Safeguarding Society: A DeepFake Video Detection Framework

Chaitali Nigade[1], Shreya Sakare[2], Shruti Sakare[3], Mayuri Salunkhe[4], Nilofar Mulla[5]
[1,2,3,4,5] *Information Technology Department, Bharati Vidyapeeth's College of Engg. for Women, Pune, Maharashtra, India*

Abstract - With the growing prevalence of deepfake media, the need for effective detection methods has become crucial to combat misinformation and preserve the integrity of digital content [1]. This project focuses on the development and implementation of a deepfake detection model using a combination of ResNet50v2 and Long Short-Term Memory (LSTM) networks [3]. The proposed model aims to identify and classify deepfake content through an intricate analysis of both spatial and temporal features in video. To enhance detection performance, transfer learning is employed by leveraging ResNet50v2 as the base model, which is pre-trained on large-scale datasets such as ImageNet. Instead of training the model from scratch, transfer learning enables the system to utilize the rich feature representations learned by ResNet50v2, making the detection process more efficient and accurate [6]. The ResNet50v2 component captures spatial patterns, such as facial features and inconsistencies in pixel structures, which are often indicative of manipulated media [4]. Meanwhile, LSTMs, known for their ability to process sequential data, analyze temporal features to detect irregularities in frame sequences and unnatural speech patterns—common indicators of deepfake videos [7]. The model is optimized for real-time detection, allowing it to be applied in various scenarios such as live video streams, social media content verification, and multimedia forensics [9]. Additionally, the system includes a user-friendly interface for monitoring and managing the detection process, providing detailed analysis and reports for end-users and content moderators [8].

Keywords – DeepFake, Long Short-Term Memory (LSTM), ResNet50v2, Transfer Learning

1. INTRODUCTION

In today's digital world, we are witnessing remarkable advancements in technology, but these developments also bring new challenges. One of the most concerning issues is the rise of deepfake videos, which use artificial intelligence to convincingly alter someone's appearance or actions. These manipulations can be exploited to spread misinformation, damage reputations, manipulate public opinion, or even facilitate fraud, making it increasingly difficult to trust digital content [1].

Our project, the Safeguarding Society: DeepFake Video Detection Framework, is designed to address this growing problem. By leveraging transfer learning, the system applies advanced machine learning techniques to analyze videos for subtle signs of manipulation, such as unnatural facial movements or pixel inconsistencies [3]. Instead of training a model from scratch, transfer learning enables us to use ResNet50v2, a pre-trained convolutional neural b ++.3network, as a base model. This approach significantly improves detection efficiency by utilizing the rich feature representations learned from large-scale datasets, reducing computational cost while enhancing accuracy [6].

The project integrates a combination of Long Short-Term Memory (LSTM) networks and ResNet50v2 to effectively detect deepfakes. LSTM, a type of recurrent neural network (RNN), is well-suited for processing sequential data, making it ideal for analyzing video content [7]. By examining the temporal relationships between frames, LSTM identifies inconsistencies in motion and facial expressions that conventional models may overlook. ResNet50v2, originally trained on the ImageNet dataset, is fine-tuned to extract spatial features within individual frames. Its residual learning framework allows for efficient training of deep architectures, improving the model's ability to detect subtle visual anomalies and pixel-level inconsistencies indicative of manipulation [4]. The synergy between these two architectures ensures a robust detection system that effectively analyzes both spatial and temporal aspects,

enabling the identification of even the most sophisticated deepfake videos [9].

II. PROPOSED METHODOLOGY AND ALGORITHMS

2.1 Proposed Methodology

The proposed framework starts with data preprocessing, where input videos are decomposed into individual frames at a fixed frame rate to ensure adequate temporal coverage. Each frame is resized to a standard dimension, for example, 256x256 pixels, and normalized for consistency, with face detection algorithms isolating facial regions to minimize noise from irrelevant areas [1].

To enhance efficiency and accuracy, the framework employs transfer learning, using ResNet50v2 as a pre-trained base model rather than training from scratch. Transfer learning allows the system to leverage rich feature representations learned from large-scale datasets, reducing computational costs while improving performance [6]. In the spatial analysis stage, ResNet50v2 is applied to extract detailed spatial features. Grouped convolutions along with residual connections help detect pixel inconsistencies, unnatural lighting, and distorted facial features, which may indicate deepfake manipulation [4].

The extracted spatial features are then processed in the temporal analysis stage using LSTM, which captures sequential dependencies across frames. LSTM focuses on motion anomaly detection, identifying mismatched lip-sync or unnatural blinking, which a CNN alone may fail to detect [7]. The next stage integrates spatial and temporal features by passing them through a dense layer and subsequently classifying them via the SoftMax function as "Real" or "Fake" [9].

The model has been trained using FaceForensics++, Celeb-DF, and DFFD datasets, which contain a diverse set of real and fake videos, ensuring robustness against various deepfake generation techniques [3]. Finally, the system outputs a classification label, providing a comprehensive approach to detecting both spatial and temporal inconsistencies in deepfake videos. Figure 1 depicts the general process flow for deepfake video detection.

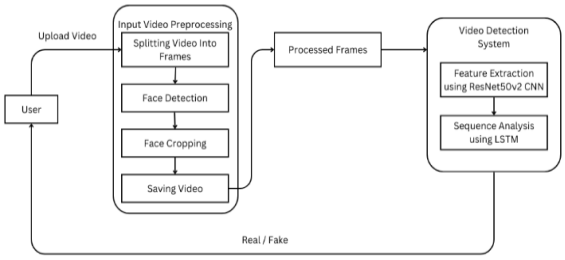

Fig. 1. Video Detection Flow

2.2 Algorithms

2.2.1 Transfer Learning

To enhance the efficiency and accuracy of deepfake detection, our project employs transfer learning by utilizing ResNet50v2 as the base model. Transfer learning is a powerful machine learning technique where a pre-trained model, originally developed for a large-scale dataset, is fine-tuned for a specific task. Instead of training a deep neural network from scratch, which requires extensive computational resources and large amounts of data, transfer learning allows us to leverage the pre-trained ResNet50v2 models feature extraction capabilities [3].

In our approach, ResNet50v2, pre-trained on the ImageNet dataset, is adapted for deepfake detection by fine-tuning its final layers. The lower layers of ResNet50v2, responsible for detecting fundamental spatial patterns such as edges and textures, remain mostly unchanged [4]. However, the higher layers are retrained on our curated deepfake dataset (FaceForensics++, Celeb-DF, and DFFD) to recognize deepfake-specific artifacts such as pixel inconsistencies, unnatural facial textures, and lighting distortions [1]. This fine-tuning process enables the model to specialize in detecting subtle manipulations while maintaining the robustness of its pre-trained features [7].

By integrating transfer learning, our system achieves faster training times, improved accuracy, and better generalization across various deepfake video styles [6]. This approach not only reduces computational overhead but also ensures that the model effectively distinguishes real videos from manipulated ones, making it a reliable solution for real-world applications in digital forensics, social media verification, and content moderation [9].

2.2.1 Long Short-term Memory

The LSTM module plays a vital role in the deepfake detection system by analyzing the temporal dynamics of videos, addressing a critical gap that standard convolutional models, which focus only on individual frames, cannot cover [1]. Long Short-Term Memory (LSTM) networks, a type of Recurrent Neural Network (RNN), are specifically designed to process sequential data by retaining long-term dependencies, making them well-suited for understanding the flow of information across video frames [3].

In this system, a video is processed frame by frame, with spatial features first extracted using the ResNet50v2 module. To enhance efficiency and accuracy, transfer learning is employed by utilizing a pre-trained ResNet50v2 model as the base, allowing it to leverage previously learned features from large-scale datasets such as ImageNet while being fine-tuned for deepfake detection [6]. This enables the model to detect visual artifacts like misaligned facial features, unnatural textures, and pixel inconsistencies [4].

The extracted spatial features are then passed to the LSTM module, which examines how these features evolve over time. This step is crucial for identifying temporal inconsistencies such as irregular blinking patterns, abrupt changes in facial expressions, or disjointed movements, which are often indicative of deepfake manipulation [7]. The LSTM effectively maintains a memory of visual patterns across frames, enabling it to classify the video as real or fake based on the naturalness of its temporal behavior [9].

By integrating ResNet50v2's spatial analysis with LSTM's sequential processing, the system provides a more robust and reliable method for detecting deepfake videos. The combination of transfer learning and deep learning architectures ensures high accuracy while optimizing computational efficiency, making it applicable for real-time deepfake detection across various digital platforms.

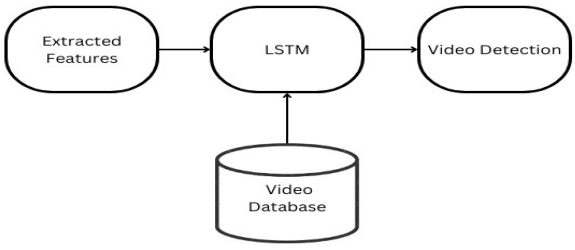

Fig. 2. Long Short-Term memory (LSTM)

2.2.2 ResNet50v2

The ResNet50v2 module is a key component of the deepfake detection system, designed to extract spatial features from individual video frames. ResNet50v2 is a highly efficient convolutional neural network (CNN) architecture known for its improved residual learning framework, which enhances gradient flow during training. This enables it to capture detailed spatial features and detect even the most subtle inconsistencies within images [1].

To improve efficiency and accuracy, the proposed system employs transfer learning by utilizing ResNet50v2 as a pre-trained base model rather than training from scratch. By leveraging the rich feature representations learned from large-scale datasets such as ImageNet, the model can focus on fine-tuning its higher layers to specialize in deepfake detection. This significantly reduces computational costs while improving performance [6].

In the deepfake detection framework, ResNet50v2 processes each frame of a video to identify visual inconsistencies that may indicate tampering. These inconsistencies include subtle anomalies such as unnatural textures, misaligned facial features, or blending artifacts—common signs of manipulation in deepfake videos [4]. The model's optimized residual connections allow it to learn complex spatial patterns, improving its ability to detect fine-grained artifacts within frames [7].

Once the spatial features are extracted, they are passed to the LSTM module, which analyzes the temporal dynamics across video frames. ResNet50v2 complements LSTM by providing a strong foundation of spatial data, enabling the system to detect both frame-level artifacts and temporal inconsistencies [9].

By integrating ResNet50v2's powerful spatial feature extraction with LSTM's sequential analysis, while leveraging transfer learning, the deepfake detection system becomes more robust and efficient. This approach ensures accurate identification of fake videos across various manipulation techniques while optimizing computational resources for real-time applications [3].

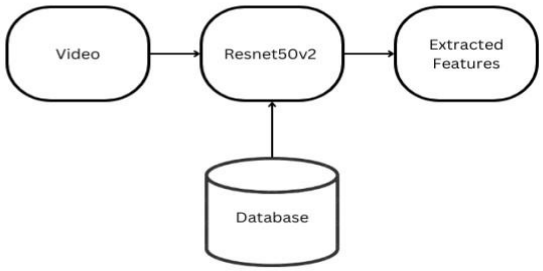

Fig. 3. ResNet50v2

III. SYSTEM ARCHITECHTURE

The system begins with the Upload Video step, where a video is provided as input. This video is compared against a dataset containing both real and fake videos. To ensure effective analysis, the dataset undergoes a Preprocessing stage, where videos are decomposed into individual frames, normalized for consistency, and key features are extracted. Face detection algorithms are applied to isolate facial regions, reducing noise from irrelevant areas. The processed information is then stored in the Processed Dataset for further use [1].

To enhance model efficiency, transfer learning is employed by using ResNet50v2 as a pre-trained base model instead of training from scratch. This allows the system to leverage previously learned spatial features while fine-tuning the higher layers for deepfake detection, improving both computational efficiency and accuracy [6].

The system divides the dataset into training and validation sets in the data splitting phase, ensuring a robust model training and evaluation process. A data loader feeds these sets into the deepfake detection model, which utilizes ResNet50v2 CNN for extracting spatial features and LSTM for analyzing temporal dynamics across frames. This combination allows the model to detect both frame-level and sequential anomalies, identifying subtle inconsistencies indicative of deepfakes [4].

Once the model is trained, it is saved through the export trained model step and can later be used for real-time predictions in the load trained model step. The final outcome of the system is a classification output, where the video is labeled as either "real" or "fake". The workflow clearly distinguishes between the training flow (solid arrows) and prediction flow (dashed arrows), ensuring a seamless and efficient pipeline for deepfake detection [7].

IV. DATASETS

The datasets used in the system are Face Forensics++, CelebDF, DFFD. These datasets are taken from kaggle[9]. All three datasets are merged together and the final dataset consists of total 6200 videos in which 4000 videos are fake videos and 2200 are real videos. These videos are preprocessed and are stored in another folder. During preprocessing the videos are first divided into frames and the frames are converted to gray scale. The face from these gray scale images is detected and cropped. These cropped images are then saved into a folder and used further[4].

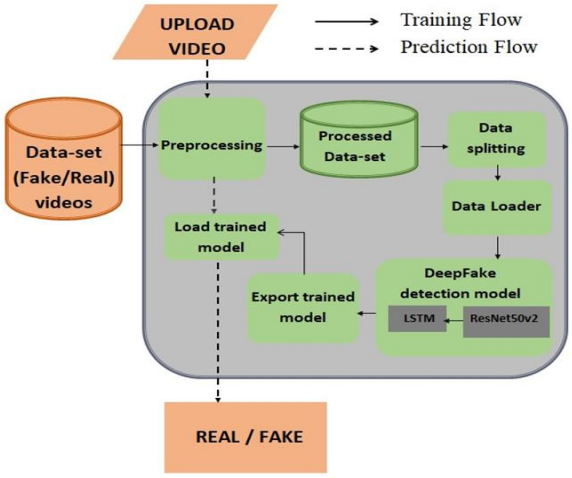

Fig. 4 System Architecture

VI. RESULTS

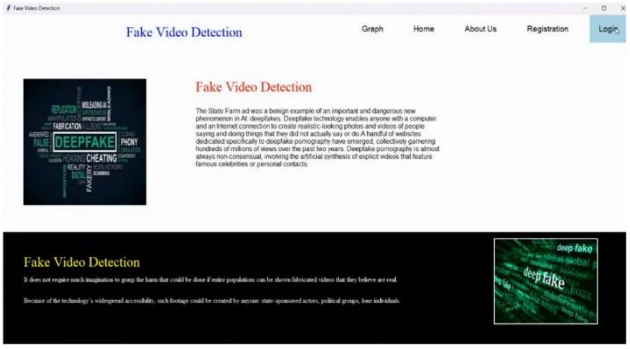

Fig.5.Home Page

Fig. 6. Login Page

Fig. 7 Video Upload Page

Fig. 8 Result (Real Video)

Fig.9. Result (Fake Video)

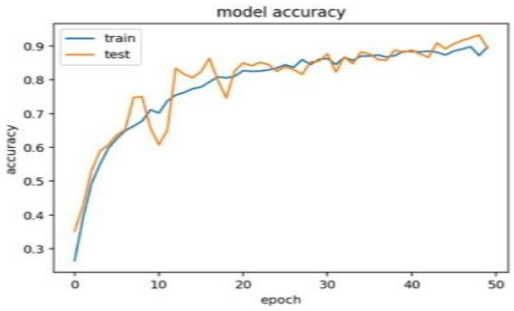

Fig.10. Model Accuracy

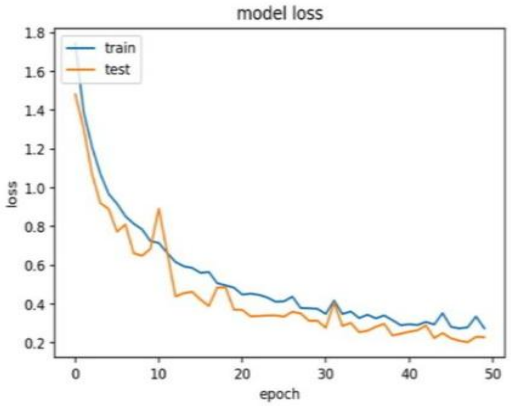

Fig.11. Model Loss

Fig.10. illustrates the accuracy progression of the deepfake detection model over 50 epochs for both training and testing datasets. The blue line represents training accuracy, while the orange line represents testing accuracy. Initially, both training and testing accuracy increase rapidly, demonstrating effective learning. Around 10 epochs, the test accuracy experiences fluctuations but remains closely aligned with the training accuracy. By epoch 50, both training and test accuracy exceed 90%, indicating that the model generalizes well to unseen data. High Generalization, The close alignment of training and validation accuracy suggests minimal overfitting.

$$Accuracy = TP+TN/TP+TN+FP+FN$$

Fig.11. presents the loss curves for both training and testing datasets over 50 epochs. The blue line represents training loss, while the orange line represents testing loss. Initially, the loss is high (~1.6) but decreases steadily, indicating proper learning. The test loss fluctuates slightly in the early epochs but follows the training loss closely. By epoch 30, the loss

stabilizes, with the test loss remaining slightly lower than the training loss, indicating that the model does not overfit. At epoch 50, both losses are below 0.3, confirming effective model optimization. - Stable Loss Curve, the decreasing and stabilizing loss indicates that the model converges effectively. Small variations in the test curves may result from dataset complexity or challenging deepfake patterns. At epoch 50, the model achieves high accuracy and low loss, proving its effectiveness in deepfake detection

$$Loss(\ y, \hat{y}\ ) = -\sum_{i=1}^{k} \quad yilog(\hat{y}i)$$

Table 1. Accuracy and Loss Metrics for Training and Testing

Metric	Training set	Testing set
Accuracy	~90%	~92%
Loss	~0.25%	~0.2%

List of parameters	Face-Forensic ++	DeepFake Detection Challenge Dataset	Celeb-DF	Hybrid Dataset (FF,DFDC ,Celeb-DF, self-created)
Learning rate	0.001	0.0005	0.0001	0.0001
Batch size	32	64	128	4
No of Epoch	100	50	200	20
Dropout rate	0.5	0.2	0.3	0.4
Weight decay	0.01	0.001	0.0001	0.003
Accuracy	91.21 %	66.26 %	79.49 %	95.83 %

Table 2. Comparative analysis of different datasets and hybrid dataset with face feature extraction

The evaluation of the proposed deepfake detection model, based on a hybrid architecture combining ResNet50V2 and LSTM, demonstrates highly effective performance in real-time video manipulation detection. The model achieved an accuracy of approximately 90% on the training dataset and 92% on the testing dataset, with a training loss of 0.25 and a testing loss of 0.2.These results highlight the model's strong ability to extract spatial features via ResNet50v2 and capture temporal dependencies across video frames using LSTM networks.The proposed method proves highly effective in detecting manipulations even under varied real-world scenarios, indicating its robustness against disinformation spread through deepfakes.Key hyperparameters influencing the model's performance included:

- Learning Rate: Controlled step size during optimization.
- Batch Size: Number of samples processed per iteration.
- Number of Epochs: Number of full training cycles.
- Dropout Rate: Applied to prevent overfitting.
- Weight Decay: Regularization to enhance generalization.
- Number of Hidden Layers and Neurons: Critical for LSTM network depth and capacity.

Comparative evaluation against standard datasets such as FaceForensics++, DFDC, and Celeb-DF shows that the hybrid model achieves higher accuracy and lower loss, supporting its capability to generalize across different types of manipulated videos. The use of a ResNet50v2 backbone significantly improves feature extraction compared to conventional CNN architectures, resulting in more reliable deepfake detection. Overall, the experimental results confirm that the ResNet50v2 + LSTM hybrid model is a promising approach to combatting disinformation spread via deepfakes.

Sr. No	Model	Architecture	Accuracy
1	Hybrid CNN-LSTM [1]	CNN+LSTM	~90 %
2	XceptionNet [2]	CNN	~85 %
3	EfficientNet [2]	CNN+Lightweight	~88 %
4	DFFMD CNN Model [3]	Deep CNN	~86-88%
5	Simple CNN [4]	Basic CNN	~82%
6	Dense CNN [5]	DenseNet CNN	~87%
7	Dynamic Difference Learning [7]	CNN+Spacio-temporal	~90-91%
8	MRE-Net [8]	Multi-Rate CNN	~91%
9	Long-Distance Attention [10]	Attention-based model	~91%
10	Hybrid ResNet50v2+LSTM	ResNet50v2+LSTM	92 %

Table 3. Accuracy Comparison of Various Deepfake Detection Approaches and the Proposed Model

Table represents a comparative analysis of various deepfake detection models reported in recent literature, highlighting their architectures and achieved accuracies. The proposed hybrid model based on ResNet50v2 and LSTM demonstrates competitive performance, achieving 92% accuracy, and shows significant improvement over traditional CNN-based methods.

VII. CONCLUSION AND FUTURE WORK

In conclusion, this project addresses the urgent need for a reliable deepfake detection system to combat the risks of AI-manipulated videos, including privacy breaches and misinformation [1]. By leveraging advanced techniques such as ResNet50v2 and LSTM, the system effectively distinguishes real content from deepfakes by analyzing both spatial and temporal inconsistencies [4].

The findings contribute to strengthening digital media security, empowering platforms and users to counter digital deception while maintaining public trust [6]. The integration of a deep learning-based detection framework enhances accuracy, making it a valuable tool for identifying deepfake content across various digital platforms [7].

Future work will focus on improving the system to analyze videos with audio, allowing for the detection of deepfakes that manipulate both visual and auditory elements [9]. Additionally, efforts will be directed toward optimizing the model for real-time detection, ensuring quicker responses to emerging threats in social media, digital forensics, and multimedia security applications [3].

REFERENCE

[1] OmarAlfarouk Hadi Hasan Al-Dulaimi 1,2, * and Sefer Kurnaz 1, “A Hybrid CNN-LSTM Approach for Precision Deepfake Image Detection Based on Transfer Learning”, Electronics 2024, 13, 1662. https://doi.org/10.3390/ electronics13091662, April 2024.

[2] Asad Malik, Minoru Kuribayashi, Sani M. Abdullahi, Ahmad Neyaz Khan, ‘‘Deepfake Detection For Human Face Images And Videos: A Survey”, IEEE ACCESS Vol. 10, February 2022.

[3] Norah M. Alnaim , Zaynab M. Almutairi, Manal S. Alsuwat , Hana H. Alalawi , Aljowhra Alshobaili, Fayadh S. Alenezi , ”DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era With Deepfake Detection Algorithms”, IEEE ACCESS Vol. 11, February 2023.

[4] Aarti Karandikar, Vedita Deshpande, Sanjana Singh, Sayali Nagbhidkar, Saurabh Agrawal, “Deepfake Video Detection Using Convolutional Neural Network”, International Journal of Advanced Trends in Computer Science and Engineering, Volume 9 No.2, March -April 2020 .

[5] Yogesh Patel, Sudeep Tanwar, Pronaya Bhattacharya, Rajesh Gupta, Turki M. Alsuwian, Innocent Ewean Davison, Thokozile F. Mazibuko, “An Improved Dense CNN Architecture for Deepfake Image Detection”, IEEE ACCESS Vol. 4, 2016.

[6] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, ‘‘Deep learning for deepfakes creation and detection: A survey’’, arXiv:1909.11573, 2019.

[7] Qilin Yin, Wei Lu, Member, Bin Li, Senior Member and Jiwu Huang,"Dynamic Difference Learning with Spatio–Temporal Correlation for Deepfake Video Detection",IEEE Transactions on Information Forensics and Security, VOL. 18, 2023.

[8] Guilin Pang, Baopeng Zhang, Zhu Teng, Zige Qi, and Jianping Fan, "MRE-Net: Multi-Rate Excitation Network for Deepfake Video Detection", IEEE Transactions on Circuits and Systems for Video Technology, VOL. 33, NO. 8, AUGUST 2023.

[9] Haya R. Hasan and Khaled Salah, "Combating Deepfake Videos Using Blockchain and Smart Contracts", Digital Object Identifier 10.1109/ACCESS.2019.2905689,April 12, 2019.

[10] Wei Lu , Member, IEEE, Lingyi Liu, Bolin Zhang , Junwei Luo, Xianfeng Zhao , Senior Member, IEEE, Yicong Zhou , Senior Member, IEEE, and Jiwu Huang, " Detection of Deepfake Videos Using Long-Distance Attention", IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING, SYSTEMS,2162-237X © 2023 IEEE.

[11] https://www.kaggle.com/datasets/reubensuju/celeb-df-v2