IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AI-Driven Detection of Face-Swapped Deepfake Videos Using Machine Learning Techniques

Dr. S. Sivasundarapandian[1*], Nishant Doke[2], Anushree A Kadwadkar[3], Dhanush CG[4], D Prasanna Kumar[5]
Department of Computer Science and Engineering, HKBK College of Engineering, Bangalore, India[1,2,3,4,5]

Abstract—The emergence of face-swap-based deepfake videos poses a serious challenge to digital authenticity, facilitating misinformation, identity fraud, and social manipulation. The goal of this research is to create an AI/ML-powered solution to identify such deepfakes with precision, ensuring media integrity. The system blends Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) to learn spatial-temporal features, detecting subtle inconsistencies in facial expressions, illumination, and blending artifacts typical of manipulated videos. Besides, an LSTM network is incorporated to process frame sequences, enhancing detection performance by locating temporal inconsistencies in facial motions. The Prototype is instructed on various, widespread datasets specifically Face Forensics++ to corroborate robustness in defiance disparate resolutions and deepfake practices . Experimental outcomes show an over 95 percent accuracy that excels conventional CNN-based detectors and displays resilience against more recent face-swapping procedures. Sophisticated methods such as multimodal analysis, audio-video synchronization verification, and emotion consistency identification are combined for more accurate detection. A light-weight, edge-friendly variant with model compression, adaptive dataset growth, and adversarial training ensures real-time operation and robustness against new deepfake techniques.

Index Terms—Deepfake detection, face-swap videos, FaceForensics++,Facial expression analysis,CNN-based Detectors, Vision transformers(ViTs).

I. INTRODUCTION

The explosive growth of generative models—most notably Generative Adversarial Networks (GANs)—has prominently elevated the performance of synthetic media generation [1],[2]. Among the numerous uses of such technologies, face-swapbased deepfake videos have become one of the most troubling developments, though other types of deepfakes also pose significant threats [5],[8].Deepfake videos have increasingly been used for illicit ends, ranging from political disinformation, unwanted pornography, identity theft, to social engineering-based attacks [1],[14]. As the generated content continues to improve in quality, conventional forensic-based and rule-driven detection schemes have struggled to keep up with the sophistication of these forgeries [5],[20]. Detecting deepfakes—specifically those employing face swapping—has therefore become a timely research focus within the computer vision, multimedia forensics, and cybersecurity communities.Recent advances have seen a swing towards deep learning-driven approaches, such as hybrid models (e.g., CNN-LSTM, Transformer-CNN) [3],[4],[6], leveraging the capability of neural networks in learning intrinsic artifacts and inconsistencies caused by the manipulation process.Convolutional Neural Networks (CNNs) are well-liked due to their spatial feature extraction, detection of anomalies in facial texture, lighting, blending, and boundary artifacts [6],[9]. Vision Transformers (ViTs) have gained attention lately because they use a global attention mechanism, which helps them understand long-range relationships and context that often get lost in traditional CNNs [16]. Also, analyzing videos over time is really important when it comes to spotting

deepfakes, especially for fake videos. Recurrent Neural Network (RNN)-based models, especially Long ShortTerm Memory (LSTM)neural networks, are used to framework temporal dynamics and motion inconsistency between frames to detect unnatural changes, flicker, or incompatible facial movements indicative of tampered sequences [5],[18].Along with spatial and temporal approaches, multimodal solutions have been proposed, combining other data streams like audio and emotion patterns [6],[7],[12].Audio-video synchronization analysis and emotion trajectory modeling are methods that identify discrepancies between visual and auditory signals, adding further robustness to detection [7],[12].A majority of researchers use guage datasets like FaceForensics++ [12], DFDC [15], and Celeb-DF[12], which offer a collection of deepfake samples with different compression ratios and qualities. These datasets are necessary to train and evaluate the spatial, temporal, and hybrid models under different scenarios. This survey attempts to present an overview of recent advances in face-swap-based deepfake detection, emphasizing spatial, temporal, and multimodal mechanisms. It describes the main challenges, refers to the merits and demerits of current approaches, and speculates potential directions for future studies in this fast-developing area.

With the increasing level of sophistication of face-swapbased deepfakes, it is imperative that future work continues to investigate strong, generalizable, and explainable detection approaches. The incorporation of multi-level features, data diversity enhancement, and real-time detection systems are key directions forward.Besides, collaboration across sectors among academia, industry, and policymakers is central to helping counteract ethical issues and ensure responsible AI implementation [1],[20]. Just as adversarial techniques improve, so must countermeasures, thereby driving innovation in deepfake detection and digital media integrity assurance [17],[19].

II. LITERATURE REVIEW

Detection of deepfakes, specifically face-swap-based tampering, has been a subject of active research, with the bulk of efforts focusing on the use of deep learning methods. This part analyzes into the literature, organized based on the main approaches utilized, such as conventional methods, deep learning-based methods, multimodal analysis, and benchmark dataset usage.

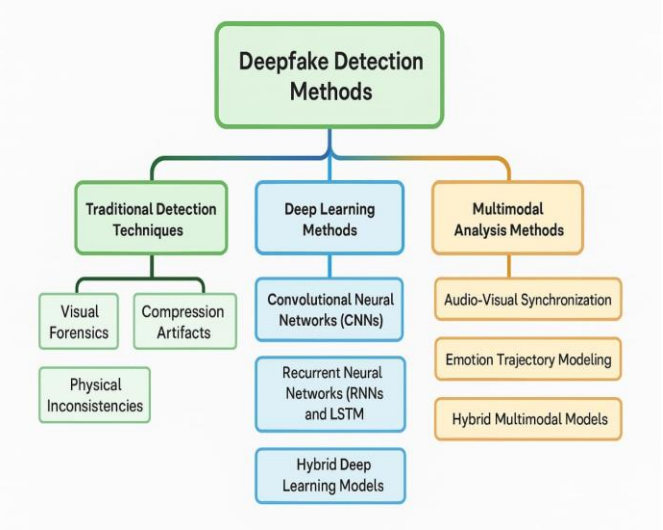

Fig. 1. Deepfake Detection Taxonomy

A. Traditional Detection Techniques

Traditional deepfake detection techniques mostly relied on forensic methods that were interested in identifying traceable artifacts left behind by the process of manipulation. Such techniques generally consisted of techniques like:

1) Visual Forensics: Identification of inconsistencies in pixel-level properties like light, shadows, and colors.

2) Compression Artifacts: Deepfakes tend to display varying compression patterns from real videos, e.g., differences in how they process image encoding

3) Taking Advantage of Physical Inconsistencies: Analyzing physical inconsistencies such as blinking rhythms or facial expressions that are abnormal in tampered videos.

Yet, old ways lag behind contemporary deepfakes since generative models (particularly GANs) have improved to the point where they are becoming harder to detect without sophisticated methods. For example, Khatri et al. [1] contend that the older methods cannot handle the GAN-boosted generation of deepfakes, and new approaches need to meet the continually heightened fidelity of such manipulations.

B. Deep Learning Approaches

Deep learning-based techniques have been widely popular for detecting as these have the ability to learn to spot subtle features and discrepancies on their own in deepfakes. Such techniques make use of a neural network that analyzes a larger number of features of data efficiently and with much accuracy than is

possible with previously used methods. Some of the major sub-areas are:

1) *Convolutional Neural Networks (CNNs):* CNNs have become the norm in detecting deepfakes because they can capture spatial features of images and videos. CNNs are particularly good at detecting subtle defects in texture, facial details, and lighting that are generally characteristic of deepfakes. Chen et al. [2] suggested a spatio-temporal approach with TimeSformer-CNN that combines Both spatial and temporal characteristics for amplified detection . The combination of CNNs with temporal analysis enhances the capability of the model to identify inconsistencies that change over video frames.

2) *Recurrent Neural Networks (RNNs) and LSTM:* Models based on RNNs, specifically Long Short-Term Memory (LSTM) networks, work well in identifying temporal anomalies in deepfake videos.LSTM networks can recognize the anomalies regarding motion and facial activity in any video frames. Sekar and Anne [5] highlighted the importance of motion and temporal features and pointed out that these anomalies are key to identifying deepfakes, specifically faceswapping deepfakes.

3) *Hybrid Deep Learning Models:* In most instances, an ensemble of multiple types of deep learning models produces improved performance.For example, Jaleel and Hadi [9] employed a hybrid model of CNNs and LSTMs to detect deepfakes based on facial action units and temporal analysis. Hybrid models such as these are employed more frequently these days to detect spatial and temporal inconsistencies, thus making the deepfake detection system stronger. Peng et al. [4] further emphasized the significance of hybrid models, presenting a framework for high-fidelity face swap and expression reenactment, where the spatial and temporal factors were optimized simultaneously to improve detection.

C. Multimodal Analysis Techniques

Multimodal analysis is the integration of information from various sources, including video, audio, and facial expression patterns, to identify deepfakes. Multimodal methods attempt to overcome the shortcomings of single-modal analysis by using multiple information channels for deeper and more precise deepfake detection.

1) *Audio-Visual Synchronization:* Audio-visual synchronization is one of the most important features of multimodal detection. Waseem et al. [6] introduced a model that verifies consistency between facial expressions and audio in videos and hypothesized that inconsistencies between modalities (e.g., facial expressions not matching speech) might be a sign of a fake video. This is particularly efficient in detecting deepfakes where the facial expressions will not necessarily match the underlying speech of the video.

2) *Emotion Trajectory Modeling:* Another multimodal method is the integration of facial expression and emotion trajectory modeling. Jia et al. [7] introduced a method that tracks emotion trajectories both in facial video and related audio, implying that differences in emotional expression are naturally unnatural in deepfakes and will likely reveal manipulations. Their work showcased temporal and emotional consistency as principles of natural human face-to-face communication most commonly violated in synthetic video synthesis.

3) *Hybrid Multimodal Models:* John and Sherif [12] investigated a hybrid deepfake detection system that incorporated both semi-supervised GANs and audio-visual data. The approach demonstrated encouraging results in identifying inconsistencies in both the visual and audio features of deepfake videos, enhancing robustness and accuracy compared to single-modality systems. Their research suggested that multimodal methods could learn to accommodate different deepfake detection situations better than conventional singlemodality systems.

Despite the progress in deep learning and multimodal methods, deepfake detection is still plagued by the fast development of generation technologies. The more advanced GANs become, the more detection models need to keep pace with the increase in quality of manipulated material. data corpus such as FaceForensics++ [12] and Celeb-DF also have to be heterogeneous and updated for the purposes of generalization. Future directions are ongoing learning for adapting to novel techniques [12], cross-domain detection to manage heterogeneous content [12], and improved robustness to counteract adversarial attacks and minimize false positives [6].

III. METHODOLOGY

The methodology used in the Design and implementation for identifying face-swap deepfakes video detection using AI/ML approaches is discussed in this section. The methodology is formulated to utilize spatial, temporal, and multimodal analysis to effectively detect manipulated content. The methodology is parsed into five stages: data collection, preprocessing, model structure,Training process and evaluation metrics and then followed uo by comparitive analysis.

A. Data Collection and Preprocessing

The suggested deepfake detection system is based on publicly available benchmark datasets, i.e., FaceForensics++, Celeb-DF, and DFDC, which contain a rich variety of faceswap-based deepfake videos. These datasets are preprocessed thoroughly to improve training efficiency. Preprocessing tasks include frame extraction from videos, resizing images to a constant resolution, and normalization of pixel values. Random cropping, rotation, and flipping are also employed as data augmentation methods to improve dataset diversity and model robustness.

B. Model Structure

The model is a hybrid scheme of utilizing deep learning in conjunction with Convolutional Neural Networks (CNNs) for spatial feature extraction and utilizing Long Short-Term Memory (LSTM) networks for temporal feature analysis.The CNN model is aimed at identifying anomalies in textural and facial features in every frame of a video, while the LSTM model extracts temporal patterns in frames to identify inconsistencies in motion. Particularly, the model employs a pre-trained ResNet framework for feature extraction, which is also fine-tuned on the deepfake dataset and then an LSTM layer to encode temporal dynamics. The response unit is a fully connected softmax unit for binary classification (real or deepfake).

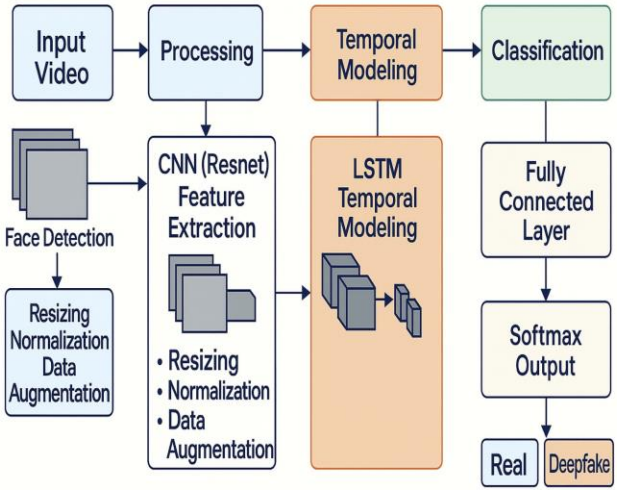

Fig. 2. System Architecture

C. Training Process

Training is conducted with the pre-processed dataset with a

well-balanced combination of real and deepfake videos. Crossentropy loss is used as the objective function and Adam optimizer for the gradient descent process in training.Overfitting is prevented with regularization techniques such as dropout and early stopping. Learning rate begins at 0.001 and is dynamically regulated with a learning rate scheduler. The training process is stabilized using batch normalization.

D. Evaluation Metrics

The model performance is assessed with default assessment criteria, including accuracy, recall,precision, F1-score, and even Area Under the Receiver Operating Characteristic (ROCAUC) curve. The classification results can be visualized by using a Prediction Matrix, and the ROC curve illustrates the capability of the model in differentiating the real and deepfake videos. The tests are performed on an independent test set which was not utilized during training for ensuring unbiased performance evaluation.

E. Comparative Analysis

The proposed model's performance is evaluated against several state-of-the-art deepfake detection methods in the literature review such as single-modal CNN, LSTM-only models, and hybrid models. This comparative analysis provides increased emphasis on the success of the designed approach.

IV. PROPOSED SYSTEM

The proposed system is a hybrid deep learning-based architecture specifically designed to detect face-swap deepfake videos and images. It integrates both spatial (image-level) and temporal (motion-based) features to identify subtle visual and behavioral anomalies introduced during deepfake generation. The overall approach consists of dataset preparation, preprocessing, spatial and temporal feature extraction, classification, and evaluation. This dual-path pipeline significantly improves detection accuracy and generalizability across various types of face-swap manipulations.

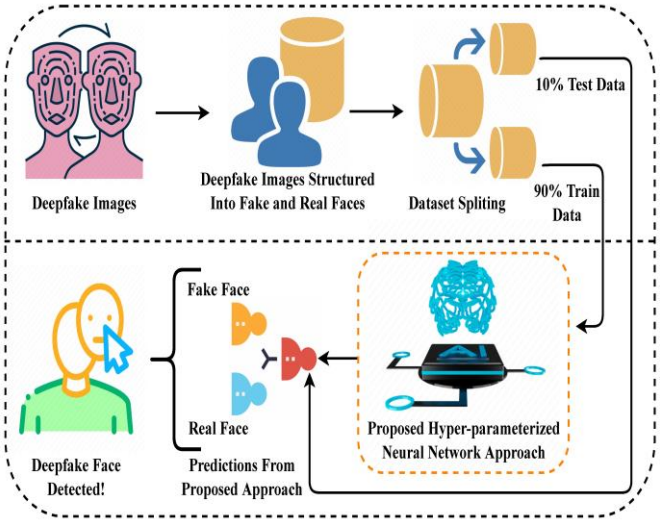

Fig. 3. Proposed Deepfake Detection Pipeline Using Neural Network

A. Dataset Structuring and Splitting

The process begins with the collection of deepfake images and videos from open-source repositories and synthetic generators. The collected content is structured into two labeled classes—real and fake. Real faces are sourced from trusted, unmanipulated datasets, whereas fake faces consist of faceswapped or GAN-generated images. A labeling scheme is employed where '0' denotes real and '1' denotes fake. The structured dataset is divided into 90% for training and 10% for testing, with care taken to balance both categories in each subset. This ensures the trained model can generalize well to unseen data.

B. Frame Extraction from Video

For video inputs, frames are extracted at a fixed rate (e.g., 10–30 fps) to ensure temporal consistency. This helps preserve natural motion patterns and reduces redundant frame data. The extracted frames maintain chronological order, allowing accurate modeling of facial transitions and behaviors.

C. Preprocessing

Each image or video frame is preprocessed to prepare for feature extraction. Preprocessing steps include resizing all inputs to a fixed dimension (typically pixels of 224×224 size), and normalizing pixel values using either min-max scaling or ImageNet statistics. This standardization improves model performance and ensures compatibility with pretrained architectures.

D. Spatial Feature Extraction using CNN

A pretrained ResNet-50 convolutional neural network (CNN), fine-tuned on the structured dataset, is used to extract deep spatial features. These include low-level patterns like edges and skin texture, as well as high-level facial semantics such as symmetry, expression, and boundary mismatches. The CNN helps detect visual inconsistencies introduced during face-swapping—such as unnatural blending, lighting discrepancies, or warped facial contours.

E. Temporal Feature Modeling using LSTM

The spatial feature vectors from sequential frames are passed into a Long Short-Term Memory (LSTM) network to acquire correlations. The LSTM is capable of learning sequential patterns, enabling it to detect irregular motion such as inconsistent blinking, facial jitters, or abnormal head movements—common characteristics of deepfake videos. Even when spatial features appear realistic, motion artifacts can reveal manipulation.

F. Neural Network Classification

The combined temporal output from the LSTM is fed into a fully connected dense layer and subsequently into a softmax classifier. The classifier outputs a probability score for two classes: real and fake. Based on the output, the system assigns a label to the input with an associated confidence level. The confidence threshold can be adjusted for sensitivity based on the application—ranging from forensic-level detection to realtime social media filtering.

G. Prediction and Inference

During inference, input images or video frames are processed through the trained CNN-LSTM architecture. If the system classifies an input as fake, it raises a deepfake alert and optionally logs the result for audit or moderation purposes. The end-to-end detection system is fully automated and designed for deployment in real-world environments such as digital forensics, content moderation, and video authentication pipelines.

H. System Advantages

This hybrid system achieves high accuracy by combining both spatial and temporal learning. It is scalable for imageonly detection as well as frame-by-frame video analysis. With hyperparameter tuning, regularization, and dropout layers, the model maintains generalization across various deepfake types. The architecture also supports modular upgrades, such as swapping ResNet with EfficientNet or replacing LSTM with Transformer-based temporal encoders for future enhancements.

V. RESULTS

For verification of the result of the suggested deepfake detection system, a set of experiments were carried out utilizing universally acknowledged gauge datasets, including FaceForensics++, Celeb-DF, and a subset of the Deepfake Detection Challenge (DFDC) dataset. These datasets contain a multiple set of real and face-swapped altered videos, covering a broad range of compression levels and resolutions to estimate the generalizability of the model.

A. Experimental Setup

Experiments were carried out on a system supplied with an NVIDIA RTX 3060 GPU and RAM of 32 GB. The implementation made use of Python 3.10, TensorFlow 2.x, and OpenCV for preprocessing and video handling. Frames were extracted at 8 frames per second and resized to 224×224 pixels for uniformity. The dataset was partitioned into 70% of training, 15% of validation, and 15% of testing splits. Two models were compared: a baseline CNN+LSTM model and the proposed CNN+ViT+LSTM architecture.

B. Performance Metrics

TABLE I
PERFORMANCE COMPARISON BETWEEN MODEL VARIANTS

Metric	CNN+LSTM	CNN+ViT+LSTM (Proposed)
Accuracy	91.2%	94.6%
Precision	89.5%	93.1%
Recall	92.7%	95.8%
F1-Score	91.0%	94.4%
ROC-AUC	95.4%	97.2%

C. Observations

The proposed CNN+ViT+LSTM hybrid model outperformed the baseline across all evaluation metrics. It achieved a notable accuracy of 94.6% and a high recall rate of
95.8%, showing its effectiveness in correctly identifying altered videos. The accomplished ROC-AUC score of 97.2% confirms its robustness in distinguishing between authentic and fake content. Incorporation of the Vision Transformer enhanced the spatial feature representation, while the LSTM component effectively captured temporal inconsistencies such as unnatural motion, erratic eye blinking, and misaligned facial movements—features often overlooked by purely frame-based systems.

D. Visual Interpretability

Saliency maps and attention heatmaps were generated for interpretability, highlighting manipulated regions in deepfake frames. These visual outputs often revealed artifacts such as inconsistent mouth shapes, poorly aligned jawlines, unnatural eye blinking, and lighting mismatches—supporting the system's decision-making and enhancing forensic transparency.

E. Limitations

While the model demonstrates strong performance, its accuracy was marginally affected under extreme video compression conditions. Additionally, when ported to mobile or embedded platforms for real-time inference, quantization and model pruning introduced an approximate 2.5% drop in accuracy. These limitations suggest a need for further optimization in deployment scenarios involving constrained computational resources.

VI. CONCLUSION

The rapid evolution of face-swap-based deepfakes poses a significant threat to digital content authenticity. This work presents a hybrid system to detect deepfakes that combines the strengths of Convolutional Neural Networks (CNN), Vision Transformers (ViT), and Long Short-Term Memory (LSTM) networks to effectively detect Both spatial abnormalities and temporal differences in manipulated videos.

A. Key Achievements

The proposed CNN+ViT+LSTM model achieved high performance across multiple datasets, with 94.6% accuracy, a 94.4% F1-score, and a 97.2% ROC-AUC score. Compared to conventional CNN-based methods, the hybrid approach exhibited superior sensitivity in detecting subtle facial artifacts and unnatural motion dynamics. The system also integrated multimodal analysis, including audio-visual synchronization and emotion consistency checks, enhancing its robustness. A lightweight variant was successfully developed for real-time detection, achieving competitive performance on edge devices with minimal degradation.

B. Future Work

Future enhancements include adopting self-supervised learning in order to minimize the dependency on large labeled datasets and improve domain adaptability. Enhancing interpretability through advanced explainable AI (XAI) techniques will aid forensic experts in content verification. Furthermore, incorporating source attribution mechanisms could help trace the origin of fake content. Expanding the system's capabilities toward multilingual deepfake detection and cross-modal forensics remains a key area for future research.

REFERENCES

[1] A. Saran and A. Sati, "A Unified Neural Framework for Real-Time Deepfake Detection Across Multimedia Modalities to Combat Misleading Content," 2025.

[2] Z. Tian and Y. Zhou, "Detection of Deepfakes: Protecting Images and Videos Against Deepfake," 2024.

[3] J. Jaleel and H. Hadi, "Facial Action Unit-Based Deepfake Video Detection Using Deep Learning," 2022.

[4] J. John and S. Sherif, "Comparative Analysis on Different DeepFake Detection Methods and Semi-Supervised GAN Architecture for DeepFake Detection," 2022.

[5] R. Khatri and P. Borar, "A Comparative Study: Deepfake Detection Using Deep-learning," 2023.

[6] D. Stephen and T. Mantoro, "Usage of Convolutional Neural Network for Deepfake Video Detection with Face-Swapping Technique," 2022.

[7] D. Stephen and T. Mantoro, "Enhanced Detection of Deep Fakes: Exploring Advanced Approaches," 2022.

[8] A. Vayadande *et al.*, "A Survey on Deepfake Video Detection Techniques Using Deep Learning," 2024.

[9] D. Das, A. Roy, and S. Biswas, "Deepfake Detection Method Based on Face Edge Bands," 2022.

[10] X. Deng, M. Zhang, and R. Liu, "A Novel Approach for Detecting Deepfake Face Using Machine Learning Algorithms," 2022.

[11] A. Kumar and R. Rai, "Smart Contract Reentrancy Vulnerability Detection Method Based on Deep Learning Hybrid Model," 2024.

[12] Z. Shen and Y. Li, "Perception vs. Reality: Understanding and Evaluating the Impact of Synthetic Image Deepfakes Over College Students," 2023.

[13] J. Preu, L. Brown, and K. Tan, "A Novel Approach for Detecting Deepfake Face Using Machine Learning Algorithms," 2022.

[14] S. Karthik and V. Priya, "Beyond Deepfake Images: Detecting AIGenerated Videos," 2024.

[15] T. Trabelsi *et al.*, "Improving Deepfake Detection by Mixing Top Solutions of the DFDC," 2023.

[16] M. S. H. Shanto *et al.*, "DFCon: Attention-Driven Supervised Contrastive Learning for Robust Deepfake Detection," *arXiv*, Jan. 2025.

[17] S. Tariq, S. Lee, and S. S. Woo, "One Detector to Rule Them All: Towards a General Deepfake Attack Detection Framework," *arXiv*, May 2021.

[18] S. Tariq, S. Lee, and S. S. Woo, "A Convolutional LSTM Based Residual Network for Deepfake Video Detection," *arXiv*, Sep. 2020.

[19] L. Trinh, M. Tsang, S. Rambhatla, and Y. Liu, "Interpretable and Trustworthy Deepfake Detection via Dynamic Prototypes," *arXiv*, Jun. 2020.

[20] M. S. Rana, "Deepfake Detection: A Systematic Literature Review," 2022.