Development of Ai/Ml Based Solution for Detection of Face Swap Based Deepfake Videos

Chand Riyaj Mulani[1], Shweta Dattatray Dugane[2], Swarup Jitendra Jadhav[3], Siddharth Madhukar Jadhav[4], Aditya Mahendra Kamble[5] , Pradeep Shinde[6]

[1]Developer/Programmer, Information Technology, SOET, DYPU, Ambi, Pune
[2]Project Leader, Information Technology, SOET, DYPU, Ambi, Pune
[3]Research Analyst, Information Technology, SOET, DYPU, Ambi, Pune
[4]Documentation Manager, Information Technology, SOET, DYPU, Ambi, Pune
[5]Tester/Quality Analyst, Information Technology, SOET, DYPU, Ambi, Pune
[6]Project Guide, Information Technology, SOET, DYPU, Ambi, Pune

Abstract:

In recent months, free deep learning-based software tools have facilitated the creation of credible face exchanges in videos that leave few traces of manipulation, known as ”DeepFake” (DF) videos. While digital video manipulation has been demonstrated for several decades through the use of visual effects, recent advances in deep learning have significantly increased the realism of fake content and the ease with which it can be created. These AI-synthesized media, popularly referred to as DF, have become widely accessible.

Creating DF using artificially intelligent tools is a simple task. However, detecting these DF videos presents a major challenge, as training an algorithm to spot them is not straightforward. We have taken a step forward in detecting DF using Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

Our system utilizes a CNN to extract features at the frame level. These features are then used to train an RNN, which learns to classify whether a video has been manipulated or not. Additionally, it detects temporal inconsistencies between frames introduced by DF creation tools. Our approach consists of a multi-step pipeline. Initially, the input video is processed by extracting frames, followed by facial detection and cropping. The processed frames are then passed to a ResNext CNN model, which extracts deep feature representations of each frame. These feature vectors are subsequently fed into an LSTM-based RNN that captures temporal relationships and inconsistencies between frames. This combination allows our system to effectively differentiate real from manipulated content.

To validate our model’s effectiveness, we utilize a diverse dataset comprising videos from sources such as YouTube, FaceForensics++, and the DeepFake Detection Challenge dataset. The dataset is split into 70.

Our expected results are evaluated against a large dataset of fake videos collected from standard sources. We demonstrate how our system achieves competitive results in this task using a simple yet effective architecture. The implementation of this system could have significant implications for combating the spread of misinformation, securing digital media integrity, and enhancing trust in online video content.

Keywords: Deepfakes, faceswap, AI/ML, Detection Depthwise Separable Convolution, Convolution Neural Network (CNN), Social Media, Compressed Video

1. INTRODUCTION

The increasing sophistication of smartphone cameras and the widespread availability of high-speed internet have significantly expanded the reach of social media and media-sharing platforms. As a result, the creation and transmission of digital videos have become easier than ever before. Simultaneously, advancements in computational power have made deep learning more powerful—achievements that would have seemed impossible just a few years ago. However, like any transformative technology, these developments present new challenges.

One such challenge is the emergence of ”Deepfakes” (DF), which are manipulated video and audio clips generated using deep generative adversarial networks (GANs). The rapid spread of deepfake content on social media has led to increased misinformation, deception, and potential threats to individuals and society. The ability to fabricate realistic-looking fake videos can be exploited for malicious purposes, making deepfake detection a crucial necessity.

To combat this issue, we propose a novel deep learning-based method to effectively distinguish AI-generated deepfake videos from authentic ones. It is imperative to develop technologies that can detect and mitigate the spread of deepfakes across the internet.

Deepfake detection requires a thorough understanding of how GANs create these manipulated videos. GANs take an input video and an image of a specific individual (the ”target”) and generate another video where the target’s face is replaced with that of another person (the ”source”). Deepfake technology relies on adversarial neural networks trained on face images and target videos to map facial expressions and movements from the source to the target. With proper post-processing, the resulting deepfake videos can appear highly realistic. The GAN model typically splits the video into frames, replaces the target face in each frame, and then reconstructs the video using autoencoders.

Our deep learning-based detection method leverages the same principles used by GANs to generate deepfakes. Due to computational limitations and time constraints, deepfake algorithms can only synthesize face images of a fixed size. To match the target’s face configuration, these images undergo affine warping, which introduces noticeable artifacts. These artifacts result from resolution inconsistencies between the manipulated face area and its surrounding context.

Our approach identifies these artifacts by analyzing the generated face regions and comparing them to their surroundings. To achieve this, we split the video into frames and extract features using a ResNeXt Convolutional Neural Network (CNN). Additionally, a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) is employed to capture temporal inconsistencies between frames—artifacts introduced by GAN-based reconstruction. To train our ResNeXt CNN model, we simulate resolution inconsistencies in affine face wrappings, allowing for a more robust deepfake detection framework..

2. Literature Survey:

The explosive growth in deep fake video and its illegal use is a major threat to democracy, justice, and public trust. Due to this there is a increased the demand for fake video analysis, detection and intervention. Some of the related word in deep fake detection are listed below:

1. Exposing Deepfake Videos by Detecting Face Warping Artifacts:

Used an approach to detects artifacts by comparing the generated face areas and their surrounding regions with a dedicated Convolutional Neural Network model. In this work there were two-fold of Face Artifacts.

Their method is based on the observations that current DF algorithm can only generate images of limited resolutions, which are then needed to be further transformed to match the faces to be replaced in the source video.

2. Exposing AI Created Fake Videos by Detecting Eye Blinking :

Describes a new method to expose fake face videos generated with deep neural network models. The method is based on detection of eye blinking in the videos, which is a physiological signal that is not well presented in the synthesized fake videos. The method is evaluated over benchmarks of eye-blinking detection datasets and shows promising performance on detecting videos generated with Deep Neural Network based software DeepFake.

Their method only uses the lack of blinking as a clue for detection. However certain other parameters must be considered for detection of the deep fake like teeth enchantment, wrinkles on faces etc. Our method is proposed to consider all these parameters.

3. Using capsule networks to detect forged images and videos:

Uses a method that uses a capsule network to detect forged, manipulated images and videos in different scenarios, like replay attack detection and computer-generated video detection.

In their method, they have used random noise in the training phase which is not a good option. Still the model performed beneficial in their dataset but may fail on real time data due to noise in training. Our method is proposed to be trained on noiseless and real time datasets.

4. Detection of Synthetic Portrait Videos using Biological Signals:

Approach extract biological signals from facial regions on authentic and fake portrait video pairs. Apply transformations to compute the spatial coherence and temporal consistency, capture the signal characteristics in feature sets and PPG maps, and train a probabilistic SVM and a CNN. Then, the aggregate authenticity probabilities to decide whether the video is fake or authentic.

Fake Catcher detects fake content with high accuracy, independent of the generator, content, resolution, and quality of the video. Due to lack of discriminator leading to the loss in their findings to preserve biological signals, formulating a differentiable loss function that follows the proposed signal processing steps is not straight forward process.

3. Proposed System:

There are many tools available for creating DeepFakes (DF), but very few tools exist for detecting them. Our approach to DF detection will be a significant contribution toward preventing their spread across the worldwide web. We propose a web-based platform that allows users to upload videos and classify them as either fake or real. This project can be expanded from a web-based platform to a browser plugin for automatic DF detection. Major applications like WhatsApp and Facebook could integrate this system into their platforms to enable pre-detection of DF before a video is sent to another user. One of the key objectives is to evaluate the system’s performance and acceptability in terms of security, user-friendliness, accuracy, and reliability. Our method focuses on detecting all types of DF, including replacement DF, retrenchment DF, and interpersonal DF. Figure 1 illustrates the simple system architecture of the proposed system.

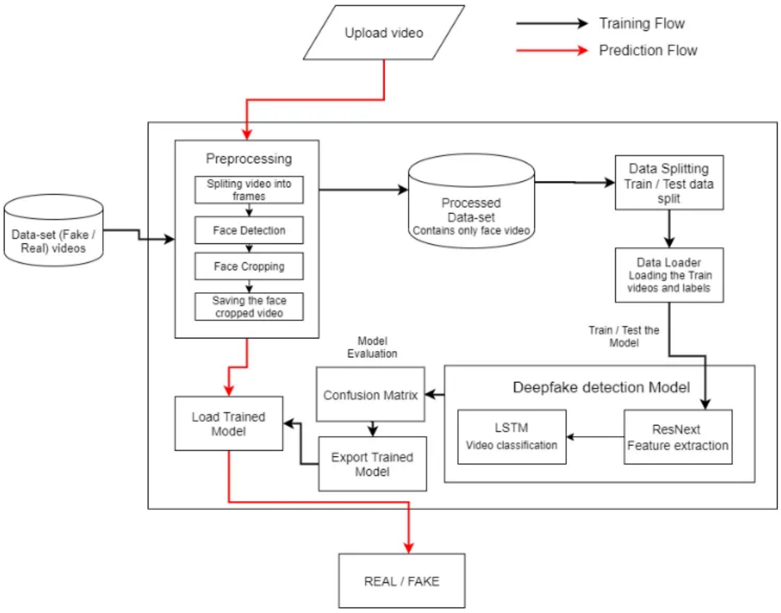

Upload video
Training Flow
Prediction Flow
Preprocessing
Spliting video into frames
Face Detection
Face Cropping
Saving the face cropped video
Data-set (Fake / Real) videos
Processed Data-set Contains only face video
Data Splitting Train / Test data split
Data Loader Loading the Train videos and labels
Train / Test the Model
Model Evaluation
Confusion Matrix
Deepfake detection Model
LSTM Video classification
ResNext Feature extraction
Load Trained Model
Export Trained Model
REAL / FAKE

Figure 1: System Architecture

1. Dataset:

We use a mixed dataset consisting of an equal number of videos from various sources, including YouTube, FaceForensics++, and the DeepFake Detection Challenge dataset. Our newly prepared dataset contains 50To ensure a balanced and effective training process, we split the dataset into 70.

2. Preprocessing:

Dataset preprocessing involves splitting the video into frames, followed by face detection and cropping the detected face from each frame.

To maintain uniformity in the number of frames, we calculate the mean frame count across all dataset videos. The newly processed dataset is then created, ensuring that each video contains a number of frames equal to this mean. Frames without detected faces are ignored during preprocessing to enhance data quality.

Processing a 10-second video at 30 frames per second results in a total of 300 frames, requiring significant computational power. For experimental purposes, we propose using only the first 100 frames for training the model to optimize resource usage while maintaining detection accuracy.

3. Model:

The model consists of ResNeXt-50 (32x4d) followed by a Long Short-Term Memory (LSTM) layer.A Data Loader is used to load the preprocessed, face-cropped videos and split them into training and testing sets. The frames from the processed videos are then passed to the model in mini-batches for efficient training and evaluation.

4. ResNeXt CNN for Feature Extraction:

Instead of designing a new classifier from scratch, we propose using the ResNeXt CNN classifier to extract features and accurately detect frame-level features.To enhance performance, we will fine-tune the network by adding necessary layers and selecting an appropriate learning rate to ensure proper convergence of gradient descent.

The 2048-dimensional feature vectors obtained from the last pooling layers of ResNeXt are then used as sequential inputs for the LSTM model, enabling temporal analysis of video frames.

5. LSTM for Sequence Processing:

We use a sequence of ResNeXt CNN feature vectors extracted from input frames as input to a 2-node neural network, which determines the probability of the sequence belonging to either a deepfake video or an authentic (untampered) video. A key challenge is designing a model capable of processing video sequences recursively and meaningfully. To address this, we propose using a 2048-unit LSTM with a 0.4 dropout rate, which effectively captures temporal dependencies and improves generalization.

LSTM enables sequential frame processing, allowing for temporal analysis by comparing the frame at time t with frames at t - n, where n is any chosen number of frames before t. This approach helps detect inconsistencies introduced by deepfake generation techniques.

6. Prediction:

When a new video is submitted for prediction, it first undergoes preprocessing to match the format of the trained model. The preprocessing steps include splitting the video into frames and cropping detected faces. Instead of storing the processed frames in local storage, they are directly passed to the trained model for deepfake detection. The model then classifies the video as either fake or authentic based on the extracted features and temporal analysis.

7. Result:

The output of the model is going to be whether the video is deepfake or a real video along with the confidence of the model. One example is shown in the figure 2.

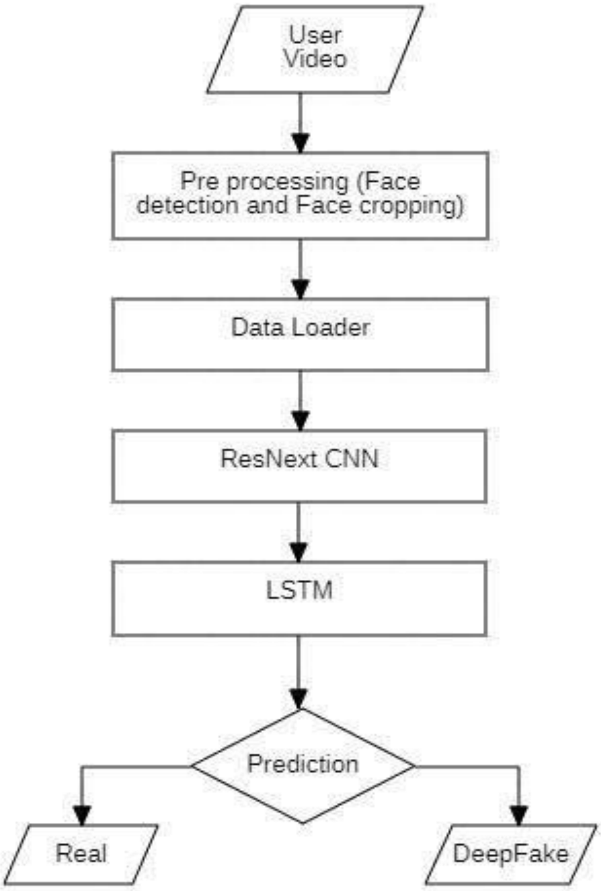
8. Conclusion:

In this work, we presented a neural network-based approach for classifying videos as either deepfake or real, along with the confidence level of our proposed model. Our method is inspired by the way deepfakes are generated using Generative Adversarial Networks (GANs) and Autoencoders.

The proposed approach performs frame-level detection using ResNeXt CNN and video classification using an RNN with LSTM. By leveraging these techniques, our model effectively identifies deepfake content based on key parameters outlined in this paper. We believe that this approach will achieve high accuracy when applied to real-time data, making it a valuable tool for deepfake detection.

9. Limitations:

Our method does not consider audio, which means it is unable to detect audio deepfakes. However, we plan to extend our approach in the future to include audio analysis for detecting deepfake content more comprehensively.



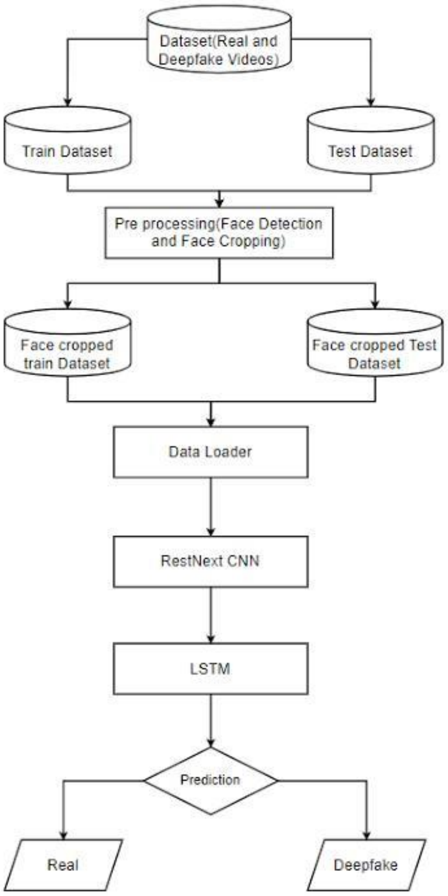

Dataset(Real and Deepfake Videos)
Train Dataset
Test Dataset
Pre processing(Face Detection and Face Cropping)
Face cropped train Dataset
Face cropped Test Dataset
Data Loader
RestNext CNN
LSTM
Prediction
Real
Deepfake

Fig 2: Prediction Flow

Fig 3: Training Flow

4. Methodology:

The proposed deepfake detection system employs a deep learning-driven pipeline to identify face-swapped deepfake videos. The methodology consists of five key stages: Data Acquisition, Preprocessing, Feature Extraction, Model Training, and Evaluation.

1) Data Acquisition :

A diverse dataset is essential to train an AI/ML-based deepfake detection system. The following datasets were used:

- FaceForensics++ – Contains real and manipulated videos generated using face-swapping techniques.
- Deepfake Detection Challenge (DFDC) – A large dataset of deepfake videos created using various face-swapping algorithms.
- Celeb-DF – High-quality deepfake dataset containing low-compression manipulated videos.
- Custom Dataset – Collected from real-world social media sources to improve model generalization.

2) Preprocessing:

Before training, the raw video data undergoes preprocessing to enhance the model’s learning efficiency:

- Frame Extraction: Videos are converted into frames at 30 FPS to capture temporal changes.
- Face Detection & Alignment: The MTCNN (Multi-task Cascaded Convolutional Networks) is used to detect and crop facial regions for analysis.

- Normalization: Pixel values are normalized between 0 and 1 for stable training.
- Data Augmentation: Techniques such as Gaussian noise addition, flipping, and rotation are applied to enhance robustness.

3) Feature Extraction:

Feature extraction is crucial for distinguishing real and manipulated faces. The proposed method leverages Convolutional Neural Networks (CNNs) and deep feature representations:

- ResNeXt-50-based Feature Extraction: A powerful CNN-based feature extractor is used to capture subtle inconsistencies in manipulated faces.
- Depthwise Separable Convolution: Used to enhance feature detection efficiency while reducing computational cost.
- Frequency Domain Analysis: Fourier transforms are applied to identify unnatural artifacts introduced during deepfake generation.

4) Model Training:

The extracted features are used to train a Recurrent Neural Network (RNN)-based deepfake classification model:

- ResNeXt + LSTM Hybrid Model:
  - ➢ ResNeXt-50 CNN extracts spatial features from video frames.
  - ➢ LSTM (Long Short-Term Memory) RNN captures temporal dependencies between frames, improving classification accuracy.
- Binary Cross-Entropy Loss Function is used to train the model for real vs. deepfake classification.
- Adam Optimizer is applied with an initial learning rate of 0.0001 for stable convergence.
- Training Pipeline:
  - ➢ 80% of the dataset is used for training.
  - ➢ 10% is used for validation to tune hyperparameters.
  - ➢ 10% is reserved for testing.

5) Model Evaluation:

To assess the performance of the proposed AI/ML-based deepfake detection system, the following metrics are used:

- Accuracy – Measures overall classification performance.
- Precision & Recall – Evaluate model effectiveness in detecting deepfake videos.
- F1-score – Ensures a balance between precision and recall.
- AUC-ROC Curve – Measures the model's ability to distinguish between real and fake videos.

5. Model Used:

1) Convolutional Neural Network (CNN) Based Models:

- MesoNet: Specifically designed for deepfake detection, focusing on mesoscopic features of images
- XceptionNet: A CNN architecture that has proven highly effective for deepfake detection
- EfficientNet: Provides a good balance between computational efficiency and detection accuracy.

2) Multi-modal Approaches:

- Face X-ray: Detects the blending boundary in manipulated facial images.
- FWA (Face Warping Artifacts): Focuses on detecting warping artifacts that are common in deepfakes.
- Capsule Networks: Can detect various inconsistencies in facial features.

3) Temporal Analysis Models:

- Recurrent Convolutional Models (CNN+LSTM): Analyze temporal inconsistencies across video frames.
- 3D CNNs: Process spatio-temporal information to detect unnatural movements or expressions
- Attention-based temporal models: Focus on specific facial regions across frames

4) Features to Analyze:

- Physiological signals: Heart rate, blinking patterns, and other biological inconsistencies.
- Visual artifacts: Blending boundaries, color inconsistencies, and unnatural textures
- Temporal coherence: Inconsistencies in movement, expression changes, and head poses between frames.

5) Recent Developments:

- Contrastive learning approaches: Self-supervised methods that learn to distinguish between real and fake faces.
- Transformer-based models: Applying vision transformers to detect subtle manipulation cues.
- Multi-task learning frameworks: Simultaneously performing face detection, landmark detection, and deepfake classification.

References:

1. Yuezun Li, Siwei Lyu, “ExposingDF Videos By Detecting Face Warping Artifacts,” in arXiv:1811.00656v3
2. Yuezun Li, Ming-Ching Chang and Siwei Lyu “Exposing AI Created Fake Videos by Detecting Eye Blinking” in arxiv.
3. Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen “ Using capsule networks to detect forged images and videos ”.
4. Hyeongwoo Kim, Pablo Garrido, Ayush Tewari and Weipeng Xu “Deep Video Portraits” in arXiv:1901.02212v2.
5. Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin “Detection of Synthetic Portrait Videos using Biological Signals” in arXiv:1901.02212v2
6. I an Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.
7. David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In AVSS, 2018.
8. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.
9. An Overview of ResNet and its Variants : https://towardsdatascience.com/an-overview-of-resnetand-its-variants5281e2f56035
10. Long Short-Term Memory: From Zero to Hero with Pytorch:https://blog.floydhub.com/long-short-termmemory-from-zeroto-hero-with-pytorch
11. Deepfake Detection Challenge https://www.kaggle.com/c/deepfake-detection-challenge/data
12. Confused about the image preprocessing in classification https://discuss.pytorch.org/t/confused-about-theimagepreprocessing-in-classification/3965
13. FaceForensics https://github.com/ondyari/FaceForensics
14. .Y. Qian et al. Recurrent color constancy. Proceedings of the IEEE International Conference on

Computer Vision, pages 5459–5467, Oct. 2017. Venice, Italy.

15. R. Muthumeenakshi, Balasubramaniam S., Charanjeet Singh, Pallavi V. Sapkale, “An Efficient and Secure Authentication Approach in VANET Using Location and Signature-Based Services”, Ad Hoc and Sensor Wireless Networks 53 (Issue 1-2), 59-83, 2022
16. Uttam D. Kolekar, “Development of Optimized and Secure Routing Algorithm using AODV, ACO and LSB Steganography for Mobile Ad-Hoc Network”, Journal of Advanced Research in Dynamical and Control Systems (JARDCS),Vol. 11, issue 9, pp. 560-568, Sept 2019.
17. Sandeep B Hake, “Design and development of universal test bench for engine aftertreatment controls system”, International journal of advanced research in electronics and communication engineering, Volume 6, Issue 4, Pages 309-312, 2017.
18. Samarjeet Powalkar, “Fast face recognition based on wavelet transform on pca” International Journal of Scientific Research in Science, Engineering anf Technology, Vol 1, Issue 4, PP 21-24, 2015.
19. U Waghmode, DP Deshmukh, S Ekshinge, A Kurund, “An Innovative Approach Using Cyber Security for Steganography for Wireless Adhoc Mobile Network Application” International Conference on Science Technology Engineering and Management (ICSTEM), Pages 1-5, 2024.
20. C Kaur, DS Rao, S Bandhekar, “Enhanced Land Use and Land Cover Classification Through Human Group-based Particle Swarm Optimization-Ant Colony Optimization Integration with Convolutional Neural Networ”, International Journal of Advanced Computer Science and Applications, Vol 14, Issue 11, 2023.
21. Divya Rohatgi, Veera Ankalu Vuyyuru, KVSS Ramakrishna, Yousef A Baker El-Ebiary, V Antony Asir Daniel, “Feline Wolf Net: A Hybrid Lion-Grey Wolf Optimization Deep Learning Model for Ovarian Cancer Detection”, International Journal of Advanced Computer Science and Applications, Vol 14, Issue 9, 2023.
22. Uttam D. Kolekar, “Trust-Based Secure Routing in Mobile Ad Hoc Network Using Hybrid Optimization Algorithm”, The Computer Journal, Oxford University Press, Vol. 62, issue 10, pp. 1528-1545, Oct 2019.
23. Uttam D. Kolekar, “E-TDGO: An Encrypted Trust based dolphin glowworm optimization for secure routing in mobile adhoc network”, International Journal of Communication Systems, Wiley publication, Vol. 33, issue 7, May 2020.
24. Dilip P Deshmukh, Abhijeet Kadam, “Efficient Development of Gesture Language Translation System using CNN”15th International Conference on Computing Communication and Networking Technologies (ICCCNT) Pages 1-6, 2024.
25. Prajwal Kote, Mounesha Zonde, Om Jadhav, Vaibhav Bhasme, Nitin A Dawande “Advanced and Secure Data Sharing Scheme with Blockchain and IPFS: A Brief Review”15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Pages 1-5, 2024.