Deep Fake Detection in Images and Video

[1]Sheethal HL, [2]Chandu Shree M, [3]Mohammed Ibrahim, [4]Vishal Patil, [5]Meena Deshpande (Prof.)
[1, 2, 3, 4, 5]Artificial Intelligence and Machine Learning, AMC engineering College
[1]sheethalhl123@gmail.com, [2]chandushreem452@gmail.com, [3]ibrahimm9139@gmail.com,
[4]Vishal6102004p@gmail.com, [5]meena.deshpande@amceducation.in

Abstract: The process of producing deepfakes has become much easier and quicker, which allows one to make rather convincing fake photos and videos. These may pose a major danger to integrity, privacy and security of our digital lives. Advanced changes are difficult to identify with traditional forensic tools such as metadata analysis and manual examination when artifacts have either been subtly or inconsistently dated. The system employs well chosen sets of real and fake images and videos of the various types of face changes, including both, features, which are spatial and temporal, and those obtained in a variety of frame sequences. The images that are uploaded are resized, standardized and undergo an organized workflow. They are divided into videos, face detection, and standardized 112112 transformations and then sequence is formed. They are two convolution-based classifiers, a simple CNN2D and an adapted model, which has more convolutional block and more effective regularization. The performance is judged based on standard measures and the modified CNN2D scored the highest 95.775 percent, which was higher than the baseline method. The approach enhances its detection capability of manipulation by using optimized preprocessing and powerful architectures alongside multi-mode analysis to ensure that the fake visual information is rightly detected.

***"Index Terms:** Convolutional Neural Networks (CNN), Deep Learning, Image Classification, Feature Extraction, Model Optimization, Pattern Recognition".*

1. INTRODUCTION

Due to advances in generative modeling and high-fidelity visual synthesis [1], production of deepfakes has become one of the most significant advances in the current artificial intelligence. With these techniques you are able to create fake faces, altered identities, fabricated scenes that appear very much like true human responses and behavior [2]. Due to this, deformed visual media have begun to disseminate using digital ecosystems. It has transformed the manner in which individuals view things, damaged the privacy of individuals, and challenged the age old notions of the trustfulness of visual evidence [3]. Since this technology has propagated so fast, deepfakes have become a significant issue in such fields as digital forensics, media security, and cyberdefense, where the veracity of visual data counts heavily on the decision-making process [4].

Although this is receiving increased focus, it remains difficult to appropriately distinguish between authentic and counterfeit pictures and videos. In the case of current synthetic generation pipelines, the classical forensic techniques such as manual analysis, metadata examination, and rule-based verification are difficult as the output of the pipelines resembles more and more natural images and complies with the laws of statistics [5]. Similarly, numerous existing methods of automatic detection are in essence concentrated on the image-based or video-based cues. This results in systems that are not robust, and adaptive, and lack cross-modal coherence when subjected to other forms of manipulation sources [6]. This disjunction demonstrates that there is a requirement in both spatial and temporal inconsistencies friendly approaches, as well as approaches that can be readily expanded to apply to the real world [7].

It is undertaking this in order to avoid these issues by establishing one detection tool, which is able to evaluate both images and videos within the same operating environment. This is to develop a user-friendly platform that enables individuals to interact in a safe way, enables multi-format deep fake evaluation and provides the output which can be understood and be helpful in actual verification scenarios [8]. This characteristic also emphasizes on architectural wholeness, end-to-end usability and extensibility such that the platform can be extended to accommodate the emerging modes of data manipulation and expansive datasets over time [9]. This holistic perspective is aimed at ensuring that the methods of deepfake identification are more useful, efficient, and applicable in more scenarios.

The larger significance of this work is that it may contribute to the enhancement of the validity frameworks of digital contents and reduce the potential risks of media manipulation to the society [10]. The system facilitates easier dealing with media environments that are increasingly becoming hard to navigate by consolidating different detection pathways and being more concerned with ease of use by the analysts, institutions and end users.

2. LITERATURE REVIEW

Research as to how to tell when media has been changed has increased in generative adversarial models and large-scale visual synthesis processes in the recent past. Zhang provides an overview of the big picture of the making and discovery of deepfakes. He dwells upon the fact that the quality of editing has changed and is becoming more difficult to notice small mistakes in fake content [11]. It is evident in this summary that recognition systems should be capable of evolving at any given moment in order to keep pace with the rapidly developing nature of generative models. Vahdati et al. dwell on artificial intelligence analysis of videos to prove this point of view. They demonstrate that recent edits are more consistent across time, which implies that appearance-only detectors cannot used individually [12]. In their work they demonstrate how significant the application of temporal cues is to achieve powerful recognition in a broad variety of manipulation conditions.

Korshunov and Marcel are the first to perform a vulnerability analysis of fake videos. It demonstrates that a significant number of mainstream biometric and face-recognition can fall out of operation when it comes to deepfaces attacks since most systems do not pay attention to the indicators of the manipulated information [13]. They demonstrate in their work that general-purpose recognition models do not perform when they are confronted with purposeful visual inconsistencies that have been introduced to deceive them. Tyagi and Yadav explore how to create fake images and video and note that under the current deepfake environment where the output of the fake generators is a counterfeit replica of natural statistics, such classical indications of forensic evidence as lighting patterns or splicing marks are not effective [14]. All these indicate that in the case of high quality simulated media, both old and early machine learning based approaches do not work well.

Other researches have focused on both integrated and behavior-based methods. Agarwal et al. introduce a new model which is founded on appearance and behavior. Some people say that because slight modifications in facial looks can reveal changes that would not be observed in the pictures that do not change [15]. Pan et al. demonstrate that detectors based on deep learning can often outperform handcrafted algorithms that were developed long ago, but also discuss the issue of cross-dataset generalization, where a detector that was trained to respond to a certain type of manipulation fails on a different one [16]. Heidari et al. provide a comprehensive description of the identification of deepfakes, and they emphasize the current issues such as bias in datasets, insufficient coverage in the real world, and susceptibility to adversarial attacks [17]. The issues demonstrate that we are yet to have flexible detection systems that would learn useful models that would not be influenced by change.

The diversity of datasets is highly significant to system generalization. Zi et al. introduce the WildDeepfake dataset that uses the uncontrolled environment of a real world and demonstrates that many of the best models fail when faced with variations in lighting, obstruction, and camera motion [18]. The idea of combining feature cues to increase the accuracy and create a smart detection system is proposed by Elpeltagy et al., but such method is still confined to testing in controlled conditions and cannot be applied on a larger scale [19]. Coccomini et al. advance the field of hybrid deep designs, which apply both convolutional and transformer-based components. Those are more effective with video datasets [20]. However these high-tech models still have issues in case they encounter new manipulation techniques or changes in distribution of which the model has never encountered.

In conclusion, the evidence provided above demonstrates that it is difficult to discover man-made material in a broad variety of different types, sources, and locations. The generalization, coverage of the dataset, time time modeling, and actual real world implementations remain big issues. The present research closes these gaps providing a unified system, which unites cross-modal analysis, user-friendly accessibility, and scalability in operation. Such an approach contributes to ensuring that the visual media can be more reliably and flexibly verified in digital spaces that keep evolving. This it does by ensuring that what is being spotted matches what is required in the real world.

3. MATERIALS AND METHODS

The proposed algorithm is grounded on the well-selected collections of real and fake images and videos containing various examples of face transformations, and it must be capable of accurately determining when visual data have been modified. The approach involves the utilization of one pipeline to receive information, standardize it, and prepare it to be subjected to spatial and temporal analysis. This allows modal consistency in processing. Processing of images is done with convolution-based classifiers i.e. a plain CNN2D and a better version that introduces more convolutional blocks as well as regularization that aid in conveying the difference between features. Video sequences are separated into frames and localized the faces to isolate the important regions then temporal modeling is applied. It consists of optimized preprocessing, structured normalization and multi-modal feature aggregation to enhance the robustness and generalization. This ensures that both the spatial cues which are still and those patterns which are dynamic are stored appropriately. Deepfakes are highly confidentially detected with ease using the overall approach. It is also fast to work with and can be applied in light-weight web based environments.

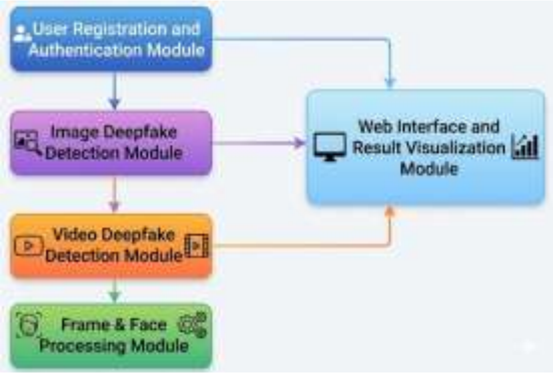

User Registration and Authentication Module
Image Deepfake Detection Module
Video Deepfake Detection Module
Frame & Face Processing Module
Web Interface and Result Visualization Module

Fig.1 Proposed Architecture

It is a full deepfake detection framework in the design of the system. It begins with the User Registration and Authentication Module that ensures that it is only legit users who access it. Image Deepfake Detection Module scans the images posted, and the Video Deepfake Detection Module scans the videos removing the frames and identifying the differences on the faces by the Frame and Face Processing Module. The entire outputs are sent to the Web Interface and Result Visualization Module that presents the recognition results in a manner that can be understood easily. Such design allows finding deepfakes in pictures and videos easy and quickly.

a) Dataset Collection:

The data of this system consists of the edited sets of real and fake pictures and modified video clips found online. It possesses numerous face-centered examples of very diverse people, lighting, and forms of manipulation. Each case is labelled as REAL or FAKE. It contains not only features of the spatial pictures but also time-ordered video frames. This renders the dataset highly helpful as it contains numerous varieties of data. It is a good dataset, which is used to train detection models because it contains many and various kinds of examples, and it presents a wide range of different ways to manipulate them. It aids the model to learn generalizing (as compared to other datasets which are limited in scope).

Fig.2 Image Dataset

Fig.3 Video Dataset

b) Pre-Processing:

Preprocessing prework prepares a protocol under which deepfakes in both images and videos are detected. This ensures that the quality of data, inputs and features representation is homogenous which is critical to good model inference and classification accuracy.

Image Data Pre-processing: Image data undergoes resizing, normalization, and formatting so that it can have the same type of input at any given time. Visual examples are rescaled to be the same size and converted into normalized numerical tensors which are used to stabilize gradient behavior throughout learning. This is a controlled change that reduces the difference that results due to resolution, devices, and lightings. This allows the network to specialise in face recognition. Such normalization simplifies the process of agreement and enhances the accuracy of detection of deepfakes in a wide variety of images.

Video Frame Extraction: Video information is decomposed into frames arranged sequentially so that the visual patterns can be studied on time and space. A change in visual state is depicted in each frame. This allows the system to detect minute errors that manifest themselves throughout the years in movies which were edited. Frame extraction ensures that time discrepancies, micro-expressions and motion errors are stored to be corrected at a later time. Replacing the continuous video streams with the structured frame sequences provides a sound foundation on the temporal models. It is also better at detecting a wide variety of manipulation styles.

Face Localization and Cropping: The most important elements of the face are identified and extracted out of the entire video frame to classify them. Finding faces causes the noise on the background to dissipate and reduces the impact of parts of the scene that are not significant, ensuring that the analysis is done on the facial features that deepfakes tend to make errors. Spatial alignment is also simpler with cropping and reduces the labor of the computer to input only those faces with clear boundaries. Such targeted extraction makes the system more robust and it is more effective in distinguishing the differences between videos in variety of settings.

Feature Transformation and Normalization: Controlled resizing and leveling procedures convert face pictures which have been taken out into standard feature representations. To reduce the influence of lighting variations, camera quality and pose, the transformation positions all samples to a constant spatial resolution and statistically normalizes them. These modifications ensure that there is stability in the distribution of features. It implies that the subsequent models will be capable of capturing the features of manipulation successfully. Unifications of change enhance generalization, enhance a model stability, and enhance the capability to observe minor variations in both fixed and sequential face data.

Temporal Sequence Formation: Frame representations are then made individual and then placed into patterns such that the sequence of the frames is maintained. This plays a critical role in the detection of inconsistencies in the changed content. When frames are placed in clips of the same length, it allows the models to view transitions, motion patterns and the level of coherence between the frames. These are the stuff that tend to transform real to fake movies. This systematic arrangement provides us with a general perspective of time, and this allows us to be aware of more of the errors that may occur in manipulation. The creation of such sequences is beneficial in terms of temporal models and video deepfake detection is more trustworthy.

c) Algorithms:

The **CNN2D** algorithm proposed is a naive feature extractor, which is expected to acquire the spatial patterns of input images at a high rate of accuracy. Through stacked convolutional and pooling layers, it gradually transforms low level visual data to high level representations which are highly significant in classification. It is more accurate, less noisy and can generalize across the different types of pictures, stronger through its hierarchical learning which is ordered. It is also a suitable architecture in the resource-constricted environment or real-time applications due to a good balance between cost and speed in terms of computing power.

The modified **CNN2D** technique alters the structure, which enhances the model ability to represent and classify more effectively. Improvement in deeper convolutions, improved filter layouts and further regularization mechanisms are some of the enhancements that allow the network to learn more intricate spatial associations without overfitting. These transformations allow distinguishing features more simply, assist the system to perform more effectively in more picture instances, as well as increase its vulnerability to miniature errors. Thus, the learning process is more stable and reliable using this longer framework than using standard convolutional models.

4. EXPERIMENTAL RESULTS

Accuracy: The specificity of a test is determined by the fact that it should be able to distinguish the sick and healthy cases. To have an idea of how accurate a test is, we will determine the percentage of true positives and the true negatives when compared to all the cases that were tested. This is mathematically expressed as.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} (1)$$

Precision: Precision is based on the percentage of examples or instances that are correctly identified out of those that are identified as positive. The formula of determining the accuracy is:

$$Precision = \frac{\text{True Positive}}{\text{True Positive } + \text{False Positive}} (2)$$

Recall: The parameter of machine learning that demonstrates the ability of the model to identify all the examples of a specific category is called recall. It is the proportion of observed positive which is actually predicted. This provides details of the extent to which a model is complete in modeling the instances of a particular class.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} (3)$$

F1-Score: The F 1 score is a measure of the accuracy of a machine learning model. It sums up the accuracy and recall scores of a model. The accuracy measure is checking the number of correct predictions that the model made on the entire dataset.

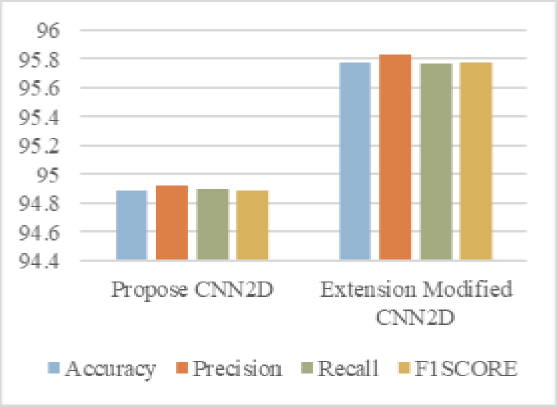
$$F1\ Score = 2 * \frac{Recall \text{ X } Precision}{Recall + Precision} * 100(1)$$

Table.1 Performance Evaluation Table

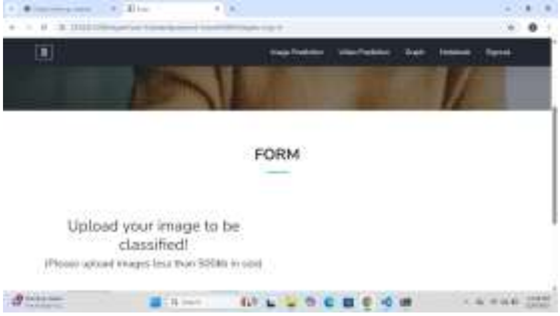
Algorithm Name	Accuracy	Precision	Recall	F1-Score
Propose CNN2D	94.890	94.921031	94.895695	94.889489
Extension Modified CNN2D	95.775	95.829520	95.768225	95.773232

According to the performance analysis, the Extension Modified CNN2D is more efficient than the base CNN2D, which is demonstrated in Table 1. The Extension Modified CNN2D is better at identifying distorted visual content and its accuracy is 95.775%.

Graph.1 Comparison Graph



The comparison graph indicates how the two convolutional models compose themselves against one another in crucial domains. It reveals that the Extension Modified CNN2D will always be more accurate, precise, recalls and has better F1-score compared to the baseline model. This demonstrates the fact that it is more efficient and dependable in detecting distorted images.

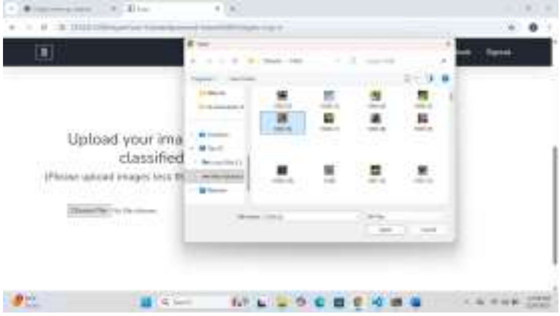


FORM

Upload your image to be classified!

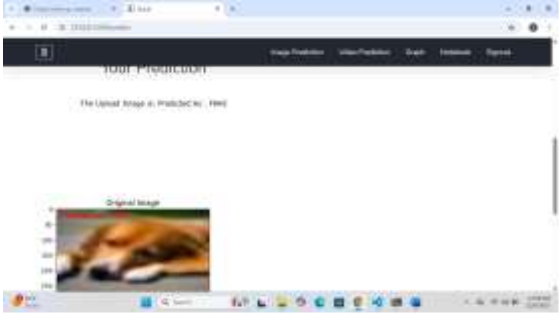
(Please upload images less than 500kb in size)

Fig.4 Home Page



Upload your ima
classified
(Please upload images less th

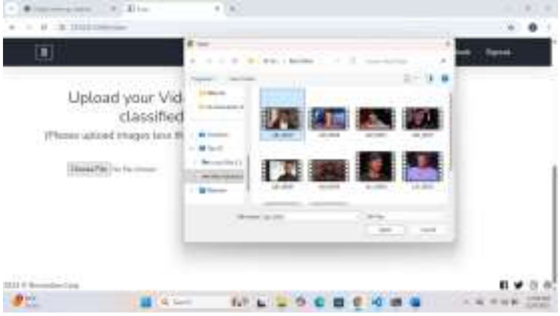
Fig.5 Upload Input Image



Your Prediction

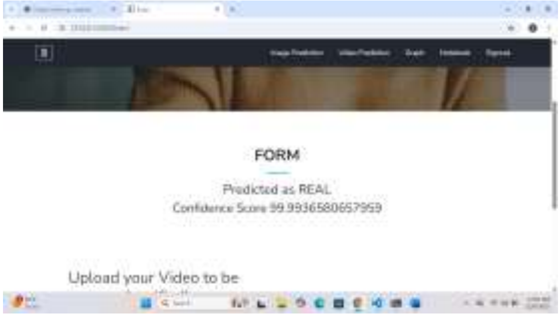
The Upload Image is Predicted as : FAKE

Fig.6 Predicted Results



Upload your Vid
classified
(Please upload images less th

Fig.7 Upload Input Video



FORM

Predicted as REAL
Confidence Score 99.9936580657959

Upload your Video to be

Fig.8 Predicted Results

5. CONCLUSION

In conclusion, the system has been designed to fulfill the increased demand of finding an effective method of identifying edited images and videos, and increase the trust in authenticating digital media. The technique involved the use of one system to integrate hand-selected real-and- fake picture sets and various types of video edits. This system is able to extract both still and moving features on content that is posted. The principal analysis sections were convolution-based classifiers that comprised of a simple CNN2D and an enhanced version of the earlier that was created to enhance the representational depth and regularization. The adapted CNN2D achieved a highest accuracy of 95.775 per cent, which showed that the architecture performed to make the distinction between actual material and counterfeit edits. The system was also expanded to facilitate support of time-informed video assessment and organized deployment with the help of light-weight web tool that enables easy usage by a number of people. In general, the approach is a stable means of uncovering the issues that have been enhanced to achieve automated verification procedures, enable individuals to decide on the information and create a more favorable digital integrity in reality.

The system may be extended in the future to support big video streams which play in real time. This would enable one to monitor digital platforms at all times to seek content that has been altered. Even more would be possible by using transformer-based topologies, multimodal fusion, and self-supervised learning to better generalize to previously unseen methods of deepfaking. Placing it on edge devices would be useful in fast verification, and applying it to explainable AI tools would make the forensic process more transparent. The model would also be strengthened in a broader scope of real-life contexts by adding additional data to the collection and cross-domain adaptation.

REFERENCES

[1] Malik, A., Kuribayashi, M., Abdullahi, S. M., & Khan, A. N. (2022). DeepFake detection for human face images and videos: A survey. Ieee Access, 10, 18757-18775.

[2] Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection: challenges and opportunities. Artificial Intelligence Review, 57(6), 159.

[3] Yu, P., Xia, Z., Fei, J., & Lu, Y. (2021). A survey on deepfake video detection. Iet Biometrics, 10(6), 607-624.

[4] Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS) (pp. 1-6). IEEE.

[5] Chadha, A., Kumar, V., Kashyap, S., & Gupta, M. (2021, May). Deepfake: an overview. In Proceedings of second international conference on computing, communications, and cyber-security: IC4S 2020 (pp. 557-566). Singapore: Springer Singapore.

[6] Ramadhani, K. N., & Munir, R. (2020, November). A comparative study of deepfake video detection method. In 2020 3rd International Conference on Information and Communications Technology (ICOIACT) (pp. 394-399). IEEE.

[7] Nasar, B. F., Sajini, T., & Lason, E. R. (2020, December). Deepfake detection in media files-audios, images and videos. In 2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 74-79). IEEE.

[8] Wodajo, D., & Atnafu, S. (2021). Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv:2102.11126.

[9] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397.

[10] Kharbat, F. F., Elamsy, T., Mahmoud, A., & Abdullah, R. (2019, November). Image feature detectors for deepfake video detection. In 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA) (pp. 1-4). IEEE.

[11] Zhang, T. (2022). Deepfake generation and detection, a survey. Multimedia Tools and Applications, 81(5), 6259-6276.

[12] Vahdati, D. S., Nguyen, T. D., Azizpour, A., & Stamm, M. C. (2024). Beyond deepfake images: Detecting ai-generated videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4397-4408).

[13] Korshunov, P., & Marcel, S. (2019, June). Vulnerability assessment and detection of deepfake videos. In 2019 International Conference on Biometrics (ICB) (pp. 1-6). IEEE.

[14] Tyagi, S., & Yadav, D. (2023). A detailed analysis of image and video forgery detection techniques. The Visual Computer, 39(3), 813-833.

[15] Agarwal, S., Farid, H., El-Gaaly, T., & Lim, S. N. (2020, December). Detecting deep-fake videos from appearance and behavior. In 2020 IEEE international workshop on information forensics and security (WIFS) (pp. 1-6). IEEE.

[16] Pan, D., Sun, L., Wang, R., Zhang, X., & Sinnott, R. O. (2020, December). Deepfake detection through deep learning. In 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT) (pp. 134-143). IEEE.

[17] Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(2), e1520.

[18] Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y. G. (2020, October). Wilddeepfake: A challenging real-world dataset for deepfake detection. In Proceedings of the 28th ACM international conference on multimedia (pp. 2382-2390).

[19] Elpeltagy, M., Ismail, A., Zaki, M. S., & Eldahshan, K. (2023). A novel smart deepfake video detection system. International Journal of Advanced Computer Science and Applications, 14(1).

[20] Coccomini, D. A., Messina, N., Gennaro, C., & Falchi, F. (2022, May). Combining efficientnet and vision transformers for video deepfake detection. In International conference on image analysis and processing (pp. 219-229). Cham: Springer International Publishing.

[21] Cozzolino, D., Rössler, A., Thies, J., Nießner, M., & Verdoliva, L. (2021). Id-reveal: Identity-aware deepfake video detection. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 15108-15117).

[22] Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. IEEE access, 10, 25494-25513.