Robust DeepFake Video Detection Using Hybrid Model

Yash Dewangan[1], Dr. Sunil B. Mane[2]
[1]*Computer Science & Engineering Department/ University Teaching Department, CSVTU Bhilai, India*
[2]*Computer Science & Engineering Department/ COEP Tech University Pune, India*

***Abstract*—Recent improvements in learning poses threat and challenges due to the building of fake images and live videos, which deteriorate the relationship and the credibility. Deep learning tools can drive people to come across content that they have never seen in this manner facilitating the emergence of deepfakes. Originally created for uses in the tech, commercial and Heyday is steering deepfakes to a more surgical level, retrofitting new dimension to what creators are now pumping out. Although, the improved accuracy also bring security risks because of their ubiquity and the way they are built. In this context we need models that can reliably discriminate between true positives and false positives. This paper presents recent research on deep content analysis using deep learning to solve unsolvable problems, demonstrates the advantages and limitations of the current system, and suggests future opportunities. Discriminators based on neural networks (CNN) are often used to detect changing deep news. We use the optical flow-based feature extraction method to extract time-related features, and then use them for classification, because these methods usually focus on the features of each frame of the video and cannot learn time-related information from the video, frames, Argument. Combining CNNs and RNNs with optical flow feature design forms the basis of this approach.**

***Index Terms*—DeepFake Detection, Deep Learning, CNN, RNN, Optical Flow**

I. INTRODUCTION

”DeepFake”, the name is directly derived from ”deep learning” and ”fake” meaning fake media where an image or a video of a person is replaced by a similar camera-shooted content. Due to its easy accessibility, the technology is frequently used to produce photo realistic fake videos of celebs, politicians and all sorts of other upstanding citizens. Deepfakes also use the deep image generated by the video stream to make the fake of video, audio recordings. One of these methods is used to generate deepfakes by learning from a very large collection of real pictures and videos.

Deepfakes are hard to detect as they often seem valid. The continued interest has spurred research into deepfake detection methods — including machine learning algorithms and neural networks. The traditional method for conducting video analysis is deep neural networks [1]. At it is the heart, the basic idea is for neural networks to ”learn” on large amounts of genuine and fake videos the standard patterns that can differentiate real from deepfakes. Technically speaking, deep learning refers to the five types of neural networks - Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Hybrid (CNN + RNN), so on and so forth, including a variety of hybrid models that combine CNN and RNN. Where CNNs particularly excel [2]–[4], is with depth perception (i.e: how we see things in three dimensions — think taking a photo of a person and then messing with their body or face). Local feature space: They try to understand color, texture, edges (gradients), etc. in parts of an image, or movie, then they learn to see patterns for the depth along these dimensions.

RNN [5]: For physical manifestations of fraud, they do this through deconstructing the sequence of frames and unfolding the events in those frames. For deep learning a model which combines CNN and RNN [6] can be even better for analyzing the video body and video interior.

This device was initially created to create movies, for the entertainment sector. Deepfakes are valuable, for producing science fiction animated videos and films as they can generate outcomes [7]. With the expansion of the industry individuals can now explore the alteration feature extensively resulting in the development of software that

provides these choices to users. While anyone who wants to use them, they can use these apps which do not produce real-world results. When deep learning combines with these technologies, the result is disastrous. Photos and videos taken by artificial intelligence and most artificially edited are called deepfakes [8].

Detecting deepfake videos involves methods each having its pros and cons. Lets compare a few known styles:

A. Convolutional Neural Networks (CNN):
For convolutional neural networks (CNNs) in deep learning as they excel at extracting key details from images and videos. By training on datasets like ImageNet (e.g., ResNet50 VGG, Inception) CNNs can later be adjusted for profound datasets and facial alterations.

B. Recurrent Neural Networks (RNN):
RNNs, known as Recurrent Neural Networks are capable of examining the flow of time, in video content since they like LSTMs [9] are tailored for handling data that unfolds in a sequence. Over time RNNs can be taught to identify trends and irregularities in videos like variations in gestures or expressions over a delayed period. Nevertheless RNNs struggle with grasping long term connections within systems [10]. Require computational resources posing a challenge, for deep learning tasks that rely on a solid understanding of physical principles.

C. Generative adversarial networks (GANs):
They are a tag team of artists: one is a creator (the generator) while the other is a critic (the discriminator) Thus, the generator creates fake data, which it is trying to make look as real as possible, and the discriminator evaluates and separates the fakes from the real data. The generator uses feedback from this adversarial process to produce more realistic outputs. The incredible tug-of-war between these two forces of creativity turns out amazingly realistic images, videos and beyond - nearly anything the artificial intelligence can think of.

However, using convolutional neural networks (CNN), RNN or LSTM [11] for deep video detection may have some disadvantages such as low performance, Expensive computation, vulnerability to attacks and poor interpretation. CNNs require a lot of processing power, are vulnerable to countermeasures, only deeply analyze the data they learn from, and can be difficult to understand or interpret.

While there has been great success in the application of deep learning algorithms for deepfake video detection, there exist certain issues and challenges in the current networks. Although we have made a lot of milestone in deep fake video detection on deep learning tools but still there are some errors and inadequacies of the current network. This makes it vulnerable to adversarial attacks which alter the input data in a way that makes the network misclassify it. Because now, the fake depth should not be impersonate-able and an attacker can try to adjust the fake depth, a desirable challenge to have in deep detection.

II. RELATED WORK

Deepfake technology is becoming increasingly common, so the risks associated with it need to be understood and minimized. There are many studies and reports on the progress of deep learning technology. In the survey, ”Detecting Deepfakes in Facial Photos and Videos” [12] they discuss how deep learning and computer vision techniques such as GANs and autoencoders can be combined to create deepfakes. DFaker, Style GAN, FaceSwap1 and other technologies can be used to perform deep matching. In addition to tools such as FaceSwap-GAN and FSGAN [13], object detection techniques are being studied using CNN and RNN.

Information on deepfake production and detection techniques can also be found in the publication ”A Survey on Deepfake Video Detection” [14]. While generative adversarial networks (GANs) have problems replicating faces with real- life behavior, methods include convolutional neural networks (CNN) for phase classification, time-based classification using recurrent neural networks (RNN). Similar methods, visual artifact-based methods, camera-based methods, and biosignal-based methods are introduced to detect deepfake videos.

In his article ”Detecting Deepfakes Using Deep Learning” researcher Rushit Dev provides a comprehensive review of the methods used by

different researchers to solve this problem. He discussed the advantages of CNN architecture and how it can be used to extract faces or other features (such as a person's lips or eyes) from images to determine the difference between depth and the original image. The second method uses multi-sequence LSTM. They wanted to use an LSTM model to find long-term dependence across the entire series, rather than looking at consecutive lines with shorter time intervals. Some in-depth research into biometrics is also introduced.

The study, "FaceForencis++ [15]: Gaining the ability to analyze facial images", discusses how current advances in deep learning, especially in neural networks (CNN), can be leveraged to obtain very good insight. Using this method, pattern analysis features are extracted and then fed into the SVM classifier for training.

There is another way. For example, Rishma Rafique's research paper [16] "Deep fake detector and error level analysis and classification using Deep Learning" uses compression methods (also known as error level analysis) to classify Deepfake images. First, the proposed framework evaluates errors in the image to determine whether the image has been corrected. The images are then sent to a convolutional neural network for deep feature extraction. Support vector machine and K nearest neighbors were used to classify the feature vectors obtained after hyperparameter tuning. The method combined with KNN and residuals has the highest accuracy with a score of 89.5%.

Automation has become much easier thanks to some major advances in computer and communications technology over the last decade. Deepfakes that use deep learning to create fake photos and videos have their pros and cons. Although they aid in visual training and simulation, they raise issues such as identity theft and misrepresentation. This work investigates deep false detection methods, focusing on the contribution of mixed models to improving accuracy. The investigation aims to resolve problems arising from the widespread use of deepfakes [17]–[20].

III. METHODOLOGY

Deep learning techniques and algorithms are necessary for the creation of deepfake videos. We will be solving our deepfake video detection challenge with a deep neural network based solution in order to recognize fake videos. Since more convolutional layers might result in a large loss of data in video frames, we have opted to use hybrid models in conjunction with standard Convolutional Neural Networks. Therefore, by using a hybrid model, we are attempting to prevent this problem. We also wish to employ the Long Short Term Memory (LSTM) approach because the video frames we work with are sequential. The existing deepfake video detection techniques try to resolve the problem precisely, but they might not be able to persuade us of their dependability.Therefore, our goal is to employ explainable AI to make the key areas of the frames that the suggested model focuses on visually evident.

In summary, the following stages are taken as part of the methodology:

Step 1: Compiling the dataset. Step 2: Preparing the data.

Step 3: Labeling Data Step 4: Splitting the data.

Step 5: DenseNet169 Model with Bi-LSTM Proposal.

Step 6: Using Resnet50 and pre-trained models to create a hybrid model.

Step 7: Presenting the suggested hybrid model's functionality. Step 8: A comparative analysis of the proposed model with the current models.

A. DATASET

760 videos are included in the unique dataset that we created. Every video includes authentic and phony versions for training, testing, and validation. After extracting the frames from the videos using optical flow network algorithms [21], each frame is cropped for facial data. After removing the frames from the videos, it has 5362 train images, 1428 valid images, and 714 test images. The dataset is assembled using several Deepfake, GAN-based, and non-learned methods.

B. Data Pre-Processing

The first step we will do is preparing the data. This turns unprocessed data from a variety of sources into one more relevant data. Pre-processing data to remove any redundant information by missing or inconsistent values in the raw data this is very much needed to ensure that model is not giving unbiased results or there will be a huge data loss. Also, most simple Deep Learning algorithms require a standardized input in terms

of the task. This is how the dataset should be before training. At the time of pre-processing the data is in form of videos and we extract the frames from the videos using optical flow network algorithms.

C. Data Labeling

The custom dataset was divided into more than 18 frames of each videos. Each of the parts was pre-labeled with a .json file inside it containing the labels of images of each video being real of fake.

D. Data Splitting

Pre-processing and data segmentation or splitting were done at the same time. Based on a 7:2:1 ratio, the total dataset was divided into 80%, 20%, and 10% for training, testing, and validation data, respectively. The target size is (5, 128, 128, 3), and the divided datasets have a batch size of 32. In total, 760 videos were utilized, and 18 frames were typically extracted from each film; this means that 13,680 frames were used for training, validation, and testing.

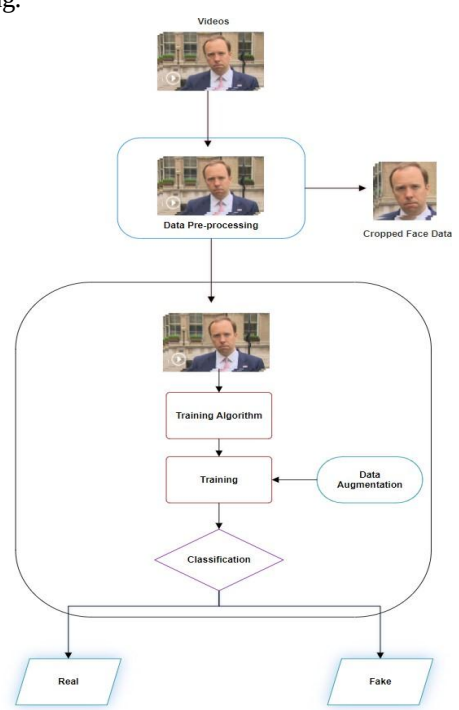

Fig. 1. Workflow of the classification model

IV. ARCHITECTURE

A. Proposed Hybrid model

The hybrid model consists of neurons that use function vectors to try to predict the position and properties of an object. The length of the vector function indicates the probability of the object appearing in the frame or image, while the index accounts for other factors such as rotation and thickness. Using the directional approach, each functional model at a given level can predict the index of higher level models.

B. Custom CNN Architecture

Specialized convolutional neural networks (CNNs) that are specifically designed for a type of problem or dataset. Typically it has the convolutional layers for feature extraction and then the pooling layers for subsampling and the whole layers for classification and so on. Decisions in its design, such as using several layers, various filter sizes and robustness, distinguish it from others. The hierarchy is an ongoing one, and consists in decisions being made under the conditions that currently prevail. Custom CNNs offer research and development ability to experiment with different network architectures and impact the way the network behaves until arriving at a result that meets our requirement. At runtime model we can extract desired features from input and make future predictions. Model parameters: Saves the parameter of model to be used in later time when trained.

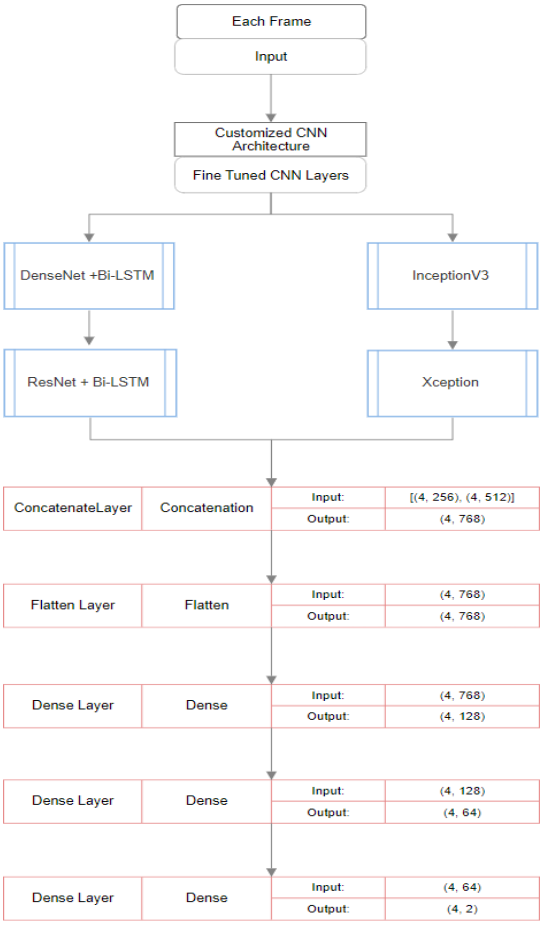

Fig. 2. Proposed Hybrid Model Architecture

C. *DenseNet$_B$iLSTMNet*

Bidirectional long-term memory (BiLSTM) networks and densely connected convolutional neural networks (CNN) form the DenseBiLSTMNet hybrid neural network architecture. DenseNet acts as a feature extractor in this design, extracting hierarchical features from the input data. The BiLSTM layer then obtains these features and uses them to represent physical relationships and data connections, thus capturing past and future context. This combination makes the model more capable to learn complex patterns both spatially and temporally, making it useful when working with connected data such as natural language processing for Lip recognizing and time estimation. DenseBiLSTMNet leverages the power of CNNs and LSTMs to be more efficient at tasks that require understanding connectivity and spatial patterns in data.

D. *Transfer Learning Model*

Transferring learning with ResNet50 - This includes a pre-trained model which was previously trained on a significant dataset (Ex:- ImageNet) so we will be fine-tuning the ResNet50 model to adjust it to our specific task or dataset. Rather than training a neural network from scratch, transfer learning makes use of the knowledge that ResNet50 has learned on ImageNet to accelerate training and improve performance on a fresh related task. During fine-tuning, the parameters of the pre-trained ResNet50 model are updated with the learning process through back-propagation using a smaller dataset aims to the target task, in order to switch the learned features to more proper ones fit the new data. This is an efficient technique as it uses the information already learned and significantly reduces training time which is very helpful when the target dataset is small or the resources on the computer are limited.

E. *Transfer Learning InceptionV3*

Transfer learning - which refers to using an InceptionV3 model, pre-trained on an entirely separate, much larger dataset, like ImageNet, to enhance the performance of a new task or dataset. This model adds layers to the InceptionV3 architecture [22] and trains the feature learning features and representations layer-wise instead of training the model from scratch. Small data: Fine-tuning defines a part of the pre-trained weight that is custom to fit to the data intricacies. This method can converge faster than training from scratch, and typically performs better, except when the new dataset is small. The architecture of InceptionV3 is especially appropriate for imaging applications such as deep feature extraction and classification, which are designed based on the advanced inception module, capturing rich spatial hierarchy in images.

F. *Transfer Learning Xception*

Transfer learning refers to applying the conclusion drawn from earlier tasks to new tasks with datasets of interest. Transfer learning involves borrowing knowledge from data rich tasks (e. g. ImageNet) to improve learning in data constraint tasks. We introduced Inception-inspired Design Xception [23] by learning differences between differences, which takes utmost advantage of horizontal propagations in hierarchies and augmented compensations for minimizing essential computations it demands. In doing research tweaking the Xception model, We use the pre trained xception weight as initial weight. It is then used on little processing data in order to tune the model, that is the incoming data can improve some of the characteristic of the model. This may allow the training and development of general resources in situations where data is scarce or computational resources are limited. By virtue of its depth and efficiency, Xception is suitable for tasks that need to be performed to enhance image classification and overall equipment.

V. RESULTS AND DISCUSSION

The performance of the hybrid model that was implemented can be summed up as follows:

Accuracy: The accuracy and effectiveness of the trained model helps to increase the efficiency of our hybrid model in validation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Our hybrid model performs well with 99.86% accuracy in training. We also found that our model achieves about 99.85% validation accuracy in terms of validation. The true accuracy of the model is represented by the true accuracy.

Precision: This is like a master archer accurately

targeting the bullseye, picking all true positives from massive data. This will make sure that every prediction will be precise and entirely improve in the relation to the false positives and then refining the direct hit, true positive with the level of accuracy like an artist. The accuracy value drops as FP rises because the denominator's value grows larger than the numerator's.

$$Precision = \frac{TP}{TP + FP}$$

The precision achieved by our model is 1.

Recall: Recall answers how accurately the true positives have been identified, by definition. It represents the percentage of a class that has been correctly classified out of a provided sample.

$$Recall = \frac{TP}{TP + FN}$$

Our Hybrid model has the recall value of 99.83%.

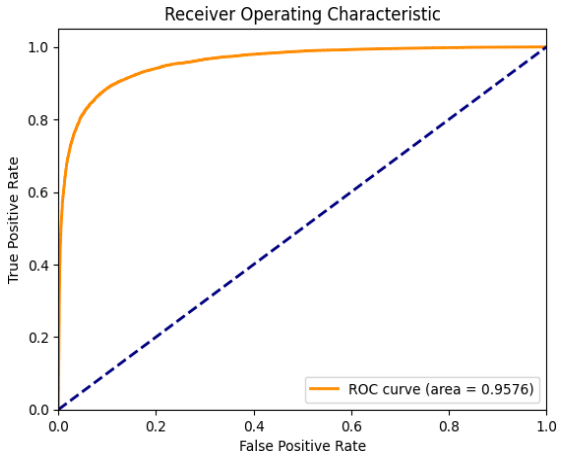

Fig. 3. AUC Curve

F1score: The F1-score is a measure that considers recall as well as precision, and its definition is as follows:

$$F1score = 2 * \frac{\text{precision} * \text{Recall}}{\text{precision} + \text{Recall}}$$

We also found that our model achieves about 98.1% F1 Score in terms of validation.

AUC: The performance of the model is the area under the curve (AUC) plot of the receiver operating characteristic (ROC) curve. On our training data, the AUC of our proposed model is 99.74%, while the generalization on our validation data is 95.76%. This causes the AUC of the model to work well.

VI. LIMITATIONS & FUTURE WORKS

While it shows great promise on existing data, we are trying to evaluate the effectiveness of our model for deep conflict learning using more data. The combination of many data sets will make the model more powerful. We plan to expand the dataset and test more frame rates to improve accuracy by adjusting things like brightness and contrast. We also plan to explore other areas such as iris tracking and audio processing. We plan to analyze the initial warm-up to detect the situation in fake video detection after using the AI annotation. By answering these questions, we can improve our models and guide future discoveries.

VII. CONCLUSION

This provides a comprehensive review of the latest developments in fake news and the threats they pose to the reliability and integrity of online information as a result of deepfake videos. We examine the techniques used to create deep, AI-driven, consistent data and how search engines deeply exploit these inconsistencies. Here, we propose a new hybrid model combining LSTM and convolutional neural network architectures. We used a series of pre-recorded images to train our model, paying attention to the facial region where most of the changes in deepfake videos occur. CNN architecture is used to identify feature vectors for both real and fake videos. The LSTM layer then used these feature vectors, which included spatial inconsistencies, to identify temporal discrepancies within these frames and predict the authenticity of the video from which these features were gathered.
In future, we will try to answer the question "why" are the videos which will be detected as fake are actually fake. In other words, what are the features that makes these videos fake by extending our model with the Explainable AI.

REFERENCES

[1] D. Pan, L. Sun, R. Wang, X. Zhang and R. O. Sinnott, "Deepfake Detec tion through Deep Learning," 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), Le icester, UK, 2020, pp. 134-143, doi: 10.1109/BDCAT50828.2020.00001.

[2] Rushikesh Potdar, Delp, "Deepfake Video Detection using Deep Learn ing", International Research Journal of Modernization in Engineering Technology

and Science, Volume:03/Issue:07/July-2021.

[3] M. T. Jafar, M. Ababneh, M. Al-Zoube and A. Elhassan, ”Forensics and Analysis of Deepfake Videos,” 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 053-058.

[4] Han Chen, Yuezun Li, Dongdong Lin, Bin Li, Junqiang Wu,Watching the BiG artifacts: Exposing DeepFake videos via Bi granularity arti facts,Pattern Recognition,Volume 135,2023,109179, ISSN 0031-3203.

[5] D. G¨ uera and E. J. Delp, ”Deepfake Video Detection Using Recurrent Neural Networks,” 2018 15th IEEE International Conference on Ad vanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639163.

[6] Suratkar, S., Kazi, F. Deep Fake Video Detection Using Transfer Learn ing Approach. Arab J Sci Eng (2022). https://doi.org/10.1007/s13369 022-07321-3.

[7] Almars, A. (2021) Deepfakes Detection Techniques Using Deep Learn ing: A Survey. Journal of Computer and Communications, 9, 20- 35. doi: 10.4236/jcc.2021.95003.

[8] Yuezun Li and Siwei Lyu. “Exposing deepfake videos by detecting face warping artifacts”. In: arXiv preprint arXiv:1811.00656 (2018).

[9] Shraddha Suratkar, Sayali Bhiungade, Jui Pitale, Komal Soni, Tushar Badgujar & Faruk Kazi (2022) Deep-fake video detection approaches using convolutional– recurrent neural networks, Journal of Control and Decision, DOI: 10.1080/23307706.2022.2033644.

[10] Akhil Sunil Kumar, Amruta Khavase, Himesh Rajendran, Deepfake Video Detection using Neural Networks, IJIRT 151259 INTERNA TIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOL OGY, Volume 7 Issue 12 — ISSN: 2349-6002, 2021.

[11] V.Phani Krishna, Deep Fake detection using LSTM and RESNEXT, Journal of Engineering Sciences, Vol 13 Issue 07,July/2022, ISSN:0377 9254.

[12] A. Malik, M. Kuribayashi, S. M. Abdullahi and A. N. Khan, ”DeepFake Detection for Human Face Images and Videos: A Survey,” in IEEE Access, vol. 10, pp. 18757-18775, 2022.

[13] S.K.Rout, B. Sahu, G.B.Regulwar, and V. Kavididevi, “Deep Learning in Early Prediction of Sepsis and Diagnosis. In 2023 International Conference for Advancement in Technology (ICONAT) (pp. 1-5). IEEE, 2023, January

[14] A Survey on Deepfake Video Detection,Peipeng Yu, Zhihua Xia, Jianwei Fei, Yujiang Lu, 2021

[15] A. R¨ossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, ”FaceForensics++: Learning to Detect Manipulated Facial Images,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 1-11.

[16] Rafique, R., Gantassi, R., Amin, R. et al. Deep fake detection and classification using error-level analysis and deep learning. Sci Rep 13, 7422 (2023).

[17] Y. Li, M.-C. Chang and S. Lyu, ”In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,” 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630787.

[18] M. T. Jafar, M. Ababneh, M. Al-Zoube and A. Elhassan, ”Forensics and Analysis of Deepfake Videos,” 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 053-058.

[19] Sharma, P., Kumar, M. & Sharma, H. Comprehensive analyses of image forgery detection methods from traditional to deep learning approaches: an evaluation. Multimed Tools Appl 82, 18117–18150 (2023)

[20] B. Malolan, A. Parekh and F. Kazi, ”Explainable Deep-Fake Detection Using Visual Interpretability Methods,” 2020 3rd International Confer ence on Information and Computer Technologies (ICICT), San Jose, CA, USA, 2020, pp. 289-293, doi: 10.1109/ICICT50521.2020.00051

[21] I. Amerini, L. Galteri, R. Caldelli and A. Del Bimbo, ”Deepfake Video Detection through Optical Flow Based CNN,” 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019, pp. 1205-1207.

[22] M. Masood, M. Nawaz, A. Javed, T. Nazir, A.

Mehmood and R. Mahum, ”Classification of Deepfake Videos Using Pre-trained Convo lutional Neural Networks,” 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2), 2021, pp. 1-6, doi: 10.1109/ICoDT252288.2021.9441519.

[23] Franc¸ois Chollet. “Xception: Deep learning with depthwise separable convolutions”. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 1251–1258.

[24] W. A. Jbara and J. H. Soud, ”DeepFake Detection Based VGG 16 Model,” 2024 2nd International Conference on Cyber Re silience (ICCR), Dubai, United Arab Emirates, 2024, pp. 1-6, doi: 10.1109/ICCR61006.2024.10533024.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. Ima geNet classification with deep convolutional neural networks. Commun. ACM 60, 6 (June 2017), 84–90. https://doi.org/10.1145/306538