LSTM-Based Deep-fake Detection in Video

Gallardo, Matthew
College of Computer and Information Sciences*(CCIS)*
Polytechnic University of the Philippines
Quezon City, Philippines

Guevarra, John Joemer
College of Computer and Information Sciences*(CCIS)*
Polytechnic University of the Philippines
: Manila City, Philippines

Rapiza, Leinard
College of Computer and Information Sciences*(CCIS)*
Polytechnic University of the Philippines
: Malabon City, Philippines

***Abstract*—The increasing computational power has enabled deep learning algorithms to create highly realistic human-synthesized videos known as deep fakes. These videos are increasingly being used in harmful ways, such as creating political unrest, fake terrorism events, revenge porn, and blackmailing people. In response, we have developed a new method to effectively distinguish AI-generated fake videos from real ones. Our web-based system uses a combination of Res-Next Convolution neural network for feature extraction and Long Short Term Memory (LSTM) based Recurrent Neural Network (RNN) to classify whether a video has been manipulated or not, thus determining if it is a deep fake or a genuine video. The method was evaluated on a large and diverse dataset that includes Face-Forensic++, Deep fake detection challenge, and Celeb-DF, showing competitive results with a straightforward and robust approach.**

Keywords—Deep fake Detection, Long Short Term Memory(LSTM), GAN Models, ResNext CNN

I. INTRODUCTION

In the age of rapidly advancing technology and social media, the digital landscape has witnessed a remarkable transformation, giving rise to innovative applications. Among these advancements, the emergence of deepfake technology has undeniably captured the world's attention. Deepfakes, a portmanteau of "deep learning" and "fake," represent a groundbreaking fusion of artificial intelligence and multimedia manipulation, allowing individuals to alter, replace, or generate hyper-realistic content in the form of images, videos, and audio. Spreading of the DeepFakes over the social media platforms have become very common leading to spamming and peculating wrong information over the platform. As billions of users engage with and share content on these platforms daily, the potential for the rapid dissemination of manipulated or fabricated information becomes a major concern. With deep fakes becoming increasingly difficult to discern from authentic content, the implications for misinformation, privacy violations, and even potential threats to national security cannot be understated.

To overcome such a situation, DeepFake detection is very important. So, we describe a new deep learning-based method that can effectively distinguish AI-generated fake videos (DF Videos) from real videos. It's incredibly important to develop technology that can spot fakes, so that the DF can be identified and prevented from spreading over the internet.

To explore the detection of Deepfakes, it is very important to understand the way Generative Adversarial Network (GAN) creates the DF. In the realm of GANs, two neural networks, the Generator and the Discriminator, engage in a competitive process to produce high-fidelity deepfakes. The Generator is responsible for generating synthetic data, while the Discriminator acts as a "detective," attempting to differentiate between real and generated data. The Generator starts with random noise as input and progressively refines its output to resemble real media, while the Discriminator improves its ability to identify the genuine from the fabricated. GANs take a video and an image of a specific individual, referred to as the 'target', as input and produce another video where the target's face is replaced with that of a different individual, known as the 'source'. The foundation of deepfakes lies in deep adversarial neural networks, which are trained on face images and target videos to automatically map the faces and facial expressions of the source onto the target. The process of generating deep fakes begins by splitting the video into individual frames. In each frame, the input image is replaced with the source's face, effectively replacing the target's appearance throughout the video. This replacement process is often accompanied by proper post-processing techniques to enhance realism, resulting in videos that can achieve a remarkably high level of authenticity. Crucial to this process are autoencoders, which facilitate the reconstruction of the video after the target's face has been replaced with that of the source.

We introduce deep learning-based method for accurately distinguishing deepfake (DF) videos from genuine ones, inspired by the process used by GANs to create deep fakes. DF videos exhibit specific properties due to the limitations in computation resources and production time during the deepfake generation. A key characteristic of DF videos is the fixed size of the synthesized face images, leading to an affinal warping process to align them with the source's face configuration. Unfortunately, this warping introduces noticeable artifacts in the output deepfake videos because of resolution inconsistencies between the warped face area and its surroundings. Our approach addresses this by employing a comprehensive framework that examines the generated face regions and their surroundings. We split the video into frames and extract relevant features using a ResNext Convolutional Neural Network (CNN). Additionally, we utilize a Recurrent Neural Network (RNN) with Long Short Term Memory (LSTM) cells to capture temporal inconsistencies introduced by the GAN during the deepfake reconstruction across frames. To train the ResNext CNN model, we directly simulate the resolution inconsistency encountered in affine face wrappings, streamlining the process while ensuring reliable results. By

combining the power of CNNs and LSTMs, our approach achieves robust deepfake detection, effectively differentiating between authentic videos and those manipulated using GAN-based techniques.

II. LITERATURE SURVEY

DeepFake Detection Through Key Video Frame Extraction using GAN

Deepfake videos, using advanced deep learning techniques, have emerged as a significant threat due to their potential for image falsification. Detecting and preventing such deceptive content on social media is crucial. This literature survey explores a novel neural network-based technique proposed in the paper titled "A Robust Neural Network-based Approach for Detecting Deepfake Videos," published in the 2022 International Conference on Automation, Computing, and Renewable Systems (ICACRS). The approach utilizes an efficient video frame extraction method to accelerate the detection process, incorporating a convolutional neural network (CNN) and a classifier network with GAN technology. The researchers opted for the Confusion Matrix over Resnet, Resnext50, and LSTM for pairing with the classifier to identify fake videos. By using feature vectors from the CNN module as input, the classifier accurately categorizes videos as either real or fake, achieving a remarkable 97.2% accuracy on the Deepfake Detection Challenge dataset. The primary objective of the study is to achieve high accuracy without requiring extensive training data. This literature survey offers insights into the model's design and its effectiveness in combating the proliferation of deepfake videos (Sooda, K. 2022).

A Robust Deepfake Video Detection Method based on Continuous Frame Face-swapping

Deepfake video detection poses significant challenges in real-world scenarios due to generalization issues. Existing methods are limited to single-frame image detection and struggle with continuous frame videos. This literature survey presents a novel approach for robust deepfake video detection based on continuous frame face-swapping. The proposed method utilizes a face-swapping dataset created with Delaunay triangulation and piecewise affine transform to achieve continuous frame face-swapping. A feature enhancement module is designed to focus on the mask fusion zone, incorporating both facial and background information. The detection model employs Efficient Net for intra-frame fusion feature extraction and LSTM for inter-frame time feature extraction. Cross-domain experiments demonstrate superior detection AUC compared to existing methods, validating the robustness and generalization capability of the proposed approach (Liu, D. 2022).

Deepfake Video Detection Using Recurrent Neural Networks

This study by Guera, D. et al, introduces a temporal-aware pipeline designed for automatically detecting deep-fake videos. The proposed system employs a convolutional neural network (CNN) to extract frame-level features, which are subsequently used to train a recurrent neural network (RNN). The RNN learns to classify videos as manipulated or authentic. The evaluation is conducted on a comprehensive dataset of deepfake videos sourced from various video platforms. The survey demonstrates that the proposed system achieves competitive results with simple architecture.

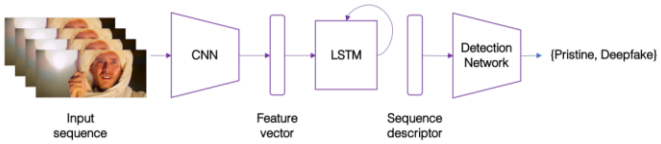

Figure 1. Overview of their detection system

Is this figure, the system learns and infers in an end-to-end manner and given a video sequence, outputs a probability of it being a deepfake or a pristine video. It has a convolutional LSTM subnetwork, for processing the input temporal sequence.

The study highlights the prevalence of deep fake videos that exhibit manipulation only in small segments. To address this, the survey extracts continuous subsequences of fixed frame length as input data for the system during training, validation, and testing phases. The performance of the system is presented in Table 1, showcasing the detection accuracy using different sub-sequence lengths (N = 20, 40, and 80 frames). Notably, with less than 2 seconds of video (40 frames at 24 frames per second), the proposed system accurately predicts whether a fragment originates from a deepfake video or not, achieving an accuracy greater than 97% (Guera, D. et al. 2018).

This literature survey introduces a temporal-aware approach for automatically detecting deepfake videos. The experimental results, based on an extensive collection of manipulated videos, demonstrate the efficacy of the proposed simple convolutional LSTM structure for accurate deepfake detection, requiring as little as 2 seconds of video data. The presented work serves as a robust initial defense against fake media generated using the described deepfake tools. The survey underscores the potential of the proposed pipeline architecture for competitive results in deepfake detection. In future research, enhancing the system's robustness against manipulated videos using novel training techniques will be explored.

Detecting Real-Time Deep-Fake Videos Using Active Illumination

The study by Gerstner, C. and Farid, H. addresses the increasing challenge of detecting real-time deepfake videos in live video-conferencing applications. The trust traditionally placed in video calls is now threatened by the emergence of sophisticated deepfakes. The survey proposes a technique that exploits the unique imaging configuration in video calls, where participants are typically illuminated by the computer display. By measuring and comparing the temporal impact of dynamic,

colored squares displayed on the screen, the authenticity of video-call participants can be verified. The study evaluates the technique on simulated and real-world datasets, showcasing its efficacy under various environmental conditions. The proposed technique involves generating active illumination, detecting faces, separating sources, and measuring and comparing illumination patterns on the face. Key methods used include Active Illumination, Face Detection, Source Separation, Measurement, Comparison, and Countermeasures (Gerstner, C. & Farid, H. 2022).

The research highlights the importance of addressing the threat posed by real-time deepfakes in video-conferencing scenarios. By leveraging the controllable light source of the computer display, the proposed technique offers a viable intervention. Further enhancements, such as incorporating 3-D estimation of lighting and fine-grained measurements, could increase the difficulty for forgers to circumvent the system. Additionally, consideration is given to potential audio correlates for deepfake detection, particularly for synthetic voices. The study emphasizes the growing significance of techniques for authenticating video and audio calls, warranting increased attention from the media-forensics community.

Detecting Real-Time Deep-Fake Videos Using Active Illumination

The study by Bondi, L. et al addresses the need for reliable detection systems to identify deepfake videos on social media and the Internet. It explores the impact of various training strategies and data augmentation techniques on CNN-based deepfake detectors. The methodology involves face detection and extraction using BlazeFace, followed by training an EfficientNetB4 CNN model for deepfake detection. Four datasets, DF, DFD, DFDC, and CelebDF, are used for evaluation, and experiments are conducted using PyTorch on a workstation with NVIDIA Titan V (Bondi, L. et al. 2020).

The study highlights two main findings. Firstly, a carefully designed data-augmentation pipeline improves the generalization of CNN models for deepfake detection across different datasets. However, not all augmentations are equally beneficial, and their usefulness should be verified during pipeline development. Secondly, using triplet loss enhances both intra-dataset and cross-dataset detection performance, especially when training data is limited. For large datasets, data augmentation on a BCE-trained CNN architecture yields the best results.

III. PROPOSED SYSTEM

In creating deep fakes (DF), numerous tools have emerged; however, the availability of effective DF detection tools remains scarce. Our proposed system aims to bridge this gap by presenting an approach for detecting deep fakes, thereby preventing their dissemination across the World Wide Web. This contribution holds the potential to safeguard against the deceptive spread of DF-generated content.To achieve this, we will develop a user-friendly web-based platform that allows users to upload videos for classification as either fake or real. This platform can serve as a powerful first line of defense against the proliferation of deep fakes, offering an easy-to-use and accessible solution for users seeking to verify the authenticity of videos.

Crucially, we prioritize evaluating our system's performance and acceptability in terms of security, accuracy, reliability, and user-friendliness. By rigorously testing and fine-tuning our approach, we aim to instill confidence in users and stakeholders that they can rely on our system to identify and combat the growing threat of deep fake misinformation. Through continuous improvement and adaptation, we strive to create a robust and indispensable tool in the battle against deceptive media on the internet.

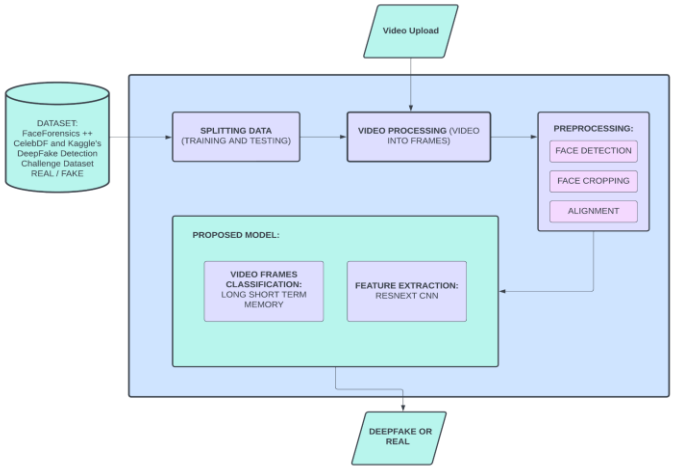

Figure 2: System Architecture of System

A. Dataset

To train and evaluate the model, a diverse and representative dataset of videos will be collected from FaceForensics ++, CelebDF and Kaggle's Deep fake Detection Challenge Dataset. The dataset should contain both authentic and manipulated videos, including various types of deepfake techniques and scenarios. Our newly prepared dataset contains 50% of the original video and 50% of the manipulated deep fake videos. The dataset is split into 70% train and 30% test set.

B. Preprocessing of Data

During dataset preprocessing, we perform several essential steps to prepare the data for deepfake detection. Firstly,splitting the video into individual frames to facilitate frame-level analysis. Subsequently, we apply face detection algorithms to identify and isolate the faces present in each frame. By cropping the frames to contain only the detected faces, we focus the analysis on the most relevant regions. To ensure consistency in the number of frames across the dataset, we calculate the mean number of frames in the entire video dataset. Then, we create a new processed face-cropped dataset containing frames equal to this mean value. Frames that do not contain detectable faces are omitted during this preprocessing phase to eliminate irrelevant data from the training process. Given the computational intensity involved in processing the entire 10-second video at 30 frames per second (resulting in a total of 300 frames), we acknowledge the need for efficient experimentation. To manage computational resources, we propose using only the first 100 frames for training the model. This allows us to gain valuable

insights and preliminary results while minimizing the computational burden.

C. Model

The proposed model consists of a ResNextCNN followed by one LSTM layer. The Data Loader is responsible for loading preprocessed face-cropped videos and splitting them into training and testing sets. Frames from the processed videos are then fed into the model in mini-batches for training and testing.

D. **ResNext CNN for Feature Extraction:**

We use the ResNext CNN classifier to extract features and accurately detect frame-level features. Afterward, we will fine-tune the network by adding the necessary additional layers and selecting an appropriate learning rate to ensure the gradient descent of the model converges effectively. The 2048-dimensional feature vectors obtained after the last pooling layers will serve as the input to the sequential LSTM.

E. **LSTM as Classifier:**

For classification, we will use an LSTM-based algorithm. The LSTM will take the sequence of 2048-dimensional feature vectors obtained from the ResNext CNN as input. The LSTM will process the sequence in a sequential manner, considering the temporal dependencies between frames, thus enabling the model to perform temporal analysis of the video. The objective of the classifier is to determine whether the input video sequence is part of a deep fake video or an untampered video.

The LSTM-based classifier will be designed with appropriate layers and dropout regularization to prevent overfitting and improve the model's generalization capabilities.

F. **Prediction**

During the prediction phase, a new video is provided as input to the trained model for analysis. This new video is first preprocessed to match the format required by the trained model. The preprocessing steps involve splitting the video into individual frames and applying face cropping techniques to isolate the facial regions within each frame.

Rather than storing the entire preprocessed video into local storage, the cropped frames are directly fed into the trained model for detection and classification. The model processes each frame sequentially, making predictions for each frame's authenticity, whether it belongs to a deep fake video or an untampered video. By analyzing the sequence of frames, the model can identify temporal patterns and variations, enhancing its ability to detect potential deep fake content accurately.

Using this approach of passing cropped frames directly to the trained model for prediction allows for efficient processing and minimizes storage requirements, making it a practical and effective solution for real-time or batch processing of videos to detect deep fake content.

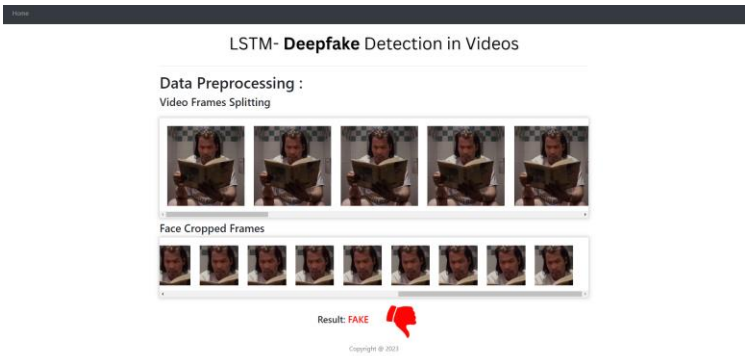

Figure 3: Model output/Result

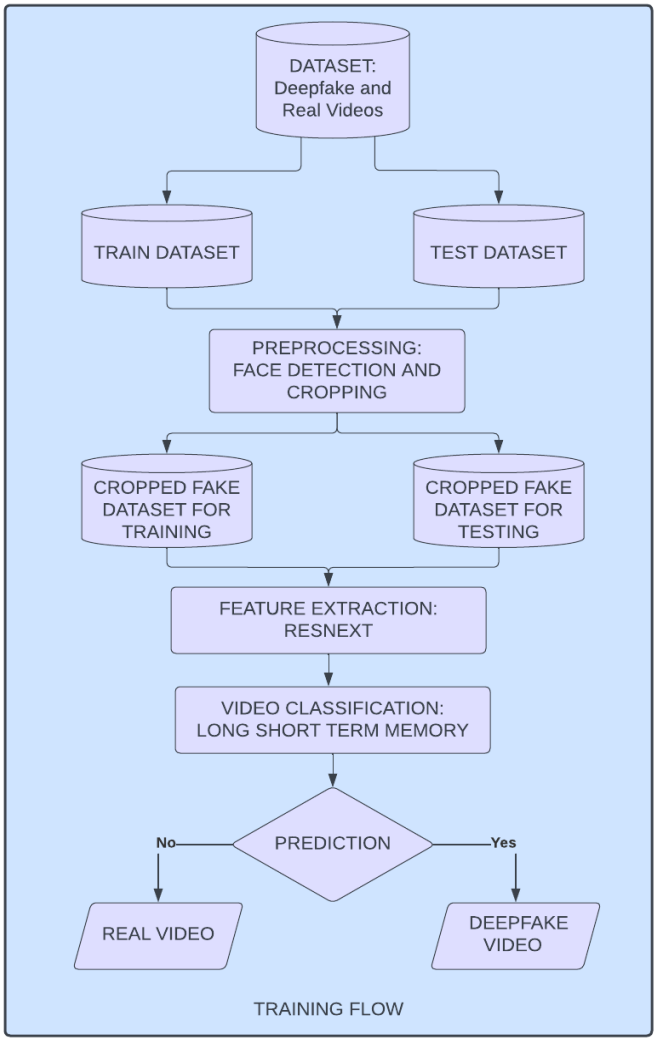

Figure 4: Training Flow

IV. RESULT

We developed web-based deep fake detection using ResNext for feature extraction and Long-Short-Term-Memory (LSTM) for video classification. Users can input a video into the system, and the model will determine whether the video is a deepfake or a genuine video. One example is shown in figure 3.

V. CONCLUSION

We developed a web-based deepfake detection system using a neural network-based approach to classify videos as either deep fakes or genuine. Our method draws inspiration from the way deep fakes are generated using Generative Adversarial Networks (GANs) .

The proposed method leverages the ResNext CNN for feature extraction and employs a Recurrent Neural Network (RNN) with Long-Short-Term-Memory (LSTM) for video classification. By analyzing the extracted features and temporal dependencies between frames, our method can accurately determine whether a video is a deep fake or real.

We have high confidence in the effectiveness of our proposed method, which utilizes state-of-the-art deep learning techniques for robust detection. The system's real-time capabilities allow for efficient and reliable deepfake detection, making it a valuable tool in identifying potentially deceptive content.

As with any detection system, continuous improvements and updates are necessary to keep up with evolving deepfake generation techniques. Nevertheless, our work lays a solid foundation for combatting the growing threat of deepfake content and contributes to the ongoing efforts to ensure video authenticity and trustworthiness in the digital age.

VI. LIMITATION

Our current method focuses solely on video content for deepfake detection and does not account for audio analysis. As a result, it may not be able to detect audio deep fakes. However, we plan to address this limitation in future iterations by exploring methods to achieve audio deepfake detection.

Additionally, our deepfake detection model relies on publicly available datasets, including FaceForensics++, CelebDF, and Kaggle's Deep Fake Detection Challenge Dataset. While these datasets offer diverse samples of deepfake and real videos, the model's performance may be influenced by the quality and diversity of the training data.

Despite these limitations, our method demonstrates promising results in accurately identifying deepfake videos based on analyzed features and temporal patterns. We acknowledge the need for continuous improvement and incorporation of advancements in the field to enhance the overall effectiveness and adaptability of our deepfake detection system.

ACKNOWLEDGMENT

We extend our sincere gratitude to the previous researchers whose work has significantly contributed to the successful completion of this research project. Their dedication, support, and valuable insights have played a vital role in shaping our approach and understanding of the subject matter. We are grateful for their contributions, which have been instrumental in the development and implementation of our deepfake detection system. Their efforts have inspired and guided us throughout this journey, and we acknowledge the impact of their research on our study.

REFERENCES

L. Bondi, E. Daniele Cannas, P. Bestagini and S. Tubaro, "Training Strategies and Data Augmentations in CNN-based DeepFake Video Detection," 2020 IEEE International Workshop on Information Forensics and Security (WIFS), New York, NY, USA, 2020, pp. 1-6, doi: 10.1109/WIFS49906.2020.9360901.

Gerstner, C. R., & Farid, H. (2022). Detecting real-time deep-fake videos using active illumination. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 53-60)

D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639163.

D. Liu, Z. Yang, R. Zhang and J. Liu, "A Robust Deepfake Video Detection Method based on Continuous Frame Face-swapping," 2022 International Conference on Artificial Intelligence, Information Processing and Cloud Computing (AIIPCC), Kunming, China, 2022, pp. 188-191, doi: 10.1109/AIIPCC57291.2022.00048.10.1109/AVSS.2018.8639163.

L. S and K. Sooda, "DeepFake Detection Through Key Video Frame Extraction using GAN," 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2022, pp. 859-863, doi: 10.1109/ICACRS55517.2022.10029095.