http://doi.org/10.35784/iapgos.6777

received: 01.11.2024 | revised: 14.02.2025 | accepted: 20.03.2025 | available online: 31.03.2025

ROBUST DEEPFAKE DETECTION USING LONG SHORT-TERM MEMORY NETWORKS FOR VIDEO AUTHENTICATION

Ravi Kishan Surapaneni, Hameed Syed, Harshitha Kakarala, Venkata Sai Srikar Yaragudipati
Velagapudi Ramakrishna Siddhartha Engineering College, Department of Computer Science and Engineering, Vijayawada, India

Abstract. *Developments achieved in recent years have propelled techniques for generating and manipulating multimedia content to attain an exceptionally high degree of realism. According toa survey, 25 percent of the videos viewers watch are fake. The increasingly blurred distinction between authentic and synthetic media presents significant security concerns, with the potential for exploitation in various domains. These threats encompass the manipulation of public opinion during electoral processes, perpetration of fraudulent activities, dissemination of disinformation to discredit individuals or entities, and the facilitation of blackmail schemes. Detecting fakes is tricky and difficult for viewers who are watching them, with studies showing that over 70 percent struggle to identify them accurately. To counter this issue, we envision this project whose primary goal is to construct a model that is capable of distinguishing between deepfake and authentic videos. Our proposed model operates at the video level, analyzing entire videos at once to provide a comprehensive assessment. The dataset utilized for training and evaluation is sourced from repositories such as DFDC, FaceForensics++ and Celeb-DF. The dataset sourced from DFDC and Celeb-Df are converted into frames from videos, in this architecture first face recognition tool is used for detecting the faces, followed by ResNext for feature extraction and then LSTM is used to classify the videos.*

Keywords: deepfake detection, deep fake video, feature recognition, deep learning, ResNext, Long Short-Term Memory

SKUTECZNE WYKRYWANIE DEEPFAKE'ÓW PRZY UŻYCIU SIECI LONG SHORT-TERM MEMORY DO AUTENTYKACJI WIDEO

Streszczenie. *W ostatnich latach techniki generowania i manipulowania treściami multimedialnymi osiągnęły wyjątkowo wysoki stopień realizmu. Według badań, 25 procent wideo oglądanych przez widzów jest fałszywych. Coraz bardziej zacierająca się różnica między autentycznymi a syntetycznymi mediami stwarza poważne obawy dotyczące bezpieczeństwa, z możliwością wykorzystania w różnych dziedzinach. Zagrożenia te obejmują manipulowanie opinią publiczną podczas procesów wyborczych, popełnianie oszustw, rozpowszechnianie dezinformacji w celu zdyskredytowania osób lub podmiotów oraz ułatwianie szantażu. Wykrywanie fałszywych wideo jest trudne, a badania pokazują, że ponad 70 procent widzów ma trudności z ich dokładną identyfikacją. Aby przeciwdziałać temu problemowi, opracowaliśmy niniejszy projekt, którego głównym celem jest skonstruowanie modelu zdolnego do odróżniania deepfake'ów od autentycznych informacji wideo. Proponowany model działa na poziomie wideo, analizując całych wideo jednocześnie, zapewnić ich kompleksową ocenę. Zbiór danych wykorzystany do szkolenia i oceny pochodzi z repozytoriów takich jak DFDC, FaceForensics++ i Celeb-DF. Zbiory danych pochodzące z DFDC i Celeb-Df są konwertowane na klatki; w tej architekturze pierwsze narzędzie do rozpoznawania twarzy jest używane do ich wykrywania, następnie ResNext do ekstrakcji cech, a LSTM jest używany do klasyfikacji wideo.*

Słowa kluczowe: wykrywanie deepfake'ów, fałszywe wideo, rozpoznawanie cech, głębokie uczenie, ResNext, Long Short-Term Memory

Introduction

Technology for editing images, videos, and audio is progressing quickly. There's been a lot of new ideas for changing images and sound recordings. Nowadays, you can make really realistic pictures with just a little bit of know-how from guides online. Deepfakes are a type of trick where you swap out one person's face in a video for another's. It's like putting one person's face on another person's body in a video. Deepfake tech has made people worried about whether digital media can be trusted. So, it's become really important to come up with good ways to tell if something is a deepfake or not, to stop false information from spreading and keep trust in media. With smartphones getting better and better and internet access being everywhere, it's easier for people to make and share digital videos on social media and other sites Deepfake videos are fake videos created using special computer programs. These programs use advanced techniques to mix real images or videos with altered sounds or images, making them look real but actually, they're not. The impact of deepfakes on society is significant and multifaceted. Firstly, deepfakes can erode trust and credibility in media and information withthe ability to fabricate realistic videos of individuals sayingor doing things they never actually did, deepfakes can beused to spread misinformation, manipulate public opinion, and damage reputations. This can have serious consequences for individuals, organizations, and even democratic processes like elections. Secondly, deepfakes pose risks to privacy and security. By superimposing someone's face onto explicit or compromising content, deepfakes can be used for harassment, extortion, or other malicious purposes. Detecting these fake videos means carefully looking for mistakes or weird things in them to tell if they're real or fake.

1. Related work

Bonettin and team [1] develop a deep learning model aimed at detecting face manipulation in video frames using an ensemble of convolutional neural networks (CNNs). The approach utilizes the DFDC dataset from Facebook and leverages the EfficientNet family of models, which are designed for automatic scaling of CNNs. This set of architectures is noted for its improved accuracy and efficiency compared to other state-of-the-art CNNs. Additionally, the model incorporates Siamese training strategies to extract more information from the data during the learning process. By combining these techniques, the model enhances its ability to identify manipulated facial content effectively.

Ismai and team [5] present a new deep learning approach fordetecting deepfake videos that uses the YOLO face detector to identify facial regions in video frames. The study utilizesthe Celeb-Df and Face Forensics++ datasets and employs the InceptionResNetV2 convolutional neural network (CNN) to extract important spatial features from these facial images. Themethod aims to identify visual artifacts that help differentiate real videos from deepfakes. These extracted features are then fed into an XGBoost classifier for effective classification of authentic versus manipulated content. By combining CNNs for feature extraction with gradient boosting for classification, this approach improves detection accuracy, though it may still encounter difficulties with advanced deepfake techniques that hide visual artifacts.

M. T. Mamtha and J. Usha [6] introduce a novel deepfake detection methodology by analyzing eye blinking patterns. Published in the International Research Journal of Modernization in Engineering Technology and Science (April 2020), this method leverages insights from medicine, biology, and brain engineering to detect deepfakes, especially focusing on the rapid, repetitive eye blinks often seen in these videos. The system employs Generative Adversarial Networks (GANs) to identify anomalies in blinking behavior, using factors like blink duration, frequency, and intervals to flag suspicious content. The autoencoder-based partitioning of latent features into "real" and "fake" categories enhances its classificationaccuracy.

Preeti and team [8] propose a deepfake detection method utilizing the Deep Convolutional GAN (DCGAN) architecture, presented at the International Conference on Machine Learning and Data Engineering. The model, designed primarily with convolutional layers and without max-pooling layers, employs convolutional strides for downsampling and transposed convolutions for upsampling. To train both the discriminator and generator, the method utilizes the Binary Cross-Entropy (BCE) loss function, which measures the difference between predicted and target labels for real and fake samples. Adam optimization is applied to enhance the training efficiency, making the model well-suited for detecting deepfakes in social media contexts.

Malik and team [10] This methodology starts with frequency-based frame extraction, where videos from FF++ and DFDC are processed to extract frames based on frequency using the OpenCV library. Frames are extracted at a frequency of 1/5th of the key frame fps to reduce data redundancy. Faces are then extracted from these frames using the Viola-Jones algorithm. A CNN is employed for object classification and image recognition, with an architecture that includes Conv2D layers with varying kernel sizes and dense layers with sigmoid activation for classification. The methodology primarily relies on frequency-based frame extraction and grayscale conversion for feature representation, although this simplistic feature extraction approach may overlook complex spatial and temporal patterns present in deepfake videos, potentially limiting the model’s ability to accurately detect sophisticated manipulations. While frequency-based frame extraction may capture certain manipulations effectively, the effectiveness of frequency-based features remains speculative.

C. Zhao and C. Wang team [17] introduce the ISTVT (Interpretable Spatial-Temporal Video Transformer) for deepfake detection in the IEEE Transactions on Information Forensics and Security. This framework uses several datasets, including FaceForensics++ and CelebDF, for training. The ISTVT extracts low-level texture information from video frames using Xception blocks, which effectively capture detailed textural features. The generated feature maps represent distinct patterns in each frame. The model is trained with a binary cross-entropy loss function, suitable for the binary classification of deepfakes. By focusing on both spatial and temporal features, the ISTVT aims to enhance accuracy and interpretability in detecting deepfake videos, although it may still face challengeswith complex manipulations.

Patel and team [12] discuss the rise of deepfake technology and its implications in their article published in IEEE Access. They explain that deepfakes are generated using advanced techniques like Generative Adversarial Networks (GANs) and are detected primarily through Convolutional Neural Networks (CNNs). The work emphasizes the need for improved detection accuracy to identify inconsistencies in deepfake content. The authors also provide an overview of various tools and models used for both generating and detecting deepfakes, highlighting the ongoing challenges in this field.

Tran, V. N., Kwon and team [13] present a meta-learning approach aimed at developing a generalized deepfake detection model that performs well across a variety of unseen domains. Utilizing the FaceForensics++, CelebDF v2, and DFDC datasets, this methodology incorporates two novel loss functions Pair-Attention Loss and Average Center Alignment Loss alongside the traditional softmax loss. These enhancements aim to improve the model’s ability to detect deepfakes effectively, even when faced with diverse manipulation techniques not encountered during training.

2. Proposed model

A. Dataset collection

The data is collected from the CelebDF and Deep Fake Detection Challenge (DFDC) datasets, both of which are available on Kaggle. These datasets are comprehensive collections of videos specifically created for deepfake detection research.

The CelebDF dataset contains a total of 2,550 videos, comprising 1,300 authentic (real) videos and 1,250 manipulated (fake) videos. This balanced distribution of real and fake content facilitates effective training and evaluation of deepfake detection models.

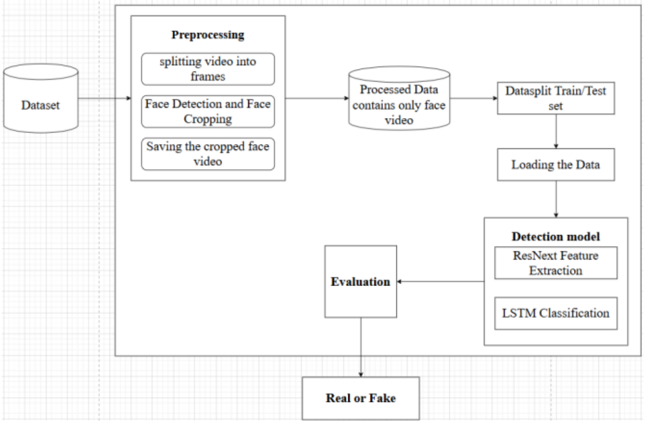

Dataset
Preprocessing
splitting video into frames
Face Detection and Face Cropping
Saving the cropped face video
Processed Data contains only face video
Datasplit Train/Test set
Loading the Data
Detection model
ResNext Feature Extraction
LSTM Classification
Evaluation
Real or Fake

Fig. 1. Proposed method

In addition, the DFDC dataset provides a diverse range of video manipulations, enhancing the robustness and generalizability of our detection approach. Together, these datasets enable us to develop a model capable of accurately distinguishing between authentic and manipulated videos.

B. Data preparation

1. Split Videos into Frames: Splitting the video into frames enhances the flexibility to independently manipulate and analyze each frame, offering significant advantages even for video-level analysis. By treating each frame as a discrete entity, researchers can apply advanced algorithms to extract temporal features and patterns, thereby enriching the understanding of video dynamics.
2. Face Detection: Face detection with the face recognition library streamlines the task of detecting faces within individual frames, reducing the need for laborious manual intervention. It employs a three-stage neural network detector that can recognize faces and facial landmarks, including the mouth, nose, and eyes.
3. Face Cropping: Face cropping involves isolating the identified faces from the surrounding image, concentrating exclusively on the most relevant visual data for subsequent analysis. This approach effectively reduces computational burden by eliminating unnecessary details from the frames.

C. Proposed method

The primary objective of the proposed system architecture isto provide a robust and efficient structure that ensures the deep fake video detection. The Process Flow Diagram, depicted in Fig. 1 visually represents the sequential steps involved in the system’s operation, facilitating a clear understanding of overall process.

The system architecture consits of essential modeules for deepfake video detection utilizing Long Short-Term Memory (LSTM) networks and ResNeXt convolutional neural networks (CNNs) is designed to leverage the strengths of both recurrent and convolutional architectures in a unified framework. The architecture is tailored to effectively capture temporal dependencies within video sequences while efficiently extracting spatial features from individual frames.

First, we start by breaking down the video into individual frames and then look for faces in each frame. If a face is found, the frame is cropped to include only the face, while frames without any faces are discarded. After detecting faces in all frames, they are reassembled to create a new video containing only the segments with faces. For sequential analysis, we input 2048-dimensional feature vectors into the LSTM. Our LSTM setup consists of one layer with 2048 latent dimensions and 2048 hidden layers, including a dropout probability of 0.4, customized for our needs. The LSTM facilitates the temporal analysis of video content by examining frames at differenttime points, and we utilize the Leaky ReLU activation function within our model. A linear layer with 2048 input features and 2 output features aids in learning the average correlation rate between input and output, with an adaptive average pooling layer (output parameter of 1) adjusting the output to an image size of H x W.

For processing sequential frames, a Sequential Layer witha batch size of 4 is employed for training in batches, and a SoftMax layer is used to measure the model's confidence during prediction. To expedite the process, we use the existing ResNext model for feature extraction, specifically the resnext50 32x4d model with 50 layers and dimensions of 32 x4. The 2048-dimensional feature vectors obtained from the final pooling layers of ResNext are subsequently used as input for the sequential LSTM.

D. Training and testing accuracy

Fig. 2 illustrates the training and testing accuracy of our video classification model across training epochs. The x-axis represents the number of epochs, signifying the number of times the model has iterated through the entire training dataset. Each epoch involves processing all the training videos. The y-axis depicts the model's accuracy, expressed as the percentage of videos correctly classified.

Fig. 3 depicts the loss function values for both the training and testing datasets across training epochs. The loss function quantifies the model's performance on a given data point, with lower values indicating better performance. The x-axis represents the training epochs, while the y-axis shows the corresponding loss value.

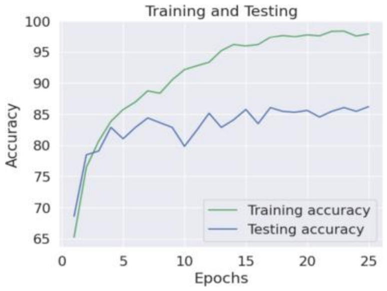


Fig. 2. Training and testing accuracy over epochs

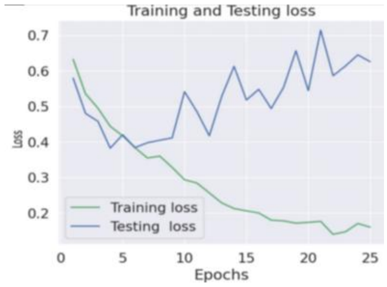


Fig. 3. Training and testing loss over epochs

E. Confusion matrix

The confusion matrix in Fig. 4 provides a visual representation of model's classification performance, allowing for a more detailed understanding of true positive, true negative, false positive, and false negative predictions. Table 1 presents the distribution of actual and predicted real and fake videos used in the evaluation. The values for our confusion matrix are as follows.

Table 1. Distribution of actual real and fake videos with model predictions

	Predicted real	Predicted fake
Actual real	232	91
Actual fake	27	309

From Table 1, we can derive the following metrics:

- True Positive (TP): 232
- False Positive (FP): 91
- False Negative (FN): 27
- True Negative (TN): 309
- Precision: 0.7725
- Recall: 0.9196
- F1 Score: 0.8397

These values highlight the model's effectiveness in distinguishing between real and manipulated videos. The confusion matrix (Fig. 4) is particularly useful for evaluating the effectiveness of the deepfake detection model in identifying real videos and minimizing false positives.

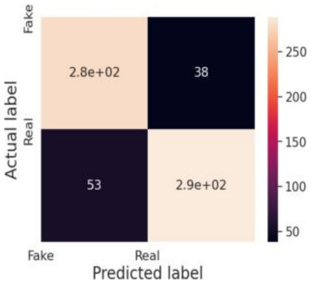


Fig. 4. Confusion matrix of video classification model

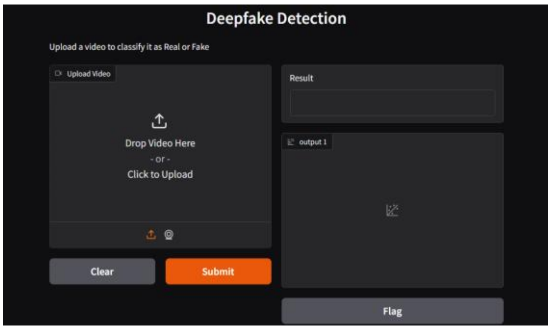
3. Results

A. Interface results

The interface of the deepfake video detection system consists of two main buttons: the first allows users to upload video through "Click to Upload" option, and the second button, click on "Submit", is for uploading the selected file. After selecting the video file and clicking the Submit button, the video is processed, and the output is displayed. Additionally, there is a "Clear" button beside the submit button, which enables the user to clear the input and output and return to the initial page after the output is printed.

B. Output results

Fig. 5 illustrates the deepfake detection system. The input video is first uploaded through the "Click to Upload" option and then "Submit" option. Afterprocessing, the video is split into individual frames, and the system displays whether the video is classified as fake or real.

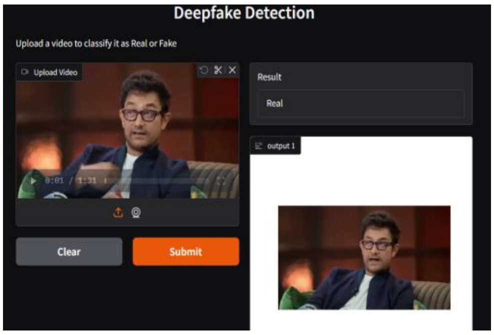


Deepfake Detection

Upload a video to classify it as Real or Fake

Fig. 5. User interface of deepfake video detection system

As shown in Fig. 6, when a real video is uploaded, the system correctly classifies it as Real and displays the corresponding output. Fig. 7 illustrates the system's response to a deepfake video, where the classification result is Fake, along with a corresponding frame from the video.



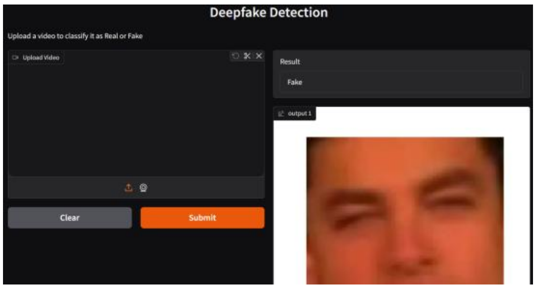
Deepfake Detection

Upload a video to classify it as Real or Fake

Result

Real

Fig. 6. Output for real video



Deepfake Detection

Upload a video to classify it as Real or Fake

Result

Fake

Fig. 7. Output for fake video

4. Conclusion

In conclusion, the analysis presented in this study demonstrates the effectiveness of our video classification model across various training epochs. The comparisons highlight the differences in training and testing accuracy, providing insights into the model's performance and generalization capabilities. The accuracy and loss trends observed suggest that the model is learning effectively, with potential areas for further improvement identified.

Our neural network based approach demonstrates promising capabilities in accurately classifying videos as either deepfake or real. By utilizing a pre-trained ResNeXt CNN model to extract features at the frame level and implementing an LSTM for processing temporal sequences, our approach attainsimpressive levels of accuracy. By processing videos at frame level granularity and analyzing temporal dynamics between frames, our model effectively captures subtle changes indicative of deepfake manipulation.

Looking ahead, several avenues for enhancing and extending our deepfake detection system are identified. Firstly, we aim to expand the scope of our algorithm to encompass voicebased analysis alongside frame-level processing. Incorporating voice-based features and analyzing audio content can provide complementary information for more robust deepfake detection. Secondly, while our current focus is on detecting face deepfakes, extending our algorithm to include voice detection is important. By broadening the detection capabilities, we can better address the evolving landscape of manipulated media content. Additionally, optimizing our model to process videos with specific frame counts, such as the average frame count per video, can further enhance its efficiency and accuracy. Finetuning our model's parameters and training methodologies to accommodate different frame rates will improve its performance across diverse video datasets.

References

[1] Bonettini N. et al.: Video face manipulation detection through ensemble of cnns. 25th international conference on pattern recognition (ICPR). IEEE, 2021.

[2] Cozzolino D. et al.: Id-reveal: Identity-aware deepfake video detection. IEEE/CVF International Conference on Computer Vision, 2021.

[3] Deng L., Suo H., Li D.: Deepfake Video Detection Based on EfficientNet-V2 Network. Computational Intelligence and Neuroscience, 2022, 3441549.

[4] Gu Z. et al.: Spatiotemporal inconsistency learning for deepfake video detection. 29th ACM international conference on multimedia, 2021.

[5] Ismail A. et al.: A new deep learning-based methodology for video deepfake detection using XGBoost. Sensors 21(16), 2021, 5413.

[6] Jung T., Kim S., Kim K.: Deepvision: Deepfakes detection using human eye blinking pattern. IEEE Access 8, 2020, 83144–83154.

[7] Khan S. A., Dai H.: Video transformer for deepfake detection with incremental learning. 29th ACM International Conference on Multimedia, 2021.

[8] Kumar M., Sharma H. K.: A GAN-based model of deepfake detection in social media. Procedia Computer Science 218, 2023, 2153–2162.

[9] Li X. et al.: Sharp multiple instance learning for deepfake video detection. 28th ACM International Conference on Multimedia, 2020.

[10] Malik M. H. et al.: Frequency-based deep-fake video detection using deep learning methods. Journal of Computing & Biomedical Informatics 4(02), 2023, 41–48.

[11] Mittal T. et al.: Emotions don't lie: An audio-visual deepfake detection method using affective cues. 28th ACM International Conference on Multimedia, 2020.

[12] Patel Y. et al.: Deepfake generation and detection: Case study and challenges. IEEE Access 11, 2023, 143296–143323.

[13] Tran V.-N. et al.: Generalization of forgery detection with meta deepfake detection model. IEEE Access 11, 2022, 535–546.

[14] Vashishtha S. et al.: Optifake: optical flow extraction for deepfake detection using ensemble learning technique. Multimedia Tools and Applications 83(32), 2024, 77509–77527

[15] Wodajo D., Atnafu S.: Deepfake video detection using convolutional vision transformer. arXiv 2102.11126, 2021.

[16] Zhang L. et al.: Unsupervised learning-based framework for deepfake video detection. IEEE Transactions on Multimedia 25, 2022, 4785–4799.

[17] Zhao C. et al.: ISTVT: interpretable spatial-temporal video transformer for deepfake detection. IEEE Transactions on Information Forensics and Security 18, 2023, 1335–1348.

[18] Zi B. et al.: Wilddeepfake: A challenging real-world dataset for deepfake detection. 28th ACM International Conference on Multimedia, 2020.

M.Tech. Ravi Kishan Surapaneni
e-mail: suraki@vrsiddhartha.ac.in

He is an associate professor in the Department of Computer Science and Engineering at Velagapudi Ramakrishna Siddhartha Engineering College (VRSEC) now known as Siddhartha Academy of Higher Education in Vijayawada, Andhra Pradesh, India. He earned his M.Tech. in Computer Science and Engineering from Jawaharlal Nehru Technological University, Kakinada, in 2007, and a B.Tech. in the same field from Madras University in 1997. Since 1999, he has been a faculty member at VRSEC, initially serving as a lecturer before becoming an associate professor in 2007. His research interests include data analytics and web designing.

https://orcid.org/0000-0001-5145-2574

Eng. Hameed Syed
e-mail: hameedsd95@gmail.com

Syed Hameed is pursuing a Bachelor of Technology in Computer Science and Engineering at Siddhartha Academy of Higher Education, Vijayawada, India.
His research interests include Machine Learning, and Web Development.

https://orcid.org/0009-0009-7979-1864

Eng. Harshitha Kakarala
e-mail: harshithakakarala6@gmail.com

Harshitha Kakarala is pursuing a Bachelor of Technology in Computer Science and Engineering at Siddhartha Academy of Higher Education, Vijayawada, India.
His research interests include Machine Learning, and App Development.

https://orcid.org/0009-0008-2320-3560

Eng. Venkata Sai Srikar Yaragudipati
e-mail: ypatisrikar@gmail.com

Venkata Sai Srikar Yaragudipati is pursuing a Bachelor of Computer Science and Engineering at Siddhartha Academy of Higher Education, Vijayawada, India. His research interests encompass Internet of Things, Machine Learning, Deep Learning, Computer Vision with a particular focus on specific areas such as, deepfake detection.

https://orcid.org/0009-0005-5814-0115