Deep-Fake Detection Using Deep Learning

Nagashree K T; Shristi ; Sania Firdaushi ; Shweta B Patil ; Shristi Singh

Information Science and Engineering, AMC College of Engineering, Bengaluru, Karnataka, India

Publication Date: 2025/02/05

Abstract: Deep-Fake Detection is a new technology which has caught extreme fashionability in the present generation. Deep-Fake has now held serious pitfalls over spreading misinformation to the world, destroying political faces and also blackmailing individualities to prize centrals. As this technology has taken over the internet in a veritably short span of time and also numerous readily apps are also available to execute Deep-Fake contents, and numerous of the individualities has made systems grounded on detecting the deepfake contents whether it's fake or real. From the DL(deep learning) – grounded approach good results can be attained, this paper presents substantially the results of our current study which is indicating the traditional machine learning (ML) fashion. This projects points in discovery of video tape deepfakes using deep literacy ways like ResNext and LSTM. We've achived deepfake discovery by using transfer literacy where the pretrained ResNext CNN is used to gain a point vector, further the LSTM subcaste is trained using the features.

Keywords: *Detection, Deepfakes, Deep Learning, ResNext and LSTM.*

How to Cite: Nagashree K T; Shristi ; Sania Firdaushi ; Shweta B Patil ; Shristi Singh (2025). Deep-Fake Detection Using Deep Learning. *International Journal of Innovative Science and Research Technology*, 10(1), 1700-1706. https://doi.org/10.5281/zenodo.14808073

I. INTRODUCTION

Deepfake technology represents a significant advancement in artificial intelligence, allowing the creation of largely realistic but fabricated vids and images. These synthetic media pieces can make individualities appear to perform conduct or speak words they noway actually did, posing serious challenges for media authenticity, sequestration, and security. Detecting these manipulations is pivotal, and deepfake analysis using ResNext and Long Short- Term Memory(LSTM) models offers a robust result by combining the strengths of both convolutional and intermittent neural networks.

ResNext is an advanced variant of the Residual Network(ResNet) armature, specifically designed to ameliorate point birth capabilities through its innovative cardinality dimension. This conception of cardinality refers to the number of resemblant metamorphoses applied at each subcaste, which allows the network to reuse the input data through multiple pathways contemporaneously. This design enhances ResNext's capability to capture complex patterns and subtle details within visual data, making it largely effective for relating inconsistencies that may indicate manipulation.

The Long Short- Term Memory(LSTM) networks, a type of intermittent neural network(RNN), exceed at handling successional data and long- term dependences . This makes them particularly suitable for tasks involving time series or videotape sequences. In the environment of deepfake discovery, LSTM models are employed to examine the temporal thickness of videotape frames.

By assaying the sequence of features uprooted by ResNext, LSTMs can identify temporal anomalies that suggest manipulation, similar as unnatural movements or transitions that would not do in genuine vids.

The integration of ResNext and LSTM models provides a comprehensive approach to deepfake discovery by using both spatial and temporal confines of videotape data. originally, vids are broken down into individual frames, which are also preprocessed to insure a harmonious size and format. Each preprocessed frame is passed through the ResNext model to prize detailed point vectors representing the visual characteristics of the frame. These features are also organized into sequences that correspond to the frames of each videotape and fed into the LSTM model. The LSTM network analyzes these sequences to descry any temporal inconsistencies, and grounded on this analysis, the system classifies the videotape as real or fake.

This combined approach offers several advantages. The delicacy of deepfake discovery is significantly bettered by considering both the detailed spatial features and the temporal dynamics of the videotape data. The robustness of the system is enhanced, making it effective against a variety of deepfake ways, including those involving subtle and complex manipulations. also, the scalability of these models allows them to handle large datasets, icing their connection in real- world scripts similar as social media monitoring, digital forensics, and content verification.

As deepfake technology continues to evolve, the need for sophisticated discovery systems becomes decreasingly critical. The use of ResNext and LSTM models in deepfake analysis represents a slice- edge result that combines advanced point birth and temporal analysis capabilities. By icing the integrity of digital media, this approach helps combat misinformation and protects sequestration, playing a vital part in maintaining trust in the digital age.

II. LITERATURE SURVEY

- Jaiswal et al.(2020) proposed a deepfake discovery system using ResNeXt for face image analysis. Their system used a pretrained ResNeXt model on face images to classify whether the face was real or synthetic. The results demonstrated that ResNeXt handed a high type delicacy compared to other models like ResNet and VGGNet.
- Cozzolino et al.(2018) explored the use of ResNeXt as part of a larger model to descry manipulated facial features in images. By training ResNeXt on a dataset containing both real and manipulated faces, they achieved a robust performance in distinguishing deepfakes from genuine images.
- Xue et al.(2020) extended the idea of ResNeXt- rested point birth and incorporated spatial attention mechanisms to concentrate on critical face regions, perfecting discovery delicacy in deepfake facial images. 4). Dolhansky et al.(2020) estimated deepfake videotape discovery using the ResNeXt model. They trained the network on a combination of frame- position CNNs and temporal modeling ways, achieving high discovery performance, especially when fine- tuned with deepfake-specific datasets like FaceForensics.
- Dang et al.(2019) employed an LSTM- rested model to descry temporal inconsistencies across frames of a videotape. Their system involved training LSTM units on sequences of frames, learning the temporal dynamics of mortal faces, and relating irregularities convinced by deepfake generation processes.
- Afchar et al.(2018) introduced the first LSTM- rested frame for detecting deepfakes by exploiting temporal features. The model used a combination of CNN for point birth and LSTM for sequence knowledge. It was shown to be effective in detecting vids with synthetic faces, indeed when high- quality GAN models were used for deepfake generation.
- Yang et al.(2020) explored the use of an LSTM model in convergence with autoencoders for deepfake discovery. The LSTM was assigned with detecting anomalies in facial movements, analogous as unnatural blinking or lip sync crimes, which are constantly present in deepfake vids. “ Enabling Robots to Understand Incomplete Natural Language Instructions Using firm sense ” by Haonan Chen et al.,(2020)

This paper introduces a new approach nominated Language- Model rested firm sense(LMCR), which empowers a robot to comprehend natural language instructions from humans, assess its girding terrain, and autonomously infer any missing details from the instruction by exercising contextual information and a new establishment sense frame. The disquisition was presented at the 2020 IEEE International Conference on Robotics and automation(ICRA). Natural language is naturally unstructured and constantly depends on common sense for interpretation, posing significant challenges for robots in directly and effectively understanding analogous language. For case, in a domestic terrain where a robot is holding a bottle of water alongside scissors, a plate, bell peppers, and a mug on a table, a mortal muscle instruct the robot to “ pour me some water. ” From the robot's viewpoint, this command lacks particularity regarding the destination for the water, whereas a human would presumably infer that the water should be poured into the mug. A robot equipped with the capability to privately address analogous inscrutability in natural language instructions, akin to mortal sense, would grease more natural relations with humans and enhance its overall functionality. The Language- Model rested firm sense(LMCR) approach enables a robot to hear to mortal instructions, observe its terrain, resolve any inscrutability, and subsequently execute the designated task autonomously.

III. METHODOLOGY

❖ *Proposed Solution*

Deepfake detection has become a critical task in the age of advanced artificial intelligence, where synthetic media is increasingly convincing and widespread. Leveraging deep learning models such as ResNeXt and Long Short-Term Memory (LSTM) networks for deepfake detection combines the strengths of both spatial and temporal feature extraction. This methodology focuses on utilizing these two deep learning models to distinguish between real and fake content in videos, aiming to improve the accuracy and robustness of detection.

Deepfake videos are created using advanced generative models like Generative Adversarial Networks (GANs), which can manipulate visual and auditory content to make it appear authentic. As deepfake technology evolves, so too must the methods used to detect it. In this context, a combination of convolutional neural networks (CNNs) like ResNeXt for spatial feature extraction and recurrent neural networks (RNNs) like LSTMs for temporal feature learning offers a potent approach.

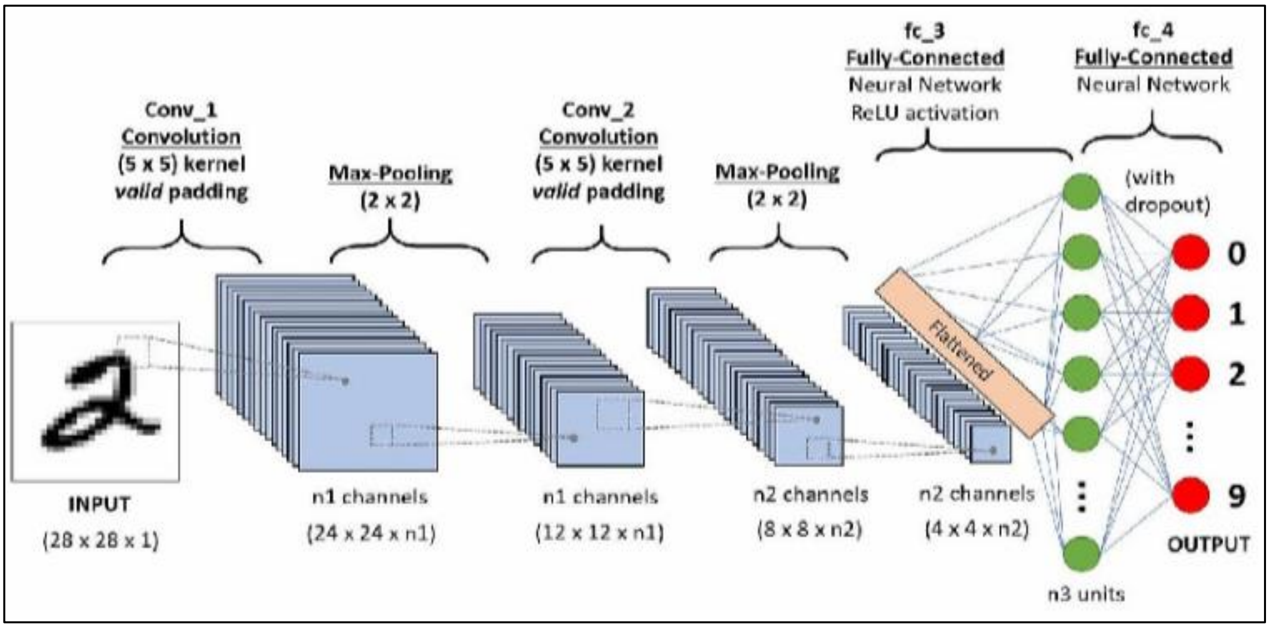

Fig 1: Working of CNN Model

A. Data Collection and Preprocessing :-

The first step in deepfake detection involves collecting a comprehensive dataset containing both real and deepfake videos. Popular datasets for this task include the DeepFake Detection Challenge (DFDC), Celeb-DF, and FaceForensics++. These datasets provide labeled video data, with the real videos representing authentic content and the deepfakes containing manipulated media. Preprocessing of this data is essential for ensuring that the model receives high-quality input.

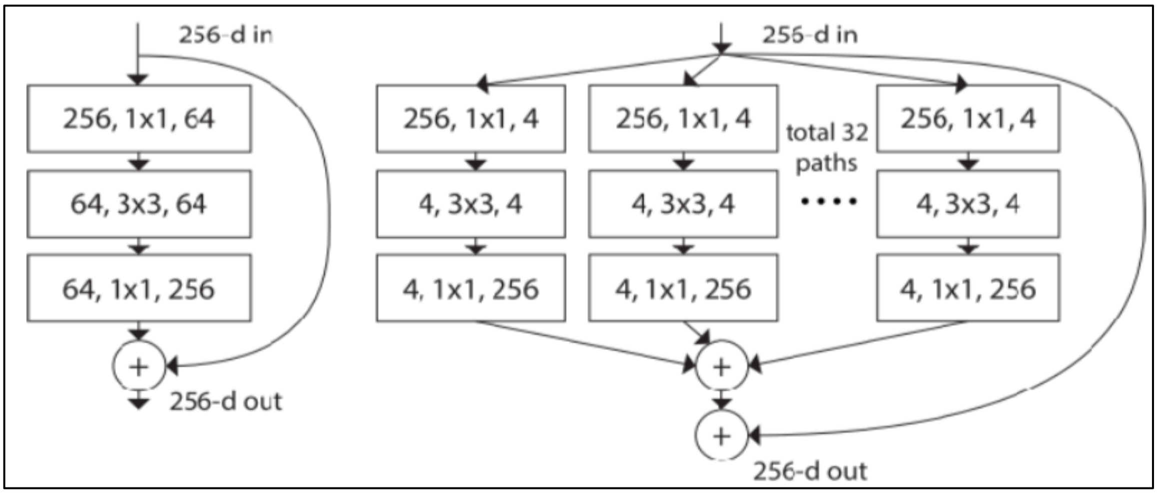

Fig 2 : ResNeXt Model Architecture

For each video, the frames are extracted and then preprocessed to facilitate effective model learning. Frame extraction involves converting the video into individual frames at a certain frame rate, typically 30 frames per second, so that each frame is treated as an independent image for analysis. Face detection algorithms such as MTCNN or Dlib are employed to locate and align faces in each frame, ensuring that the model focuses on the relevant portions of the video. After face alignment, the frames are normalized to a consistent size and pixel value range, typically rescaled between 0 and 1 or standardized to have zero mean and unit variance.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14808073

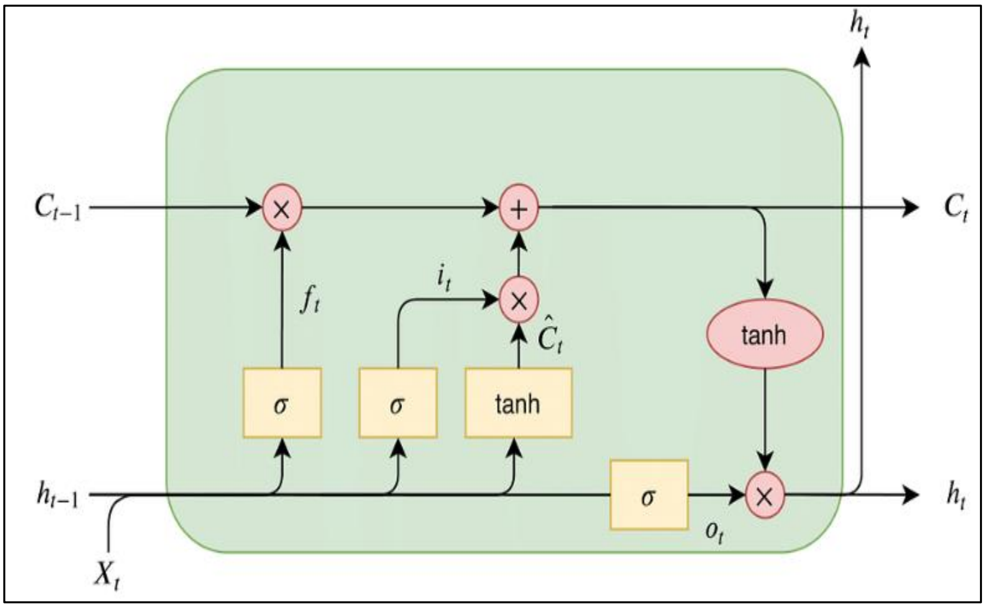

Fig 3 : LSTM Model Work Flow

Additionally, data augmentation techniques, such as random rotations, flips, and scaling, are often applied to increase the variety of training examples and prevent overfitting. This ensures that the model learns to generalize better and does not become overly sensitive to specific facial angles or positions.

B. Feature Extraction Using ResNeXt :-

Once the data is preprocessed, the next step is to extract relevant features from the frames using a deep convolutional neural network (CNN). In this case, the ResNeXt architecture is used for feature extraction. ResNeXt is a variant of the traditional ResNet model that incorporates a concept known as "cardinality" – the number of independent paths through the network. By increasing the cardinality, ResNeXt is able to capture richer, more diverse features without a significant increase in computational complexity.

The ResNeXt model is pretrained on large datasets like ImageNet and then fine-tuned for deepfake detection. Fine-tuning involves updating the weights of the network to better adapt it to the specific characteristics of real and fake video frames. During this process, the model learns to identify subtle discrepancies between authentic and manipulated faces, such as unnatural facial expressions, lighting inconsistencies, and irregularities in facial texture and details that are often present in deepfakes.

At the end of the ResNeXt network, the output is typically a high-dimensional feature vector that encodes spatial information about each frame. These feature vectors are crucial as they contain the learned representation of the frame's content, which is then passed on to the temporal model for further analysis.

C. Temporal Feature Learning Using LSTM:-

While ResNeXt excels at extracting spatial features, detecting deepfakes in videos requires understanding how these features evolve over time. Deepfake videos often contain subtle artifacts in the way faces move or interact with their surroundings. To capture these temporal dynamics, a Long Short-Term Memory (LSTM) network is employed.

LSTMs are a special type of recurrent neural network (RNN) designed to capture long-range dependencies in sequential data. This is particularly useful for video analysis, where the relationships between consecutive frames hold important clues to the authenticity of the video. Unlike traditional RNNs, LSTMs are equipped with gates that regulate the flow of information, allowing the network to remember or forget information over long sequences, which is essential when analyzing the entire video.

In this methodology, the feature vectors produced by ResNeXt from each frame are fed into the LSTM network as a sequence. The LSTM processes these sequential features, learning the temporal dependencies between frames. It is through this analysis that the LSTM can detect anomalies such as unnatural movement patterns, inconsistencies in facial expressions, or other temporal artifacts that are characteristic of deepfakes.

To enhance the temporal modeling, a bidirectional LSTM can be used. A bidirectional LSTM processes the sequence of frames in both forward and backward directions, which helps capture both past and future context in the video. This further improves the model's ability to detect subtle inconsistencies that may not be evident when only considering frames in a single direction.

D. *Classification Layer and Model Output:-*

After processing the video sequence through both ResNeXt and LSTM, the next step is to classify the video as either real or fake. The output of the LSTM network is a sequence of hidden states that summarize the temporal information learned from the frames. These hidden states are passed through a fully connected layer, which is followed by a softmax or sigmoid activation function to produce a final classification.

In binary classification tasks, such as deepfake detection, the output layer typically generates two classes: one indicating the video is real and the other indicating it is a deepfake. The model is trained using a loss function like binary cross-entropy, which measures the difference between the predicted and actual labels. Optimization techniques such as Adam or stochastic gradient descent (SGD) are used to minimize the loss function and update the model's parameters during training.

E. *Model Evaluation and Deployment:-*

Once the model is trained, it is important to evaluate its performance. Common evaluation metrics for binary classification tasks include accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic curve (AUC-ROC). These metrics give insights into how well the model distinguishes between real and fake videos.

To ensure the model generalizes well to unseen data, techniques like cross-validation or hold-out validation are used. Cross-validation involves splitting the dataset into multiple subsets and training the model on different subsets while testing on others. This ensures that the model is not overfitting to any particular portion of the data.

After evaluation, the trained model can be deployed for real-time deepfake detection. The system can be used to automatically analyze incoming video content, providing a label for each video, indicating whether it is real or fake.

IV. RESULTS AND DISCUSSIONS

❖ *Existing Methods*

In the realm of deepfake detection, leveraging advanced machine learning techniques such as ResNeXt and Long Short-Term Memory (LSTM) networks has demonstrated significant promise. Both of these models are chosen for their unique strengths in handling the specific challenges posed by deepfake content, which involves subtle manipulations of image or video data that often require the consideration of both spatial and temporal patterns.

ResNeXt is a deep convolutional neural network (CNN) architecture that is designed to be both highly efficient and effective in capturing spatial features from images. Its primary strength lies in its ability to learn features from input data with a high degree of accuracy while maintaining computational efficiency. By introducing a cardinality dimension to the network architecture, ResNeXt can handle more complex feature extraction tasks with fewer parameters than traditional CNNs, thus improving both its performance and generalization capabilities. In the context of deepfake detection, ResNeXt is particularly adept at recognizing inconsistencies in facial features, textures, and other visual artifacts that are commonly introduced in manipulated images or videos. These artifacts may be subtle and difficult for the human eye to detect, but deep learning models like ResNeXt can efficiently identify them, thus improving the accuracy of deepfake detection.

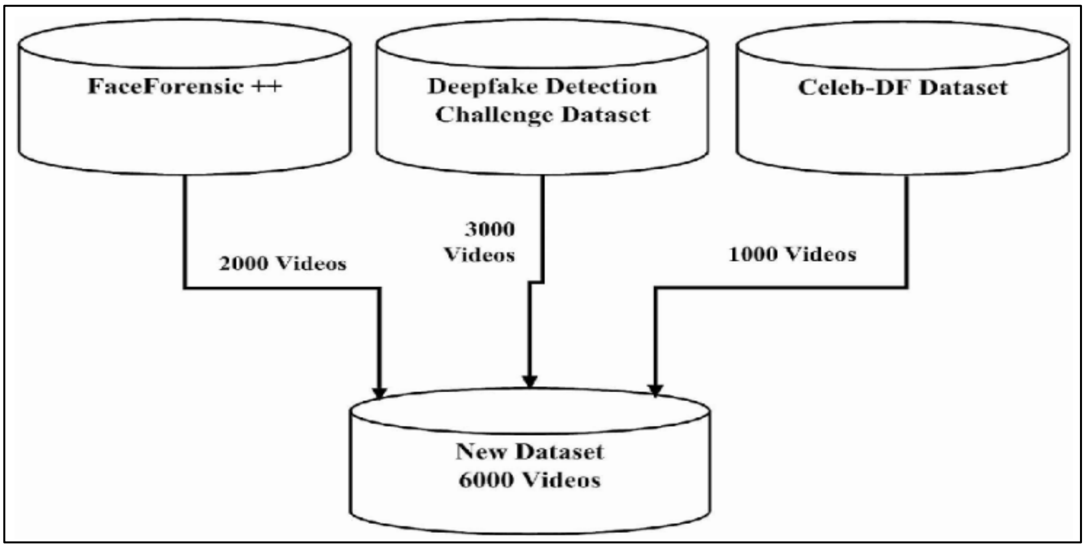

Fig 4 : Dataset of Deep-Fake

On the other hand, **LSTM networks**, a type of recurrent neural network (RNN), are designed to model sequential data and capture long-range dependencies within such data. LSTMs are particularly useful in deepfake detection when analyzing video content, as they can learn the temporal relationships between consecutive video frames. Deepfakes often exhibit abnormalities in facial movements, lip synchronization, or eye motion that become evident only when considering the flow of video data over time. LSTMs are capable of identifying these temporal inconsistencies by

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14808073

modeling the sequence of frames in a video, detecting irregular patterns that signal potential manipulation.

When used in tandem, ResNeXt and LSTM form a powerful combination for deepfake detection. ResNeXt's role is to extract robust spatial features from each individual frame of the video, identifying artifacts that may not be apparent in the raw data. The extracted features are then passed to the LSTM, which analyzes the temporal sequence of frames to detect any abnormal transitions or inconsistencies that emerge across the video. This dual approach — combining spatial and temporal analysis — allows the system to catch a wide range of deepfake manipulations, from those that alter static facial features to those that cause unnatural movements over time.

The integration of these two models enables a more comprehensive understanding of deepfake content. ResNeXt ensures that each frame is properly analyzed for visual discrepancies, while LSTM provides the necessary context to understand how those discrepancies evolve over time. This two-step process not only improves the detection accuracy but also enhances the model's ability to generalize across different types of deepfakes, making it more robust to diverse manipulation techniques.

However, while this approach has proven effective, it is not without its challenges. The computational demands of training ResNeXt and LSTM together can be quite high, especially when dealing with large datasets or high-resolution video content. Furthermore, as deepfake generation techniques continue to evolve, new challenges emerge, such as the creation of deepfakes designed to evade detection models. Despite these hurdles, the combined use of ResNeXt and LSTM offers a promising pathway forward for deepfake detection, providing a solid foundation for future advancements in the field.

In conclusion, the integration of ResNeXt and LSTM for deepfake detection is a powerful approach that leverages the strengths of both spatial feature extraction and temporal sequence modeling. While challenges remain, particularly in terms of computational resources and the adaptability of detection models to new deepfake techniques, this combination represents a significant step forward in the fight against digital content manipulation.

V. CONCLUSION

Combining **ResNeXt** and **LSTM** for deepfake detection effectively captures both spatial and temporal features. ResNeXt excels at extracting detailed spatial information from individual frames, identifying artifacts and inconsistencies in facial features. LSTM analyzes the temporal dynamics across video frames, detecting unnatural movements or discrepancies in facial expressions, lip-syncing, or eye movement. This synergy improves the model's ability to detect even subtle deepfakes that may not be apparent in single-frame analysis. The approach demonstrates strong accuracy and robustness, offering a powerful solution for identifying advanced, high-quality deepfakes, though challenges in generalization remain.

REFERENCES

[1]. Botha J, Pieterse H. Fake news and deepfakes: A dangerous threat for 21st century information security, in ICCWS 2020 15th International Conference on Cyber Warfare and Security, 2020, pp. 1–57.

[2]. Wagner TL, Blewer A. The word real is no longer real': deepfakes, gender, and the challenges of Ai-altered video. Open Inform Sci. 2019;3:32–46.**Article Google Scholar**

[3]. Hancock JT, Bailenson JN. The social impact of deepfakes, Cyberpsychology, Behavior, and Social Networking, vol. 24, no. 3, pp. 149–152, 2021.

[4]. Mirsky Y, Lee W. The creation and detection of deepfakes: a survey. ACM Comput Surv (CSUR). 2021;54(1):1–41. **Article Google Scholar**

[5]. Masood M, Nawaz M, Malik KM, Javed A, Irtaza A, Malik H. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward, Applied Intelligence, vol. 53, no. 4, pp. 3974–4026, 2023.

[6]. Kietzmann J, Lee LW, McCarthy IP, Kietzmann TC. Deepfakes: Trick or Treat? Bus Horizons. 2020;63(2):135–46.**Google Scholar**

[7]. Yu Y, Liu X, Ni R, Yang S, Zhao Y, Kot AC. PVASS-MDD: predictive visual-audio alignment self-supervision for Multimodal Deepfake Detection. IEEE Trans Circuits Syst Video Technol. Jan. 2023;1(99):1–2.

[8]. Mukta MSH, Ahmad J, Raiaan MAK, Islam S, Azam S, Ali ME, Jonkman M. An investigation of the effectiveness of Deepfake models and tools. J Sens Actuator Networks. 2023;12(4):1–61.**Google Scholar**

[9]. Salvi D, Liu H, Mandelli S, Bestagini P, Zhou W, Zhang W, Tubaro S. A Robust Approach to Multimodal Deepfake Detection. J Imaging. 2023;9(6):1–22. **Article Google Scholar**

[10]. Ismail A, Elpeltagy M, Zaki MS, Eldahshan K. A new deep learning-based methodology for video deep fake detection using xgboost, Sensors, vol. 21, no. 16, pp. 13–54, 2021.

[11]. França RP, Monteiro ACB, Arthur R, Iano Y. An overview of deep learning in big data, image, and signal processing in the modern digital age. in Trends Deep Learn Methodologies, 2021, pp. 63–87.

[12]. Castillo Camacho I, Wang K. A comprehensive review of deep-learning-based methods for image forensics. J Imaging. 2021;7(4):6–9.**Article Google Scholar**

[13]. Rashid MM, Lee SH, Kwon KR. Blockchain technology for combating deepfake and protect video/image integrity. J Korea Multimedia Soc. 2021;24(8):1044–58.**Google Scholar**

[14]. Yang J, Sun Y, Mao M, Bai L, Zhang S, Wang F. Model-agnostic method: exposing deepfake using pixel-wise spatial and temporal fingerprints. IEEE Trans Big Data. 2023;9(6):1496–509.**Article Google Scholar**

[15]. Ganiyusufoglu I, Ngô LM, Savov N, Karaoglu S, Gevers T. Spatio-temporal features for generalized detection of deepfake videos, in Computer Vision and Image Understanding, vol. 1, pp. 1–11, 2022.

[16]. Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network, Physica D: Nonlinear Phenomena, vol. 404, art. 132306, 2020.

[17]. Tariq S, Lee S, Woo SS. A convolutional lstm based residual network for deepfake video detection, Conference'17, Washington DC, 2020, pp. 1–11.

[18]. Oyetoro A. Image Classification of Human Action Recognition Using Transfer learning in Pytorch. Int J Adv Res Ideas Innovations Technol. Apr. 2023;9(2):1–6.

[19]. Jha M, Tiwari A, Himansh M, Manikandan VM, Face Recognition: Recent Advancements and Research Challenges, in. 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2022, pp. 1–6.

[20]. Li B, Lima D. Facial expression recognition via ResNet-50. Int J Cogn Comput Eng. 2021;2:57–64 **Google Scholar**

[21]. Shad HS, Rizvee MM, Roza NT, Hoq SM, Monirujjaman Khan M, Singh A, Zaguia A, Bourouis S. Comparative analysis of deepfake image detection method using convolutional neural network. Comput Intell Neurosci. 2021;1:1–20.**Article Google Scholar**

[22]. Amerini I, Galteri L, Caldelli R, Del Bimbo A. Deepfake video detection through optical flowbased CNN, in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 1–3.

[23]. Kohli A, Gupta A. Detecting deepfake, faceswap and face2face facial forgeries using frequency CNN. Multimedia Tools Appl. 2021;80:18461–78.**Article Google Scholar**

[24]. Saikia P, Dholaria D, Yadav P, Patel V, Roy M. A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features, in 2022 International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1–7.

[25]. Tran VN, Lee SH, Le HS, Kwon KR. High performance deepfake video detection on CNN-based with attention target-specific regions and manual distillation extraction, Applied Sciences, 11, 16, pp. 76–8, 2021.

[26]. Patel Y, Tanwar S, Bhattacharya P, Gupta R, Alsuwian T, Davidson IE, Mazibuko TF. An Improved dense CNN Architecture for Deepfake Image Detection. IEEE Access. 2023;11:22081–95.**Article Google Scholar**

[27]. Masud U, Sadiq M, Masood S, Ahmad M, El-Latif A, Ahmed A. LW-DeepFakeNet: a lightweight time distributed CNN-LSTM network for real-time DeepFake video detection, Signal, Image and Video Processing, pp. 1–9, 2023.

[28]. Warke K, Dalavi N, Nahar S. DeepFake Detection through deep learning using ResNext CNN and LSTM. IEEE Trans Neural Networks Learn Syst. 2023;10(5):1–10.**Google Scholar**

[29]. Botha J, Pieterse H. Fake news and deepfakes: A dangerous threat for 21st-century information security, in ICCWS 2020 15th International Conference on Cyber Warfare and Security, Academic Conferences and Publishing Limited, March 2020, pp. 1–57.

[30]. Tariq S, Lee S, Kim H, Shin Y, Woo SS. Detecting both machine and human created fake face images in the wild, in Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, January 2018, pp. 81–87.

[31]. Liu H, Li X, Zhou W, Chen Y, He Y, Xue H et al. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 772–781, 2021.

[32]. Sun Z, Han Y, Hua Z, Ruan N, Jia W. Improving the efficiency and robustness of deepfakes detection through precise geometric features, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3609–3618.

[33]. Li J, Xie H, Li J, Wang Z, Zhang Y. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6458–6467.