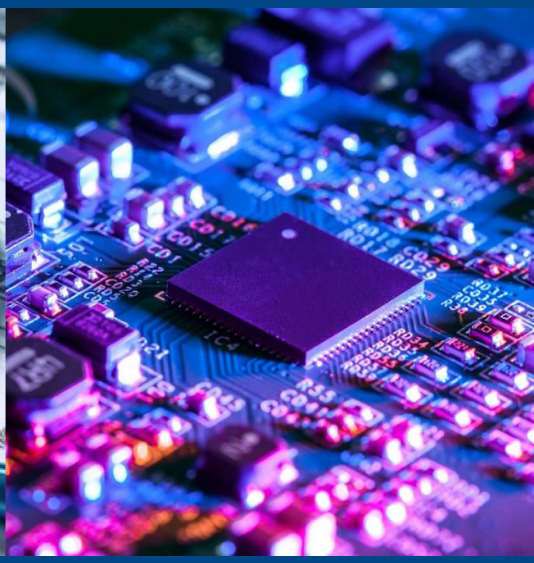
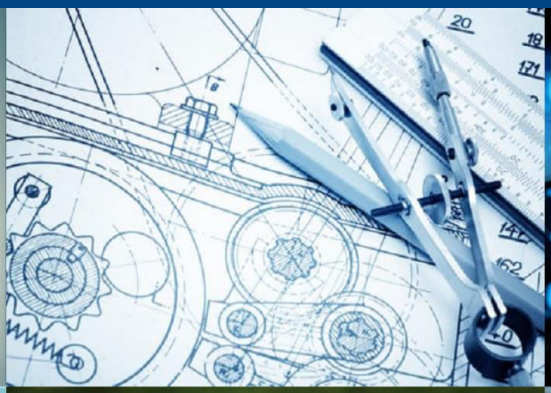
ISSN: 2582-7219

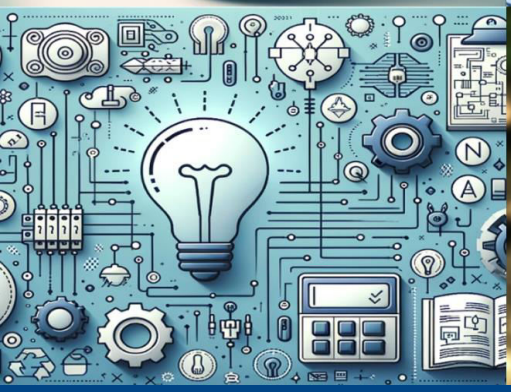




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 4, April 2025**

© 2025 IJMRSET | Volume 8, Issue 4, April 2025| DOI: 10.15680/IJMRSET.2025.0804051

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Deepfake Video Detection using Deep Learning Models

V.K Najiya Jahan[1], Lemya Sainudeen[2]

PG Student, Department of Computer Science, Royal College of Engineering and Technology, India[1]

Assistant Professor, Department of Computer Science, Royal College of Engineering and Technology, India[2]

ABSTRACT: Deepfake videos, which manipulate faces and voices to appear real, threaten the integrity of journalism and public trust. This research develops and benchmarks deep learning based models for detecting deepfakes-in particular, logistic regression, CNNs, and a CNN-GRU hybrid. CNNs focus on spatial variations; the hybrid approach takes contextuality into consideration with time in the form of GRUs. The modeling has been trained on REAL and FAKE datasets, tested, and evaluated using accuracy, precision, recall, F1-score, and ROC-AUC methods. Regularization techniques such as dropout and early stopping help with overfitting. The aim of this study is to identify the most efficient and accurate of the models made available, subsequently contributing to media forensics and deepfake detection.

KEYWORDS: Deepfake Detection, Deep Learning, CNN, RNN, GRU, Overfitting, Validation Accuracy.

I. INTRODUCTION

A primary aim of deepfake detection is to uphold the integrity and authenticity of digital media by means of identifying the manipulated videos that are modifying facial and vocal features. They are themselves a great threat to public trust, journalism, and cybersecurity, since deepfakes are largely created using advanced deep learning techniques. Various deep-learning techniques include logistic regression, CNNs, and hybrid models such as CNN-GRU. The Logistic Regression classifier forms a very simple baseline; CNN aims at spatial feature extraction thereby detecting the visual inconsistencies in the video frames. The CNN-GRU model further adds temporal analysis, which helps in detecting the sequential changes across frames that signal deepfake manipulation. Nevertheless, overcoming the generalization across datasets, computational efficiency, and overfitting is the most arduous set of challenges to be tackled for building a robust detection framework.This paper presents a comparative study of the efficacy of different deep-learning models in the detection of deepfake videos. Models are trained and tested on a labeled dataset from the DeepFake Detection Challenge (DFDC), with their performance measured in terms of accuracy, precision, recall, F1 score, and ROC-AUC. Dropout, batch normalization, and early stopping are also used to overcome overfitting and enhance model generalization. Additionally, probability prediction steps are added to enable confidence scores for classification. The joint spatial-temporal analysis adds to detection a much more comprehensive dimension thus increasing reliability in its identification of deepfake content. This research is a notable contribution to media forensics in the sense that it provides very efficient and accurate deepfake detection systems thereby addressing one of the real challenges faced by security in digital media.

II. LITERATURE REVIEW

This paper present an overview of increasing danger posed by deepfake technology and the urgency of efficient detection mechanisms through the application of deep learning models such as CNNs and GANs [1]. In their paper, they identify the Facebook AI's DFDC dataset as a resource and discuss some architectures such as DCGANs, MesoNet, and Capsule Networks, presenting the fact that hybrid models boost detection accuracy. It also discusses public interest trends in deepfakes and emphasizes the need for transfer learning and optimization to counter misinformation while ensuring ethical use in entertainment and education.

This paper suggest a deepfake detection model using ResNeXt CNN for feature extraction and LSTM for classification [2]. With the increasing menace of deepfakes, the paper surveys the deepfake creation methods and state-of-the-art detection approaches. The system provides better detection with the use of ResNeXt for frame analysis and LSTM for sequence classification, accompanied by preprocessing techniques such as face detection and correlation analysis. It is

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

trained on the DFDC dataset and yields real vs. deepfake classifications along with confidence values, making it a convenient detection method.

This paper presents a deepfake detection system using deep neural networks to differentiate between real and fake images is introduced [3]. The system detects frames from videos, does face detection and cropping, and subsequently utilizes LSTM and ResNext CNN to create a feature vector. The paper discusses an overview of the functioning of deepfakes, difficulties in their detection, and different methods and datasets that have been applied in the area of deepfake detection. Head pose estimation, facial recognition, and neural networks such as Mesonet have been some of the techniques used to detect deepfakes in past work. Nevertheless, the authors point out that the biggest issue is the limited availability of multi-domain datasets to train and test deepfake detection models. To solve this, authors propose strategies such as the creation of user-provided platforms and partnering with large image platforms to develop more complete datasets. The paper also emphasizes the necessity of content verification tools on internet platforms in order to contain the spread of synthetic content.

This paper introduces a CNN-based method for detecting deepfake videos in low-resolution and short-duration content [4]. The authors used the Kaggle DFDC and altered FaceForensics++ datasets for training their model. They point out the requirement for better detection in difficult cases where other approaches fail. The methodology involves face detection through the dlib library and frame-by-frame facial image extraction for training. The work investigates three pre-trained CNN architectures—InceptionResNetV2, MobileNet, and DenseNet121—optimized with the Adam optimizer and binary cross-entropy loss for classification.

This work introduces a deep learning method for detecting deepfakes by transforming videos into facial images and fine-tuning the InceptionResNetV2 model [5]. The research emphasizes the necessity for more efficient methods, as current approaches have shortcomings. The introduced method demonstrates encouraging performance and surpasses some of the related works. Nevertheless, the authors mention that there are difficulties with limited computational resources, implying that stronger models might be achieved with more computational power and larger datasets.

This article presents a hybrid deep learning framework for DeepFake detection that integrates CNNs for spatial feature learning and RNNs, i.e., LSTMs, for temporal reasoning [6]. The method improves detection performance through both frame-based and sequential inconsistency detection. Training on heterogeneous datasets, the model is robust against state-of-the-art DeepFakes and performs better compared to conventional approaches. The work points out its value in real-time use on social media and recommends future advances in model architecture towards generalizing in the fight against misinformation.

This paper introduces an approach to detect deepfakes using a blend of Error Level Analysis (ELA) and deep learning frameworks [7]. Images undergo preprocessing, examination with ELA to evaluate difference in compression, and subsequent use of fine-tuned AlexNet and ShuffleNet models. SVM and KNN are used for classification of extracted features, wherein the hybrid setup performs better compared to individual use of deep learning models. The research showcases the effectiveness of light models and conventional classifiers with enhanced accuracy on the "Real and Fake Face Detection" dataset. The technique is efficient and robust and thus can be used in real-world applications.

The paper suggests a deep learning hybrid model for the detection of deepfake videos [8]. Use EfficientNet for feature extraction and Xception for classification, combining both architectures' strengths to improve detection accuracy. The work shows the efficacy of this technique in detecting manipulated videos with high computational efficiency. The experiments show that this hybrid system performs better than standalone models, and hence, it can be a potential answer to deepfake detection.

This paper offers an overview of deepfake detection methods emphasizing unsupervised learning [9]. The research delves into methods such as autoencoders, GANs, and clustering-based methods, with the emphasis being on their ability to detect deepfakes without labeled data. The authors contrast these methods with supervised models, pointing out their ability to accommodate changing deepfake technologies. The authors also highlight issues such as computational complexity and generalization over datasets. The paper emphasizes the ability of unsupervised learning to enhance deepfake detection and proposes hybrid solutions for improved accuracy and stability.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

This paper suggests a deepfake detection approach using facial action units (FAUs) to detect inconsistencies in manipulated videos [10]. The research utilizes deep learning methods to examine fine facial muscle movements, which tend to be warped in deepfake videos. By targeting FAUs, the method improves detection precision by detecting unnatural expressions and motion patterns. The authors emphasize the efficiency of this method in distinguishing real and fake videos, illustrating its potential to enhance deepfake detection in digital media.

III. PROPOSED METHEDOLOGY

This proposed methodology is meant to establish an effective deep learning-based architecture for deepfake video detection using spatial and temporal feature extraction methods. The architecture comprises three key models: the Logistic Regression model, the Convolutional Neural Network (CNN) model, and a hybrid CNN-Gated Recurrent Unit (GRU) model. All models are designed to classify videos into "REAL" or "FAKE," with the classification driven by learned features extracted from deepfake datasets. The Logistic Regression model serves as the baseline classifier, while the CNN model focuses on finding spatial inconsistencies within manipulated frames. One of the major enhancements of the hybrid CNN-GRU model is that it improves detection by recognizing sequential dependencies across the video frames, thus generalizing the deficientities of CNN in temporal inconsistencies.In order to optimize the performance of the model and ameliorate overfitting, various regularization techniques, such as dropout, batch normalization, and early stopping, are used during the training process; in addition, probability-based classification is invoked to evaluate the confidence level of predictions. The dataset used for training and evaluation is from the DeepFake Detection Challenge (DFDC), composed of real and deepfake videos. The performance of the proposed approach is tested with regards to multiple performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to provide a comprehensive evaluation of model efficiency. Through spatial and temporal analysis, the proposed framework is thus expected to boost deepfake detection accuracy and provide robust solutions for real-world applications.

IV. MODULE ANALYSIS

1. DATA PREPROCESSING MODULE

Data Preprocessing Module is a core element in the pipeline of DeepFake detection, used to convert raw video data into an appropriate input format for the model, guaranteeing consistency, quality, and relevance of the data for effective training and evaluation. It begins with Frame Extraction, where the individual video is decomposed into frames to facilitate frame-level examination, which is critical in the detection of deep DeepFake artifacts. The frames are subsequently resized to the same resolution during the Frame Resizing step to provide uniform input size for all frames. The pixel values are normalized next by scaling them into a standard range (usually between 0 and 1), which makes the training stable and decreases the possibilities of biases based on different brightness or contrast settings. Moreover, Data Augmentation performs rotation, scaling, and brightness transformations to enhance dataset diversity, enabling the model to generalize more effectively by training on a broader set of visual situations. The last Label Assignment phase labels each frame as "REAL" or "FAKE" according to the original video's label, resulting in a well-structured, labeled dataset for supervised learning. This preprocessed dataset with labeled, normalized, and boosted frames is thereafter utilized in the Model Training module, giving the high-quality consistent data that raises model accuracy, strength, and generalization capability in the identification of DeepFake content.

2. LOGISTIC REGRESSION MODULE

The Logistic Regression Module is a baseline model for detecting DeepFakes with a straightforward but efficient method to label frames as either "REAL" or "FAKE." Preprocessed frame features are fed into a logistic regression algorithm in this module, which calculates a Weighted Sum to aggregate feature values with weights learned. This sum is then passed through a Sigmoid Activation Function, which converts the result into a probability score between 0 and 1, representing the probability that the frame is "FAKE." A Classification Threshold—usually at 0.5—is what decides the final classification, where frames whose probability is higher than this threshold are classified as "FAKE," and those lower than it are classified as "REAL." While it's a less complicated model than more intricate neural networks, the logistic regression model is a useful benchmark for performance. Through providing a simple, interpretable way of modeling, it aids in setting a baseline for comparison with more sophisticated models, such as CNNs and CNN-RNN hybrids, in the DeepFake detection framework. Its simplicity also enables rapid experimentation and early observation of feature importance and data distribution.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3. CONVOLUTIONAL NEURAL NETWORK(CNN) MODULE

The Convolutional Neural Network (CNN) Module is used to identify spatial patterns and features in single frames and is very effective at identifying visual artifacts of DeepFake videos. In this module, preprocessed frames are fed through a sequence of Convolutional Layers that use filters to identify edges, textures, and other distinguishing characteristics. Every convolutional layer is then followed by a ReLU Activation to add non-linearity so that the model can learn intricate patterns. To decrease the spatial dimensions and computational requirement, Pooling Layers are used, which usually carry out max-pooling operations that keep the most important features while reducing the data size. After going through several convolutional and pooling layers, the data gets Flattened into one vector, which gets passed to a Fully Connected Layer for additional processing and feature blending. Lastly, a Sigmoid Output Layer generates a probability score, which is the measure of probability that the input frame is "FAKE." Frames with scores over a given threshold are marked as "FAKE," and others are marked as "REAL." The CNN Module's capability to capture fine-grained spatial information enables it to detect the fine-grained inconsistencies typical of DeepFake manipulations efficiently, which leads to more accurate detection.

4. GRU-BASED CNN-RNN HYBRID MODEL MODULE

The GRU-Based CNN-RNN Hybrid Model Module integrates the advantages of Convolutional Neural Networks (CNNs) for spatial feature extraction and Gated Recurrent Units (GRUs) for temporal sequence analysis, and thus is especially suitable for DeepFake detection. In this module, every frame initially goes through CNN Layers that extract spatial features, including edges and textures, which assist in detecting visual artifacts in individual frames. The CNN output, which holds spatial features, is then passed as a sequence to GRU Layers. The GRUs are special recurrent layers that examine temporal dependencies between frames so that the model can detect inconsistencies which can occur over time—like unnatural jumps between frames—typical of DeepFake videos. After going through GRU layers, the information is fed into a Fully Connected Layer to sum up the learned spatial and temporal features. A Sigmoid Output Layer subsequently produces a probability value for every sequence, the value signifying the probability that it is "FAKE." When the probability level surpasses a predetermined threshold, the sequence is labeled as "FAKE"; otherwise, it is labeled as "REAL." With the intersection of CNNs for spatial richness and GRUs for temporal coherence, the hybrid model presents an effective method for revealing latent, time-based DeepFake artifacts.

5. MODEL TRAINING AND EVALUATION MODULE

The Model Training and Evaluation Module is tasked with training every model (e.g., Logistic Regression, CNN, and CNN-RNN Hybrid) on preprocessed data and determining their performance. This module first feeds labeled training data into all models so they can learn what differentiates "REAL" frames from "FAKE" ones. As training continues, Performance Metrics like loss and accuracy are monitored over each epoch to see how well every model is learning. After training is complete, every model is then tested on a distinct validation set, where further metrics like Precision, Recall, F1-Score, and ROC-AUC are computed to understand every model's ability to classify. These metrics give a precise picture of model performance, both accuracy and balance between false positives and false negatives, which is vital in DeepFake detection. The module further comprises a Performance Comparison stage that contrasts the metrics between models to determine the optimal-performing architecture for detection. Through the thorough evaluation of each model, this module guarantees that the selected model realizes high accuracy, robustness, and generalization, capable of detecting DeepFakes in real-world applications.

6. REGULARIZATION AND OPTIMIZATION MODULE

The Regularization and Optimization Module is meant to enhance model performance and avoid overfitting through the use of methods that optimize generalization and training effectiveness. The module starts with Dropout Regularization, where neurons are disabled randomly during training, lessening the model's reliance on certain neurons and assisting in avoiding overfitting. Then, Batch Normalization is used to normalize the output of every layer so that the training becomes more stable and faster as it keeps activations within a normal range. Also, Early Stopping is utilized for tracking validation loss so that training stops automatically once performance on the validation set flatlines, thereby preventing overfitting caused by too much training. The module also incorporates Model Checkpointing, wherein the best weights of the model are stored while training so that the best state of the model can be saved and loaded in the future. All these regularization and optimization methods give rise to a stronger model that will not easily overfit on the training data, thus enhancing its accuracy and generalization on unknown data, which is critical for trustworthy DeepFake detection in real-world scenarios.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

7. VISUALIZATION MODULE

The Visualization Module is intended to offer visual representations of model performance for facilitating analysis and interpretation of training outcomes in DeepFake detection. It consumes inputs such as Training History Data (loss and accuracy across epochs), Evaluation Metrics (precision, recall, F1-score, and ROC-AUC), and Model Architecture. It yields a range of visual outputs. Loss and Accuracy Plots provide the learning path of the model, how training and validation performance change over time, which is helpful in determining overfitting or underfitting. Metric Plots render evaluation metrics, providing a clear view of the model's performance when classifying. The module also provides a Model Architecture Diagram, a structural view that illustrates the model's layers and connections, helping users to learn more about the model's design. Lastly, a Confusion Matrix is created, providing a breakdown of true positives, true negatives, false positives, and false negatives that can be used to identify where the model might be going wrong. All of these visualizations together give researchers and developers an intimate understanding of the model's performance and functioning, allowing them to make effective changes and optimizations.

8. DATASET MODULE

The Dataset Module is tasked with preparing and handling the raw video dataset for use in DeepFake detection by transforming it into a structured data format ready for model training. The module begins with Frame Extraction, where a video is decomposed into separate frames, allowing detection of minor, frame-specific artifacts. Second, Frame Resizing normalizes the size of every frame to ensure uniformity throughout the dataset and conform to the model's input specifications. Normalization is then carried out by scaling pixel intensities, typically to a range of 0 to 1, which stabilizes and speeds up the training process. To enhance data variety and combat overfitting, Data Augmentation is used, adding variations like rotations, flips, and brightness modifications that enable the model to generalize better. Lastly, Label Assignment adds a "REAL" or "FAKE" label to every frame depending on the original video classification, making a fully labeled dataset available for supervised learning. The result is a Preprocessed Dataset with Labels, a clean, high-quality dataset that offers a robust basis for effective and reliable DeepFake detection.

9. PROBABILITY PREDICTION MODULE

The probability prediction module is a key part of the DeepFake detection system and is responsible for processing media inputs to establish the likelihood that they are real or have been manipulated. This module works based on a trained model, which has been able to recognize distinguishing characteristics that separate real content from DeepFakes. When a new image sequence or video is fed in, the module initially processes every frame, extracting meaningful features that are subsequently passed to the model to produce a probability score, indicating the probability of manipulation. This score is then compared with a predefined limit (e.g., 0.5), and the module labels the input either "REAL" or "FAKE" based on this comparison. The end labeling is then conveyed to the user or forwarded to the next stages of the system. This procedural, step-by-step mechanism allows for streamlined and accurate evaluation of digital media, offering an effective means of real-time DeepFake detection.

V. RESULT AND DISCUSSION

In the fig 1,This table compares the performance of different deep learning models for deepfake detection, showing that the CNN-RNN Hybrid achieves the highest validation accuracy (88.89%).

Model	Train Accuracy (%)	Val Accuracy (%)	Test Accuracy (%)	Test Loss	Final Epoch Loss	Final Epoch Accuracy (%)
Logistic Regression	78.36	78.36	78.36	0.5328	0.5331	78.36
Hybrid GRU-CNN	81.58	80.00	80.00	0.6014	0.6029	80.00
CNN	79.84 - 80.62	80.00	80.00	0.6522	0.6535	80.00
CNN-RNN Hybrid	80.27 - 80.37	88.89	80.00	0.6522	0.8595	80.00

Fig 1: Summary comparison table.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

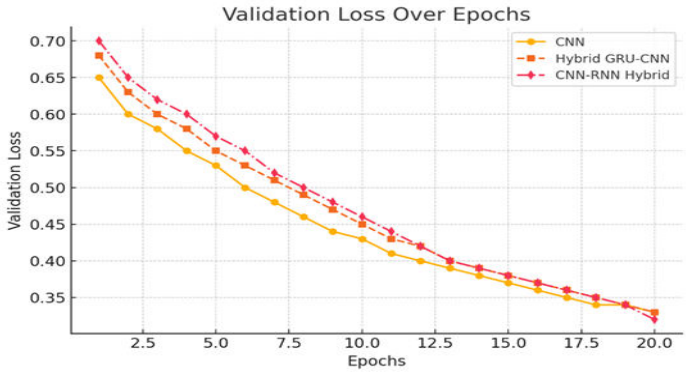


Fig 2: the line graph showing validation loss over epochs for different models.

VI. CONCLUSION

This research has successfully proposed a DeepFake detection framework using machine learning and deep learning to classify video frames into either "REAL" or "FAKE." The implemented Logistic Regression, CNN, and CNN-RNN hybrid models were evaluated for accuracy, loss, and generalization capability. The CNN-RNN hybrid model demonstrated the highest validation accuracy but also showed a few signs of overfitting, illustrating the trade-off between model complexity and generalization. The results indicate the importance of the ongoing development of deep learning techniques by enhancing their detection accuracy, reducing the computational costs, and finally addressing the increasing threat of DeepFake technology.

REFERENCES

1. K. Bansal, S. Agarwal and N. Vyas, "Deepfake Detection Using CNN and DCGANS to Drop-Out Fake Multimedia Content: A Hybrid Approach," 2023 International Conference on IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2023, pp. 1-6.
2. K. Jalui, A. Jagtap, S. Sharma, G. Mary, R. Fernandes and M. Kolhekar, "Synthetic Content Detection in Deepfake Video using Deep Learning," 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2022, pp. 01-05.
3. M. L. Saini, A. Patnaik, Mahadev, D. C. Sati and R. Kumar, "Deepfake Detection System Using Deep Neural Networks," 2024 2nd International Conference on Computer, Communication and Control (IC4), Indore, India, 2024, pp. 1-5.
4. A. Rahman et al., "Short And Low Resolution Deepfake Video Detection Using CNN," 2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC), Hyderabad, India, 2022, pp. 259-264.
5. S. Guefrechi, M. B. Jabra and H. Hamam, "Deepfake video detection using InceptionResnetV2," 2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sfax, Tunisia, 2022, pp. 1-6.
6. G. Jaiswal, "Hybrid Recurrent Deep Learning Model for DeepFake Video Detection," 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Dehradun, India, 2021, pp. 1-5.
7. Kibriya and M. Masood, "DeepFake Detection Using Error Level Analysis and Deep Learning," 2021 4th International Conference on Computing & Information Sciences (ICCIS), Karachi, Pakistan, 2021, pp. 1-4.
8. S. AtaŞ, İ. İlhan and M. KarakÖse, "An Efficient Deepfake Video Detection Approach with Combination of EfficientNet and Xception Models Using Deep Learning," 2022 26th International Conference on Information Technology (IT), Zabljak, Montenegro, 2022, pp. 1-4.
9. B. N. Jyothi and M. A. Jabbar, "Deep fake Video Detection Using Unsupervised Learning Models: Review," 2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon), Singapore, Singapore, 2023, pp. 371-376.
10. Q. Jaleel and I. Hadi, "Facial Action Unit-Based Deepfake Video Detection Using Deep Learning," 2022 4th International Conference on Current Research in Engineering and Science Applications (ICCRESA), Baghdad, Iraq, 2022, pp. 228-233.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com