Deepfake Creation and Detection of Multimedia Data

Sameeksha A B[1], Nikhitha S Naik[1] and Prof. Hemanthchandra N[3]

Undergraduate, Department of Information Science and Engineering[1,2]

Assistant Professor, Department of Information Science and Engineering[3]

Global Academy of Technology, Bengaluru, Karnataka, India

Abstract: *As the increasing number of deepfake content poses a growing threat to multimedia integrity, this paper proposes a robust deepfake detection approach based on a hybrid architecture. The proposed framework combines the power of Residual Networks (ResNet) for spatial feature extraction and Long Short-Term Memory (LSTM) with Convolutional Neural Networks (CNN) for temporal dependency modeling. The ResNet component captures complicated patterns in facial and contextual information, whereas the LSTM-CNN module identifies dynamic facial expressions and movements across multiple frames. Transfer learning strategies are used to improve model generalization by combining pre-training on a large dataset with fine-tuning on deepfake data. Experimental evaluations on a variety of deepfake datasets show superior accuracy, precision, and recall, demonstrating the hybrid architecture's efficiency in dealing with the evolving challenges posed by more advanced deepfake generation techniques. In conclusion, our novel deepfake detection framework employs an effective combination of ResNet and LSTM-CNN, demonstrating a promising solution that not only advances the state-of-the-art in multimedia forensics but is also resistant to adversarial attacks. This hybrid model, which effectively combines spatial and temporal information, has the potential to significantly improve the accuracy and reliability of deepfake detection systems in the face of emerging digital threats.*

Keywords: Block chains

I. INTRODUCTION

The rapid advancement of artificial intelligence, particularly in the field of deep learning, has resulted in the emergence of deepfake technology, which enables the creation of hyper-realistic multimedia content that can fool individuals. Malicious deepfakes can cause misinformation, reputational damage, and even threats to national security. As a response to this growing problem, the development of effective deepfake detection methods has become critical. This paper presents a novel approach to deepfake detection that combines the strengths of Residual Networks (ResNet), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN) to create a hybrid architecture that excels at capturing both spatial and temporal features.The proposed hybrid model aims to overcome the limitations of existing deepfake detection techniques by utilizing ResNet's comprehensive spatial understanding and the complicated temporal dependencies represented by LSTM-CNN fusion. With the exponential increase in deepfake complexity, traditional methods frequently struggle to detect hidden manipulations in facial features and fail to capture the temporal dynamics present in video sequences. In response, our approach not only combines these two powerful neural network architectures, but also employs transfer learning strategies to improve generalization, allowing the model to effectively adapt to the diverse and constantly evolving landscape of deepfake creation techniques.

II. LITERATURE REVIEW

According to author Luca Guarnera[1] Deepfakes are created using techniques based on Goodfellow et al.'s Generative Adversarial Networks (GAN).Previous research has focused on using Convolutional Neural Networks (CNNs) for deepfake detection, but these methods differentiate semantics more than GAN-specific traces and are extremely time-consuming.The proposed approach in this paper uses

DOI: 10.48175/IJARSCT-15308

the Expectation-Maximization algorithm to detect and extract Convolutional Traces (CT) left by GANs during image generation. The CT has high discriminative power and beats state-of-the-art deepfake detection methods, even in the presence of multiple attacks. Tests on FACEAPP-generated deepfakes demonstrated the effectiveness of the proposed technique in a real-world scenario, with 93% accuracy in fake detection.

The statement highlighted by SudeepTanwar [2],the proposed model uses a deep-CNN (D-CNN) architecture for deepfake detection, which is trained on images from multiple sources and utilizes binary-cross entropy and the Adam optimizer to improve learning rate. The proposed model is trained on seven different datasets, including 5000 deepfake images and 10,000 real images, to improve its generalizability. The model alters the captured images and feeds them into the D-CNN architecture to detect deepfakes. The proposed model's performance and generalizability capabilities are compared with existing models, MesoNet and MesoInception network, on the CelebDFdataset. The proposed model could be improved by testing with variable input neural networks and global pooling, as well as studying the effects of efficient image upscaling algorithms on performance. The proposed architecture achieves 97.2% accuracy with images from 5 different data sources for deepfakes and 2 sources for real images.

Various deepfake detection techniques have been developed, but detecting all types of deepfake images using a single model remains difficult.Jihyeon Kang et al.[3], propose a method for detecting different types of deepfake images based on three common traces produced by deepfakes: residual noise, warping artifacts, and blur effects.They use a network designed for structural analysis to detect pixel-wise residual-noise traces and landmarks to capture high-level information.They also use features from different image quality measurement tools to detect blurring.On datasets of various deepfake types, the proposed network exceeds existing detection networks in terms of stability and performance.Tripathy et al. proposed a two-stage GAN with a facial attribute vector and built a network using image quality measurement features and warping artifacts extracted from facial landmarks.They concentrated on identifying the common traces of deepfakes, which are difficult to avoid, as well as detecting image features.Their approach is directly applicable to deepfake video detection pipelines based on frame-by-frame detection.The experiments used input images with a resolution of 128 x 128 and deepfake images from various datasets.

Deepfake technology has grown, which has made it possible to create high-quality fake videos that are difficult to detect using traditional artifacts. Video editing applications and tools have improved, allowing for more powerful editing capabilities when creating fake videos.Transfer learning has demonstrated excellent performance in prediction and detection tasks, particularly when there is insufficient data to train models. It involves creating a model for one task using massive data and then reusing and adjusting the trained model for another task.Norah M. Alnaimdescribes[4], the most common metrics for Deepfake detection are accuracy, area under the ROC curve (AUC), and F1-score. These metrics are used to judge the effectiveness of detection models.This paper presents a Deepfake Face Mask Dataset (DFFMD) and a novel Inception-ResNet-v2 model with preprocessing stages, feature-based, residual connection, and batch normalization. The results show that the proposed model detects face-mask deepfakes with a 99.81% accuracy, compared to traditional methods such as InceptionResNetV2 and VGG19.

Bilal Ashfaq Ahmed[5], suggests a comprehensive study on existing methods of creating deepfake images and videos for face and expression replacement.It discusses various methods proposed for face swapping in images and videos, including the use of deep learning models such as GANs and autoencoders.The paper also presents an overview of publicly available deepfake datasets for benchmarking, such as the Korean Deepfake Dataset (KoDF) and UADFV dataset.It discusses the detection methods used to identify deepfake face and expression swaps, including the use of models such as FSGAN and DeepFaceLab. The study goes beyond identifying current barriers and outlines future research directions to address concerns about deepfake detection methods. The purpose of this paper is to help with the development of robust and effective deepfake detection solutions for facial and expression swaps.

The author Zitong Yu[6],provides an in-depth analysis of recent advances in deep learning-based face anti-spoofing (FAS) methods, highlighting the limitations of previous reviews that primarily focus on handcrafted features.The authors look into recent pixel-wise supervision methods and compare them to methods designed specifically for domain generalization and open-set FAS.The paper also covers deep learning applications for multi-modal or specialized sensors, such as depth and infrared sensors, light field, and flash sensors. In addition, the paper discusses common face spoofing attacks and examines existing FAS datasets, evaluation metrics, and protocols.The authors present a taxonomy of deep learning-based FAS methods and discuss the advantages and disadvantages of different FAS methods and

DOI: 10.48175/IJARSCT-15308

sensors. The paper concludes by discussing current open issues and potential research directions in the field of deep learning-based FAS.

The paper provides a literature review of forensic approaches related to their work and introduces the manipulation environment considering real-world scenarios where various types of manipulations and image compressions coexis. The author Heung-Kyu Lee[7], state that techniques for detecting multiple manipulations applied to uncompressed images have been reported, but no forensic approach for JPEG images compressed with varying qualities has been proposed. They propose the manipulation classification network (MCNet) to exploit multi-domain features in the spatial, frequency, and compression domains, and show its effectiveness through thorough tests with cutting-edge baselines. TheMCNet is made up of two sub-networks (VANet and CANet) and a single network that learn forensic features in multiple domains and then analyze the combined characteristics for manipulation classification.The fine-tuned model based on the multi-class manipulation task has also proven effective for different forensic tasks, such as DeepFake detection or integrity authentication of JPEG images.

The paper addresses the issue of data scarcity in medical imaging and suggests the use of DEEPFAKE image synthesis for data augmentation. The author NawafWaqas [8],provide a novel DEEPFAKE image synthesis framework for the knee profile, that uses a hierarchical structure and a self-attention layer to capture fine details and patterns in high-resolution images. The proposed framework extends the Progressive Growing Generative Adversarial Network (PGGAN) by including a self-attention layer, spectral normalization in the discriminator, and pixel normalization in the generator. The performance of the Enhanced-GAN, an improved version of PGGAN, is measured using the AM Score and Mode Score parameters and compared to that of PGGAN. The study evaluates the effectiveness of Enhanced-GAN and PGGAN synthesized data for segmentation tasks using the U-net supervised deep learning model. The Dice Coefficient metric is used for evaluation. The findings indicate that U-net trained on Enhanced-GAN DEEPFAKE data optimized with real data exceeds U-net trained on PGGAN DEEPFAKE data with real data.

According to author Nawal A. Zaher [9],Gabor filters have been studied in the past as a way to improve deep feature representations and recognition performance in convolutional neural networks (CNNs).In certain studies, learnable weighted filters have been modulated to incorporate Gabor filters into CNNs, which has reduced the architecture size and enhanced recognition performance.Applications for gabor filters include person re-identification, hyperspectral image classification, facial expression recognition, and finger vein recognition. In order to decrease the number of parameters and enhance performance in hyperspectral image classification, naive Gabor networks rigorously learn conventional Gabor filters. However, prior methods have mostly concentrated on linear Gabor filters, which has limited the architectures' diversity and ability to adapt to complex data. This paper's suggested architecture introduces a unified Gabor function to overcome these drawbacks.

The rise of realistic-looking artificial intelligence (AI)-manipulated fake face media, like DeepFake or Face2Face, highlights the significance of developing models to detect fake faces in media. Author Sungzoon Cho describes[10],face image forensics and general-purpose image forensics are two categories of prior research in this field. While face image forensics is based on convolutional neural networks inspired by object detection models specialized to extract images' content features, general-purpose image forensics concentrates on extracting hand-crafted features of traces left in the image after manipulation. In order to improve manipulation detection performance, a hybrid face forensics framework based on a convolutional neural network is proposed in this paper. It combines face image forensics and general-purpose image forensics techniques.

The significance of the first proposed GANs model and some of its flaws, including the saddle problem and the minimax problem, are discussed in the paper.

Alec Radford et al.'s 2016 proposal [11], DCGAN (Deep Convolutional Generative Adversarial Networks), enhanced the GANs model by applying CNN (Convolutional Neural Network) models, enabling latent vector-based arithmetic operations with filters between images .

Author Ki-Ryong Kwon describes[12],fake videos and images are created using deep generative models, and increasingly common algorithms include face swapping, face manipulation, and expression reenactment .To increase the model's generalizability, the study suggests a deepfake detection technique called Meta Deepfake Detection (MDD).The MDD algorithm improves the model's capacity to produce accurate representations by using meta-weight learning to transfer information from source domains to target domains.To improve the system's detection capabilities,

DOI: 10.48175/IJARSCT-15308

the paper adds pair-attention loss and average-center alignment loss .Using evaluation benchmarks, the suggested method is contrasted with related research to demonstrate the model's generalization in unexplored domains.The use of block shuffling transformation to enhance performance and lessen overfitting is also discussed in the paper.

Author Ching-Chun Chang [13], explains on facial manipulation, a significant issue in the context of deepfakes, and tackles the problem of deepfakerestoration. In order to enable automatic immunity acquisition, the study suggests a novel neural network-based cyberimmune system framework. It is made up of an adversary for deepfake simulation, a neutralizer, a validator, and a vaccinator. In order to combat cybercrimes, the paper highlights the significance of disclosing the original media content and offers helpful forensic hints. With limited computational resources, the study considers a single overwhelming adversary model, the masked-face model, with the goal of developing attack-agnostic capability. The suggested immune system shows efficacious resistance against face replacement and different kinds of corruption. The study reviews and emphasizes the need for additional countermeasures.

The survey on DeepFake detection techniques in face photos and videos presented in the paper summarizes the findings, effectiveness, approach, and kind of detection. It examines the various DeepFake creation methods currently in use and groups them into five main groups. The training of DeepFake models on DeepFake datasets and experimental testing of these models are covered in this paper. Additionally, author Asad Malik[14], highlights the advancements in the DeepFake datasets by summarizing their trends.

In addition to discussing the difficulties in creating and detecting DeepFakes, the paper analyzes the problem of developing a generalized DeepFake detection model .Additionally, the paper highlights the importance of DeepFake tools, datasets, and temporal coherence in DeepFake detection, as well as the need for further study in journalism or law enforcement-based forgery cases .

Author MdShohelRana[15], presents an updated overview of research works in Deepfake detection through a systematic literature review (SLR).In order to conduct the SLR, the authors summarized 112 pertinent articles published between 2018 and 2020 that offered a range of Deepfake detection techniques.The articles were divided into four categories: methods based on deep learning, methods based on traditional machine learning, methods based on statistics, and methods based on blockchain.

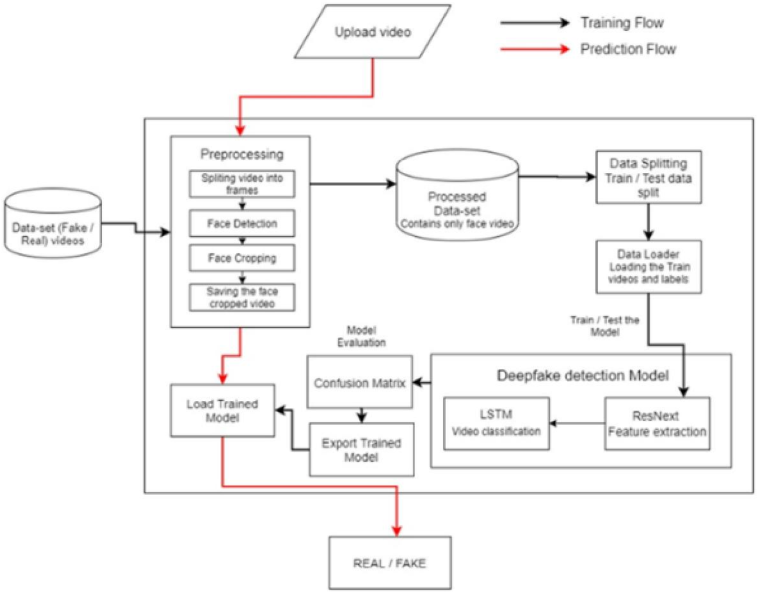
Upload video
Training Flow
Prediction Flow
Preprocessing
Spliting video into frames
Face Detection
Face Cropping
Saving the face cropped video
Data-set (Fake / Real) videos
Processed Data-set Contains only face video
Data Splitting Train / Test data split
Data Loader Loading the Train videos and labels
Train / Test the Model
Model Evaluation
Confusion Matrix
Deepfake detection Model
LSTM Video classification
ResNext Feature extraction
Load Trained Model
Export Trained Model
REAL / FAKE

Figure 1:Archietecture diagram

Deep learning-based approaches outperform other methods in Deepfake detection, according to an evaluation of the performance of the detection capabilities of the various methods using various datasets.The paper presents a novel and first-of-its-kind taxonomy that divides Deepfake detection methods into four categories .

Although Li and Lyu's approach has limitations when it comes to identifying high-resolution and high-quality deepfake videos, they did propose a way to detect deepfake videos using AI algorithms.A US-based startup called Truepic created a system to protect image integrity and identify attempts at forgery by utilizing blockchain technology and mobile apps.Similar to Truepic, Serelay is a UK-based startup that only stores the distinct fingerprint of photos and videos on its servers.Ablockchain-based online startup called PROVER uses distinctive hashes to confirm the legitimacy of user-generated videos.Gipp et al.[16] concentrate on hashing the video and storing the hash on the blockchain to use blockchain technology to guarantee the integrity of video content.A Brazilian startup called OriginalMy registers and authenticates digital documents, contracts, and identities using public blockchain .The paper's suggested framework for safe and reliable history tracking and tracing of digital content is based on blockchain's transparency, traceability, and time-sequenced logs.

III. CONCLUSION

In conclusion, deepfake creation adds an interesting yet concerning dimension to multimedia content by allowing the creation of attractive but false videos and images. This raises serious issues about misinformation and trust in digital media. On the other hand, ongoing efforts to develop detection methods are critical for detecting and mitigating the effects of deepfakes. Establishing a balance between innovation and authenticity is critical to navigating the challenges posed by deepfake technology. As we move forward, we must invest in research, education, and working together to stay ahead of false advertising and ensure a safer digital environment. In conclusion, the dual nature of deepfake technology highlights the need for a proactive and various approach. By raising awareness, improving detection capabilities, and encouraging responsible use, we can maximize the benefits of multimedia innovation while reducing the risks associated with misleading content. This effort will be critical in preserving the integrity of digital information and encouraging trust in a world that is constantly evolving.

REFERENCES

[1]Guarnera, L., Giudice, O., &Battiato, S. (2020). Fighting deepfake by exposing the convolutional traces on images. *IEEE Access*, *8*, 165085-165098.

[2]Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., &Mazibuko, T. F. (2023). An Improved Dense CNN Architecture for Deepfake Image Detection. *IEEE Access*, *11*, 22081-22095.

[3]Kang, J., Ji, S. K., Lee, S., Jang, D., &Hou, J. U. (2022). Detection enhancement for various deepfake types based on residual noise and manipulation traces. *IEEE Access*, *10*, 69031-69040.

[4]Alnaim, N. M., Almutairi, Z. M., Alsuwat, M. S., Alalawi, H. H., Alshobaili, A., &Alenezi, F. S. (2023). DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era WithDeepfake Detection Algorithms. *IEEE Access*, *11*, 16711-16722.

[5]Waseem, S., Abu-Bakar, S. R., Ahmed, B. A., Omar, Z., Eisa, T. A. E., &Dalam, M. E. E. (2023). DeepFake on Face and Expression Swap: A Review. *IEEE Access*.

[6]Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z., & Zhao, G. (2022). Deep learning for face anti-spoofing: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *45*(5), 5609-5631.

[7]Yu, I. J., Nam, S. H., Ahn, W., Kwon, M. J., & Lee, H. K. (2020). Manipulation classification for jpeg images using multi-domain features. *IEEE Access*, *8*, 210837-210854.

[8]Waqas, N., Safie, S. I., Kadir, K. A., Khan, S., &Khel, M. H. K. (2022). DEEPFAKE image synthesis for data augmentation. *IEEE Access*, *10*, 80847-80857.

[9]Khalifa, A. H., Zaher, N. A., Abdallah, A. S., &Fakhr, M. W. (2022). Convolutional neural network based on diverse Gabor filters for deepfake recognition. *IEEE Access*, *10*, 22678-22686.

[10]Kim, E., & Cho, S. (2021). Exposing fake faces through deep neural networks combining content and trace feature extractors. *IEEE Access*, *9*, 123493-123503.

DOI: 10.48175/IJARSCT-15308

[11]Jung, T., Kim, S., & Kim, K. (2020). Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, *8*, 83144-83154.

[12]Tran, V. N., Kwon, S. G., Lee, S. H., Le, H. S., & Kwon, K. R. (2022). Generalization of Forgery Detection With Meta Deepfake Detection Model. *IEEE Access*, *11*, 535-546.

[13]Chang, C. C., Nguyen, H. H., Yamagishi, J., &Echizen, I. (2023). Cyber Vaccine for Deepfake Immunity. *IEEE Access*.

[14]Malik, A., Kuribayashi, M., Abdullahi, S. M., & Khan, A. N. (2022). DeepFake detection for human face images and videos: A survey. *Ieee Access*, *10*, 18757-18775.

[15]Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE access*, *10*, 25494-25513.

[16]Hasan, H. R., & Salah, K. (2019). Combating deepfake videos using blockchain and smart contracts. *Ieee Access*, *7*, 41596-41606.

DOI: 10.48175/IJARSCT-15308