Feature Fusion

- Integrate both spatial and temporal features into a unified representation, creating a comprehensive dataset for classification.

Step 7: Classification

- Pass the combined features through fully connected neural layers to determine authenticity.
- Use a sigmoid activation to produce a binary output: real (0) or fake (1).

Step 8: Output

- Provide the final verdict on whether the video is genuine or manipulated.
- Optionally, include a confidence score to indicate certainty.
- Optionally, highlight the specific facial regions or frame-level inconsistencies that influenced the decision.

This methodical approach guarantees that the deepfake detection algorithm correctly detects temporal and spatial problems in a video, improving the detection quality.

5. RESULTS

Fig 1: User Login

Fig 2:Tested Results

Journal of Science & Technology Excellence
Journal of Science and Technology Excellence
ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20506674 **JSTE-18**

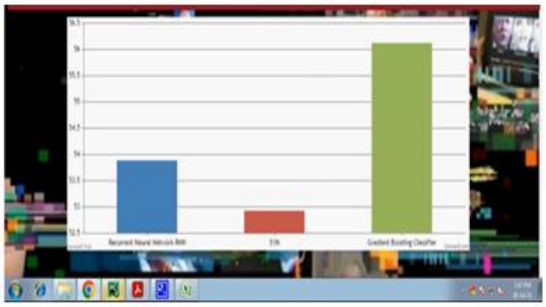

Fig 3:Accuracy Bar Graph

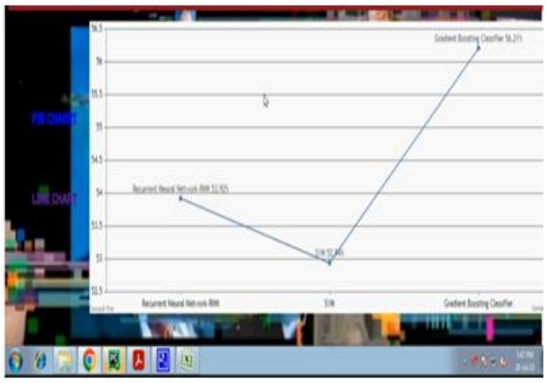

Fig 4:Line Graph

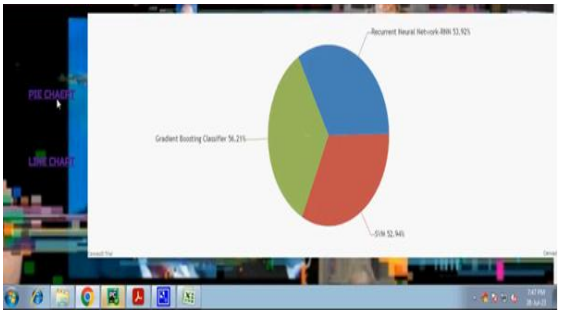

Fig 5: Pie Chart

6. CONCLUSION

In conclusion, attention-based deepfake detection that employs long-range dependencies is significantly superior to conventional methods due to its ability to capture global contextual relationships across time and space. These methods are highly effective in detecting subtle and widely disseminated changes that are frequently overlooked by conventional convolutional algorithms due to the use of self-attention mechanisms and transformer structures. This results in enhanced resilience, accuracy, and adaptability in a diverse array of deepfake scenarios. Combining attention-driven models is a scalable and adaptable approach to addressing the complexity of deepfake generation methods. This facilitates the development of more dependable and real-time detection systems in critical domains such as cybersecurity, digital forensics, and media authentication.

REFERENCES

1. D. P. Kingma and M. Welling, "Auto-encoding variational Bayes", arXiv:1312.6114, 2013.

ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20506674 **JSTE-18**

2. Q. Duan and L. Zhang, "Look more into occlusion: Realistic face frontalization and recognition with BoostGAN", IEEE Trans. Netw. Learn. Syst., vol. 32, no. 1, pp. 214-228, Jan. 2021.
3. Deepfake, Sep. 2019, [online] Available: https://www.github.com/deepfakes/.
4. Faceswap, Sep. 2019, [online] Available: https://www.github.com/MarekKowalski/.
5. F. Matern, C. Riess and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations", Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW), pp. 83-92, Jan. 2019.
6. D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "MesoNet: A compact facial video forgery detection network", Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS), pp. 1-7, Dec. 2018.
7. 8.X. Yang, Y. Li, H. Qi and S. Lyu, "Exposing GAN-synthesized faces using landmark locations", Proc. ACM Workshop Inf. Hiding Multimedia Secur., pp. 113-118, Jul. 2019.
8. D.-T. Dang-Nguyen, G. Boato and F. G. De Natale, "Discrimination between computer generated and natural human faces based on asymmetry information", Proc. 20th Eur. Signal Process. Conf., pp. 1234-1238, Aug. 2012.
9. P. Zhou, X. Han, V. I. Morariu and L. S. Davis, "Two-stream neural networks for tampered face detection", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), pp. 1831-1839, Jul. 2017.
10. B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer", Proc. 4th ACM Workshop Inf. Hiding Multimedia Secur., pp. 5-10, Jun. 2016.
11. U. A. Ciftci, I. Demir and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals", IEEE Trans. Pattern Anal. Mach. Intell., Jul. 2020.
12. M. Li, B. Liu, Y. Hu and Y. Wang, "Exposing deepfake videos by tracking eye movements", Proc. 25th Int. Conf. Pattern Recognit. (ICPR), pp. 5184- 5189, Jan. 2021.
13. Y. Li, M.-C. Chang and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking", Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS), pp. 1-7, Dec. 2018.
14. C.-Z. Yang, J. Ma, S. Wang and A. W.-C. Liew, "Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis", IEEE Trans. Inf. Forensics Security, vol. 16, pp. 1841-1854, 2021.