Deepfake Media Detection Framework Using Machine Learning with Multimodal Feature Extraction for Real and Synthetic Content

Shraddha Veer
Department of computer science & Engineering
Bapurao Deshmukh College of Engineering
Wardha , Maharashtra ,India 442104
shraddhapveerbd27@gmail.com

Dr.Sudhir Mohod
Department of computer science & Engineering
Bapurao Deshmukh College of Engineering
Wardha , Maharashtra ,India 442104
sudhir_mohod@rediffmail.com

***Abstract*—The emergence of contemporary deepfakes has at tracted significant attention in machine learning research, as artificial intelligence (AI) generated synthetic media increases the incidence of misinterpretation and is difficult to distinguish from genuine content.Techniques for creating and manipulating multimedia information have progressed to the point where they can now ensure a high degree of realism. DeepFake is a generative deep learning algorithm that creates or modifies face features in a superrealistic form, making it difficult to distinguish between real and fake features. This technology has greatly advanced, promoting a wide range of applications in cinema, such as improving visual effects in movies, as well as various criminal activities, such as misinformation generation by mimicking famous people. To identify and classify DeepFakes, research in DeepFake detection using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) has attracted increased interest. Essentially, DeepFake is regenerated media obtained by injecting or replacing some information within CNN and RNN models. This paper summarizes the DeepFake detection methods for face images and videos based on their results, performance, methodology used, and detection type. The challenges in generating a generalized DeepFake detection model are also analyzed.**

Keywords—Deepfake, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Resnet50, Long Short Term Memory (LSTM).

I. INTRODUCTION

Deepfakes are artificially generated media, such as images, videos, or audio recordings, created using deep learning algorithms. They combine existing content with new audio or video recordings, resulting in a fake but realistic-looking and sounding outcome. Deepfakes can be used for entertainment, education, or malicious purposes, such as spreading misinformation, identity theft, or reputation damage. deepfake technology advances, detecting and preventing their misuse becomes increasingly important. Deepfake detection for images and videos has become a crucial task in maintaining trust in digital media. To combat the spread of deepfakes, this paper employed deep learning-based approaches using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs are effective in detecting manipulated images, while RNNs are suitable for analyzing temporal features in videos. Methodologies such as two-stream networks, attention mechanisms, and multi-task learning have been utilized to improve detection accuracy. Despite promising results, challenges such as evasion attacks, generalizability, and explainability remain, highlighting the need for continued research in this field.

In recent years, the rise of DeepFake technology has led to the generation of highly realistic synthetic images, posing significant challenges to security, digital forensics, and media authenticity. DeepFake images, created using advanced generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) can seamlessly alter or fabricate human faces, making them nearly indistinguishable from real photographs. Recently, diffusion models have greatly enhanced the generation capability of images and videos. While these synthetic images have applications in entertainment and creative industries, the pro liferation of deepfake technology poses escalating risks to biometric security, online misinformation, identity fraud, and manipulation of public perception.

The rapid advancement of large language models such as ChatGPT has fundamentally changed the way written content is produced. These models generate text that is fluent, coherent, and contextually aligned, often rivaling the quality of human authorship. Their practical value is already evident in diverse applications, including report drafting, document summarization, and conversational agents that support everyday tasks. However, alongside these benefits come significant risks. Synthetic text can be misused in disinformation campaigns, the spread of fabricated news, online manipulation, and sophisticated phishing attempts. Such possibilities introduce serious ethical and societal challenges that demand careful attention. In contrast to the significant progress made in detecting manipulated images and videos research on deepfake text detection is still in its early stages. Recent surveys indicate that current methods struggle with robust- ness, generalization across domains, and resilience against adversarial manipulation. Equally concerning is the observation that human evaluators often perform only marginally better than chance when asked to distinguish between human authored and machine-generated text.

The field of deepfake detection has seen considerable progress in recent years with a number of sophisticated techniques being proposed in the area . However, in the context of detecting manipulated content in videos, many existing methods primarily focus on spatial features extracted from individual frames. This approach can lead to the overlooking of temporal dynamics that evolve throughout video sequences. This strategy can result in lim itations, as temporal artifacts

such as flickering and motion discontinuities, are common indicators of deepfake manip ulation. Furthermore, sophisticated deepfakes may exhibit subtle spatial inconsistencies that manifest over time, neces sitating an integrated analysis of both spatial and temporal information. Moreover, we hypothesize that capturing sub tle spatiotemporal inconsistencies that are often caused by different deepfake generation methods, could significantly enhance performance by learning representations that gen eralize to unseen forgery methods, which is often a chal lenging problem in this area.

II. RELATED WORK

Deepfake detection has traditionally been addressed as a binary classification task , where the objective is to dis cern between authentic and manipulated media. The appli cation of deep learning models, particularly convolutional neural networks (CNNs), has been central to achieving this objective. Authors of FaceForensics++ dataset used Xception network, which was one of the best performing architectures at the time, to perform deepfake detection via transfer learning . Researchers proposed a method in that utilizes residual-based descriptors in the form of a constrained CNN for image forgery detection. This approach aims to capture and analyze the residual noise present in manipulated im ages, which can be a strong indicator of forgery. In con trast, another method introduced a deep learning approach that focuses on the mesoscopic properties of images. In this context, 'mesoscopic' refers to properties or features that fall between the small scale (microscopic) and the large scale (macroscopic). By concentrating on mesoscopic fea tures, the model can capture subtle artifacts and inconsis tencies in manipulated images, potentially making it more effective in detecting forgeries. Various studies have utilized frequency analysis to detect inconsistencies that arise during deepfake creation. In the researchers employed the phase spectrum for forged face image detection, showing that CNNs can iden tify additional implicit phase spectrum features that are ad vantageous in detecting face forgeries. Concurrently, the study in developed a multi-scale patch similarity module to specifically model second-order relationships between distinct local regions, forming a similarity pattern through pairwise cosine measurements. These patterns distinguish real from forged regions by recognizing differences such as irregular textures and high-frequency noise.

There are also subjective studies analyzing human ability to distinguish between authentic and synthetic videos, and comparing human performance with AI models. The study in shows that both human and AI models exhibit compa rable performance, but with different errors, and that subjects with access to model predictions perform better than subjects without access to model predictions. The study in shows that while human perception is very different from machine perception, both are successfully fooled by deepfakes but in different ways. Specifically, algorithms have difficulty detect ing deepfake videos that are easily spotted by humans. Accord ing to people cannot reliably detect deepfakes. Raising awareness and financial incentives do not improve people's de tection accuracy. People tend to mistake deepfakes as authentic videos (rather than vice versa) and overestimate their own deepfake detection abilities. Therefore, several studies have attempted to analyze the factors that influence viewers' ability to perceive deepfake videos. The authors of conducted an investigative user study and analyzed existing AI detection al gorithms to uncover the unknown factors behind the detection of deepfakes. A similar study in shows that two-thirds of participants were unable to accurately detect a sequence of four videos as either genuine or deepfake, and that familiarity with the subjects in the videos had a statistically significant impact on the individuals' perception ability. In addition, the studies in investigated psychological and social factors that influence people's ability to detect fake videos also contain interesting subjective studies, with the former investigating the role of system-generated cues, such as video quality and technical imperfections, and the latter focusing on developing and testing strategies to enhance human's ability to identify deepfakes. Another study in delved into how humans and machines perceive deepfake videos, suggesting that incorporating emotional factors and leveraging specialized visual processing may be promising strategies to enhance deepfake detection. The study in explored the difference in detection performance between human observers and automated systems by comparing their abilities to identify deepfake videos in the presence of noisy channels. The interesting work in examined tourists' visit intention by watching deepfake destination videos, and found that the factors that affect the tourists' visit intention after watching deepfake videos include information manipulation tactics, trust and media richness, while perceived deception does not influence tourists' visit intention. The study in shows that people are more likely to feel uncertain than to be misled by political deepfakes, but this resulting uncertainty, in turn, reduces trust in news on social media. The authors of explored the impact of deepfake videos' informative cues on individuals' perceived accuracy of claims and their intentions to share non-political deepfakes, and found that high cognitive individuals were less likely to trust news and share content, and that people were more likely to believe that deepfake claims were true when informative cues were absent. The results in indicate various challenges in deepfake user perception that technology developers need to address before the potential of deepfake applications can be realized for human-computer interaction.

In this section, we review prior work on universal deepfake detection, frequency-domain analysis, and masked image modeling, and discuss the approaches related to efficiency and Green AI, providing the context for our proposed frequency-domain masking strategy.

Year	Authors	Title	Methodology	Dataset Used	Key Features	Results / Findings	Limitations
2020	Afchar et al.	MesoNet: A Compact Facial Video Forgery Detection Network	CNN-based deepfake detection	FaceForensics++	Facial texture artifacts	Achieved ~84% accuracy	Limited generalization to unseen datasets
2021	Zhou et al.	Learning Temporal Inconsistency for Deepfake Detection	RNN + CNN hybrid	Celeb-DF	Temporal inconsistencies	Effective for video-based deepfakes	Poor performance on images
2023	Khalid et al.	FakeAVCeleb: Audio-Visual Deepfake Detection	Multimodal (Audio + Video)	FakeAVCeleb	Lip-sync + audio mismatch	Robust multimodal detection	Dataset bias
2025	Sharma et al.	Hybrid Deepfake Detection Framework	CNN + Transformer + LSTM	Multiple datasets	Multimodal fusion (image, video, audio)	Accuracy > 97%	Training complexity
2025	Zhang et al.	Generalized Deepfake Detection	Domain adaptation + ML	Cross-dataset	Domain-invariant features	Better generalization	Performance drop in real-world cases

III. METHODOLOGY

A. Flowchart

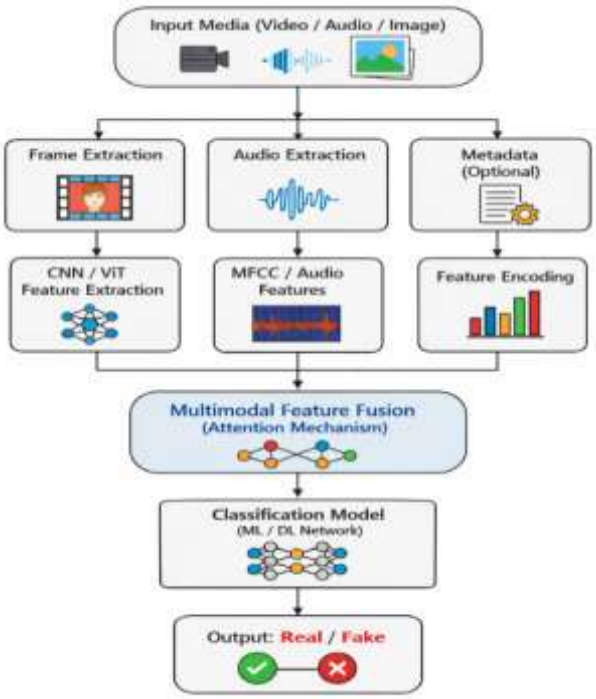


Figure No. 1 Flowchart

Our methodology begins with an input image being pro cessed by the RetinaFace detector , which identifies all faces within the image and extracts five key facial landmarks for each face : two for the eyes, one for the nose, and two defining the boundaries of the mouth. These landmarks serve as critical reference points for precise facial alignment and normalization. The bounding box of the face serve as our middle view. To obtain local face view, we construct a convex hull around the five detected landmarks, expanding the region by a 15-pixel margin to fully encompass the eyes and mouth while preserving essential facial details (see Fig. 1 for illustration). This localized crop helps focus on fine grained features, such as eyes, noise and mouth alignment and blending artifacts, that are crucial for DeepFake detection. Then, the global view is obtained by expanding the middle view in each direction by 20 pixels. This view will contain the neck and the ears and some of the background region. Finally, both the global and local face images are resized to 224 × 224 pixels, with zero-padding applied to maintain the aspect ratio and preserve original content. This pre-processing step ensures that the extracted facial features remain consistent and optimally structured for downstream analysis by the multi view encoders.

The proposed deepfake media detection framework is designed to accurately distinguish between real and synthetic content by leveraging a multimodal machine learning approach that integrates visual, temporal, and audio features. The system takes multimedia input in the form of videos, images, or audio clips and processes them through multiple stages, including preprocessing, feature extraction, feature fusion, classification, and deployment. Unlike traditional unimodal approaches that rely only on visual artifacts, this framework captures inconsistencies across multiple modalities, thereby improving robustness and generalization.

B. Feature extraction

Following preprocessing, the framework performs feature extraction across multiple modalities. For visual feature extraction, deep learning models such as Convolutional Neural Networks (CNNs), including ResNet or EfficientNet, and Vision Transformers (ViT) are employed to capture spatial features like facial textures, blending artifacts, and inconsistencies introduced during deepfake generation. To

capture temporal dependencies and motion-based anomalies, sequential models such as Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU) are used, which analyze frame-to-frame variations such as unnatural eye blinking, lip movements, and facial distortions. In parallel, audio features are extracted using CNN-based models applied to MFCC or spectrogram inputs to identify anomalies in speech patterns, pitch variations, and voice cloning artifacts. This multimodal feature extraction enables the system to detect subtle inconsistencies that may not be visible in a single modality.

Once the features are extracted, they are combined using a multimodal feature fusion strategy. In this work, an attention-based fusion mechanism is proposed, which assigns different weights to visual, temporal, and audio features based on their importance in identifying deepfakes. The extracted features are concatenated and passed through an attention layer that enhances relevant features while suppressing less significant ones. This approach ensures that the model focuses on the most discriminative aspects of the data, improving detection performance compared to traditional early or late fusion techniques.

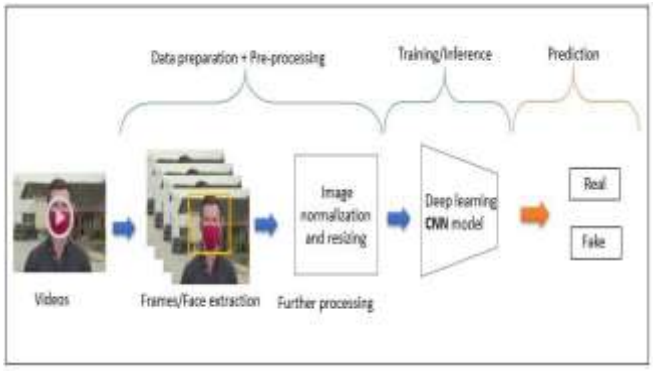


Figure No 2. Model Working

C. Masked ImageModeling

Masked image modeling (MIM) has emerged as a powerful self-supervised learning paradigm in computer vision, with increasing relevance to generalization and robustness. He et al. proposed Masked Autoencoders (MAE), demonstrating that reconstructing a high proportion of masked image patches using an asymmetric encoder-decoder architecture can achieve strong visual representations, rivaling or surpassing supervised pre-training on downstream tasks. Subsequently, Huang et al. introduced MaskedGAN, which uses randomized spatial and frequency masking to stabilize GAN training under data-scarce conditions, improving the robustness of generative models. Moving from representation learning to the domain of out-of-distribution detection, Li et al. argued that reconstruction-based objectives are inherently better at learning in-distribution representations and thus more effective at detecting out-of-distribution samples, outperforming recognition-based methods. Xie et al. extended MIM to the frequency domain with Masked Frequency Modeling (MFM), where the model learns to predict masked frequency components instead of spatial patches, capturing structured global image priors more efficiently. In the context of deepfakedetection, Dasetal.leveragedmaskedautoencodingwithspatiotemporaltransformers for videos, showing that masked modeling can facilitate strong generalization across datasets by encoding both spatial and temporal inconsistencies in fake content. More recently, Chen et al. applied masking in conjunction with a conditional diffusion model to augment training data, demonstrating improved generalization to unseen deepfake forgeries.

D. Hybrid classification model

The fused features are then fed into a hybrid classification model, which consists of fully connected neural network layers. Activation functions such as ReLU are used in hidden layers, while a Softmax or Sigmoid function is applied in the output layer to classify the input as either real or fake. The model is trained using the Binary Cross-Entropy loss function and optimized using the Adam optimizer. During training, performance is evaluated using standard metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to ensure reliability and robustness. Cross-validation and cross-dataset evaluation are also performed to test the generalization capability of the model across unseen data. After training, the model undergoes testing and validation, where it is evaluated on both benchmark datasets and real-world samples to assess its effectiveness in practical scenarios. Special emphasis is given to cross-dataset testing to address the challenge of domain shift, which is common in deepfake detection tasks. The final trained model is then deployed using lightweight frameworks such as Flask or FastAPI, allowing users to upload media files and receive real-time predictions. For user interaction, a simple interface can be developed using Streamlit or web technologies like HTML, CSS, and JavaScript.

The entire system is implemented primarily using the Python programming language, supported by libraries such as TensorFlow or PyTorch for deep learning, OpenCV for image and video processing, Librosa for audio analysis, and NumPy and Pandas for data handling. Visualization tools like Matplotlib and Seaborn are used for performance analysis. The proposed framework is designed to be scalable and adaptable, making it suitable for real-world applications such as social media monitoring, digital forensics, and misinformation detection.

Overall, this methodology introduces a robust, hybrid, and multimodal deepfake detection system that overcomes the limitations of existing approaches by combining spatial, temporal, and audio features with an attention-based fusion mechanism. The expected outcome is a highly accurate and generalizable model capable of detecting even advanced deepfake content, with potential for real-time deployment and future enhancements such as explainable AI and mobile-based detection systems.

IV. AUDIO AUTHENTICITY DETECTION

In the proposed deepfake detection framework, the analysis of audio authenticity (real vs fake) plays a crucial role as a complementary modality to visual and temporal features, enabling more reliable identification of synthetic media. Audio deepfakes are typically generated using advanced voice cloning and speech synthesis techniques, which can

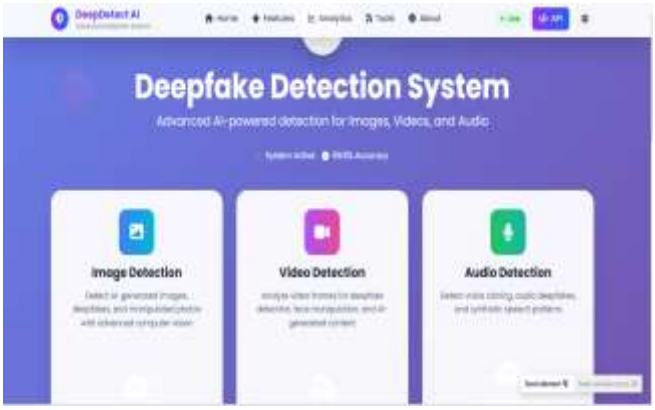
replicate a person's tone, pitch, and speaking style with high accuracy; however, these generated signals often contain subtle inconsistencies that can be captured through systematic signal processing and machine learning methods. In this system, the audio component is first extracted from the input media and preprocessed to remove background noise and normalize amplitude levels, ensuring consistent quality across samples. The cleaned audio signal is then transformed into representative features such as Mel Frequency Cepstral Coefficients (MFCCs), spectrograms, chroma features, and pitch-related attributes, which effectively capture both temporal and frequency-domain characteristics of the speech signal. These features are fed into deep learning models, such as Convolutional Neural Networks (CNNs) or hybrid CNN-Recurrent architectures, to learn distinguishing patterns between real and synthetic audio. Real audio typically exhibits natural variations in pitch, pauses, breathing patterns, and emotional expressiveness, while fake audio often shows unnatural smoothness, repetitive patterns, or mismatches in prosody due to limitations in generative models. Additionally, the system performs audio-visual consistency analysis, where the synchronization between lip movements in video frames and the corresponding speech signals is evaluated; discrepancies in timing, phoneme alignment, or articulation strongly indicate manipulated content. The extracted audio features are further integrated into the multimodal fusion layer, where their importance is weighted using an attention mechanism alongside visual and temporal features, enhancing overall detection performance. By systematically combining signal processing, deep learning, and cross-modal verification, the proposed approach ensures robust identification of fake audio, even in cases where visual artifacts are minimal, thereby significantly improving the accuracy and reliability of the entire deepfake detection system.

V. IMPACT OF THE COMBINATION OF CLASSICAL AND DEEPFAKE METHODS

In the next stage of the experiments, the effect of simultane ously using classical augmentation methods with advanced deepfake techniques on the generalization and quality of the developed image watermarking system was investigated. The primary motivation for this approach was the observation that although the successive addition of more deepfake models to training improved the robustness of the labeling method, it simultaneously began to lead to overtraining (overfitting) problems, manifested by visible artifacts in the generated images. During the study, it was found that the use of a single deepfake method (e.g., Bottleneck) together with classical intensive augmentation techniques (Gaussian Blur, Gaussian noise, filters, removal of watermark fragments) did not provide sufficient generalization- the model, despite the use of classical methods, quickly achieved overfitting on the chosendeepfaketechniqueandperformedpoorlywithothers. In contrast, introducing the same classical augmentations to training, combined with several different deepfake methods, significantly improved the results. This approach provided better generalization and effectively reduced previously observed visual artifacts. As shown in Table 1, classical augmentation with a set of deepfake methods negatively affected the metrics for watermarked images. However, it significantly improved reconstruction stability and overall quality. It also enables extended training without negatively affecting the validation cost function.

VI. EXPERIMENTL RESULT

A. Home Page



B. Dashborad

C. Input Page

D. Detection Result

Figure No . 3 Result of our system

VII. CONCLUSION

In this paper, we present the results of a subjective study aimed at analyzing human behavior in detecting audiovisual deepfake videos and comparing human performance with that of five AI models. We tested 80 human subjects and five AI models using the same set of 59 manually selected videos (34 real and 25 fake). The findings showed that subjects performed better than chance but generally found detecting deepfakes challenging. They performed marginally better than random chance, even when explicitly informed of the pres ence of deepfakes. The concern is that deepfake technology has reached a level of reality that could confuse much of the public, especially social media users, while people are overconfident in their ability to determine the authenticity of a video. Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar. This study has important implications for forensic analysis and adaptive countermeasures that improve the accuracy of digital forensic investigations, the attribution and tracking of manipulated content, the validation and verification of audio visual evidence, and the development of forensic tools.

VIII. FUTURE SCOPE

To further expand and refine our understanding of human perception of audiovisual deepfakes, we can implement a test design that measures subjective performance in a more discreet and unbiased way. For example, the applicability of our findings can be generalized by including a broader and more diverse population. Detailed analysis of the auditory and visual components of audiovisual signals will reveal how individuals perceive, interpret and process audio and visual cues in multimodal deepfake videos. One potential area for future research is analyzing and understanding the cognitive biases that hinder the human ability to identify deepfakes. Furthermore, exploring the impact of the popularity of people in videos (e.g., celebrities or politicians) may provide valu able insights into enhancing human detection capabilities and understanding how prior knowledge affects the perception of corresponding deepfake content.

REFERENCES

[1] Multimedia Security, 2016. 6 [5] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do you really mean that? content driven audio visual deepfake dataset and multimodal method for tempo ral forgery localization. International Conference on Digital Image Computing: Techniques and Applications, pages 1 10, 2022.

[2] Roberto Caldelli, Leonardo Galteri, Irene Amerini, and Al berto Del Bimbo. Optical flow based cnn for detection of unlearnt deepfake manipulations. Pattern Recognition Let ters, 146:31–37, 2021.

[3] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction classification learning for face forgery detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4113–4122, 2022.

[4] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial exam ples: Towards good generalizations for deepfake detec tions. IEEE/CVF conference on computer vision and pattern recognition, 2022.

[5] R. Burle, S. Gaurkhede, P. Gourshettiwar, S. Izankar, S. Gundewar and U. Pacharaney, "Hybrid Machine Learning Based Predictive Model To Improve Cloud Application Performance In Cloud Saas," *2024 (ICETEMS*2024,) pp. 167-172, doi: 10.1109/ICETEMS64039.2024.10965099.

[6] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. AAAI Conference on Artificial Intelligence, pages 1081–1088, 2021.

[7] F. Chollet. Xception: Deep learning with depthwise separa ble convolutions. IEEE Conference on Computer Vision and Pattern Recognition, pages 1800–1807, 2017.

[8] UmurAybarsCiftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological sig nals. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–1, 2020. 6

[9] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, M.-H. Yang, Diffusion models: A comprehensive survey of methods and applications, ACM Computing Surveys 56 (4) (2023) 1–39.

[10] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in: Proceedings of the Advances in Neural Information Processing Systems, 2020, pp. 6840–6851.

[11] A. AV, S. Das, A. Das, et al., Latent flow diffusion for deepfake video generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2024, pp. 3781–3790.

[12] H. Huang, P. S. Yu, C. Wang, An introduction to image synthesis with generative adversarial nets, arXiv preprint arXiv:1803.04469 (2018).

[13] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, StarGAN: Unified generative adversarial networks for multi-domain image-to image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.

[14] S. Suwajanakorn, S. M. Seitz, I. Kemelmacher-Shlizerman, Synthesizing Obama: Learning lip sync from audio, ACM Transactions on Graphics 36 (4) (2017) 1–13.

[15] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, Face2Face: Real-time face capture and reenactment of RGB videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2387–2395.

[16] S. M. Nadeem Jabbar Sheeraz Bhatti, Rashid, and A. Jaffar, "Single-layer kan for deepfake classifica- tion: Balancing efficiency and performance in resource constrained environments," PLOS ONE, vol. 20, no. 7, pp. 1–26, 2025. DOI: 10.1371/journal.pone.0326565.

[17] A. Nadeem CH* Saghir, A. A. Meer, S. A. Sahi, B. Hassan, and S. Muhammad Yasir, "Media forensics and deepfake - systematic survey," RS Open Journal on Innovative Communication Technologies, vol. 3, no. 8, 2023. DOI: 10.46470/03d8ffbd.7351a3bb.

[18] J. Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong, and L. S. Chao, "A survey on llm-generated text detection: Necessity, methods, and future directions," Computational Linguistics, vol. 51, no. 1, pp. 275–338, 2025.

[19] Damian Ibanez, Ruben Fernandez-Beltran, Filiberto Pla, and NaotoYokoya. Maskedauto-encodingspectral–spatial trans former for hyperspectral image classification. IEEE Transac tions on Geoscience and Remote Sensing, pages 1–14, 2022.

[20] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. Av fakenet: A unified end-to-end dense swin transformer deep learning model for audio-visual deepfakes detection. Applied Soft Computing, 136:110124, 2023.

[21] Zhiquan Jiang, Pengsen Zhao, and Zhonglong Zheng. Opti cal flow-attention fusion model for deepfake detection. In ternational Conference on Algorithms, Computing and Arti f icial Intelligence, 2023.

[22] R. Burle, S. Gundewar, S. Gaurkhede and B. Fulkar, "Conversational Symptom Checker Chatbot for Disease Prediction and Suggesting Nearby Hospitals using Machine Learning and Location Services," *(ICEARS)*, Tuticorin, India, 2025, pp. 775-780, doi: 10.1109/ICEARS64219.2025.10941338.

[23] D. Cooke, A. Edwards, S. Barkoff, K. Kelly, As good as a coin toss human detection of AI-generated images, videos, audio, and audiovisual stimuli, arXiv preprint arXiv:2403.16760 (2024).

[24] T. Weikmann, H. Greber, A. Nikolaou, After Deception: How falling for a deepfake affects the way we see, hear, and experience media, The International Journal of Press/Politics (2024) 19401612241233539.

[25] S. Ahmed, H. W. Chua, Perception and deception: Exploring individual responses to deepfakes across different modalities, Heliyon 9 (10) (2023).