DeepFakeGuard: A Deep Transfer Learning Framework for Accurate Video Deepfake Detection

Veeramani. R[1], *Sundarababu* Maddu[2], *Thiyagu* T[3] , *Asha* D[4], *Dhivya* R[5], and *Rathana Sabapathy* S[6]

[1]Associate Professor, Department of Computer Science and Engineering ,Chennai Institute of Technology, Sarathy Nagar, Kundrathur,Chennai-600069, TamilNadu, India.
[2]Assistant Professor, Department of CSM ,Anil Neerukonda Institute of technology and sciences, Vishakapatnam, Andhrapradesh.
[3]Assistant Professor, School of Computing, Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India.
[4]Assistant Professor, Department of Artificial Intelligence and Data science, St. Joseph's Institute of Technology (Autonomous), OMR,Chennai-600119, Tamil Nadu, India
[5]Associate Professor, Adhiparasakthi College of Engineering, Kalavai, Tamilnadu, India.
[6]Assistant Professor, Computer Science and Business Systems, Rajalakshmi Institute of Technology.

Abstract. The idiom "Seeing is believing" probably doesn't work in this digital age, where technology is becoming increasingly open to manipulation and creation of deepfakes can be made right on a smart phone. These AI-created deepfakes, significantly complicate the identification by the naked eye of real and false information. As such, deepfake detection has emerged as a challenging, impactful problem in many domains. This work presents a recent technique for the detection of video deepfake based on deep learning models InceptionResNextV2 and Long Short-Term Memory (LSTM). Utilizing transfer learning, the pre-trained InceptionResNextV2 Convolutional Neural Network (CNN) is used to extract feature of the video data, while the extracted feature is further processed by the LSTM network for accurate classification. Using an 80:10:10 train–validation–test split, experiments were carried out on the FaceForensics++ (FF++) dataset (c23 and c40 compression levels). The proposed method achieved as high as 98.97% and outperformed the state-of-the-art methods in detecting the deepfake videos. This express the efficacy of the joint LSTM and InceptionResNextV2 model for combating the increasing threat of deepfake in the current era of modern computing.

I. Introduction

Deepfake technology using AI and deep learning can generate lifelike synthetic media capable of undermining national security, personal privacy and social order [1]. Face swaps and lip-sync manipulations prove especially challenging to detect, causing for the necessity of robust detection models that can uncover even the most subtle cues [2].With the increasing sophistication of deep-fake generation methods, It getting increasingly challenging to tell aparbetween real and fake data, hus demanding for advanced methods in forgery detection [10]. Deepfakes, which are generated via techniques such as GANs, pose ethical issues, especially in fields of privacy, consent, and fake news. As this technology has improved, it has develop into more complicated to separate real and fake media, posing potential threats to democracy and security. Present work is aimed at the detection of deepfake implemented with the algorithms involving Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), for addressing these issues.

Conventional computer vision methods are unable to capture such fine-grained distortion and consider deeplearning-based methods essential for enhancing the power of deepfake detection [5]. This has driven much exploration on the deep learning-based solutions of facial forgery detection to develop a powerful tool for authenticating the forged content. In this study, we introduce DeepFakeGuard, a transfer learning system designed especially to increase video deepfake detection efficiency and accuracy. This model uses pre-trained CNNs, which have proved to be successful in many computer vision applications [6], that can be further used in extracting strong features from video frames and hence overcomes the limitation of classical methods in [21]. The goal of the proposed framework is to overcome the limitations of the methodology based on evolving deepfake techniques, with a focus on the submission of the deep transfer learning to learn the generalizable detection form against various deepfake representations [7].

1.1 Contributions:

The following are the contributions of the paper

- We propose an advanced approach for detecting

Corresponding author: thiyagu.57@gmail.com

video deepfakes using InceptionResNextV2, specifically Long Short-Term Memory (LSTM) and deep learning neural networks.

- We utilize transfer learning, where pre-trained InceptionResNextV2 CNNs extract features from video data, which are then classified by the LSTM network for more accurate detection.
- We achieve an impressive detection accura te98.97%, outperforming existing methods in deepfake detection.

1.2 Convolution Neural Networks:

A subset of deep learning models known as Convolutional Neural Networks (CNNs) finds widespread use in picture and video recognition. Convolutional layers, which capture spatial hierarchies and characteristics by applying filters to input data, are one of several layers that make up these models. CNNs excel at automatically learning patterns, textures, and edges, among other significant properties. Computer vision applications including object identification, face recognition, and picture categorization make extensive use of these networks.

***Convolutional layer*:** The convolutional stage of a neural network uses filters on input data to find spatial properties like edges, textures, and structures. The output of a convolution operation is calculated as:

$$Y(i,j) = \sum_{m} \sum_{n} X(i+m, j+n) \cdot K(m,n)$$

where $Y(i,j)$ is the output, X is the input, and K is the filter (kernel).

Pooling layer: A pooling Layer cuts down on the size of the input while keeping vital features and making calculations easier. The most common operation, max pooling, is defined as:

$$Y(i,j) = max\{X(m,n) \mid (m,n) \in \text{ pooling window }\}$$

where $Y(i,j)$ is the output and $X(m,n)$ is the input within the pooling window.

Fully connected layer: A fully connected (FC) layer connects every neuron in the layer to every neuron in the previous layer, enabling the network to learn complex patterns.

Each output neuron computes a weighted sum of all inputs followed by an activation function. The output of a fully connected layer is calculated as:

$$Y = f(WX + b)$$

where Y is the output, W is the weight matrix, X is the input, b is the bias, and f is the activation function.

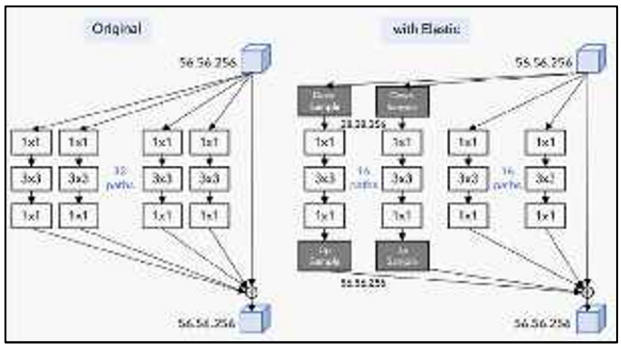

Fig 1. General architecture of CNN

1.3 Residual Next Convolutional Neural Network (ResNeXt)

The ResNeXt model is proposed as an extension of the Residual Network (ResNet) architecture to improve the effectiveness and efficiency of very deep learning architectures. It introduces the so-called cardinality, which means the number of paths (or groups) in each residual block. By increasing cardinality, ResNeXt can learn more informative features with smaller number of parameters than increasing both of depth and width.

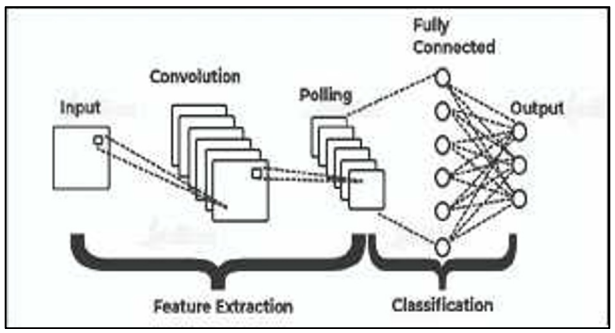

Fig 2. Res NeXt architecture

This symmetry is, therefore, used to significantly increase network's ability to **Figurez 2**. ResNeXt architecture capture complex patterns while at the same time ensuring computational efficiency. As a newly emerged model architecture, ResNeXt achieves superior performance on image classification task, and it has higher accuracy and better generalization memory in deep learning model–ResNeXtVs. VGG16.

1.4 Long Short-Term Memory (LSTM)

Long Short-Term Memory, usually referred to as LSTM, is a type of one kind of recurrent neural network that was designed to model sequences and sequence relations in data. To store and refresh information, it leverages specialised memory cells, hence avoiding one of the fundamental problems of regular machine-learning networks, the vanishing gradient problem. LSTMs consist of gates that control the information flow through the network. These gates are input, forget and outpur gates. Because they are capable of effectively handling sequential data, these types of networks are used across a range of tasks, including speech recognition, time series prediction, and natural language processing.

For each video, the LSTM received a temporal sequence of 100 consecutive feature vectors, corresponding to the first 100 detected face frames,

allowing the model to capture short- and mid-range temporal inconsistencies characteristic of deepfakes.

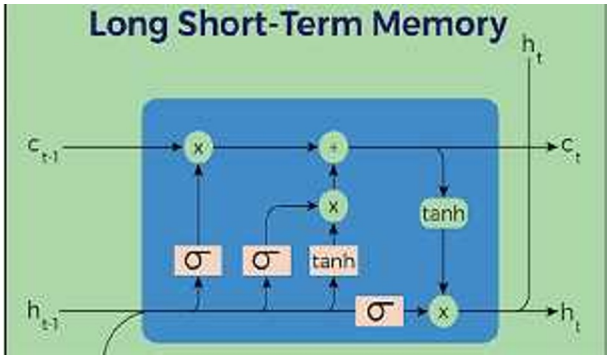

Fig 3. LSTM Architecture

2. Literature Survey

Ranjan et al. [8] introduced a Transfer Learning-based CNN system to detect deepfakes using a variety of datasets and showed that transfer learning can increases the single domain classification accuracy and the generalization of the learned model in need of facing new data. The proposed framework makes good use of pre-trained models to finetune on new datasets and significantly improves the detection performance in different scenarios. However, the most important limitation in their work is that the well-known problem of dataset shift [13] comes into play when the target domain distribution is different than the source domain distribution, which can significantly decrease the performance of the model, i.e., it is difficult for the model to generalise and adapt on unseen deepfakes variations. This limitation motivates the development of strong solutions from both source and target domains to deal with domain shift problems in real-world deepfake detection scenarios. Suratkar et al. [9] designed the deep transfer learning method for detecting deepfake video, in which autoencoder is used to pre-train a hybrid CNN-RNN model. The robustness of the model was tested with new samples, and the effect of residual image input on the accuracy was investigated, indicating that residual image input may improve the detection performance in dynamic environment.

Karandikar et al. [10] developed a CNN-based deep learning algorithm in conjunction with transfer learning for the recognition of the deepfakes in video frames based on facial features and inconsistencies. The model had an accuracy of 70 percent and had difficulties processing low-quality images and videos, so the authors suggested that it would be worth testing the methods on large high-resolution scenes. They also identified concerns such as audio synchronization errors, and suggest more work is needed to detect audio and temporal mismatch, and to investigate ensemble learning to increase precision. Anuhya et al. [11] and Kumar et al. [12] deepfake video detection by transfer learning has limitations. The model can perform poorly with low quality or compressed video and thus its applicability in real-world comes in question. Moreover, their method may not be adequate to detect the subtle temporal or audio inconsistencies present in deepfake videos.

2.1 CHALLENGES AND LIMITATION IN THE EXISTINGSYSTEM

Although current works have shown great progress towards distinguishing deepfakes via transfer learning, they still face some limitations. The models find it difficult to distinguish deepfakes generated with different generation methods and models, especially when they are manipulated in a delicate manner, which may lower the accuracy. In addition, the collection of a large diverse dataset is challenging, which restrict the effects of these models. Furthermore, the computational cost of training complex models from scratch remains impractical despite the reduction in complexity offered by transfer learning. Finally transfer learning improves detection result and saves training time but the problem that deepfakes with minor temporal and audio mistakes are hard to detect with a transfer learning method have not been solved completely.

3. Proposed methodology

The proposed workflow for deepfake detection consists of several steps: (i) preprocessing videos by dividing them into frames and by applying face detection and cropping to form a processed dataset comprising face-cropped videos, (ii) using deep learning algorithms to train on the processed dataset to construct detection models, and (iii) further testing the pre-trained detection models on new data for performance evaluation. Then we split the dataset into training and test datasets and load the train videos and labels. The deepfake detection model is a model based on Long short-term memory (LSTM) is used for video classification, and Inception ResNextV2 is used for feature extraction. These features are extracted, and model is trained on them and finally test is conducted to identify whether the video is real or fake. This approach incorporates state-of-the-art deep learning methods to achieve effective and accurate detection on deepfakes.

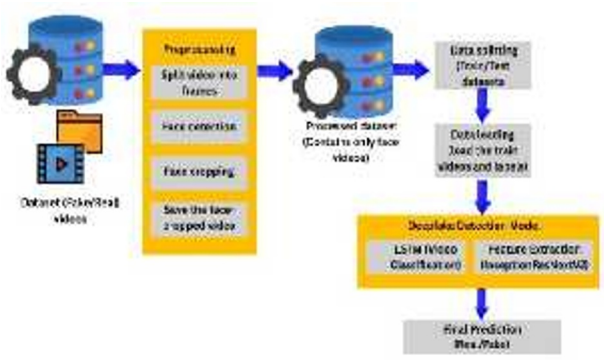

Fig 4. Proposed system architecture

3.1. Dataset

In this work, we created a dataset by aggregating a number of public available sources such as FaceForensic++, Deepfake Detection Challenge (DFDC), and Celeb-DF. The dataset was balanced with 50 percent real and 50 percent fake videos to avoid biasing the training. A total of 6000 videos were

sampled: 1500 actual and 1500 fake videos from DFDC, 1000 real and 1000 fake videos from FaceForensic++, and 500 real and 500 fake videos from Celeb-DF. Video with audio modified and publicly available in DFDC was excluded from pre-processing to narrow the scope of the dataset to visual manipulations.

3.2 Data pre-processing

Several pre-processing stages were executed on the video data before training the model, to normalize it and increase model performance. First, video files were frame-extracted, which was essential for subsequent processing. A face detection method was subsequently applied on every frame to search for faces on top of which the detected faces were cropped to contain the parts of the face that relate to features. To standardize the dataset, the average frames per video was calculated, and the videos were re-sized such that each video now contains an average number of frames. Frames with no detected face are not used to verify only real motion was handled. Because of computational limitations, the first 100 frames of each video were considered for the analysis, and our preliminary studies revealed that the first 140 frames were enough for model construction. This strategy achieved a good trade-off between computation efficiency and the keep/probability for the training.

3.3 Dataset split

The dataset was separated into training and test datasets containing 70% of the video (4200 of the video) for training of the proposed system and the remain 30% (1800 videos) was kept for testing purpose. Both training and testing set were balanced to keep the fairness and avoid the bias, each of which was composed of half real videos and half fake videos. This balanced dataset guarantees that the model has been trained and tested on a representative sample, reducing the risk of over fitting new, hitherto unknown data.

3.4 Model construction

The deepfake detection model is based on the combination of LSTM and CNN architectures. We use a pre-trained ResNeXt-50-32x4d CNN to generate 2048-dimensional frame-level features, which are furtherly encoded and processed through an LSTM network. The LSTM processes the frames temporal dynamics and has one layer of the hidden size 2048. It also uses Leaky ReLU activation function, linear layer for classification and a SoftMax layer for performing the prediction. Fine tuning is carried out by adding layers and tuning the learning rate and model is trained with a batch size of 4. The architecture makes the model possible to classify the video to a video as being real or deepfake according to the temporal and spatial features extracted from the video frames.The reported detection accuracy of 98.97% corresponds to the video-level classification performance measured on the 30% held-out test set, where the LSTM produces a single prediction from a sequence of 100 frame-level feature vectors for each video.

Table 1. Model parameters setting

Parameter	Value
InceptionResNextV2 Architecture	InceptionResNextV2 + LSTM
LSTM Hidden Units	2048
Dropout Rate	0.4
Optimizer	Adam
Batch Size	4
Learning Rate	0.001

4. Results and Discussion

A consolidated dataset of 6000 films evenly divided between actual and fake classes, obtained from FaceForensics++, DFDC, and Celeb-DF, was used to assess the suggested InceptionResNextV2 + LSTM hybrid architecture. To guarantee a fair and impartial assessment, the dataset was split between 70% training and 30% testing after preprocessing and face-centric frame extraction. The results of the proposed hybrid Inception-ResNextV2 + LSTM model showed the best statistical performances on all models with the highest accuracy (98.97%), along with best precision, recall and F1 score. We find that taking advantage of highly sophisticated CNN architectures with LSTM can significantly boost the accuracy of deepfake detection. It is clearly observed that the proposed model performs much better than the conventional models (e.g., VGG16 and ResNet) in terms of all the evaluation criteria. Figure 5 compares trends of model performance in accuracy, precision, recall, and F1 score, between various architecture of VGG16, ResNet, LSTM, our proposed Inception ResNextV2 + LSTM. The experimental results (accuracy, precision, recall and F1 score) are recorded in Table 2 across the four deep-fake detection models, i.e. VGG16, ResNet, LSTM and Inception ResNextV2 + LSTM. Figure 6 shows the model accuracy and loss for both training and validation sets over 10 epochs by the Inception ResNextV2 + LSTM model. The solid lines represent accuracy, and the dashed lines loss.

Additionally, the model demonstrated consistent performance across evaluation criteria with high precision (97.77%), recall (98.53%), and F1-score (98.15%). Figure 6 displays steady convergence with little over fitting over ten epochs, while Figure 5 clearly outperforms baseline models. The model's ability to distinguish altered face regions and detect minute temporal distortions is further demonstrated by the detection outputs shown in Figure-7.The result of deepfake detection is illustrated in Figure-7, the top row in Figure-7 are the extracted frames from the input video, face detection is performed on each frame. The last row shows the extracted faces, i.e. the isolated facial areas for the deep-fake detection, to show the difference between the authentic and tampered video segments. The system ultimately issues a classification to indicate whether the uploaded video is real or a deepfake. This decision is made using the facial structure and temporal incoherence analysis. Compared with widely used baseline methods such as VGG16, ResNet, and an LSTM-only model, the proposed InceptionResNextV2 +

LSTM framework achieves a clear performance gain—improving accuracy by 8.76% over VGG16**,** 5.30% over ResNet**,** and 2.43% over the standalone LSTM model—thereby demonstrating a consistent margin of improvement across all major evaluation metrics

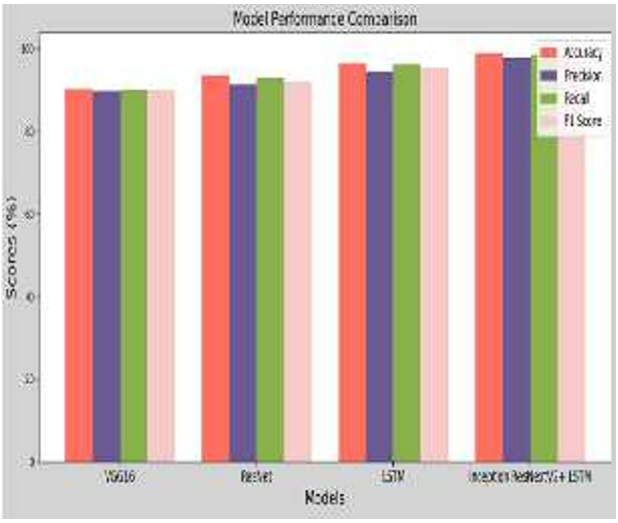

Fig 5. Comparison of model performance

Table 2.Results comparison of the proposed model

Model performance				
Models	**Accuracy**	**Precision**	**Recall**	**F1 score**
VGG16	90.21	89.78	90.14	89.96
ResNet	93.67	91.32	92.78	92.05
LSTM	96.54	94.45	96.12	95.28
Inception ResNextV2+ LSTM	98.97	97.77	98.53	98.15

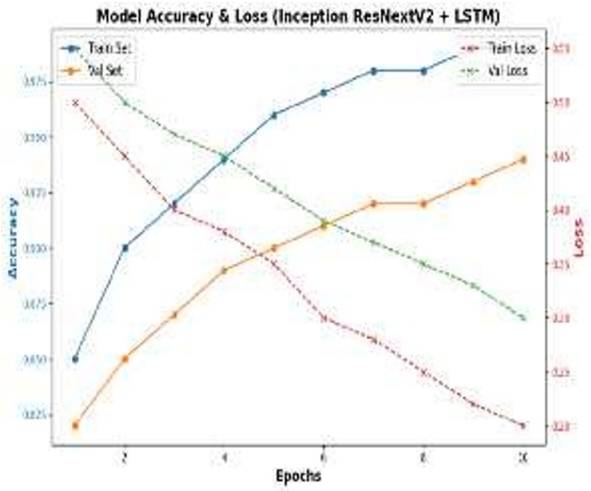

Fig 6.Training and validation model performance

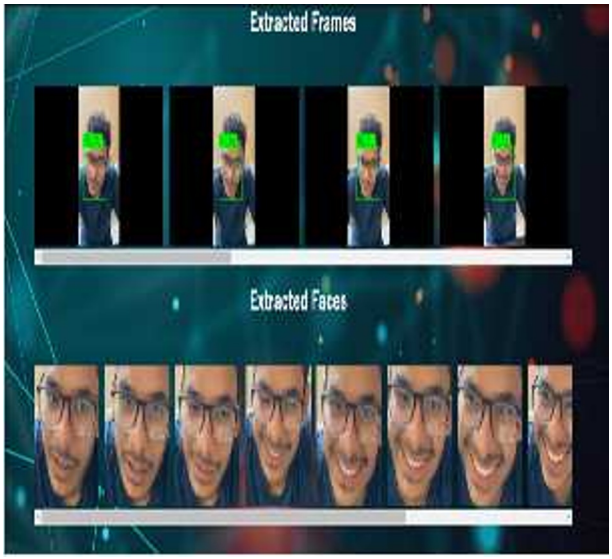

Fig 7. Deepfake Detection Output

5. Conclusion

Finally, with the advent of deepfakes, public confidence in digital content is heavily shaken, so there is an urgent need for detection models. In this paper, we introduced an automatic deepfake detection model based on deep learning such as InceptionResNextV2 for feature extraction and LSTM structure for temporal analysis, which achieved high precision and reliability for identifying the forgery media. As the development of deepfakes and the efforts to combat them continue, it may be beneficial to future research to extend this methodology to new architectures and detection methods to enhance the effectiveness of detection. At the end of the day, the introduction of RNN (particularly LSTM) for deepfake detection is a significant stride in attempting to mitigate synthetic media and its risks, and towards the continuous quest to strengthen digital content in the modern world.

References

1. A. Al-Adwan, H. Alazzam, N. Al-Anbaki,E. & Alduweib, Detection of Deepfake Media Using a Hybrid CNN–RNN Model and Particle Swarm Optimization (PSO) Algorithm. Computers, 13(4), 99,(2024).
2. T. T .Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T .Nguyen, T. Huynh,S. Nahavandi,&C.M. Nguyen,Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, **223,** 103525, (2022).
3. L.Ma, P.Yang,Y. Xu,Z.Yang, P.Li, & H.Huang, Deep learning technology for face forgery detection: A survey. Neurocomputing, 129055, (2024).
4. H.Wu, Y.Liao, R.HadiMogavi, P.Hui & P.Y.Zhou, Deepfake in the metaverse: An outlook survey. In International Conference on Human-Computer Interaction Springer, Cham (pp. 253-267), (2025).
5. E.Altuncu, V.N.Franqueira & S.Li, Deepfake: definitions, performance metrics and standards,

datasets, and a meta-review. Frontiers in Big Data, *7*, 1400024, (2024).
6. U.A.Ciftci, I.Demir & L.Yin, Fakecatcher: Detection of synthetic portrait videos using biological signals. IEEE transactions on pattern analysis and machine intelligence, (2020).
7. F.Juefei-Xu, R.Wang,Y.Huang,Q.Guo,L.Ma,& Y.Liu, Countering malicious deepfakes: Survey, battleground, and horizon. International journal of computer vision, **130(7),** 1678-1734, (2022).
8. P.Ranjan, S.Patil, & F.Kazi Improved generalizability of deep-fakes detection using transfer learning-based CNN framework. In 2020 3rd international conference on information and computer technologies (ICICT) IEEE, pp. 86-90, (2020, March).
9. S.Suratkar&F.Kazi,Deepfake video detection using transfer learning approach. Arabian Journal for Science and Engineering, ***48*(8),** 9727-9737, (2023).
10. A.Karandikar,V.Deshpande,S.Singh,S.Nagbhidkar& S. Agrawal, Deepfake video detection using convolutional neural network. International Journal of Advanced Trends in Computer Science and Engineering, ***9*(2),** 1311-1315, (2020).
11. K.Anuhya, D.Shivaram, D.Karthikeya,B.N.Jyothi, M.A.Jabbar & Z.Akhtar, Deep Fake Video Detection Using Transfer Learning Techniques. In International Conference on Advanced Informatics for Computing Research, Cham: Springer Nature Switzerland, pp.236-249, (2023, December).
12. N.Kumar, P.Pranav, V.Nirney&V.Geetha, Deepfake image detection using CNNs and transfer learning. In 2021 International Conference on Computing, Communication and Green Engineering CCGE, IEEE, pp. 1-6, (2021, September).