Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

Deepfake Video Content Detection Using ResNext-101 and LSTM

Kallyan Singha
Dept. of Computer science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Biswarghya Biswas
Dept. of Computer science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Debarati Karmakar
Dept. of Computer science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Dr. Chandra Das
Dept. of Computer science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Abhick Ghosh
Dept. of Computer science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Dr. Shilpi Bose
Dept. of Computer science and Engineering
Netaji Subhash Engineering College
Kolkata, India

***ABSTRACT* — Problem: The Video manipulation techniques called deepfakes that alter people's facial identities have become more advanced and faster as artificial intelligence (AI) and cloud computing have progressed. Finding these deep fakes is a difficult task. There are numerous instances in recent history of fakes being used as influencer to incite political unrest, stage terrorist acts, produce revenge porn, and extort people and many more. Thus, stopping the widespread proliferation of fakes on social media platforms requires the ability to recognize them. Here, we propose a deepfake detection model based on ResNext-101 and LSTM which can detect any type of fake video on the web. The model has been trained with DFDC, FF++, and CELEBDF datasets and tested accordingly to show the efficiency of the model. A general application along with a chrome extension has also been prepared for effective utilization of the model.**

I. INTRODUCTION

Deepfakes are media, often videos, altered using artificial intelligence (AI) technologies. This technique, based on deep learning involves overlaying an actor's performance onto an image or video of a different person, making it seem as though the target individual is performing the actions of the actor. The development of deepfakes has been driven by recent advancements in AI and machine learning, resulting in highly realistic outputs that are almost indistinguishable from genuine footage to the human eye as shown in Fig. 1.

Deepfakes are often used to spread false information or for harmful purposes. They can be crafted to harass, intimidate, or discredit individuals, causing confusion and spreading misinformation on important issues. The impact of this technology is profound, as it can fabricate scenarios like politicians giving speeches they never made, alter historical footage, or place celebrities in inappropriate contexts [1]. By altering faces and voices, deepfakes can create realistic but fake statements, such as a CEO announcing a company's bankruptcy or a data breach. Malicious actors might use these fabricated media files for blackmail, threatening to release them to the public or social media.[2] Therefore, developing robust algorithms to distinguish authentic content from fake content is crucial.

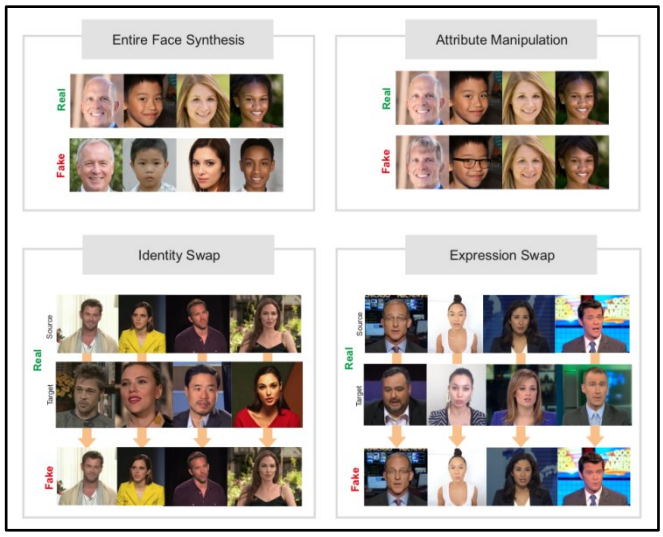

Fig. 1. Some deep fake manipulated frames of videos

The detection of deepfakes is important due to their potential for misuse and the negative consequences they entail. Deepfakes leverage artificial intelligence (AI) and machine learning to fabricate realistic but entirely synthetic content, often in the form of videos or other media formats. These synthetic materials can be wielded to manipulate information, propagate misinformation, and deceive individuals [3][4].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

Objective: Our project aims to uncover the distorted truths behind deepfakes. We strive to minimize the abuse and misinformation affecting people on the internet. Our project will identify and classify videos as either deepfakes or authentic. We will also provide an easy-to-use extension to help users determine if a video is real or fake.

Statement of Scope: While numerous tools exist for creating deepfakes, there are very few tools available for detecting them [5]. Our approach to deepfake detection will significantly help prevent the spread of deepfakes across the internet. Major platforms like WhatsApp and Facebook could integrate our project for easy pre-detection of deepfakes before they are shared with other users.

II. LITERATURE SURVEY

Deepfake video is becoming more and more prevalent, posing a serious threat to democracy, justice, and public trust. This has prompted an increase in the demand for video analysis, detection, and intervention.

Deepfake detection techniques for videos have evolved significantly since 2018 [6]. Early methods concentrated on facial feature analysis as shown in Fig. 2, scrutinizing expressions and movements [7]. As deepfake technology progressed, researchers embraced advanced machine learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to discern visual and temporal patterns. Deepfake detection challenges played a crucial role, providing datasets for model refinement [8].

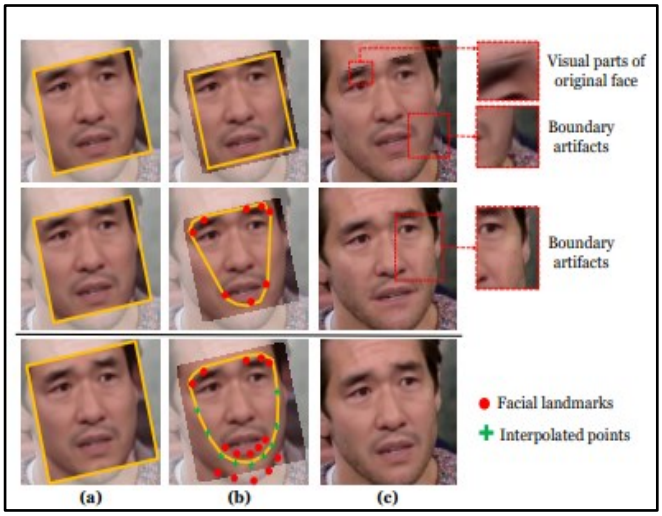

Fig. 2. Various aspects of the frame during Deepfake Detection

Artifact analysis became prominent, focusing on subtle visual distortions introduced during the deepfake generation process. Some approaches incorporated optical flow analysis for motion-based detection.

Generative Adversarial Networks (GANs), central to deepfake creation, posed both a challenge and opportunity for detection, leveraging adversarial relationships between genuine and manipulated content. Recent efforts emphasize explainability and interpretability in detection methods, aiming to provide transparent insights into flagged deepfakes [9].

Ongoing research underscores the need for adaptive and resilient detection techniques in response to evolving deepfake technology as exampled in Fig. 3. With a growing emphasis on trustworthiness, researchers explore methods to enhance the interpretability of detection systems, ensuring a robust defence against the continual evolution of video-based deepfake threats.

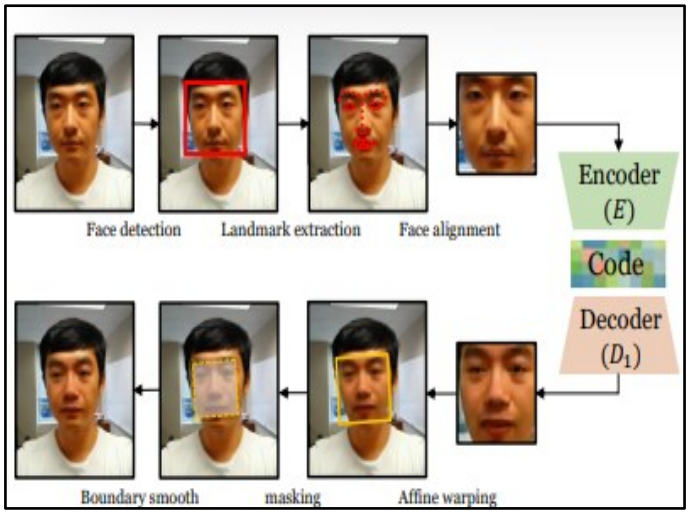

Fig. 3. Process of Deepfake Detection

Researchers have been working on various ways of deepfake detection, as it has become a major threat in this modern time [10][11]. There has been an exponential growth in the number of works done in this field in the last few years, aiming to provide better efficiency than before. A growth chart has been shown Fig. 4.

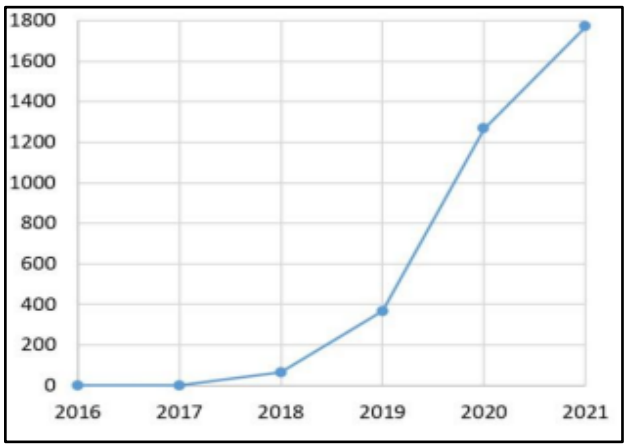

Fig. 4. Graph showing the count of papers published from 2016-2021 based on deep fake detection project

"Deepfake Video Detection using the Ensemble of Neural Networks" (2020): This research employs a combination of neural Xception-type networks for deepfake video detection. It emphasizes the use of artificial intelligence solutions and was awarded a bronze model with an optimal 7% false positive rate for deepfake videos. The study looks into addressing audio content for enhanced forgery detection in future work [12-14].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

"Exposing AI-Created Fake Videos by Detecting Eye Blinking" (2020): Focused on detecting AI-generated fake videos, this study analyzes eye blink patterns. It not only improves the quality and efficiency of recognizing production videos with fake content but also outperforms
CNN and EAR methods, achieving a superior Area Under ROC (AUC) of 0.99. The research aims to explore other physiological signals neglected in artificial intelligence [15].

"An Audio-Visual Deepfake Detection Method using Affective Cues" (2020): This research presents a multimodal approach by extracting and analyzing similarities between audio and visual effects within the same video. Using both audio (speech) and video (face) methods simultaneously, the study achieves an AUC of 84.4% on DFDC and 96.6% on DF-TIMIT databases. Future work involves expanding the approach for audio forgery detection [16, 17].

"MesoNet: A Compact Facial Video Forgery Detection Network" (2018): In the study titled "MesoNet," researchers tackle the task of identifying facial video forgeries while maintaining a low computational burden [18]. By incorporating deepfake and face2face videos, the research achieves impressive accuracy rates of 98% for deepfake videos and 95% for face2face videos. The research prioritizes enhancing accuracy, particularly concerning facial features such as eyes and mouth.

"How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals" (2020): This ground breaking research is the first to incorporate biological signals for deepfake source detection. Conducting an in-depth study of biological signals, the research achieves an impressive 93.39% accuracy on the FaceForensics++ database [19]. Future work includes creating a new authentic ground database (PPG) and exploring diverse spatiotemporal signals.

"Fakecatcher Detection of synthetic portrait videos using biological signals" (2020): Utilizing the Butterworth filter and Welch method, this research addresses threat security through the detection of synthetic portrait videos. Achieving a high paired separation accuracy of 99.39% in Face Forensics, the study aims to discover formulations for other spatiotemporal signals that can be faithfully extracted from original videos [20].

"Video face manipulation detection through an ensemble of CNNs"(2020): This research enhances face manipulation detection by increasing the attention span for simultaneous network learning capabilities. The study successfully tackles face manipulations and scores in the top 3% on a public dataset. Emphasizing the analysis of multiple frames for detection, the research contributes to robust video face manipulation detection [21].

"Detecting both Machine and Human created Fake face image in the wild" (2018): The focus of this research is on detecting both machine and human-created fake face images in real-world scenarios without accessing metadata. Achieving a commendable 74.9% AUROC dataset score, the study plans to expand its work to encompass other datasets in future research [22].

"DeepFakes: A New Threat to Face Recognition?" (2018): Addressing the emerging threat of DeepFakes, this research works on both low-quality and high-quality videos. It provides the first publicly available database of 650 videos and achieves an accuracy of 85.64% for low-quality videos and 95% for high-quality videos. Future work aims to achieve less than an 8.97% error rate [23].

"Deep Learning Based Computer-Generated Face Identification" (2017): Focused on detecting fake images generated by computers, this research deals with complex duplicate images. With an impressive 98% accuracy, the study suggests future work involving the combination of deep learning-based computer-generated face identification with other hidden layers for improved results [24].

III. PROPOSED WORKS

The proposed project aims to tackle the escalating concern of DeepFake videos proliferating on the internet by developing a Chrome (or Chromium-based) extension and a DeepFake classification model for real-time detection. The objective is to create a web-based plugin or extension seamlessly integrated with popular video and social media platforms such as YouTube, Facebook, and Instagram. This extension will empower users to detect fake videos and verify their authenticity before sharing them across various applications like WhatsApp.
Initially, the video undergoes pre-processing as shown in Fig 5 wherein it is divided into frames. Subsequently, face recognition is conducted, followed by the trimming of frames containing the detected faces. To maintain consistency within the total frame count, the mean of the dataset video is calculated, and a new dataset comprising cropped faces is generated, ensuring the frame count matches the calculated mean [20][21].
The crux of the project involves a pre-trained deepfake detection model utilizing the ResNext101_32X8D variant, which has 101 layers and a cardinality of 32 with a dimension of 8, providing robust feature representations. The final pooling layer of the ResNeXt model outputs 2048-dimensional feature vectors.
The extracted 2048-dimensional feature vectors are then fed into an LSTM network. The LSTM network consists of a single LSTM layer with 2048 hidden units and a dropout rate of 0.4 to prevent overfitting. This architecture is designed to capture temporal dependencies by comparing the frame at time t with frames at previous time steps **(t-n)**, where n represents the number of frames before **t**.

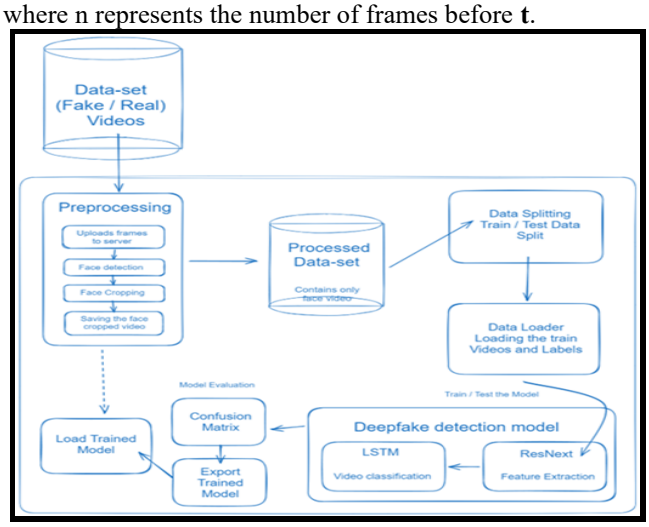

Fig. 5. Video Classification Model Architecture

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

To enable adaptive learning rate optimization, we use the Adam optimizer with the model parameters. The learning rate is set to 1e-5 (0.00001) to effectively reach a better global minimum during gradient descent. Additionally, a weight decay of 1e-5 is applied to regularize the model.
For this classification problem, we use the cross-entropy loss function to calculate the loss.

On the product side, a backend server with API endpoints will be established to process captured 10s video sent by the extension and detect its authenticity. The Chrome extension's user-friendly interface will include a content script for video capture. Secure communication with the backend through HTTPS as shown in Fig. 6 will be implemented to uphold privacy and security standards. We have used celery task queue for handling prediction request more efficiently without blocking the main request thread.

To enhance usability, accuracy, and reliability, the system will specialize in identifying various types of DeepFake, such as retrenchment DeepFake and replacement DeepFake. Rigorous testing and optimization will precede the extension's deployment on the Chrome Web Store.

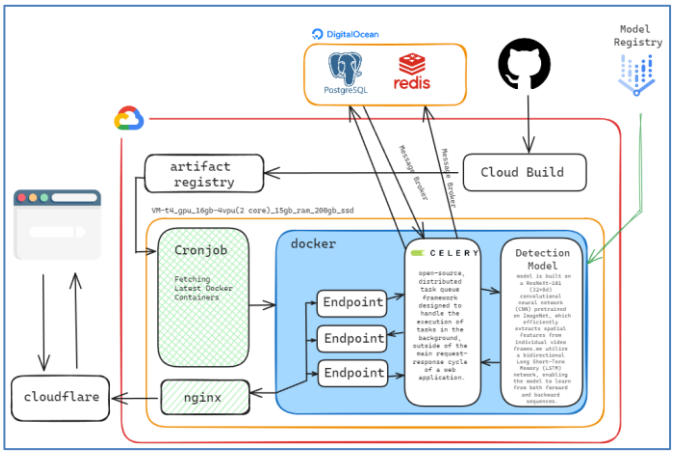

Fig. 6. Backend Server Architecture

Continuous feedback loops will facilitate ongoing updates and improvements, enabling users to report issues and ensuring compliance with platform terms of service, user privacy, and legal considerations.
Ultimately, the project aims to deliver a robust video detection system, classifying videos as either authentic or DeepFake, thereby fostering a safer online environment.

Algorithms /Models :
1) ResNeXt:

ResNeXt, an abbreviation for Residual Networks with an extension, represents a development of the ResNeXt architecture, aiming to boost the effectiveness and efficiency of deep neural networks. Initially presented by researchers at Facebook AI Research, ResNeXt introduces the innovative notion of cardinality, intended to enhance feature representation [25].

Cardinality in ResNeXt:
The cardinality parameter stands out as a crucial aspect of ResNeXt architecture. It denotes the number of parallel pathways integrated within a residual block. Diverging from conventional ResNet blocks that follow a single path, ResNeXt permits the incorporation of multiple pathways to capture varied features [24]. This parallelism is accomplished through grouped convolutions, facilitating the network to acquire more comprehensive and intricate representations.

Formula for Residual Blocks in ResNeXt:
In ResNeXt, the output of a residual block is calculated through the following formula:
Output = Fused path(input) + Residual path(input)
Here's the breakdown:
Fused path: This represents the outcome of applying a 1x1 convolution to the input.
Residual path: This denotes the result of employing grouped convolutions, wherein the output of each group is concatenated before being added to the input [20].

$$\text{Residual path}(\text{input}) = \sum_{i=1}^{C} \text{Group}_i(\text{input})$$

$$\text{Group}_i(\text{input}) = \text{Conv}_{\frac{1}{C}}(\text{input})$$

Here, $\text{Conv}_{\frac{1}{C}}$ denotes a 3×3 grouped convolution with $\frac{1}{C}$ filters in each group.

Application of ResNeXt:
ResNeXt finds application across diverse computer vision tasks, spanning from image classification and object detection to image segmentation. Its scalability and efficiency render it apt for deep neural networks tasked with capturing a broad spectrum of complex features. [26] ResNeXt is commonly employed by researchers and practitioners striving for cutting-edge performance in visual recognition tasks. Adjusting the cardinality parameter to suit the task's specific demands and the computational resources at hand holds paramount importance.

2) LSTM for Video Classification:
Long Short-Term Memory (LSTM) networks belong to the family of recurrent neural networks (RNNs) and are crafted to tackle the vanishing gradient issue while grasping prolonged dependencies in sequential data. In the realm of video classification, LSTMs showcase prowess in comprehending temporal connections within a succession of frames, rendering them ideal for endeavors such as action recognition and scene comprehension in videos.[26]

The formula for LSTM:
The LSTM cell has three gates: an input gate (i), a forget gate (f), and an output gate (o). The LSTM updates its internal state (c_t) and produces an output (h_t) at each time step t. The key equations governing the LSTM are as follows:

- Input gate: $i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi})$
- Forget gate: $f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf})$
- Cell update: $\tilde{c}_t = \tanh(W_{ic}x_t + b_{ic} + W_{hc}h_{t-1} + b_{hc})$
- Internal state update: $c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t$
- Output gate: $o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho})$
- Output: $h_t = o_t \cdot \tanh(c_t)$

These equations govern how information is input, forgotten, and updated over time, allowing LSTMs to capture sequential patterns.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

Application of LSTM for Video Classification:
LSTMs play a pivotal role in video classification by processing sequences of video frames as input. In this method, each frame is sequentially fed into the LSTM network, and the resultant final output is utilized for classification purposes [23]. LSTMs are particularly effective in capturing temporal dependencies, which are crucial for identifying actions or patterns unfolding across multiple frames.
In video classification tasks, practitioners and researchers frequently leverage LSTMs in combination with various other neural network architectures. For instance, 3D convolutional networks are commonly employed alongside LSTMs to directly capture spatial and temporal features from video frames, thereby enhancing overall performance [27].
Fine-tuning hyperparameters such as the number of LSTM units, learning rates, and sequence lengths are imperative for optimizing the LSTM network's performance in specific video classification tasks.

IV. RESULT

Our deepfake detection model is trained on 8000 videos from below pre-processed video datasets

1. Celeb-DF Fake processed videos
2. Celeb-DF Real processed videos
3. FaceForensics++ Real and fake processed videos
4. DFDC Fake processed videos
5. DFDC Real processed videos
6. Kaggle DFD Challenge dataset videos

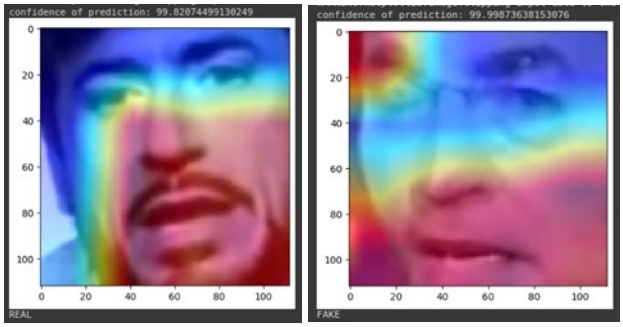

Fig. 7. Model Results

The user workflow commences with the activation of a browser plugin designed for initiating video detection. This plugin assumes responsibility for identifying the video and subsequently dispatches the video segment to the backend for thorough analysis. Evaluations indicate that the selected model achieves an accuracy rate of approximately 94.28%. as shown Fig 7.

Table 1: Model Performance with different video frames

Model Name	No of Frames	Accuracy
model_89_acc.pt	40	89.56
model_90_acc.pt	60	90.68
model_91_acc.pt	80	91.23
model_94_acc.pt	100	94.28

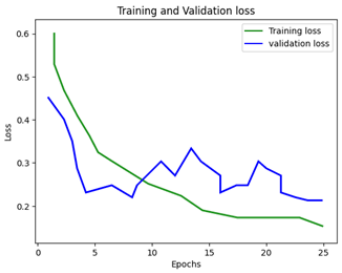

Fig. 8. Training and Validation Loss

Fig. 9. Training and Validation Accuracy

Table 2: Comparison with Other Models

Model Name	Description	Accuracy (%)
EfficientNet-B5	The novel scaling technique proposed by Efficient Net, a mobile-friendly pure convolutional model (ConvNet), uses a straightforward yet incredibly powerful compound coefficient to scale all three dimensions of depth, width, and resolution equally.[18]	**87.48%**
DenseNet-121	A variation on convolutional neural networks (CNNs) that employs 121 layers is called Densenet-121.DenseNet-121 comprises the following layers, in summary: • 1 7x7. Convolution • 58 3x3 Convolution • 61 1x1 Convolution • 4 AvgPool • 1 Layer with Full Connectivity	**90.7%**
Optical Flow+ VGG-16	Optical flow and VGG-16 together present an intriguing method for deepfake detection. Optical flow provides dynamic context by predicting pixel movement and capturing motion between video frames. A deep convolutional neural network called VGG-16 is very good at extracting spatial data from pictures.[27][28]	**91.21%**
Proposed Model (ResNeXt-101+LSTM)	Our model combines Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to analyze both spatial and temporal features of video data. Using the pre-trained resnext101_32x8d model, we extract 2048-dimensional feature vectors from video frames. These features are then processed by a Long Short-Term Memory (LSTM) network with 2048 hidden units and a dropout rate of 0.4 to capture temporal dependencies. The model is fine-tuned with the Adam optimizer, utilizing a learning rate of 1e-5 and a weight decay of 1e-5. Cross-entropy loss is used for training, aiming to achieve robust and accurate deepfake detection.[29][30]	**94.28%**

The backend meticulously scrutinizes each frame extracted from the designated video. Following the comprehensive examination of all frames, a notification appears in the user's browser window if the video is determined to be fake or AI-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

generated, presenting the findings of the analysis. This notification aims to inform the user whether the video is authentic or classified as a DeepFake. The seamless integration of user interaction, backend processing, and result presentation ensures a straightforward and informative experience in discerning the legitimacy of encountered videos online. The model performance, the validation loss and accuracy have been shown in Table 1, Fig. 8 and 9 respectively. Table 2 shows the performance comparison with the existing methods. The results show the superiority of the proposed mode.

V. CONCLUSION

The ability of deep-fake technology to fool vast numbers of people has made it a serious worry. Even though not all deepfake content is harmful, considering the possible risks presented by some deepfake materials, it is clear that trustworthy detection techniques are necessary. In several domains, including virtual reality, robotics, advanced media, and education, the dangers of video face alteration are well known. Deepfake technology has the potential to be innovative, but it also has drawbacks that could compromise society's norms. Early identification is becoming more and more important in order to prevent potential harm as fake films continue to spread throughout the world. With the use of a variety of feature sets and machine learning and deep learning approaches, researchers are currently investigating methods for categorizing films as real or fraudulent

When it comes to video categorization, convolutional neural networks (CNN) and long short-term memory (LSTM) networks have been shown to be two of the most accurate combination methods. For training and evaluation, researchers have used a variety of datasets that include both real and fake videos. The research makes it clear that the CNN-LSTM fusion routinely yields reliable results and high accuracy when distinguishing between authentic and fraudulent films.This technical development has the potential to have a significant social impact. People who are targeted by deep-fake content can quickly confirm its legitimacy with the help of this technology. People will be better able to recognize and stop the spread of fake materials if they are more vigilant and attentive.

REFERENCES

[1] Amato, G., Bolettieri, P., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: Visione at video browser showdown 2021. In: International Conference on Multimedia Modeling. pp. 473–478. Springer (2021)

[2] Amato, G., Ciampi, L., Falchi, F., Gennaro, C., Messina, N.: Learning pedestrian detection from virtual worlds. In: International Conference on Image Analysis and Processing. pp. 302–312. Springer (2019)

[3] Amerini, I., Galteri, L., Caldelli, R., Del Bimbo, A.: Deepfake video detection through optical flow based cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)

[4] Buslaev, A., Parinov, A., Khvedchenya, E., Iglovikov, V.I., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. ArXiv e-prints (2018)

[5] Chen, C.F., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. arXiv preprint arXiv:2103.14899 (2021)

[6] Choi, Yunjey, e.a.: Stargan: Unified generative adversarial networks for multi-domain image- to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)

[7] Chesney, B., Citron, D.: Deep fakes: A looming challenge for privacy, democracy, and national security. Calif. L. Rev. 107, 1753 (2019)

[8] Ciampi, L., Messina, N., Falchi, F., Gennaro, C., Amato, G.: Virtual to real adaptation of pedestrian detectors. Sensors 20(18), 5250 (2020)

[9] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.:The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020)

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A.,Weissenborn, D., Zhai, X., Unterthiner, T.,Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)

[11] Dufour, N., Gully, A.: Contributing data to deep-fake detection research (2019)

[12] Fagni, T., Falchi, F., Gambini, M., Martella, A., Tesconi, M.: Tweepfake: About detecting deepfake tweets. Plos one 16(5), e0251415 (2021)

[13] Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412 (2020)

[14] Giudice, O., Guarnera, L., Battiato, S.: Fighting deepfakes by detecting gandct anomalies. arXiv preprint arXiv:2101.09781 (2021)

[15] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S.,Courville, A., Bengio, Y.: Generative adversarial networks. In: Advances in neural information processing systems 27 (2014)

[16] Guarnera, L., Giudice, O., Battiato, S.: Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2020)

[17] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A survey on visual transformer. arXiv preprint arXiv:2012.12556 (2020)

[18] Heo, Y.J., Choi, Y.J., Lee, Y.W., Kim, B.G.: Deepfake detection scheme based on vision transformer and distillation. arXiv preprint arXiv:2104.01353 (2021)

[19] Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2889–2898 (2020)

[20] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)

[21] Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. arXiv preprint arXiv:2101.01169 (2021)

[22] Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: International Conference on Machine Learning. pp. 1857–1865. PMLR (2017)

[23] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprintarXiv:1312.6114 (2013)

[24] Korshunov, P., Marcel, S.: Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685 (2018)

[25] Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3207–3216 (2020)

[26] de Lima, O., Franklin, S., Basu, S., Karwoski, B., George, A.: Deepfake detection using spatiotemporal convolutional networks. arXiv preprint arXiv:2006.14749 (2020).

[27] MacAvaney, S., Nardini, F.M., Perego, R., Tonellotto, N., Goharian, N., Frieder, Efficient document re-ranking for transformers by precomputing term representations. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 49–58 (2020)

[28] Messina, N., Amato, G., Carrara, F., Falchi, F., Gennaro, C.: Testing deep neural networks on the same-different task. In: 2019 International Conference on Content-Based Multimedia Indexing (CBMI). pp. 1–6. IEEE (2019)

[29] Messina, N., Amato, G., Carrara, F., Gennaro, C., Falchi, F.: Solving the same different task with convolutional neural networks. Pattern Recognition Letters 143, 75–80 (2021)

[30] Messina, N., Amato, G., Esuli, A., Falchi, F., Gennaro, C., Marchand-Maillet, S.:Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. arXiv preprint arXiv:2008.05231 (2020)