SCIENTIFIC CULTURE

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 12527-12533
Open Access. Online & Print

DOI: 10.5281/zenodo.20475581www.sci-cult.com

DEEP FAKE DETECTION WITH DYNAMIC TEMPORAL WARPING

Gurpreet Kour Khalsa*[1], Rattan Deep Aneja[2], Rakesh Ahuja[3]

[1*]*Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India*
er.kaurgurpreet123@gmail.com
[2]*Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India*
rattananeja@gmail.com,
[3]*Department of Computer Science and Engineering, SRM University Delhi-NCR, Sonipat, Haryana, India*
ahuja2305@gmail.com

Abstract- The rapid advancement of deep learning technologies has led to the creation of incredibly lifelike deepfake films, which raises serious concerns about misinformation, security, and digital trust. Even though current deepfake detection techniques which are based on Long Short Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) have given better results by assuming the videos to be temporally consistent but perform appallingly in the frames where frame misalignment is very frequent. The frame misalignment which is induced because of variations in the frame rate, compression artifacts and inconsistent preprocessing has significant impact on the performance of the sequence based models. To overcome this challenge, this paper proposes a robust deepfake detection framework that combines CNN for spatial feature extraction, LSTM for temporal sequence modeling, and Dynamic Temporal Warping (DTW) for temporal alignment. The proposed method extracts frame-level features using a pre-trained CNN, addresses temporal inconsistencies by aligning feature sequences using DTW, and then describes temporal relationships for accurate classification using LSTM. Experimental results on benchmark datasets show that the proposed approach outperforms conventional CNN and CNN–LSTM models, particularly when there are misaligned video frames.

Keywords: Deep Fake Detection, Deep Learning, Frame Misalignment, Video Forensics, Computer Vision

Introduction

Due to the rapid advancements in the domain of artificial intelligence the Deepfakes or the forged media which appears to be quite realistic is becoming more approachable. The most prevalent method through which the pragmatic human features and audios can be created is Generative Adversarial Networks (GANs). Despite of the fact that this technology has found its wide usage in the field of broadcasting & entertainment, there are certain risks linked with it which includes dissemination of the erroneous information, impersonation and political maneuvering. This has led into the detection of deepfakes as a critical field of study in terms of digital media forensics. In recent years the deepfake approach has gained widespread popularity as a comprehensive field of examining images for manipulation [1]. The deep learning techniques are put into use to replace one person's face with another in order to produce manipulated images. It was referred to as "deep fakes" by Reddit users in 2017. It uses deep adversarial models, such as Generative Adversarial Networks (GAN), to turn celebrity faces into pornographic videos [2]. The primary problems with fake pornography are hoaxes, financial fraud, and fake news in the content. In order to produce phony digital images or videos, deep fakes primarily use deep learning and machine learning techniques to merge, replace, combine, and superimpose images [2]. Many software tools are available through which deep fake images are created without a programming knowledge and technical side background information. Usually the profile pictures from the social media are taken and fake images or videos are developed with a help of the expert. Security enhancement in the detection of face swap and the accuracy are very low. Generally there are three types of deepfake videos: face-swap, synthesis, and facial feature manipulation. In face-swap deepfakes [3] [4], an individual's face is replaced

with the source person's to produce a phony video that targets an individual for actions they haven't committed, potentially damaging their reputation. Another kind of deepfake known as lip-synching involves manipulating the target person's lips to change their movements in accordance with a specific audio track. The goal of lip-syncing is to have someone speak in the victim's attacker's voice. Deep fakes are created using puppet-master by mimicking the target's head, eye, and facial expressions [5].This is done to spread misleading information on social media by using fictitious profiles. Finally, deep audio fakes, also known as voice cloning, are used to alter a person's voice so that it sounds like the speaker they haven't actually spoken. As a result, finding the truth in the digital sphere has become more crucial. Since deepfakes are primarily used for malicious purposes and almost anyone can now create them using the tools already available, dealing with them is significantly more challenging. To date, numerous approaches have been developed to identify deepfakes. A conflict between poor and good deep learning applications has emerged since the majority are also deep learning-based. In order to address this issue, the US Defense Advanced Research Projects Agency (DARPA) initiated a media forensics research plan to create techniques for detecting fraudulent digital media [6] . In order to identify deepfakes from audiovisual information, a number of academics have investigated Machine Learning and Deep Learning (DL) fields in recent years [12]. Prior to the classification stage, the ML-based algorithms employ labor-intensive and inaccurate human feature extraction. Consequently, these systems' performance is erratic when handling larger databases. Nevertheless, these jobs are automatically completed by DL algorithms, which have been incredibly beneficial in a number of applications, including deep fake detection. Irrespective of the substantial analysis on deepfake detection there is always a greater margin of enhancement in terms of efficiency. The emphasis should be on the fact that these deepfake generation methods are quickly progressing which in turn leads to their reduced accuracy on the more complex datasets which are also increasing at the same rate. The deepfake content can have significant negative impact on human beings in terms of financial scams, political upheaval and character assassination. On the other hand the production of such contents with the rightful purpose on the social media has the potential to improve different types of fields & industries. Thereby these approaches increase the plausibility & reliability of internet and information disseminated on media [13]. Thus it is the need of hour to have such tools which can be relied upon for distinguishing between real and tampered information as deepfake technology id advancing on a greater scale and becoming easily available. Hence in this time of digital media it has become very crucial to have such reliable system for deepfake detection

Literature Survey

In order to address the issues in the current state of art , here the author introduced a multimodal deepfake detection framework. The sequential, time domain and spatial analysis are amalgamated by utilizing sophisticated deep learning architectures. Thus the framework based on multimodal deep fake detection is implemented by making use of the deep learning methods such as EfficientNetV2, 3D CNN & Bi-LSTM along with the transport layers which in turn enhances the detection accuracy , stability and efficiency for real world applications in AI generated content verification. In conclusion, the experimental results presented in this paper demonstrate that the multimodal deep fake detection framework greatly outperforms current approaches in terms of accuracy, robustness, and efficiency, thereby bolstering digital security against deep fake technology [7]. A hybrid deep learning model is given for deepfake detection by the author [8]. The integration of content & trace feature extraction improves overall robustness and detection accuracy. It demonstrates the ability to identify subtle manipulations, even when media has undergone challenging conditions such as video compression. The effectiveness of the model is further supported by Class Activation Maps (CAMs), which help to pinpoint critical areas in the media that are indicative of manipulation [8]. This paper [9] introduces a novel audio-visual aware multimodal Deepfake detection framework designed to enhance the identification of forgery clues. The framework addresses the challenges posed by advanced Deepfake techniques that modify both video and audio information, and the limitations of existing multimodal methods in exploring intra-modal and cross-modal forgery clues [9]. This module is proposed to magnify temporal intra-modal defects. It achieves this by focusing on sequence-level relationships within the data. The DDIC module utilizes Jensen–Shannon divergence to adaptively align multimodal information. Its purpose is to further magnify

cross-modal inconsistencies, which are crucial indicators of Deepfake content. The framework also delves into spatial artifacts by connecting multi-scale feature representations. This provides comprehensive information for detection. Experiments conducted demonstrate that the proposed framework surpasses the performance of independently trained models Furthermore, the framework exhibits superior generalization capability when tested on unseen types of Deepfake content [9]. The author presented an Explainable Deepfake Detection Framework utilizing a multi-stage deep learning pipeline with spatial, temporal, and frequency feature extraction, achieving state-of-the-art results on the FaceForensics++ dataset, enhancing both detection accuracy and interpretability for forensic experts [10]. In order to fill in current research gaps, the paper [11] suggests a new framework for DeepFake detection using deep learning techniques, integrating CNNs, Transformers, and Bi-LSTM along with Explainable AI (XAI) to improve accuracy and interpretability in detecting both images and videos.

Existing Systems

The prevailing systems have found out that Convolutional Neural Networks (CNNs) are the main tool used for feature extraction and classification in the existing deep fake detection systems. Regardless of it, the deficiency in terms of temporal consistency makes it very complex for the systems to identify forged content. While traditional CNN-based models, like Xception Net, ResNet, and VGG-16, have demonstrated excellent accuracy for fake images, they are unable to recognize the sequential dependencies between frames in deepfake films. Another kind of detection model uses Long Short-Term Memory (LSTM) or Recurrent Neural Networks (RNNs) to examine temporal disparities in videos. Nevertheless, stand-alone LSTM-based techniques are less successful for frame-by-frame deepfake detection since they frequently lack reliable spatial feature extraction.While some hybrid models combine CNNs and RNNs to increase detection accuracy, their high computational costs limit their real-time usefulness. Adversarial attacks, in which deepfake models are optimized to avoid detection techniques, also pose a challenge to most CNN-based detection systems. The performance of detection has been improved by recent developments in transformer-based models and attention mechanisms. The performance of detection has been improved by recent developments in transformer-based models and attention mechanisms. Deepfake detection systems primarily use CNN-based models, such as XceptionNet and ResNet, to extract features, but they struggle with temporal consistency in videos. RNNs and LSTMs are not very good at extracting spatial features, but they are helpful for analyzing sequential discrepancies. Although hybrid CNN-RNN models increase accuracy, their computational cost restricts their use in real-time. Current models that were trained on big datasets frequently overfit to particular kinds of deepfakes, which limits their ability to generalize. Detection systems are further challenged by adversarial attacks, and transformer-based models are promising but resource-intensive. Additionally, when an automated system's capacity to explain itself is inadequate, trust in it declines. A reliable, scalable hybrid model that balances accuracy, efficiency, and real-time detection is required. Furthermore, the inability of current deepfake detection systems to explain their decisions makes it difficult to interpret and validate them, which lowers confidence in automated detection. In conclusion, the different deepfake detection techniques currently in use lack a balanced trade-off between accuracy, computational efficiency, real-time applicability, and generalization across different deepfake techniques, necessitating the development of a hybrid model that is both more robust and scalable.

Proposed System

The proposed system aims to improve deepfake detection by incorporating a Hybrid MobileNet LSTM model for real-time image and video analysis. The widespread production of extremely realistic deepfake films as a result of the quick development of deep learning algorithms poses major risks to the security, authenticity, and public confidence of information. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid CNN–LSTM architectures have been the basis of many deepfake detection systems, but these techniques mainly rely on learning spatial and temporal features from video frames under the assumption of consistent and well-aligned frame sequences. However, frame misalignment frequently occurs in real-world video data because of things like asynchronous video processing, fluctuating frame rates, frame losses, and compression artifacts. Due to this mismatching pattern of frames the temporal continuity of facial expressions, synchronization of the lips , blinking of eyes & head motion are all affected by this and they prove to be very important for deepfake detection system. In continuation to it the models which are completely sequence based like that of LSTM are not in a

position to exhibit these temporal dependencies which in turn results into the generalization problem across the different datasets and thereby decreases performance & increase in the rate of incorrect classification. The current state of art methods when applied on real world domain do not provide much accurate results as they lack the ability to handle these temporal anomalies and misalignment sequences. This deficiency of the current systems requires a much more robust framework which can handle the discrepancies due to these frame misalignment and modeling the spatiotemporal characteristics. The most pivotal and regular matter in deepfake detection systems which completely rely on deep learning methods taking into consideration the spatiotemporal analysis of video data is the misalignment of frames. Basically the frame misalignment is the disruption in the actual pattern, synchronization or the frame continuity in the given video due to temporal inconsistencies or sequencing which is uneven. In real world domain the most classical basis of this problem are frame rate variations, deletion or duplication of the frames while compression and transmission, preprocessing activities like that of detection of face, cropping and rescaling. In GAN based systems the generation of deepfake system treats frames individually thereby decreasing the chance to maintain coherent temporal dynamics across these frames. This in turn results into the continuous distortions in the motion of the input frames before the start of detection process. The models based on deep learning specially focusing on the Long Short Term Memory (LSTM) networks & transformers based systems which require the input frames following a regular temporal progression are drastically impacted by frame misalignment. If the model is not capable to learn the important temporal dependencies then it results into inaccurate interpretation of motion patterns. When the frames are misaligned then the minor indication like that of synchronization in the lips, the blinking rate of eyes and micro expressions appear to be warped or conflicting with each other. In the mean time when noise is induced in the feature space then the models ability to capture these temporal patterns is reduced. This results into the generation of false positives and negatives which ultimately results in the degradation of overall detection accuracy. The multimodal deepfake detection algorithms which depend on synchronization of these audio and visual streams have greater issue in terms of dealing with these frame misalignments. Because of the artificial desynchronization between lip and voice movements brought on by misaligned frames, inconsistencies in edited footage may be hidden or actual videos may seem suspicious. The another major crucial problem is the generalization problem where the actual videos have notable temporal abnormalities because in present times the most of the deepfake detection algorithms are trained on the datasets which have been intentionally trained on these frames which have well framed alignment whereas real world videos have noticeable temporal abnormalities. Thereby this problem of mismatching limits the usage of the model in real world videos leading to poor cross dataset performance. Deepfakes themselves may produce minute temporal distortions when combined with spontaneous frame misalignment, making it more difficult for detection systems to distinguish between actual errors and altered content, leading to a confusing effect. Irrespective of these much discrepancies many current systems assume the temporal conditions to be perfect and making use of sequence base models by concentrating on spatial feature extraction using CNN and completely ignoring the issue of frame misalignment. This suggests a clear research gap in the development of trustworthy deepfake detection frameworks that can effectively handle temporal distortions. Therefore, improving the robustness, reliability, and usefulness of deepfake detection systems requires fixing frame misalignment. Even in the presence of irregular frame sequences the technique like Dynamic Temporal Warping, adaptive frame sampling and attention based sequence modeling make the model capable to learn consistent temporal relationships. In order to improve detection accuracy and reliability in the presence of misaligned video frames, this research attempts to develop an improved deepfake detection system that combines robust temporal alignment techniques, like Dynamic Temporal Warping (DTW), with spatial feature extraction.

This study is based on integration of spatial feature extraction, temporal sequence modeling & alignment techniques to handle the misalignment of frames in video data to provide a robust deepfake detection framework. The approach followed here comprises of several steps which includes pre processing, then feature extraction, temporal modeling, alignment of sequences and classification. In the first step which is preprocessing the image is processed to make it clean and free of distortions. The videos are broken down into individual frames at a given sample rate of 10 FPS in order to reduce the impact of

fluctuating frame rates on these videos. Further face identification methods are applied which helps in extracting facial frame features which in turn resizes and normalizes it to standard input dimensions best suitable for deep learning models. Then these frames are presented into the fixed length sequences to reduce the inconsistency. In terms of real world scenarios where such discrepancies such as missing or redundant frame do exist there still can be the problem of temporal misalignment which can be further handled in future iterations. Then Convolutional Neural Network (CNN) is applied on the individual frames to extract the spatial information out of it. In order to make it sure that the training time is shorter and feature represent is efficient enough the pre trained architectures like XceptionNet or ResNet are utilized. Then each frame is processed independently by CNN to learn discriminative features having reference to facial textures, artifacts and inconsistencies. Then CNN produces a high dimensional feature vector for each frame which acts as an input for temporal modeling. The extracted feature vectors received in this stage are passed on to Long Short-Term Memory (LSTM) which checks in for temporal dependencies between frames. There are certain patterns which are related to motion dynamics, lip synchronization, eye blinking & head movements which are being learnt by LSTM by processing these frames. Inspite of learning process still LSTM is vulnerable to the misalignment in frames because its potency is heavily reliant on the proper temporal order of frames. Thus to solve this problem of misalignment Dynamic Temporal Warping (DTW) is utilized as a contrivance to solve these frame misalignment problems. The feature vector sequences generated in the upper phases vary in length then DTW aligns them by bringing down the distance between temporally distorted sequences. Irrespective of the fact that there are inconsistencies in frame order or timing this approach makes it sure to accurately match the temporal patterns. This approach gradually decreases the temporal noise & further enhances the uniformity of the input sequences which are fed into the LSTM model by using DTW. This step improves the model's ability to learn precise temporal correlations and fortifies its resistance to real-world video distortions. To achieve the binary classification the LSTMs output which is aligned temporal characteristics are passed on to fully connected dense layer which applies sigmoid activation function on it. In order to train the model, the loss functions like that of binary cross entropy are employed & Adam optimizer method is used to update the network weights associated to it. Using the training validation test split the given model is trained on labeled datasets. The metrics which have been used to evaluate the performance are accuracy, precision, recall , F1 – score and Area Under the Curve(AUC). In order to check the models capacity for generalization under different circumstances the cross dataset evaluation is also carried out. To further increase the productivity and performance the data normalization, drop out regularization and data augmentation is used.

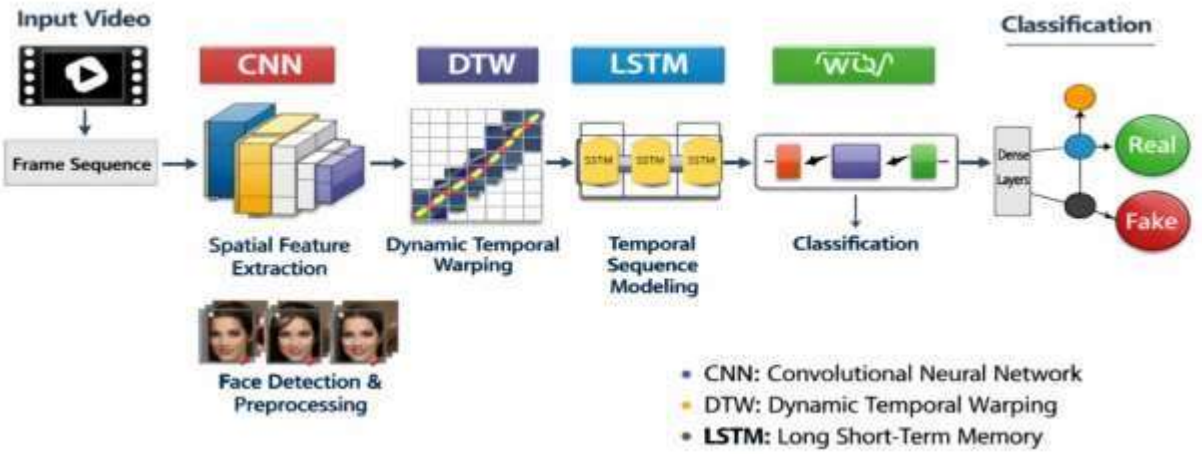

Fig.1. Proposed framework

Results

In order to access the given deepfake detection methodology the benchmark datasets like FaceForensics++, Celeb-DF & DFDC are used. The dataset is split into training, validation and testing sets to ensure an impartial evaluation of the model performance. The main preprocessing techniques which have been used for all video samples are face

detection, normalization & fixed length frame sequencing. In the models implementation CNN is make into use for feature extraction further LSTM is responsible for temporal modeling and Dynamic Temporal Warping (DTW) is used for alignment of sequences. In terms of all the datasets the proposed model achieves high detection performance exhibiting the effectiveness of hybrid approach of spatial , temporal and alignment features.

Model	Accuracy	Precision	Recall	F1Score
CNN	88%	86%	85%	85%
CNN+LSTM	92%	90%	91%	90%
CNN+LSTM+DTW	96%	95%	96%	95%

Table.1.Overall Performance of the proposed model

To analyze the effect of frame misalignment, experiments were conducted on aligned video sequences and misaligned video sequences

Condition	CNN+LSTM	Proposed Model
Aligned Frames	93%	96%
Misaligned Frames	82%	94%

Table.2. Frame misalignment rates

The findings unequivocally show that frame misalignment has a major impact on how well conventional deepfake detection methods work. Although CNN-only models are good at catching spatial artifacts, they are unable to identify temporal discrepancies. Similar to this, CNN + LSTM models enhance performance by discovering temporal patterns; however, when input sequences are mismatched, their efficacy is diminished. In order to overcome this restriction, Dynamic Temporal Warping (DTW) integration is essential. DTW guarantees that identical motion patterns are accurately matched even when frames are unevenly spaced or temporally deformed by matching feature sequences prior to temporal modeling. This results into the LSTM network learning temporal features more reliably and precisely. The enhanced capacity for generalization is also one of the significant finding in the given proposed model. This method has proven to give constant performance across different datasets with different video quality effects as compare to the traditional methods which work well on certain limited datasets. Thereby making sure that this model works best for practical implementations where there is greater rate of uncertainty in the given video settings. The model exemplifies against typical video deformation such as variations in the frame rate & constricted glitches. In terms of social media platforms the content which is being disseminated on these platforms and which goes through several steps of transitions also is considered to be very crucial. But at the same time the alignment step based on DTW makes it more computationally complex thereby increasing the processing time for the video clips which are quite lengthy. For the purpose of minimizing the computational cost involved the future research may focus on the more enhanced alignment methods or attention base lightweight procedures. This incorporation of temporal alignment along with DTW assures the enhanced working of deepfake detection systems. The proposed CNN–LSTM–DTW framework effectively addresses the challenges posed by frame misalignment and provides a more reliable method of detecting deepfake videos in real-world scenarios.

Conclusion and Future Scope

In this study a novel technique for deepfake detection based on one of the critical problem of frame misalignment is given. As compare to the traditional approaches which generally focus on spatial or sequential feature learning under given conditions this method integrates the spatial feature extraction using CNN, then temporal sequence modeling which is done using LSTM and finally temporal based alignment using Dynamic Temporal Warping (DTW). This hybrid approach allows the model to look into the inconsistencies and peculiarity based on temporal & spatial domain and thereby reducing the adverse effects of misaligned frame sequences. In terms of experimental findings it has been found out that misalignment in frames has a greater impact on validity and generalization of deepfake detection algorithms. The temporally distorted sequences in the scenario are effectively aligned by using DTW which in turn allows LSTM model to uncover the notable temporal relationships. In order to evaluate it on cross-dataset performance, robustness and detection accuracy the system surpasses the

baseline models. The main focus is on the alignment step due to which the model becomes computationally complex but still managing its efficiency. The future work could be focused on improving efficiency without compromising the performance of these models and specially focusing on optimization strategies, structures which are quite light weighted or attention based models. In the existing systems the detection capabilities can be enhanced by exaggerating it to multimodal deepfake detection by adding textual and audio information into it. Thus the proposed study shows that to create dependable and efficient deepfake detection systems the frame misalignment is addressed and the suggested approach based on CNN-LSTM-DTW framework offers a possible a path for additional research in this field.

References

[1] Citron DK. 2019. How deepfakes undermine truth and threaten democracy. Available at https://www.Ted.com/talks/danielle_citron_how_deepfakes_undermine_truth_and_threaten_democracy?language=en.

[2] Maras M.-H, Alexandrou A. 2019. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. *The International Journal of Evidence & Proof 23(3):255–262 .*

[3] Boylan, J. F. Will Deep-Fake Technology Destroy Democracy (Te New York Times, 2018).

[4] Harwell, D. Scarlett Johansson on fake AI-generated sex videos:'Nothing can stop someone from cutting and pasting my image'. J. Washigton Post 31, 12 (2018).

[5] Masood, M. et al. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Appl. Intell. 53, 1–53 (2022).

[6] 5. Turek, M.J. Defense Advanced Research Projects Agency. *https://www.darpa.mil/program/media-forensics*. Media Forensics (MediFor). Vol. 10 (2019).

[7] Duraimurugan, N., Kumar, P., & Jayaprakash, R. (2025, July). A Multimodal Deep Fake Detection Framework for Robust and Efficient Synthetic Media Identification. In *2025 International Conference on Information, Implementation, and Innovation in Technology (I2ITCON)* (pp. 1-7). IEEE.

[8] Vinod, C. (2025). A Hybrid Deep Neural Network for Multimodal Deepfake Detection. Indian Scientific Journal Of Research In Engineering And Management, 09(04), 1–9. https://doi.org/10.55041/ijsrem45707

[9] Liu, X., Yu, Y., Li, X.-L., & Zhao, Y. (2023). Magnifying multimodal forgery clues for Deepfake detection. Signal Processing-Image Communication, 117010. https://doi.org/10.1016/j.image.2023.117010

[10] Scientific, L. L. (2025). Expendable Deepfake Detection Using Multi-Stage Deep Learning With Spatial, Temporal, And Frequency Forensic Pipelines. *Journal of Theoretical and Applied Information Technology*, *103*(7).

[11] Wagh, A., Rane, R., & Pinjarkar, L. (2025, August). Enhancing DeepFake Detection through Deep Learning and Explainable AI. In *2025 International Conference on Artificial Intelligence and Machine Vision (AIMV)* (pp. 1-6). IEEE.

[12] 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.

[13] Suganthi, S. T., Ayoobkhan, M. U. A., Bacanin, N., Venkatachalam, K., Štěpán, H., & Pavel, T. (2022). Deep learning model for deep fake face recognition and detection. *PeerJ Computer Science*, *8*, e881.