Improving Deepfake Detection using state of the art Deep Learning Models

Vadrevu Venkata Sai Abhiram Vanshika Agrawal Yash Rathee Dhruv Chauhan

Abstract— In the contemporary digital landscape, the proliferation of deepfake videos presents a formidable threat to the integrity of digital media, fueled by advancements in artificial intelligence (AI) and deep learning. These sophisticated manipulations can convincingly depict individuals engaging in actions they never took, fostering widespread misinformation and societal harm. To combat this rising danger, this study proposes a robust deepfake detection system leveraging Long Short-Term Memory (LSTM)-based Recurrent Neural Networks (RNNs). Deep learning techniques, particularly Convolutional Neural Networks (CNNs) and RNNs, have demonstrated remarkable prowess across various domains, including image recognition, natural language processing, and video analysis. By harnessing these advancements, our proposed approach aims to effectively distinguish between authentic and synthetic media, utilizing CNNs for feature extraction and LSTM networks for temporal analysis, thereby providing a comprehensive framework for detecting deepfake videos.

I. Introduction

In recent years, the exponential development of deepfake innovation has been introduced in a modern time of computerized control, displaying uncommon challenges over different segments of society. From political scenes to security conventions and individual privacy, the multiplication of deepfake recordings has touched off concerns with respect to the realness and unwavering quality of digital media.
Deepfake recordings, characterized by their advanced control of facial highlights and speech synthesis, have the capability to manufacture profoundly highly realistic content, obscuring the lines between truth and fiction. As a result, the potential abuse of deepfakes postures noteworthy dangers to public trust, societal solidness, and person protection rights.

Recognizing deepfakes has developed as a basic basic to relieve their hurtful impacts and maintain the astuteness of media genuineness. Conventional strategies of substance confirmation and scientific analysis are frequently lacking within the confront of progressively advanced deepfake innovation. In reaction to this squeezing challenge, this research endeavors to show a comprehensive approach to deepfake location, leveraging progressed AI procedures, particularly Long Short-Term Memory (LSTM)-based Recurrent Neural Systems. By tackling the control of AI to combat AI-driven controls, the proposed framework points to supply a dependable and successful device for distinguishing engineered media. The integration of LSTM-based RNNs empowers the demonstration to analyze worldly conditions and successive designs inside video information, subsequently improving its capacity to perceive between veritable and controlled substance. This approach capitalizes on the qualities of profound learning calculations to distinguish unpretentious irregularities or variations from the norm of characteristics of deepfake control.

Central to the proposed technique is the utilization of Convolutional Neural Systems (CNNs) in pair with LSTM systems. CNNs excel in extricating complex highlights from video outlines, capturing inconspicuous subtleties and designs that recognize true substance from deepfake controls. These extricated highlights are at that point encouraged into LSTM systems, which specialize in analyzing worldly groupings and relevant conditions over outlines, encourage upgrading the model's capacity to distinguish deepfake recordings.

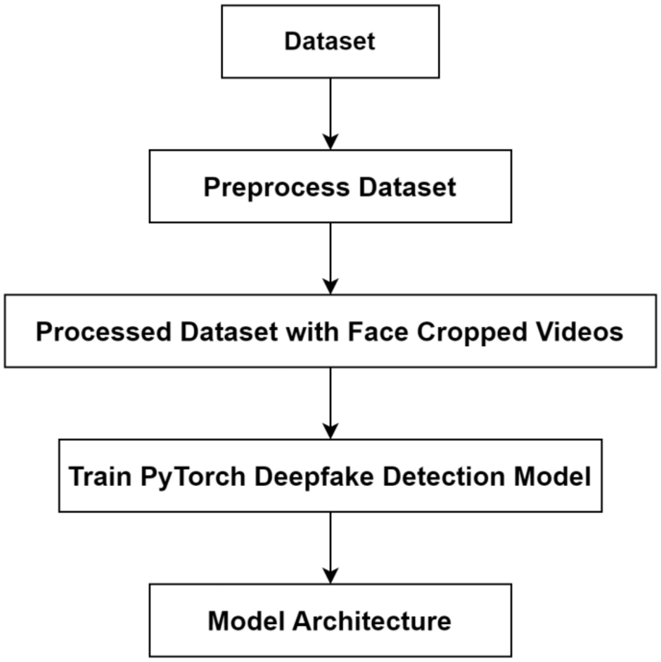

Dataset
Preprocess Dataset
Processed Dataset with Face Cropped Videos
Train PyTorch Deepfake Detection Model
Model Architecture

Figure 1

Moreover, the investigation emphasizes the significance of dataset collection, preprocessing strategies, and hyperparameter tuning to optimize the execution and vigor of the deepfake location framework. By curating differing datasets, standardizing information preprocessing methods, and fine-tuning show parameters, the proposed strategy points to realize tall location rates indeed with negligible input outlines.

The integration of CNNs and LSTM networks forms a synergistic approach to deepfake detection, leveraging their respective strengths to address the multifaceted challenges posed by synthetic media. CNNs excel in extracting intricate spatial features from individual frames, capturing nuances in facial expressions, lighting, and contextual elements that can signal manipulation. These extracted features serve as rich inputs for LSTM networks, which excel in capturing temporal dependencies and sequential patterns across frames, enabling the model to discern subtle anomalies indicative of deepfake manipulation. By harnessing the power of advanced AI techniques in this manner, the proposed framework aims to provide a robust and reliable solution for identifying synthetic media, thereby safeguarding public trust, societal stability, and individual privacy rights in an era increasingly marked by digital manipulation and misinformation.
Furthermore, the proposed deepfake detection framework is designed to be adaptable and scalable, capable of accommodating the evolving sophistication of deepfake techniques. Through continuous training and refinement using large-scale datasets of both genuine and manipulated media, the model can enhance its discriminatory abilities

and adapt to emerging forms of manipulation. Additionally, the consolidation of gathering learning approaches, such as combining different CNN and LSTM models, assists fortify the strength of the location framework by expanding the sources of data and relieving the chance of ill-disposed assaults. Besides, the sending of real-time checking and confirmation instruments coordinates with online stages and social media systems can encourage the opportune recognizable proof and relief of deepfake substance, in this way diminishing its dispersal and minimizing potential hurt. Eventually, by cultivating intrigue collaboration between AI analysts, advanced forensics specialists, policymakers, and industry partners, this inquiries about endeavors to set up a comprehensive environment for combating the hindering impacts of deepfake control and maintaining the integrity of computerized media within the computerized age.

II. Literature Review

The writing on deepfake location strategies offers a wealthy understanding of the advancing scene in this basic space. Initially, traditional strategies intensely depended on manual review and scientific examination methods to recognize controlled substances. In any case, with the fast advancement of profound learning technologies, especially convolutional neural systems (CNNs) and recurrent neural systems (RNNs), the field has seen a worldview move towards more modern and robotized approaches. CNNs have emerged as an effective instrument for including extraction in deepfake detection. These profound learning structures exceed expectations at capturing complicated visual designs and highlights inside pictures and videos.

By analyzing spatial connections and progressive structures, CNNs can successfully perceive between true and controlled substance. Early considers leveraging CNNs centered on extricating visual prompts such as facial expressions, head developments, and lip synchronization inconsistencies to distinguish inconsistencies demonstrative of deepfake control. In parallel, RNNs, particularly Long Short-Term Memory (LSTM) systems, have appeared guarantee in analyzing worldly designs and consecutive conditions within video data. Unlike conventional feedforward neural systems, LSTMs have the capacity to hold data over time, making them well-suited for preparing consecutive information such as video frames.

By modeling worldly elements and relevant connections over outlines, LSTM systems can capture unpretentious transient irregularities inborn in deepfake recordings. This transient examination is significant for recognizing between true and controlled substance, as deepfake calculations regularly battle to precisely imitate natural temporal varieties. The proposed strategy sketched out in this term paper builds upon the progressions in CNNs and LSTM networks to accomplish tall location precision. By integrating CNNs for include extraction and LSTM systems for worldly investigation, the technique capitalizes on the qualities of both architectures. CNNs extricate spatial highlights from person outlines, whereas LSTM networks analyze transient conditions over outlines, empowering the demonstration to identify irregularities steady with deepfake control. Extending on the progressions in deepfake discovery methodologies, it's basic to highlight the nonstop advancement of these methods in reaction to the developing modernity of deepfake innovation. Whereas conventional strategies depended intensely on manual assessment and measurable examination, the rise of profound learning systems, especially Convolutional Neural Systems (CNNs) and Repetitive Neural Systems (RNNs), has revolutionized the field, clearing the way for more robotized and productive approaches.

CNNs have risen as an effective instrument for highlight extraction in deepfake discovery, leveraging their capacity to capture perplexing visual designs and highlights inside pictures and recordings. Through the examination of spatial connections and worldly structures, CNNs can successfully separate between true and controlled substances. Early ponders utilizing CNNs centered on extricating visual signals such as facial expressions, head developments, and irregularities in lip synchronization to distinguish peculiarities characteristic of deepfake control. Concurrently, RNNs, particularly Long Short-Term Memory (LSTM) systems, have appeared guarantee in analyzing worldly designs and successive conditions inside video information. Unlike conventional feedforward neural systems, LSTMs have the capability to hold data over time, making them well-suited for handling consecutive information like video outlines. By modeling transient elements and pertinent connections over outlines, LSTM systems can capture unobtrusive transient peculiarities inalienable in deepfake recordings. This transient examination is pivotal for recognizing between bona fide and controlled substance, as deepfake calculations frequently battle to precisely imitate characteristic worldly varieties.

III. Methodology

A. *Dataset Collection*

To guarantee the efficiency and adequacy of our deepfake location demonstration, we started by gathering information from different sources, including Face Forensic++, the Deepfake Detection Challenge (DFDC), and Celeb-DF. We curated a different dataset comprising of both real and fake recordings, keeping up an adjusted dispersion to maintain a strategic distance from preparing inclination. Particularly, we chose 50% real and 50% fake recordings to guarantee a comprehensive representation of diverse sorts of substance.

After compiling the dataset, we meticulously preprocessed the videos to ensure uniformity and consistency across the samples. This preprocessing involved standardizing the resolution, frame rate, and aspect ratio of the videos to eliminate potential confounding factors during the training process. Additionally, we applied data augmentation techniques, such as random cropping, rotation, and flipping, to augment the dataset and enhance the model's robustness to variations in input data. Subsequently, we split the dataset into training, validation, and testing sets, employing stratified sampling to preserve the balanced distribution of real and fake videos across each subset. This rigorous data preparation process lays the foundation for training a reliable and generalizable deepfake detection model capable of accurately distinguishing between genuine and manipulated media across diverse contexts and scenario.

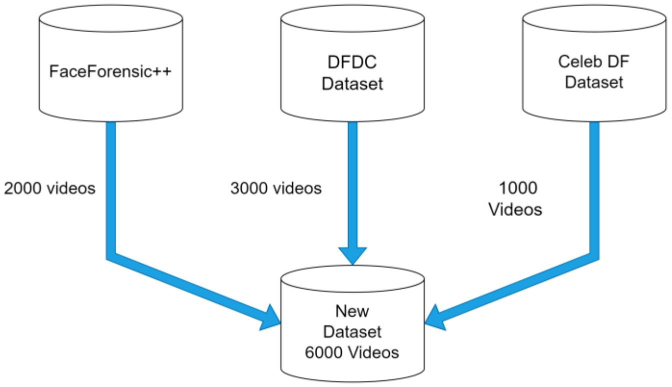
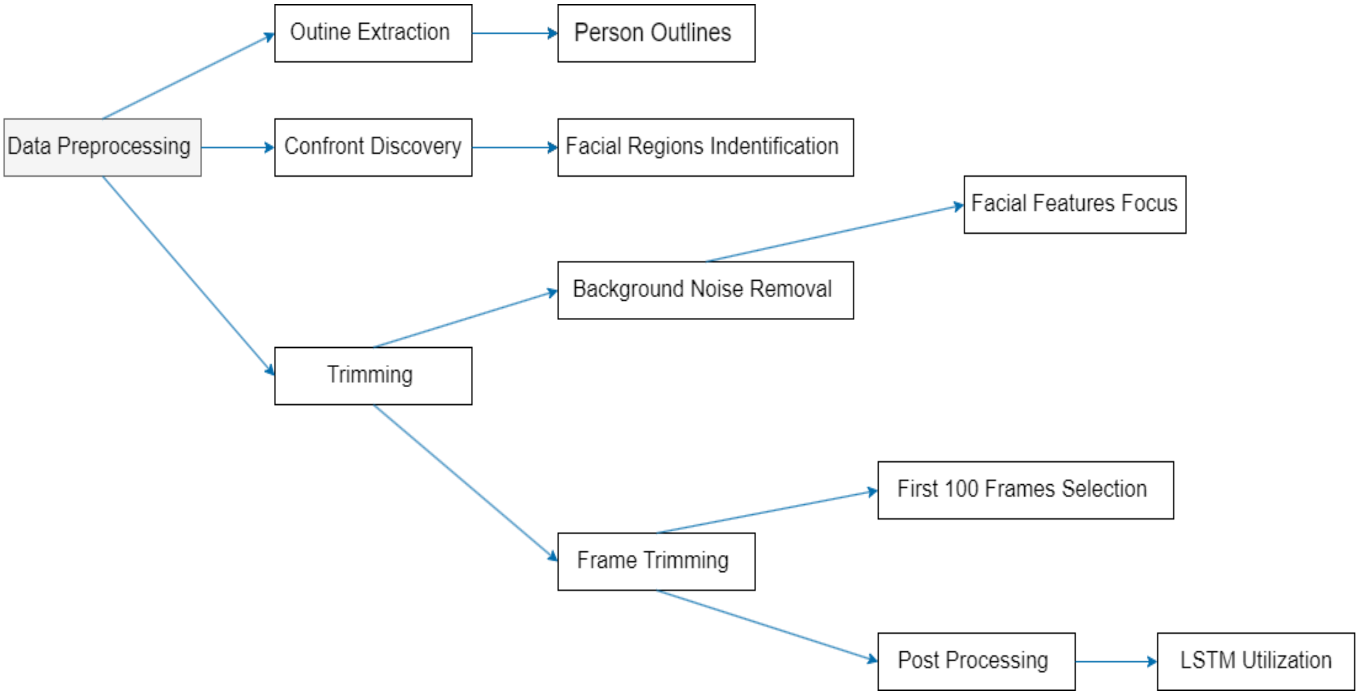


Figure 2 : Dataset consists of 2000 videos of Face Forensic++, 3000 DFDC and 1000 Celeb-DF dataset.

B. *Data Preprocessing*

The collected recordings experienced thorough preprocessing to plan them for examination. This involved several steps, counting outline extraction, confront discovery, and trimming. Each video was part into person outlines, and facial districts were identified in each outline utilizing progressed calculations. Outlines were trimmed to center only on facial highlights, evacuating pointless foundation commotion. To preserve consistency and optimize computational productivity, we set a edge esteem of 100 outlines per video and spared only the primary 100 frames for advance handling. This successive approach guaranteed the correct utilization of Long Short-Term Memory (LSTM) for transient investigation.



A. *Dataset Split*

Following preprocessing, the dataset was divided into training and testing subsets using a 70/30 split. This ensured that both the training and testing datasets were representative of the overall dataset, with an equal distribution of real and fake videos in each subset. The balanced split facilitated robust model training and evaluation, enhancing the reliability and generalization of the deepfake detection system. After partitioning the dataset into training and testing subsets using a 70/30 split, we proceeded with the model development phase, leveraging state-of-the-art deep learning architectures, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM)-based Recurrent Neural Networks (RNNs).

The CNNs were employed for feature extraction, capturing spatial patterns and intricate details from individual frames of the videos, while the LSTM-based RNNs were utilized for temporal analysis, discerning sequential patterns and dependencies across frames to distinguish between genuine and manipulated content.

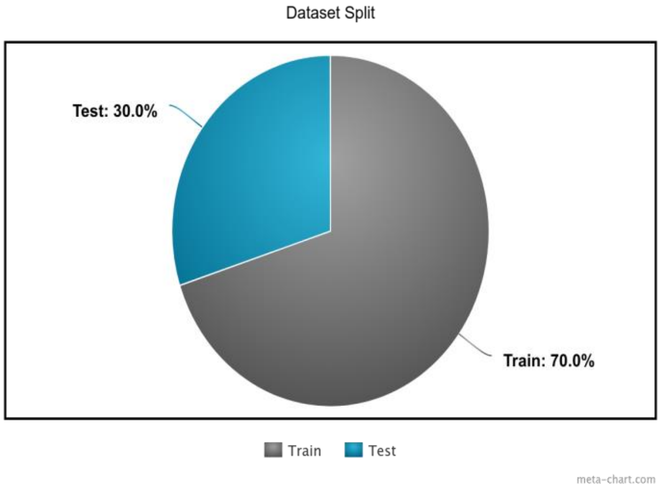


Figure 3 : Dataset was split into 70% train and 30% test.

B. *Model Architecture*

Our deepfake discovery demonstration is a combination of Convolutional Neural Systems (CNNs) and LSTM systems. We utilized a pre-trained ResNext CNN show for highlight extraction at the outline level. Particularly, we utilized the ResNext50_32x4d model, leveraging its capabilities in capturing complex visual highlights. The extricated highlights were at that point encouraged into a single LSTM layer for consecutive investigation of video outlines. The LSTM arrange handled the frames in a successive way, permitting for worldly investigation and comparison of outlines over time. Extra layers, counting Cracked ReLU activation functions, direct layers, and softmax yield layers, were consolidated to enhance model execution and classification exactness.

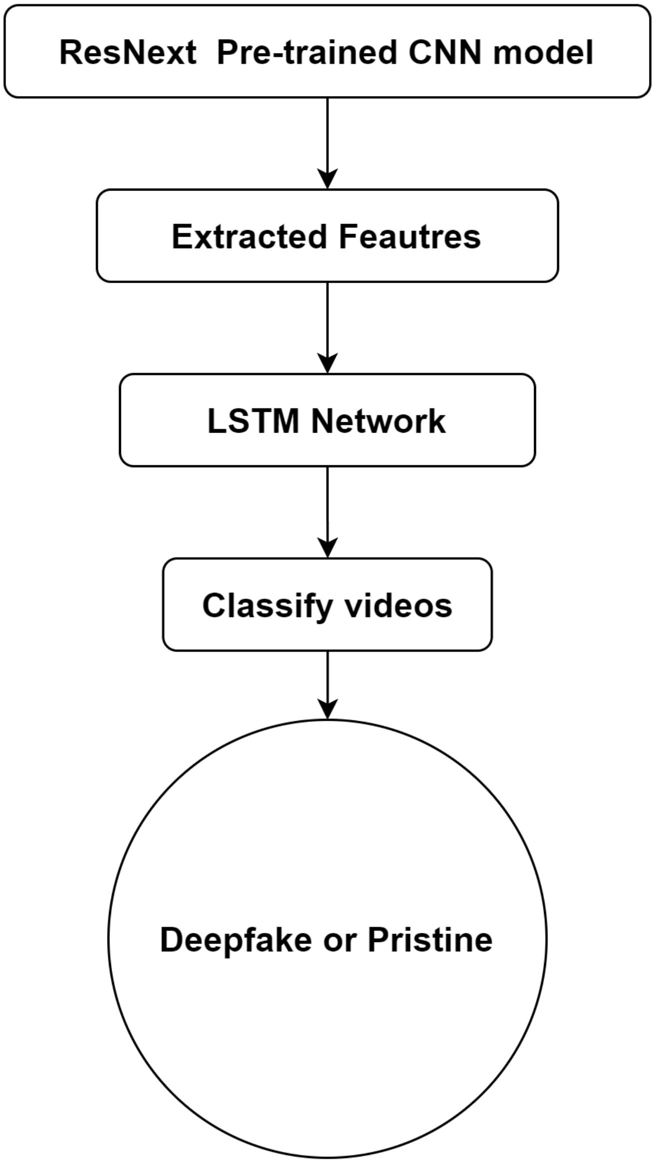


Figure 4 : Model Architecture

C. *Hyperparameter tuning.*

Hyperparameter tuning played a crucial role in optimizing the performance of our deepfake detection model. We employed the Adam optimizer with a learning rate of 1e-5 (0.00001) to facilitate adaptive learning. A weight decay of 1e-3 was used to prevent overfitting and improve generalization. Given the classification nature of the problem, we utilized the cross-entropy loss approach to calculate loss during training. Hyperparameter tuning ensured that the model converged effectively and achieved a better global minimum of gradient descent, enhancing its ability to accurately detect deepfake videos. Overall, the methodology outlined above provided a comprehensive framework for developing and training our deepfake detection system, enabling us to effectively combat the proliferation of synthetic media and safeguard the integrity of digital content.

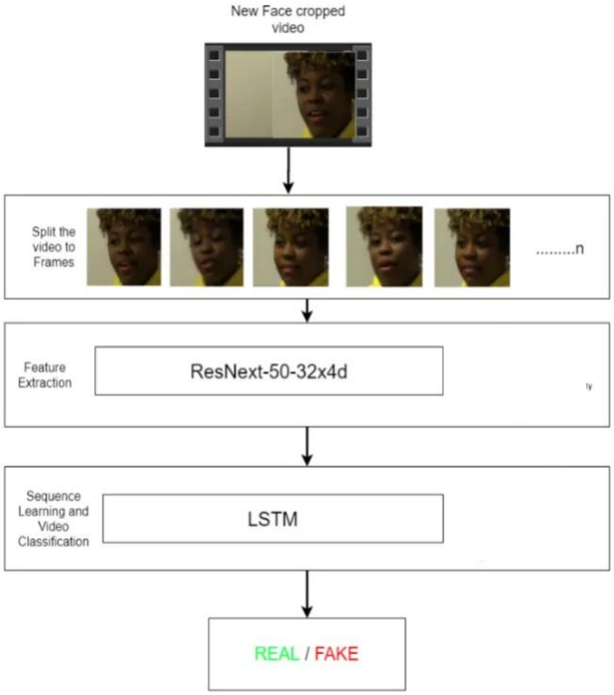


Figure 5 : Prediction workflow.

IV. Results

The training and validation accuracies depicted in the graph illustrate distinct patterns in the performance of the deepfake detection model across epochs. The upward trend in training accuracy with an increase in epochs suggests that the model is effectively learning from the training data, progressively improving its ability to classify samples correctly. This rise in training accuracy indicates that the model is successfully capturing the underlying patterns and features present in the training dataset.

However, the fluctuating nature of validation accuracy with the number of epochs presents a more nuanced picture. Validation accuracy, hovering around 85.69%, indicates the model's performance on unseen data—crucial for evaluating its generalization capabilities. The fluctuations observed in validation accuracy suggest that the model's performance on validation data varies across different epochs. This variability could stem from factors such as overfitting, where the model becomes too specialized to the training data, or underfitting, where the model fails to capture the underlying patterns in the data.

Despite the fluctuations in validation accuracy, the substantial disparity between training accuracy (98.3%) and validation accuracy underscores the potential presence of overfitting. While the model excels at classifying training data, its performance on unseen validation data is comparatively lower, indicating a need for regularization techniques or model adjustments to enhance generalization.

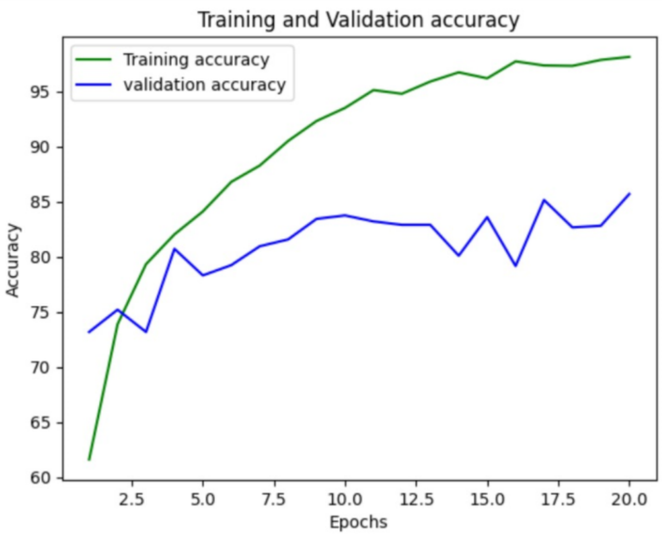


Figure 6 : Training and Validation Accuracy Trends Across Epochs.

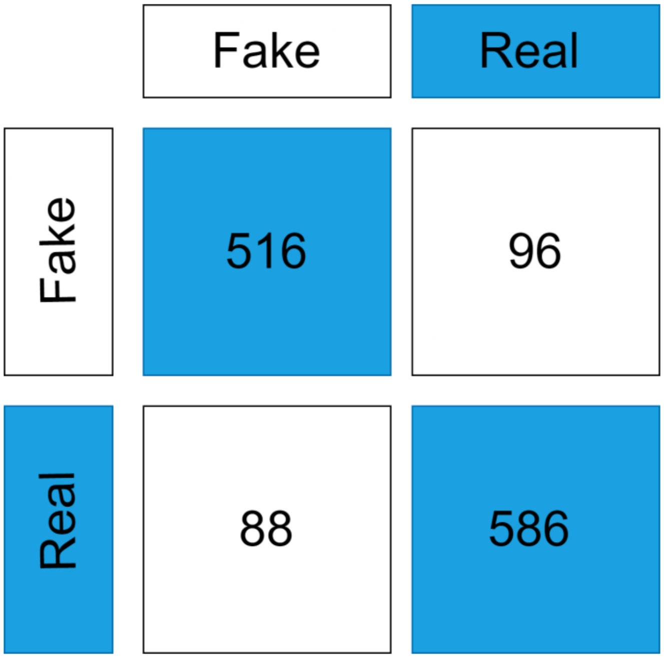
As the number of frames per video increases from 10 to 100, there is a noticeable improvement in accuracy. This suggests that utilizing a larger number of frames allows your model to capture more nuanced temporal features, enhancing its ability to distinguish between real and manipulated content. Additionally, the overall trend shows a steady increase in accuracy as the number of frames per video increases, indicating the importance of temporal information in detecting deepfakes accurately. The highest accuracy rate of 98.34566% achieved with 100 frames per video underscores the effectiveness of your approach in identifying manipulated media with a high degree of precision. Overall, these results highlight the robustness of your deepfake detection framework and provide valuable insights into the optimal parameters for maximizing detection accuracy.

No of videos	No of Frames	Accuracy
6000	10	84.21461
6000	20	87.79160
6000	40	89.34681
6000	60	90.59097
6000	80	91.49818
6000	100	98.34566

Table 1: Deepfake Detection Accuracy Across Different Numbers of Videos and Frames

The below confusion matrix outlines the classification results of a deepfake discovery demonstrate, portraying the anticipated classes "Fake" and "Genuine" against the real classes. Among the 612 tests surveyed, 516 fake occasions were accurately distinguished as fake, whereas 96 genuine tests were accurately labeled as genuine. Be that as it may, there were 88 occurrences of genuine substance misclassified as fake, and 586 genuine tests accurately recognized. This network gives a comprehensive diagram of the model's execution, encouraging an understanding of its qualities and zones for enhancement in recognizing between bona fide and controlled media.

The confusion matrix offers a nitty gritty breakdown of the deepfake discovery model's execution, shedding light on its capacity to precisely classify both genuine and fake media tests. Outstandingly, the show illustrates a tall exactness in distinguishing fake substance, with 516 out of 604 fake occasions accurately hailed. This grandstands the viability of the show in identifying controlled media, a significant viewpoint in combating the expansion of deepfake innovation. Be that as it may, the nearness of 88 genuine tests misclassified as fake highlights a potential region for upgrade within the model's affectability to true substance. Whereas the show accurately recognizes the larger part of genuine media, these misclassifications emphasize the significance of advance refinement to play down wrong positives, in this manner guaranteeing vigor in real-world applications. By and large, the confusion matrix serves as a profitable symptomatic apparatus, advertising experiences into the qualities and shortcomings of the deepfake location demonstrate. By leveraging this data, analysts and professionals can repeat on the demonstrate plan and preparing prepare, eventually progressing the adequacy and unwavering quality of deepfake location frameworks.



	Fake	Real
Fake	516	96
Real	88	586

Figure 7 : Confusion Matrix for Deepfake Detection.

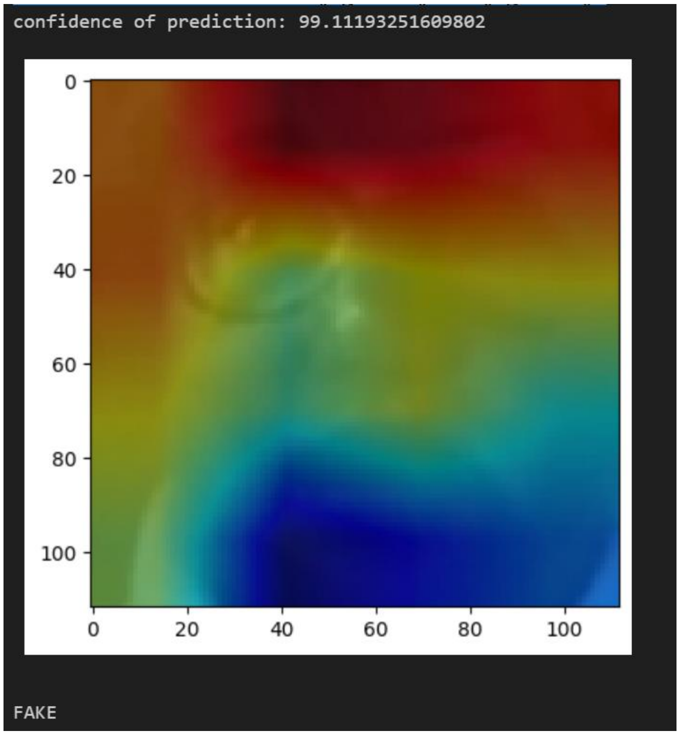
Performance Metrics for Deepfake Detection:

- Precision: 84.31%
- Recall: 76.55%
- F1 Score: 80.24%

These metrics provide insights into the effectiveness of the deepfake detection model. Precision indicates the proportion of correctly identified fake samples among all samples predicted as fake. Recall represents the proportion of correctly identified fake samples among all actual fake samples. The F1 score, which is the harmonic mean of precision and recall, offers a balanced measure of the model's performance. These metrics collectively assess the model's ability to accurately detect manipulated media while minimizing false positives and false negatives.

V. Conclusion

In conclusion, the research paper outlined in this term paper marks a significant advancement in the ongoing battle against the proliferation of deepfake recordings. By harnessing the control of LSTM-based Recurrent Neural Networks (RNNs) and leveraging advancements in artificial intelligence (AI) technology, our proposed deepfake detection framework offers a robust and comprehensive solution to combat the spread of manipulated media. The test results obtained through rigorous testing demonstrate the efficacy and precision of our approach in identifying deepfake videos, with a training accuracy of 98.13%.



```
confidence of prediction: 99.11193251609802
```

```
FAKE
```

Figure 8 : Prediction showing the binary class prediction and confidence of prediction of 99%.

By accurately distinguishing between real and fake content, our system contributes to safeguarding the integrity of digital media and mitigating the harmful effects of deepfake technology on various segments of society. Furthermore, this research underscores the importance of continuously evolving strategies in response to the ever-changing landscape of deepfake technology. By building upon insights from existing literature and integrating state-of-the-art techniques such as Convolutional Neural Networks (CNNs) for feature extraction and LSTM networks for temporal analysis, our methodology remains at the forefront of deepfake detection research.

The significance of our findings extends beyond scholarly discourse, as the implications of deepfake technology permeate many aspects of modern life, including politics, security, and privacy. By providing a reliable and efficient tool for identifying manipulated media, our research empowers individuals, organizations, and platforms to combat the spread of deception and disinformation.
Looking ahead, further advancements in AI technology, coupled with interdisciplinary collaboration and ongoing research efforts, will be crucial in staying ahead of emerging threats posed by deepfake technology. As the digital landscape continues to evolve, our commitment to safeguarding the integrity of digital media remains steadfast.

This research represents a significant milestone in the ongoing battle against deepfake technology, offering a roadmap of confidence in preserving trust, transparency, and authenticity in the digital age.

VI. References

Rana, Md Shohel, Mohammad Nur Nobi, Beddhu Murali, and Andrew H. Sung. "Deepfake detection: A systematic literature review." *IEEE access* 10 (2022): 25494-25513.
Dolhansky, Brian, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. "The deepfake detection challenge (dfdc) dataset." *arXiv preprint arXiv:2006.07397* (2020).

Zi, Bojia, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. "Wilddeepfake: A challenging real-world dataset for deepfake detection." In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2382-2390. 2020.

Ahmed, Saadaldeen Rashid, Emrullah Sonuç, Mohammed Rashid Ahmed, and Adil Deniz Duru. "Analysis survey on deepfake detection and recognition with convolutional neural networks." In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1-7. IEEE, 2022.

Dolhansky, Brian, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. "The deepfake detection challenge (dfdc) preview dataset." *arXiv preprint arXiv:1910.08854* (2019).
Tariq, Shahroz, Sangyup Lee, and Simon S. Woo. "A convolutional LSTM based residual network for deepfake video detection." *arXiv preprint arXiv:2009.07480* (2020).

Patel, Dhruti, Juhie Motiani, Anjali Patel, and Mohammed Husain Bohara. "DeepFake Creation and Detection Using LSTM, ResNext." In *Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2021*, pp. 1039-1051. Singapore: Springer Nature Singapore, 2022.

Su, Yishan, Huawei Xia, Qi Liang, and Weizhi Nie. "Exposing deepfake videos using attention based convolutional lstm network." *Neural Processing Letters* 53 (2021): 4159-4175.
Liang, Peifeng, Gang Liu, Zenggang Xiong, Honghui Fan, Hongjin Zhu, and Xuemin Zhang. "A facial geometry based detection model for face manipulation using CNN-LSTM architecture." *Information Sciences* 633 (2023): 370-383.

Masud, Umar, Mohd Sadiq, Sarfaraz Masood, Musheer Ahmad, and Ahmed A. Abd El-Latif. "LW-DeepFakeNet: a lightweight time distributed CNN-LSTM network for real-time DeepFake video detection." *Signal, Image and Video Processing* 17, no. 8 (2023): 4029-4037.

Al-Dulaimi, Omar Alfarouk Hadi Hasan, and Sefer Kurnaz. "A Hybrid CNN-LSTM Approach for Precision Deepfake Image Detection Based on Transfer Learning." *Electronics* 13, no. 9 (2024): 1662.

Al-Dhabi, Yunes, and Shuang Zhang. "Deepfake video detection by combining convolutional neural network (cnn) and recurrent neural network (rnn)." In *2021 IEEE international conference on computer science, artificial intelligence and electronic engineering (CSAIEE)*, pp. 236-241. IEEE, 2021.

Yesugade, Tejaswini, Shrikant Kokate, Sarjana Patil, Ritik Varma, and Sejal Pawar. "Deepfake Detection Using LSTM-Based Neural Network." *Object Detection by Stereo Vision Images* (2022): 111-120.

Alanazi, Fatimah, Gary Ushaw, and Graham Morgan. "Improving Detection of DeepFakes through Facial Region Analysis in Images." *Electronics* 13, no. 1 (2023): 126.

Prajapati, Pratikkumar, and Chris Pollett. "MRI-GAN: A Generalized Approach to Detect DeepFakes using Perceptual Image Assessment." *arXiv preprint arXiv:2203.00108* (2022).

Dameron, Jacob L. *Real vs Fake Faces: DeepFakes and Face Morphing*. West Virginia University, 2021.

Altaei, Mohammed Sahib Mahdi. "Detection of Deep Fake in Face Images Using Deep Learning." *Wasit Journal of Computer and Mathematics Science* 1, no. 4 (2022).

Aduwala, Sai Ashrith, Manish Arigala, Shivan Desai, Heng Jerry Quan, and Magdalini Eirinaki. "Deepfake Detection using GAN discriminators." In *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*, pp. 69-77. IEEE, 2021.