

Real-Time Multi-Modal Deepfake Detection for Edge Devices Using Lightweight CNNs

Sameerunnisa Shaik¹ Miriyala Rakesh², B. Sai Chaitanya Reddy³, M. L. Srikanth Reddy⁴, Manda Seshu⁵

¹Assistant Professor, Department of CSE (IoT, Cybersecurity including Blockchain Technology), VVIT University, Guntur, Andhra Pradesh, India, Email: sameerask.25@gmail.com

²Student, Department of CSE (IoT, Cybersecurity including Blockchain Technology), VVIT University, Guntur, Andhra Pradesh, India, Email: rakeshmiriyala04@gmail.com

³Student, Department of CSE (IoT, Cybersecurity including Blockchain Technology), VVIT University, Guntur, Andhra Pradesh, India, Email: Saichaitanyareddy0013@gmail.com

⁴Student, Department of CSE (IoT, Cybersecurity including Blockchain Technology), VVIT University, Guntur, Andhra Pradesh, India, Email: srikanthsrikanth94705@gmail.com

⁵Student, Department of CSE (IoT, Cybersecurity including Blockchain Technology), VVIT University, Guntur, Andhra Pradesh, India, Email: 22BQ1A4748@vvit.net

ABSTRACT

The rapid-fire increase of generative artificial intelligence has led to the wide creation of largely realistic manipulated images, vids, and audio, generally known as deepfakes. While being deepfake discovery styles have shown promising results, numerous depend on computationally high-end models and pall- grounded processing, limiting their felicity for real- time and sequestration-sensitive operations. This paper proposes a feather light multi-modal deepfake discovery frame designed for real- time working on edge bias. The system integrates visual and audio analysis using effective CNN- grounded models and applies point- position emulsion to find cross-modal inconsistencies. Preprocessing ways similar as face- region birth, videotape frame slice, and Mel- spectrogram generation are recommended to regularize different media types. The proposed frame is optimized for CPU- grounded prosecution and performs all conclusion locally, barring pall reliance. Experimental evaluation on standard datasets demonstrates that this approach provides dependable deepfake discovery while maintaining low quiescence, making it suitable for practical edge- grounded media authenticity verification for privacy and security.

Keywords: Deepfake Detection, Multi-Modal Analysis, Lightweight CNN, Edge Devices

1. INTRODUCTION

1.1 Background and Motivation

Deep Learning and generative models have progressed to a great extent in recent times, making it may be induced to a large degree realistic media, which is usually referred to as manipulated deepfakes. These are counterfeit pictures, vids, and audio recitations which can almost be real personalities and actual circumstances. Although the useful operations with similar technologies are present. similarities like entertainment, availability, and so on their abuse has brought about grave content creation, their content creation identity pitfalls associated with misinformation. fraud and mistrust in online media. past researches have emphasized the importance of deepfakes turns out to be the means of controlling the mind of the population distort facts, and one can rely on discovery a problematic issue of exploration A. Rössler., 2019, B. Dolhansky et al., 2020.

The social media sites are used extensively and are associated with widespread use digital content, the tools of online communication are participated at veritably high and big speed scale. This performs verification that is homemade impossible and adds to the liability which distorted media may go far before being detected. As a result, automatic discovery systems of deepfaking have arrived key field of investigation. Large standard datasets that are comparable to FaceForensics, DFDC, and others. ASVspoof have been very instrumental in empowering the construction and assessment of discovery styles deepfake discovery on visual and audio modalities A. Rössler., 2019, B. Dolhansky et al., 2020, T. Kinnunen et al., 2019. still, indeed with these

datasets, estimating deepfakes in real- The world situation is still excruciating because of differences in quality, contraction and invisible manipulation ways J. Zi, 2020.

Numerous being deepfake discovery systems calculate on computationally ferocious deep literacy models and pall-grounded architectures. While these approaches frequently achieve high discovery delicacy, they generally bear important GPUs and remote processing. This introduces advanced quiescence, functional costs, and sequestration enterprises, as druggies must upload sensitive media to external waiters. Several studies have noted that pall-dependent and heavyweight infrastructures limit the practical deployment of deepfake discovery systems on particular bias and edge platforms F. Chollet., 2017, D. Güera and E. J. Delp., 2016. These limitations motivate the development of featherlight and locally deployable discovery results.

1.2 Problem Statement

Utmost current deepfake discovery systems are designed to work on a single type of media, similar as images or vids, or only on audio spoofing. still, ultramodern deepfake generation ways frequently manipulate multiple modalities at the same time. For illustration, a fake videotape may include both altered facial expressions and modified speech. When discovery systems dissect only one modality, they may fail to descry similar complex manipulations and cross-modal inconsistencies, which reduces their effectiveness in real- world scripts Y. Mittal., 2020, J. S. Chung., 2017.

Another major limitation is computational complexity. numerous state- of- the- art styles calculate on heavy CNNs or motor- grounded infrastructures that bear high- end GPUs and pall coffers. This makes them infelicitous for deployment on edge bias similar as laptops, mobile phones, and bedded systems. In addition, several being approaches perform discovery only after content has formerly been uploaded or participated online. This reactive approach allows fake content to spread before it can be linked and controlled, reducing the practical utility of similar systems A. Dosovitskiy et al., 2021.

Privacy is also a significant concern. pall- grounded discovery systems bear druggies to upload particular or sensitive media to external waiters, which may not be respectable in disciplines similar as particular verification, enterprise surroundings, and forensic analysis. previous exploration has emphasized the significance of original, on- device processing to save stoner sequestration and reduce network reliance A. G. Howard et al., 2017, M. Tan and Q. Le., 2019. These challenges easily indicate the need for a unified, effective, and sequestration- conserving deepfake discovery frame.

1.3 Objectives and Research Contributions

The main ideal of this exploration is to develop a practical and featherlight multi-modal deepfake discovery frame that can operate in real time on edge devices. The proposed system is designed to support image, videotape, and audio analysis within a single unified channel while maintaining low computational outflow and conserving stoner sequestration.

This work contributes a unified frame that integrates visual and audio deepfake discovery using featherlight CNN- grounded models, making it suitable for CPU- grounded edge deployment. point- position emulsion is applied to combine visual and audio information, which improves robustness against complex deepfake attacks that involve both facial and voice manipulation.

The frame is designed to operate entirely on- device without reliance on cloud structure, addressing sequestration and quiescence enterprises generally reported in pall- grounded discovery approaches. In addition, the system is estimated using extensively espoused standard datasets similar as FaceForensics, DFDC, and ASVspoof, which are generally used in deepfake discovery exploration for visual and audio evaluation. This ensures that the proposed approach is assessed under realistic and standardized conditions.

2. LITERATURE SURVEY

Deepfake discovery has entered significant exploration attention in recent times due to the rapid-fire increase in the vacuity and literalism of manipulated digital media. Early work in this area substantially concentrated on visual deepfake discovery using convolutional neural networks trained on standard datasets. The FaceForensics

dataset has been extensively used to train and estimate visual deepfake sensors, and several studies have shown that CNN- grounded models can effectively learn vestiges related to facial manipulation, blending crimes, and contraction inconsistencies A. Rössler., 2019. also, the Deepfake Detection Challenge(DFDC) dataset was

introduced to give large- scale and different videotape samples, enabling the development of further robust visual discovery models. still, numerous styles trained on DFDC calculate on heavy infrastructures and show reduced performance when applied to unseen or real- world data B. Dolhansky et al., 2020, J. Zi, 2020.

In addition to visual analysis, audio deepfake and voice spoofing discovery has come an important exploration area. The ASVspoof challenge series has handed standardized datasets and evaluation protocols for detecting synthetic and converted speech. previous studies have demonstrated that CNN- grounded and cold-blooded CNN – LSTM models trained on spectrogram representations can successfully identify voice manipulation vestiges similar as pitch deformation and spectral smoothing T. Kinnunen et al., 2019. Although these approaches achieve good delicacy in controlled settings, they frequently bear significant computational coffers and are generally estimated singly from visual deepfake discovery, limiting their effectiveness in scripts where both audio and visual manipulation are present.

Recent exploration has emphasized the significance of multimodal deepfake discovery, where both visual and audio information are concertedly anatomized. Several studies have shown that combining facial features with speech and lip- sync information can significantly ameliorate discovery performance by landingcross-modal inconsistencies that are delicate to fake constantly Y. Mittal., 2020, J. S. Chung., 2017. Audio-visual emulsion styles and synchronization- grounded approaches have been particularly effective in relating mismatches between lip movements and speech signals. still, numerous of these multimodal systems are grounded on motor infrastructures and complex emulsion networks, which introduce high computational outflow and make them infelicitous for deployment on resource- constrained edge bias A. Dosovitskiy et al., 2021.

Another important challenge stressed in the literature is conception to real- world and in- the-wild deepfakes. Datasets similar as WildDeepfake and other real- world collections have shown that models trained on clean, standard datasets frequently witness a significant drop in delicacy when applied to low- quality, compressed, or preliminarily unseen manipulations [4]. These findings indicate that discovery systems must be designed with robustness in mind and should be estimated beyond controlled laboratory datasets.

Lightweight and edge-oriented deepfake detection has recently gained attention due to the need for real-time and privacy-preserving applications. Studies based on efficient CNN architectures such as MobileNet and EfficientNet have demonstrated that it is possible to achieve a reasonable balance between discovery accuracy and computational effectiveness A. G. Howard et al., 2017. These models reduce the number of parameters and floating-point operations, making them more suitable for CPU-grounded conclusion on edge platforms. However, many lightweight approaches still concentrate on a single modality and do not fully exploit the benefits of multimodal fusion.

The prior work has also discussed privacy and deployment issues. The deepfake detection systems implemented in the form of cloud-based services mean that users need to submit media to remote services, which brings up data security concerns, trust in the system, and regulatory compliance. A number of works have highlighted how on-device processing and local inference are important as a way of maintaining privacy and latency of use in specific sensitive services like personal verification, enterprise, and forensic analysis, among others M. Tan and Q. Le., 2019. These publications inspire the creation of local, edge based detectors of deepfakes, which are not cloud dependent.

Therefore according to the literature, although extensive advances have been made in visual and audio deepfake discovery and multimodal deepfake discovery, many systems are still reliant on heavy infrastructures to compute their delicacy, and featherlight systems often require multimodal services. These shortcomings drive the suggested publication that is aimed at creating a feather light, multi-modal deepfake discovery framework that is optimized to real- time, on- device execution without sacrificing the discovery performance across both visual and audio expression.

Table 1: Comparison of Related Deepfake Detection Work

Method / Study	Input Type	Core Approach	Research Gap
FaceForensics++ (A. Rössler., 2019)	Image/Video	CNN-based visual artifact detection on FaceForensics++	Limited performance on real-world data

Challenge (DFDC) (Dataset B. Dolhansky et al., 2020)	Video	Large-scale CNN models on DFDC dataset	High computational cost
ASVspoof 2019 (T. Kinnunen et al., 2020)	Audio	Spectrogram-based audio spoof detection (ASVspoof)	Audio-only analysis
Method / Study	Input Type	Core Approach	Research Gap
Audio-Visual Deepfake (Y. Mittal., 2020)	Audio + Video	Audio-visual fusion for lip-sync and feature consistency	Complex fusion models
Transformers for Image Recognition A. Dosovitskiy et al., 2021)	Audio + Video	Transformer-based multimodal detection	Not suitable for edge devices
MobileNets: Efficient CNN for Mobile (A. G. Howard et al., 2017)	Image/Video	Lightweight CNNs for efficient detection	Mostly single-Input

3. PROPOSED METHODOLOGY

A general scheme of the offered Multimodal Deepfake Detection System is presented in Figure 1. The system will be structured into sequential modules with the first being the Input Module which receives image, video data, and audio. Such inputs are initially subjected to a pre-processing phase which is used to normalize the image frames and prepare them to be analyzed, and audio signal is subjected to feature extraction in the form of Mel-spectrogram. After preprocessing, scale is established using a separate visual and audio analysis on deepfake Detection module which is lightweight CNN-based to be optimized on real-time inference. The visual and audio features that are extracted are then passed on to Fusion and Classification stage where feature level fusion is carried out to integrate cross-modal data. This will allow the system to record a discrepancy between speech patterns and facial movements. Lastly, the Result Output module produces a binary input showing whether the input media is Real or Fake. The modular architecture provides high processing via scaling and much better suited to be deployed on CPU edges and the privacy with comprehensively local inference.

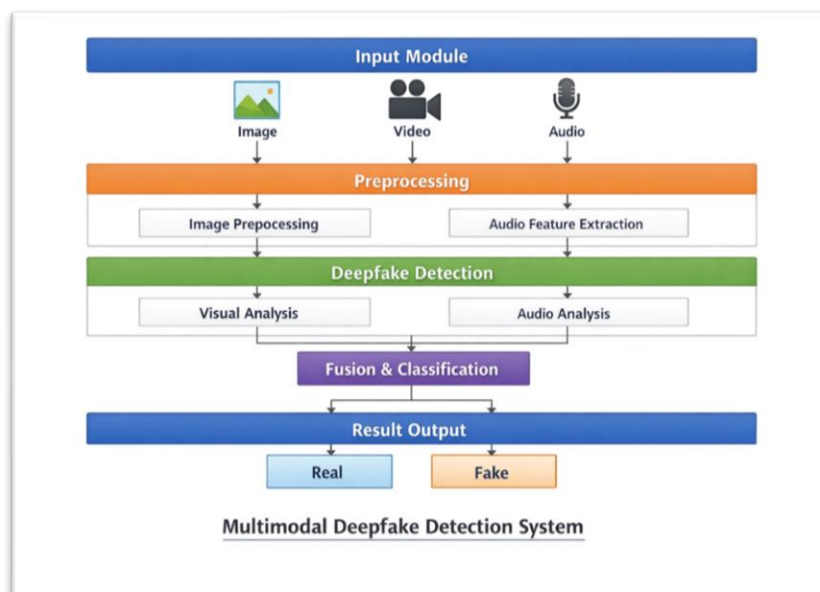


Figure 1: Deepfake System Architecture

3.1 System Architecture and Operational Overview

The suggested system is planned to be a featherlight, Multi-modal, deepfake discovery. frame that is to be operated in real- time in the edge. bias. The system facilitates image, videotape, and receives audio inputs and does all processing locally devoid of counting on cloud. infrastructure. The main ideal of the between employing factual and fictional methodology it is to find a balance computational accuracy discovery accuracy, computational effectiveness, and privacy preservation. The channel in general includes the media accession, preprocessing, featherproof light literacy. conclusion, multimodal, point, emulsion and final bracket. This is a modular design and enables each of them. CPU-effective element to work effectively founded platforms yet allowing simple extension and conservation.

3.2 Visual Data Processing

In the case of visual inputs, such as video and images frames, a lightweight CNN is employed in the system based model optimized on fast inference. Video First of all the inputs are broken down into frames sampled at fixed intervals. Each frame is scaled and normalised to correspond with the input conditions of the visual CNN. The visual model deals with the study of local arts as such as artificial textures, blending errors of the face, and patterns or tendencies of manipulation typical. proposed by the deepfake generation methods.

Features For video-level representation, features are extracted out of each frame and are. aggregated by the use of mean pooling to get a. single compact representation. This helps overall visual characteristics of the video with maintaining computational overhead low. The summed up visual feature vector may be expressed as:

$$\mathbf{F}_v = \frac{1}{N} \sum_{i=1}^N \mathbf{f}_v^{(i)} \quad \dots (1)$$

where $\mathbf{f}_v^{(i)}$ represents the feature vector extracted from the i -th video frame and N is the total number of sampled frames. This simple aggregation strategy allows the system to capture temporal variations without introducing heavy temporal models.

3.3 Audio Data Processing

In the case of audio, raw audio signals are given out first transformed into Mel-spectrogram. representations. This metamorphosis turns around invokes converting the signal, in time domain to two dimensions. which, time7 frequency representation, which, is well also fitted to CNN analysis. The Mel spectrogram is computed as:

$$S_{mel} = \text{MelFilterBank}(|\text{STFT}(x(t))|) \quad \dots (2)$$

where $x(t)$ is the audio input signal, and STFT. refers to the Short-Time Fourier Transform. The The frequency components are mapped to Mel filter bank the Mel scale, which is perceptually oriented relevant frequency bands.

A CNN model with a lightweight is then applied to the. Pseudo random-sampling to extract audio features associated with voice spoofing, e.g. spectral removing, intonation anomalies, and unnaturalistic frequency patterns. These extracted features create a small audio feature vector which reports the traits of the input. speech signal.

3.4 Multimodal Feature Fusion

To enhance resiliency to complexities. the system is being incorporated with deepfake attacks. feature-level multimodal fusion. In this stage, the image feature and the sound feature. membranes are joined up to become one. representation. This enables the system to co-ordinate. study image and sound representation and memorize depending on the mode, e.g., cross-modal inconsistencies, etc. speech-Lip Movement Inefficiencies. patterns. The fusion operation is effected with the help of Feature concatenation that can be expressed. as:

$$\mathbf{F}_{fusion} = [\mathbf{F}_v \parallel \mathbf{F}_a] \quad \dots (3)$$

where $\mathbf{F}_v$ represents the aggregated visual feature vector and $\mathbf{F}_a$ represents the audio feature vector. This simple fusion strategy is computationally efficient and suitable for edge deployment, while still enabling the model to learn joint audio-visual patterns.

3.5 Classification and Decision Making

The amalgamated feature vector is fed into a lightweight classification network with output. a probability score showing whether the input media is real or fake. The final prediction is was achieved by means of softmax-based decision. function:

$$P_{fake} = \text{Softmax}(\mathbf{W}\mathbf{F}_{fusion} + \mathbf{b}) \quad \dots (4)$$

where $\mathbf{W}$ and $\mathbf{b}$ mean the learnt weights and bias of the classifier. Based on the predicted probability, the system

gives a final designation. (True or False) and a score on confidence. This is a light decision cross stage that is intended to continue quick inference and offer trustworthy.

3.6 Edge Deployment and Explainability

The system as a whole is CPU optimized. grounded featherlight model prosecution based on featherlight models. infrastructures and smaller parameter sizes. This facilitates real time performance of without the standard laptops and edge platforms. requirement of GPU acceleration or pall connectivity. The frame can be performed locally by doing all conclusion at the same time. sustains stoner sequestration and diminishes network reliance.

Besides discovery, the system is useful in support of explainability by testing introductory modality- position benefactions. Grounded on the power of visual and sound characteristics, the system may be to denote visual vestiges, audio abnormalities, or cross-modal abnormalities. had a greater role in the ultimate decision. This enhances stoner trust and translucency, making the system is more appropriate in the real world. CC Self-pre-upload verification Similarity operations. authenticate original media originality checking.

4. RESULTS AND DISCUSSIONS

In this chapter, the authors provide the outcomes of the implementation of the suggested real-time multi-modal deepfake detector system. Image, audio, video, and audio-visual inputs have been used to evaluate the system. The observations analyzed include detected behavior, modal consistency, and multimodal fusion.

4.1 Detection Performance Across Modalities

Figure 2 summarizes the performance of the proposed system in terms of detection performance in various modalities. The figure makes a comparison of the detected accuracy of the image based, audio based, video based and multi-modular inputs.

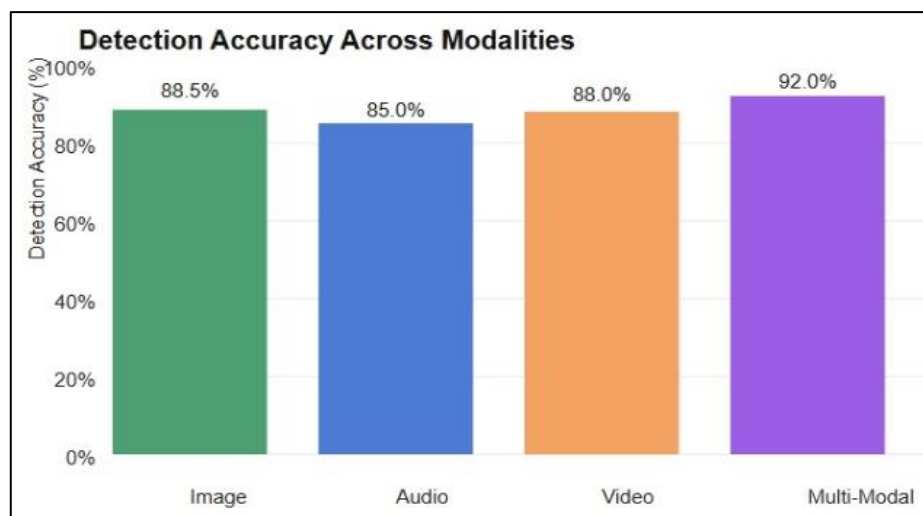


Figure 2. Detection accuracy

As Figure 2 shows, image based detection module was more stable and audio based detection was somewhat less accurate, which is an indication of the sensitivity of audio analysis to noise and recording conditions. Video-based detection showed similar results as image-based one and the advantage of the video-based detection is that it allows the combination of information in numerous frames. The most accurate detection was with the multi-modal approach where a union of audio and visual features can enhance the totality of overall detection accuracy.

4.2 Inference Time Analysis

The proposed system usability in real-time. was assessed with inference time customers on. different input modalities. Figure 3 illustrates the mean time taken to process image, audio, video and multi-modals inferred

using the benign hardware. CPU-based platforms.

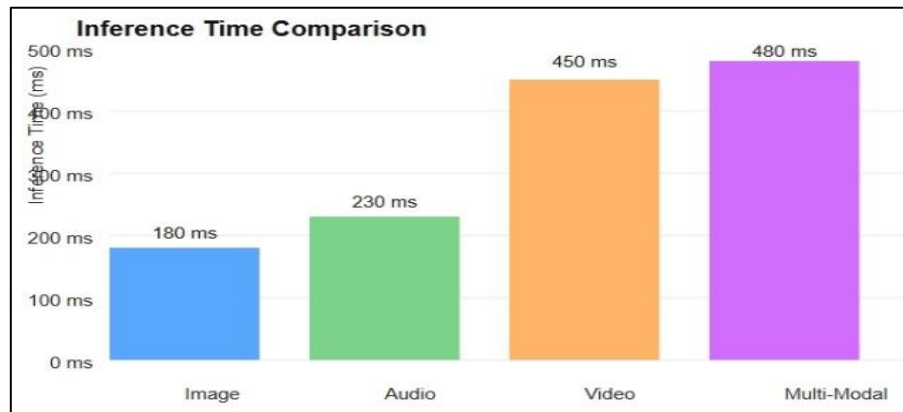


Figure 3. Inference time

Based on Figure 3, it is possible to note that inferences take less time in image and audio inputs. in comparison to the video and multi-modes. Video processing adds some new frame extraction and calculation because of frame extraction aggregation, multi-modal inference. integrates audio and video lines. All the modalities nonetheless work in sub second latency, showing the fact that the the suggested framework is real time appropriate deployment on edge devices.

4.3 Fake vs Real Output Distribution

In order to observe the behavior of the system better, it is necessary to understand it the distribution of the manually curated test set, the manually curated test set. Fake and Real Accuracies in various Figure 4 displays modalities.

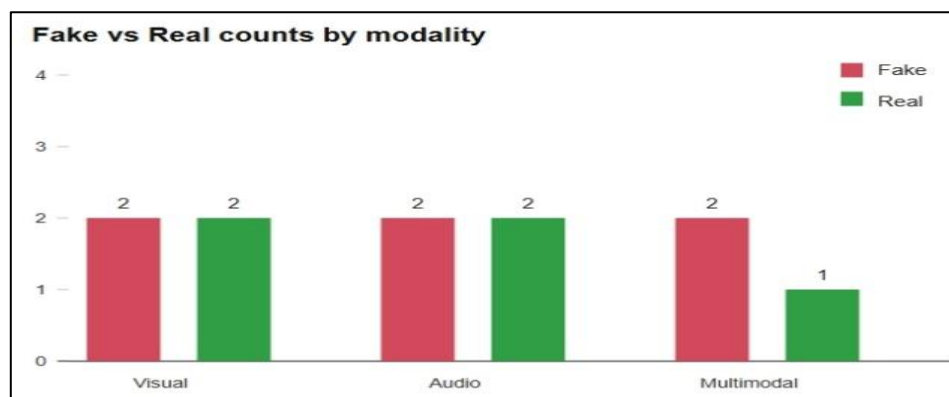


Figure 4. Fake vs Real counts

It is depicted in figure 4 where Fake has a balanced distribution. and Real classifications of visual and of audio. The multi-modal cases show a a whereas the inputs do not a slight difference in Fake detections. This It has been observed that multi-modal analysis facilitates the system to detect other manipulation cues that can be not apparent. when single modality is used individually. These counts are descriptive and are but part to depict system behavior than give some statistical measure of performance.

4.4 Observed Deepfake Artifact Characteristics

The categories of the reported artifacts and observations. The summaries of the system during testing are in. Figure 5. A primary was given to each test case category of observation depending on the most. characteristic that was highly visible.

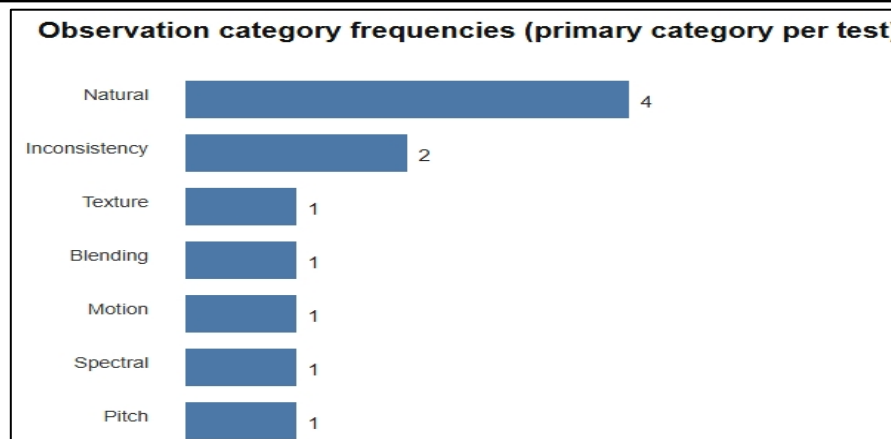


Figure 5. Frequency of primary observations.

As indicated in Figure 5, natural facial and speech. Naturally, characteristics were most common. observed category which corresponds to genuine samples With respect to manipulated samples, cross modal inconsistencies as synchronization. Characters made more mistakes than singular visual or audio artifacts. Texture irregularities, and by mixing up the artefact, discrepancy of movement, the anomalies of spectral, and the deviations of pitch were also frequented, though the less often. These results point to it as being that multimodal fusion is specifically good with indicating synchronization associated inconsistencies, which are hard to in single-modality analysis.

In general, the findings indicated that the suggested multi-modal deepfake detection. system is always consistent in various conditions. input types. Image and videotext recognition have sound visual analysis, and audio based detection records speech-related manipulation cues. The integration of audio and Because of the implication of fusion in multimodal visual information results in better accuracy of detection and more observations made in cases and informative in particular. involving cross-modal inconsistencies. The analysis of inference time proves that the system runs under real time CPU limitations beneficial aspect, based edge devices applicability.

5. CONCLUSION

This work proposed a featherlightmulti-modal deepfake discovery frame acclimatized for real- time deployment on edge bias. The system integrates image, videotape, and audio analysis within a unified channel using effective CNN- grounded models and point- position emulsion. By concertedly assaying visual vestiges and audio inconsistencies, the framework is suitable to detect complex deepfake manipulations that involve multiple modalities. All processing is performed locally on the device, barring reliance on cloud infrastructure and addressing crucial enterprises related to latency, data privacy and security, and deployment feasibility. Experimental analysis demonstrates that the multimodal approach provides more stable and dependable discovery behavior compared to single- modality styles while maintaining low computational outflow suitable for CPU- grounded surroundings.

The lightweight design of the proposed framework enables practical use in real-world scenarios such as pre-upload media verification, local content moderation, and personal media authenticity checking. While the system shows strong performance under controlled evaluation settings, further improvements can be explored to enhance robustness in challenging real-world conditions, including low-resolution media, heavy compression, and diverse manipulation techniques. Future work may also investigate adaptive learning strategies and the use of larger, more diverse datasets to improve generalization against emerging deepfake generation methods, ensuring long-term effectiveness as manipulation technologies continue to evolve.

REFERENCES

- [1]. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1–11.

- [2]. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The Deepfake Detection Challenge (DFDC) dataset. arXiv preprint arXiv:2006.07397.
- [3]. Kinnunen, T., Todisco, M., Delgado, H., Evans, N., Yamagishi, J., Lee, K. A., & Sahidullah, M. (2020). ASVspoof 2019: A large-scale public database of synthetic, converted and replayed speech. *Computer Speech & Language*, 64, 1–19.
- [4]. Zi, J., Liao, X., Li, W., Wei, X., & Li, J. (2020). WildDeepfake: A challenging real-world dataset for deepfake detection. *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.
- [5]. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1251–1258.
- [6]. Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. *Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*.
- [7]. Mittal, Y., Agarwal, M., & Gupta, S. (2020). Audio-visual deepfake detection using multimodal feature fusion. *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*.
- [8]. Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. *Proceedings of Interspeech*.
- [9]. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.
- [10]. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
- [11]. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the International Conference on Machine Learning (ICML)*.
- [12]. Thies, J., Zollhöfer, M., & Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*.
- [13]. Korshunov, P., & Marcel, S. (2018). Deepfakes: A new threat to face recognition? Assessment and detection. arXiv preprint arXiv:1812.08685.
- [14]. Ramanaharan, R., & Kannan, S. (2025). Deepfake video detection: Insights into model generalisation. *Journal of Visual Communication and Image Representation*, 98, 103835.
- [15]. Alrashoud, M., & Alromaih, H. (2025). Deepfake video detection methods, approaches, and trends: A comprehensive survey. *Journal of Visual Communication and Image Representation*, 105, 103735.
- [16]. Singh, V. K., & Sharma, A. (2023). Audio deepfake detection: A survey. arXiv preprint.
- [17]. M. (2025). Deepfake detection: A multimodal survey. *Multimodal Deepfake Detection Survey*.
- [18]. Korshunov, P., & Marcel, S. (2025). Deepfake detection using deep learning: A review. *Advances in Research*, 26(4), 555–564.
- [19]. Heidari, A. (2024). Deepfake detection using deep learning methods: A review. *WIREs Data Mining and Knowledge Discovery*.
- [20]. Gupta, G. (2023). A comprehensive review of deepfake detection using multimodal and machine learning techniques. *Electronics*, 13(1).