International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 7.2 Website: https://ijarmt.com ISSN No.: 3048-9458

Deepfake face Detection using Deep Learning Technique: A Symmetric Review

Shelke Sanket Babasaheb[1], Prof. Suresh S. Gawande[2]
M. Tech. Scholar, Department of Electronics and Communication, Bhabha Engineering Research Institute, Bhopal[1]
Guide, Department of Electronics and Communication, Bhabha Engineering Research Institute, Bhopal[2]

Abstract: The rapid advancement of artificial intelligence and deep learning has enabled the creation of highly realistic synthetic facial content, commonly known as deepfakes. Although deepfake technology has beneficial applications in entertainment, education, and digital media, its misuse poses serious challenges related to misinformation, identity theft, privacy, and digital security. Deepfake face detection has therefore emerged as an important research area in computer vision and multimedia forensics. This symmetric review provides a comprehensive analysis of deep learning-based approaches developed for detecting manipulated facial images and videos. The review examines conventional Convolutional Neural Networks (CNNs), transfer-learning models, recurrent architectures, and advanced networks such as ResNet, DenseNet, EfficientNet, and Vision Transformer-based approaches. It systematically compares these techniques in terms of detection accuracy, robustness, computational complexity, dataset requirements, and generalization capability. Particular attention is given to facial preprocessing, frame extraction, feature representation, spatial and temporal artifact analysis, data augmentation, and evaluation metrics. The review also discusses widely used deepfake datasets and highlights challenges associated with compression, unseen manipulation techniques, cross-dataset performance, and adversarial attacks. Finally, emerging research directions, including multimodal learning, explainable artificial intelligence, lightweight detection models, and hybrid spatial-temporal architectures, are discussed. The study provides a structured perspective on the current state of deepfake face detection and identifies promising opportunities for developing more reliable, generalizable, and real-time detection systems.
Keywords: Deepfake Detection, Face Forgery, Deep Learning, Convolutional Neural Network, Transfer Learning, Computer Vision, Video Forensics, Artificial Intelligence.

I. INTRODUCTION

The rapid development of Artificial Intelligence (AI), particularly Deep Learning (DL) and Generative Adversarial Networks (GANs), has significantly transformed the field of digital media generation. One of the most prominent outcomes of this technological advancement is deepfake technology, which enables the creation or manipulation of facial images and videos by replacing, synthesizing, or modifying a person's identity and facial characteristics.

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 7.2 Website: https://ijarmt.com ISSN No.: 3048-9458

Deepfakes can generate highly realistic content that is often difficult to distinguish from authentic media through visual inspection alone. Although this technology has legitimate applications in entertainment, education, filmmaking, and digital communication, its malicious use has raised serious concerns related to misinformation, identity theft, privacy violations, cybercrime, and reputational damage.

Deepfake face detection has consequently become an important research area within computer vision and multimedia forensics. Traditional image-processing and forensic techniques generally depend on manually designed features, such as image artifacts, inconsistencies in facial regions, abnormal texture patterns, or compression characteristics. However, modern deepfake-generation techniques are becoming increasingly sophisticated and can minimize many visible artifacts. As a result, conventional approaches may fail when manipulated content is generated using previously unseen methods. Deep learning provides an alternative by automatically learning complex and discriminative representations from large collections of authentic and manipulated facial images or video frames. Convolutional Neural Networks (CNNs), in particular, have demonstrated strong capabilities in extracting spatial features associated with facial manipulation.

Recent research has explored several deep learning architectures for improving the reliability of deepfake detection. Raza et al. [1] investigated a deep learning-based approach for identifying deepfake images, demonstrating the potential of automatically learned visual features for distinguishing genuine and manipulated content. Zobaed et al. [2] examined machine learning approaches for detecting forged and synthetic media and emphasized the cybersecurity implications of deepfakes. Thambawita et al. [3] further demonstrated the broader capability of generative models by producing synthetic ECG signals, highlighting that AI-generated content can extend beyond conventional visual media and create significant privacy concerns. Ahmed et al. [4] emphasized public awareness as an additional resistance mechanism against the harmful use of deepfakes. Nirkin et al. [5] introduced a context-aware detection approach that identifies discrepancies between manipulated faces and their surrounding visual environment, showing that contextual information can provide important evidence for detecting facial manipulation.

Despite significant progress, deepfake detection remains challenging because manipulated content can differ substantially in resolution, compression level, illumination, facial pose, generation technique, and source dataset. A detector trained on one type of manipulation may also experience reduced performance when applied to unseen deepfake-generation methods. Furthermore, real-world social-media videos are frequently compressed and resized, which can remove subtle manipulation artifacts. Therefore, a robust detection system should achieve not only high classification accuracy but also strong generalization, computational efficiency, robustness, and cross-dataset performance.

In this context, this review presents a systematic overview of deepfake face detection using deep learning techniques. It examines existing approaches based on CNNs, transfer learning, spatial feature extraction, temporal analysis, and context-aware detection. The review also compares the major datasets, evaluation metrics, advantages, and limitations reported in

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 7.2 Website: https://ijarmt.com ISSN No.: 3048-9458

previous studies. Finally, emerging directions such as hybrid spatial-temporal networks, Transformer-based architectures, explainable AI, multimodal detection, and lightweight real-time models are discussed to identify opportunities for developing more accurate and generalized deepfake detection systems.

II. LITERATURE REVIEW

Raza et al. [1], approach for detecting deepfake images. The study focused on identifying visual inconsistencies introduced during the generation or manipulation of facial images using deep learning techniques. The authors emphasized that conventional image-forensic methods often struggle with highly realistic synthetic images, whereas deep learning models can automatically learn discriminative features from manipulated and authentic samples. The proposed approach analyzes facial image characteristics and extracts relevant visual patterns to distinguish genuine images from deepfake content. The study demonstrated the potential of deep learning for automated deepfake detection and highlighted the importance of robust feature learning for improving classification performance. The work provides a foundation for developing more reliable deepfake detection systems, although challenges related to dataset diversity, unseen manipulation methods, and generalization across different image sources remain important areas for further investigation.

Zobaed et al. [2] presented a comprehensive study on the detection of forged and synthetic media using machine learning techniques. The authors discussed the increasing threat posed by deepfakes and other manipulated media in the context of cybersecurity. Their work examined different machine learning approaches that can be employed to identify artificial modifications in multimedia content. The study emphasized the importance of extracting meaningful features from manipulated media and using appropriate classification algorithms to differentiate genuine content from forged content. In addition to technical detection approaches, the authors discussed the broader cybersecurity implications of deepfakes, including misinformation, identity manipulation, fraud, and privacy violations. The study concluded that machine learning provides an effective mechanism for detecting synthetic media, but the continuous evolution of deepfake-generation techniques requires detection models to be regularly improved and trained on diverse datasets.

Thambawita et al. [3] investigated the generation of synthetic electrocardiograms (ECGs) using Generative Adversarial Networks (GANs), demonstrating that deepfake technology is not restricted to facial images or conventional multimedia. The researchers showed that GAN-based models can generate highly realistic synthetic medical signals that may resemble authentic ECG recordings. This work highlighted a significant privacy concern because sensitive medical information can potentially be reconstructed, synthesized, or manipulated using generative models. Although the study focused on ECG signals rather than face images, it is highly relevant to deepfake research because it demonstrates the broader capability of

generative artificial intelligence to create realistic synthetic data. The findings emphasize the need for reliable detection and authentication mechanisms capable of distinguishing genuine data from AI-generated content, particularly in privacy-sensitive applications such as healthcare.

Ahmed et al. [4] investigated awareness of deepfake technology and proposed awareness as a potential resistance mechanism against its harmful consequences. The study highlighted that the rapid development of deepfake technology has created significant social and security concerns, particularly because ordinary users may have difficulty distinguishing manipulated content from authentic media. The authors emphasized that technical detection mechanisms alone may not be sufficient to address the deepfake problem and that user awareness and digital literacy are also important components of resistance. Increasing public knowledge regarding deepfake characteristics, potential risks, and verification practices can help reduce the impact of manipulated media. The study therefore provided a complementary perspective to deep learning-based detection research by focusing on the human and societal aspects of deepfake mitigation.

Nirkin et al. [5] proposed a deepfake detection method based on discrepancies between a person's face and its surrounding context. Unlike approaches that primarily analyze facial regions for visual artifacts, the proposed method considers the relationship between the manipulated face and the contextual information surrounding it. Deepfake generation processes can produce inconsistencies in factors such as facial appearance, geometry, lighting, and the relationship between the face and its environment. By learning these discrepancies, the proposed approach can identify manipulated facial content more effectively. The study demonstrated that contextual information can provide valuable forensic evidence for deepfake detection and can complement conventional face-based analysis. This work is particularly significant because it moves beyond analyzing isolated facial artifacts and considers the broader visual context, providing a promising direction for improving the robustness and generalization of deepfake detection systems.

Dolhansky et al. [6] introduced the Deepfake Detection Challenge (DFDC) Preview Dataset as an initial dataset for developing and evaluating deepfake detection algorithms. The dataset was designed to address the growing requirement for large-scale and diverse facial manipulation data that could be used to train machine learning and deep learning models. It contains authentic and manipulated facial videos generated using different face-manipulation techniques, allowing researchers to study the visual characteristics associated with deepfake content. The authors highlighted the difficulty of detecting manipulated videos because modern synthesis techniques can produce highly realistic facial content. The DFDC Preview Dataset provided an important benchmark for the research community and encouraged the development of automated detection systems. However, its relatively limited scale compared

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 7.2 Website: https://ijarmt.com ISSN No.: 3048-9458

with the subsequent full DFDC dataset motivated the development of larger and more diverse datasets for robust deepfake detection.

Dolhansky et al. [7] presented the full DeepFake Detection Challenge Dataset (DFDC), developed to support large-scale research on deepfake video detection. The dataset contains a substantial collection of real and manipulated videos involving numerous subjects and diverse recording conditions. Multiple manipulation methods were incorporated to make the dataset more representative of real-world deepfake scenarios. This diversity enables deep learning models to learn a broader range of facial manipulation artifacts rather than relying on a small number of specific generation patterns. The DFDC dataset became an important benchmark for evaluating deepfake detection algorithms and was particularly useful for CNN-based, transfer-learning, and video-based detection approaches. The study demonstrated the importance of large and heterogeneous datasets in developing robust detection models. Nevertheless, variations in video quality, compression, facial pose, lighting, and manipulation techniques continue to make generalization across datasets a challenging problem.

Rana et al. [8] conducted a systematic literature review of deepfake detection techniques, providing a comprehensive overview of existing research in this rapidly developing field. The authors examined different approaches for detecting deepfake content, including traditional machine learning, deep learning, image-based techniques, and video-based methods. Particular attention was given to the use of CNN architectures and other deep learning models for automatically learning manipulation-related features from facial images and video frames. The review also discussed commonly used datasets, performance evaluation metrics, and the major challenges associated with deepfake detection. According to the study, differences in datasets, manipulation techniques, compression levels, and experimental settings make direct comparison between existing methods difficult. The authors emphasized the need for more generalized detection models capable of identifying previously unseen manipulation techniques. This systematic review provides an important foundation for understanding the development of deepfake detection and identifying gaps for future research.

Jin, Cruz, and Gonçalves [9] proposed a deep transfer learning approach for facial diagnosis, demonstrating how knowledge learned from large-scale face-recognition tasks can be transferred to specialized facial analysis applications. The study explored the use of deep neural networks to extract meaningful facial representations and subsequently adapt these representations for facial diagnosis. Although the primary focus was not deepfake detection, the work is relevant because transfer learning is widely used in face-related computer vision tasks where obtaining sufficiently large task-specific datasets can be difficult. Pre-trained deep architectures can provide strong low-level and high-level facial features, reducing the computational and data requirements of training a model from scratch. This concept can be applied to deepfake face detection by adapting pre-trained CNN architectures to distinguish

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 7.2 Website: https://ijarmt.com ISSN No.: 3048-9458

authentic facial content from manipulated content. The study therefore highlights the potential of transfer learning as an effective strategy for developing accurate and computationally efficient facial analysis systems.

III. DEEPFAKE FACE DETECTION

Deepfake face detection is an important research area within artificial intelligence, computer vision, and digital media forensics that focuses on identifying artificially generated or manipulated facial images and videos. The term *deepfake* generally refers to synthetic media created using deep learning techniques, particularly Generative Adversarial Networks (GANs), autoencoders, and other generative models. These techniques can replace a person's face, modify facial expressions, synthesize realistic facial identities, or generate completely artificial facial content. Because modern deepfake-generation methods can produce highly realistic results, distinguishing manipulated content from genuine content has become increasingly difficult through conventional visual inspection.

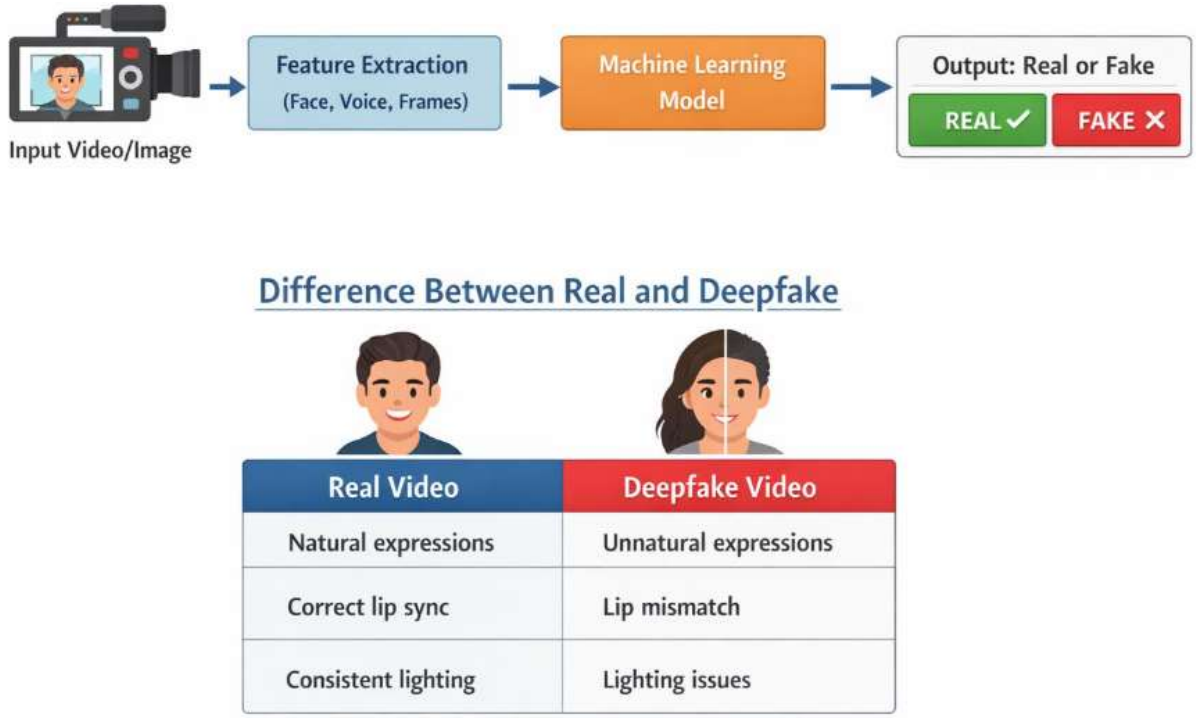

Figure 1: Deepfake Detection

Deepfake face detection techniques generally analyze spatial, temporal, and contextual characteristics of facial content. Spatial-based approaches examine individual images or video frames and identify inconsistencies in facial texture, skin boundaries, eyes, teeth, lighting, and facial geometry. CNN-based models are particularly effective for this task because they automatically learn hierarchical features from facial images. Temporal approaches analyze changes between consecutive video frames and identify unnatural variations in facial movements, blinking, expressions, or head poses. Context-based methods

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 7.2 Website: https://ijarmt.com ISSN No.: 3048-9458

additionally examine the relationship between the manipulated face and its surrounding environment, including inconsistencies in illumination, background, pose, and facial-to-context geometry.

The deepfake detection process generally consists of several stages: data collection, face detection and extraction, preprocessing, feature extraction, model training, classification, and performance evaluation. Large datasets containing both authentic and manipulated images or videos are required for training deep learning models. The extracted facial regions are commonly resized and normalized before being provided to the detection network. Transfer-learning architectures such as ResNet, DenseNet, EfficientNet, and other advanced deep neural networks can then be fine-tuned for deepfake classification. The final model generally produces a probability indicating whether the input content is genuine or manipulated.

A major challenge in deepfake face detection is generalization. A model trained on a particular dataset or manipulation technique may perform well under controlled conditions but experience significant performance degradation when exposed to unseen deepfake-generation methods. Video compression, changes in resolution, illumination, facial pose, occlusion, and social-media processing can further eliminate subtle manipulation artifacts. Therefore, an effective detection system should not only achieve high accuracy on a known dataset but should also demonstrate robustness against different manipulation techniques and real-world distortions.

Deepfake detection has applications in social-media security, digital forensics, cybersecurity, identity verification, journalism, law enforcement, and online content authentication. As generative AI continues to improve, detection methods must also evolve toward more robust and generalizable architectures. Current research is increasingly focusing on hybrid spatial-temporal networks, Transformer-based models, multimodal analysis, explainable artificial intelligence, and lightweight architectures for real-time deepfake detection. Thus, deepfake face detection represents a continuously evolving field in which the development of reliable and generalized deep learning models is essential for maintaining the authenticity and trustworthiness of digital media.

IV. DEEP LEARNING

Deep Learning (DL) is a subfield of Artificial Intelligence (AI) and Machine Learning (ML) that uses multi-layered artificial neural networks to automatically learn complex patterns and representations from large amounts of data. Unlike traditional machine learning techniques, which often require manually designed features, deep learning models can perform automatic feature extraction and classification directly from raw or minimally processed data. The basic building block of deep learning is the artificial neuron, and multiple interconnected neurons are organized into layers, including an input layer, hidden layers, and an output layer. As the number of hidden layers increases, the network can learn increasingly complex representations of the input data.

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 7.2 Website: https://ijarmt.com ISSN No.: 3048-9458

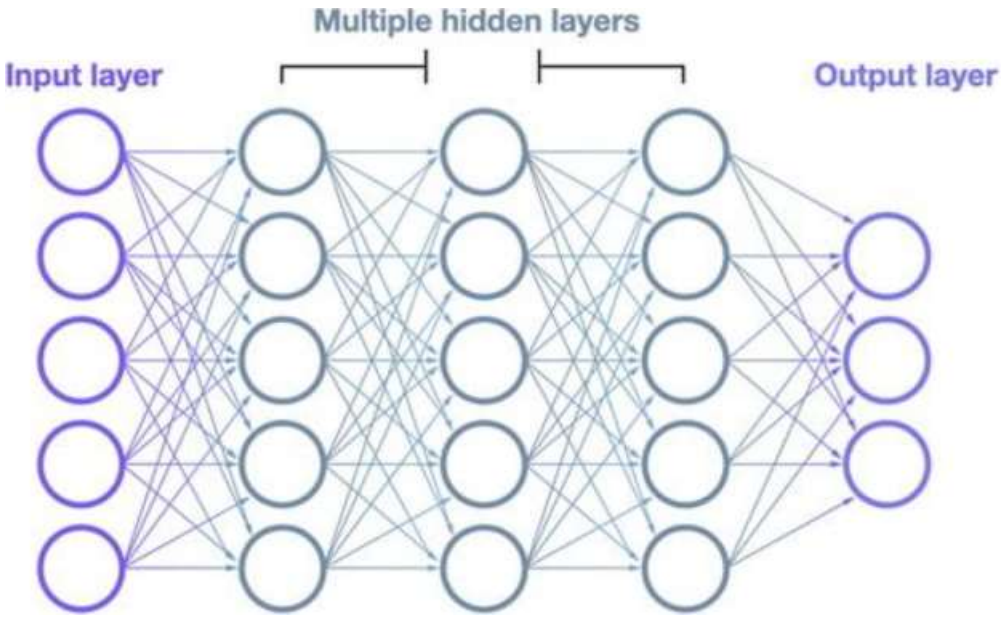

Figure 2: Deep Learning

In deepfake face detection, deep learning plays a crucial role because manipulated facial images and videos often contain subtle patterns that are difficult to identify using conventional image-processing techniques. Deep learning models can learn differences in facial texture, edges, skin patterns, lighting, facial boundaries, eyes, mouth regions, and other visual artifacts from authentic and manipulated samples. A typical deepfake detection framework first extracts faces from images or video frames, performs preprocessing such as resizing and normalization, and then provides the processed images to a deep neural network. The network learns discriminative features during training and subsequently classifies the input as either *real* or *deepfake*.

Among different deep learning architectures, Convolutional Neural Networks (CNNs) are widely used for image-based deepfake detection because they are capable of learning spatial features through convolution and pooling operations. Advanced CNN architectures such as ResNet, DenseNet, and EfficientNet can provide deeper and more discriminative feature representations. Transfer learning is also commonly employed, where a model pre-trained on a large image dataset is fine-tuned using deepfake datasets. This approach can reduce training time and improve performance when the available deepfake dataset is relatively limited.

For deepfake videos, deep learning can additionally capture temporal information by analyzing multiple consecutive frames. Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and CNN-LSTM architectures can be used to identify inconsistencies in facial movements and temporal patterns. More recently, Transformer-based architectures have also been investigated because of their ability to model long-range relationships between image regions and video frames.

The performance of a deep learning-based deepfake detector is generally evaluated using metrics such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating

Characteristic Curve (AUC-ROC). However, high accuracy on a single dataset does not necessarily indicate reliable real-world performance. Factors such as dataset bias, compression, image resolution, unseen manipulation methods, and adversarial attacks can significantly affect detection capability. Therefore, robust deep learning-based deepfake detection requires diverse training data, appropriate augmentation, cross-dataset evaluation, and models capable of generalizing to previously unseen manipulations.

V. CONCLUSION

Deepfake face detection has emerged as an important research area due to the rapid development of generative artificial intelligence and the increasing availability of highly realistic manipulated facial content. This review examined the development of deepfake detection techniques, with particular emphasis on deep learning-based approaches. Existing studies demonstrate that deep learning models, especially Convolutional Neural Networks (CNNs), transfer-learning architectures, and advanced models such as ResNet, DenseNet, and EfficientNet, can effectively learn subtle visual characteristics associated with manipulated facial content. In addition, contextual and temporal information can improve detection by identifying inconsistencies that may not be visible from individual facial regions.

The reviewed literature also highlights that the availability of large and diverse datasets, such as the DeepFake Detection Challenge (DFDC) datasets, is essential for training and evaluating reliable detection models. However, achieving high accuracy on a particular dataset does not guarantee strong performance in real-world environments. Variations in compression, resolution, illumination, facial pose, manipulation techniques, and unseen deepfake-generation methods remain major challenges. Furthermore, the increasing sophistication of generative models continuously creates new forms of synthetic content that can bypass existing detectors.

Overall, deep learning provides a strong foundation for developing automated deepfake face detection systems, but future solutions should focus on robustness, generalization, cross-dataset performance, computational efficiency, and real-time detection. Emerging approaches based on CNN-Transformer architectures, spatial-temporal feature learning, multimodal analysis, explainable AI, and lightweight networks can further improve detection reliability. Continued research in these directions is essential for protecting digital media authenticity, privacy, cybersecurity, and public trust in an increasingly AI-generated media environment.

REFERENCES

[1] Ali Raza, Kashif Munir and Mubarak Almutairi, “A Novel Deep Learning Approach for Deepfake Image Detection”, Apply Science, pp. 01-15, 2022.

[2] Zobaed, S.; Rabby, F.; Hossain, I.; Hossain, E.; Hasan, S.; Karim, A.; Hasib, K.M. Deepfakes: Detecting forged and synthetic media content using machine learning. In Artificial Intelligence in Cyber Security: Impact and Implications; Springer: Berlin/Heidelberg, Germany, 2021; pp. 177–201.

[3] Thambawita, V.; Isaksen, J.L.; Hicks, S.A.; Ghouse, J.; Ahlberg, G.; Linneberg, A.; Grarup, N.; Ellervik, C.; Olesen, M.S.; Hansen, T.; et al. DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. Sci. Rep. 2021, 11, 21869.

[4] Ahmed, M.F.B.; Miah, M.S.U.; Bhowmik, A.; Sulaiman, J.B. Awareness to Deepfake: A resistance mechanism to Deepfake. In Proceedings of the 2021 International Congress of Advanced Technology and Engineering (ICOTEN), Taiz, Yemen, 4–5 July 2021; pp. 1–5.

[5] Y. Nirkin, L. Wolf, Y. Keller, T. Hassner, Deepfake detection based on discrepancies between faces and their context, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (10) (2021) 6111–6121.

[6] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The Deepfake Detection Challenge (DFDC) Preview Dataset. arXiv eprints, page arXiv:1910.08854, 2019.

[7] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake Detection Challenge Dataset. arXiv e-prints, page arXiv:2006.07397, 2020.

[8] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. IEEE access, 10:25494–25513, 2022.

[9] Jin, B.; Cruz, L.; Gonçalves, N. Deep facial diagnosis: Deep transfer learning from face recognition to facial diagnosis. IEEE Access 2020, 8, 123649–123661.

[10] Lewis, J.K.; Toubal, I.E.; Chen, H.; Sandesera, V.; Lomnitz, M.; Hampel-Arias, Z.; Prasad, C.; Palaniappan, K. Deepfake Video Detection Based on Spatial, Spectral, and Temporal Inconsistencies Using Multimodal Deep Learning. In Proceedings of the 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington DC, DC, USA, 13–15 October 2020; pp. 1–9.

[11] Lee, H.; Park, S.H.; Yoo, J.H.; Jung, S.H.; Huh, J.H. Face recognition at a distance for a stand-alone access control system. Sensors 2020, 20, 785.

[12] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Attgan: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing, 28(11): 5464–5478, 2019.

[13] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. Deep learning for deepfakes creation and detection, 2019, arXiv preprint arXiv:1909.11573.

[14] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. pages 1526–1535, 2018.

[15] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Fr´edo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[16] Anuj Badale et al., "Deep fake detection using neural networks", 15th IEEE international conference on advanced video and signal-based surveillance (AVSS), 2018.

[17] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Icface: Interpretable and controllable face reenactment using gans. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020.

[18] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In Proceedings of the European Conference on Computer Vision (ECCV), 2018.