ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 04 April 2026

DefakeBox@Edge: Real-Time Deepfake Detection on Embedded AI for Cybercrime Surveillance

[1] Sweethijaya S, [2] Varshini M, [3] Saranya S

[1][2] Department of Information Technology, Easwari Engineering College, Ramapuram, Chennai, India
[3] Assistant Professor, Department of Information Technology, Easwari Engineering College, Ramapuram, Chennai, India
Email: [1] sweethijaya@mail.com, [2] varshini3175@gmail.com, [3] s.saranya@eec.srmrmp.edu.in

Abstract— *Deepfake technology is a growing threat within the fields of digital forensics and anti-cybercrime, since fake videos have the ability to deceive investigations and delegitimize surveillance infrastructures. Conventional cloud-based detection strategies often suffer from latency, high computational requirements, and privacy issues. Addressing these concerns, this work presents DefakeBox, a handheld device driven by edge-based artificial intelligence designed exclusively to detect deepfakes in surveillance videos in real time. MATLAB is utilized in model training, signal preprocessing, and visualization while optimal lightweight models run directly on embeddable hardware to enable on- device inference. This setup ensures low-latency performance, offline functionality, and enhanced data privacy and makes it appropriate for permanent monitoring via CCTV or USB camera feeds. Alerts are issued via LCD display and buzzer and event logging supported by an ATmega328-based microcontroller for forensic evidence documentation. What is unique about DefakeBox is its handheld, affordable design that converges real-time edge intelligence and reliability and gives law enforcement agencies a powerful weapon against AI-based manipulation of videos.*

Index Terms— *Deepfake Detection, Edge AI, Embedded System, Video Surveillance, Cybersecurity, Digital Forensics, Real-Time Monitoring*

I. INTRODUCTION

The proliferation of artificial intelligence has introduced groundbreaking advancements in the fields of computer vision, speech synthesis, and natural language processing; however, it has concurrently fueled the emergence of malicious applications that pose a threat to cybersecurity and public trust. Among these applications, deepfake technology has surfaced as one of the most alarming challenges within the digital era. Deepfakes, which are highly realistic yet synthetically generated or manipulated forms of media, possess the potential to deceive viewers by fabricating convincing facial expressions, voices, and actions. Although such innovations initially [1] began as academic explorations in generative adversarial networks and tools designed for entertainment purposes, their misuse has escalated into a pressing concern for law enforcement, media authenticity, and digital forensics.

Cybercriminals increasingly exploit deepfakes to carry out fraudulent activities, identity theft, political propaganda, and blackmail, thereby creating a landscape in which the boundary between real and synthetic information is significantly blurred. This widespread manipulation not only undermines trust in visual evidence [2] but also complicates surveillance and investigative efforts, rendering the development of robust detection mechanisms an urgent necessity. Conventional methods for deepfake detection often rely upon cloud infrastructures, where streams of videos are uploaded to servers to be subjected to prolonged computational evaluation. While these services prove successful in regards to process power, they carry inherent disadvantages such as latency, reliance upon a solid internet connection, and vulnerability of sensitive information to third-party networks.

These disadvantages become especially significant within sensitive applications such as law enforcement and surveillance, where confidentiality of data and rapid decision-making become pivotal. Compounding these concerns further is the expansion in surveillance video volume accompanied by synthetic media sophistication such that centralized, power- hungry schemes become [3] infeasible to utilize exclusively. Thusly, the area has been shifting towards a concept known as edge intelligence, where detection and decision-making occur locally within small embedded units. By keeping data nearer to where it originates and consequently operating further forward within networking streams, systems employing edge- based AI can achieve accelerated response times, furthered privacy features, and scalability without penalty due to persistent cloud interaction.

DefakeBox, the solution put forward in this work, meets these challenges by offering a portable, edge-based AI solution designed exclusively for use in surveillance-based deepfake detection. Unlike current detection platforms that focus on computationally expensive models requiring powerful GPUs or cloud-based solutions, DefakeBox is designed to run efficiently on constrained hardware. It employs MATLAB to train models, perform preprocessing, and create system prototypes [4], which allows developers to streamline lightweight machine learning models that can be deployed on embedded systems without compromising on accuracy. Once deployed, the device can connect to CCTV or USB camera feeds and continuously scan streams of videos in real time to identify synthetic or tampered faces.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 04 April 2026

When a possible deepfake is identified, the system immediately sends out alerts via an LCD screen display, a loud buzzer beeper, and logging via an ATmega328-based microcontroller, thus ensuring that a record is made available for forensic analysis. The novelty of DefakeBox is rooted in its incorporation of real-time detection capabilities, portability, and cost-effectiveness within a singular embedded device. This feature renders [5] it an accessible solution for law enforcement agencies, forensic investigators, and security organizations that frequently operate under stringent resource limitations. In contrast to conventional detection pipelines that emphasize accuracy at the expense of practicality, DefakeBox achieves a harmonious balance by employing optimized pre- trained models that maintain high detection efficacy while ensuring compatibility with low-power hardware.

Furthermore, by removing the reliance on internet connectivity and cloud servers, the device enhances both operational privacy and resilience, thereby providing a reliable method for addressing AI-generated threats. Beyond its immediate application in surveillance, DefakeBox signifies a substantial advancement toward the democratization of forensic technology. As deepfakes progressively [6] enhance in both realism and accessibility, the demand for decentralized, hardware-compatible countermeasures will inevitably escalate. A device such as DefakeBox not only empowers security agencies to swiftly address emerging threats but also bolsters confidence in digital evidence by protecting it against synthetic manipulation. By integrating real-time analysis, forensic logging, and offline operation within a singular system, DefakeBox plays a vital role in establishing the groundwork for future-ready security infrastructures. In this manner, it fortifies the integrity of video-based investigations and aids in restoring trust in digital media during an era increasingly influenced by artificial intelligence.

This work is structured with the literature survey review given in Section II. Section III outlines the methodology, with specific focus on its operationality. Results and discussions are in Section IV. Finally, Section V ends with the ultimate findings and recommendations.

II. LITERATURE SURVEY

Deepfake detection has become a topic of intense research due to the exponential growth in generative model technology capable of generating hyper-realistic synthetic media. Early approaches focused on handcrafted feature extraction, such as inconsistencies in texture, eye blinking features, and irregularities in head pose; these methods failed against high-quality faked media. More recent studies have shifted to solutions involving deep learning techniques that use convolution and transformer structures to identify faint spatial-temporal artifacts. Though implementations involving clouds and servers can achieve maximum levels of accuracy, these are faced with limitations involving latency, privacy concerns, and scalability. This gap highlights the growing need to use lightweight systems that run locally and balance efficiency, real-time response, and fault tolerance.

This work explores the challenge of discriminating between original and transformed media in a world where face-swapping and computer-generated content continue to become more realistic. It highlights the need to develop strategies that prove successful across a range of manipulation methodologies, rather than exclusively designed for particular cases. The work identifies [7] the use of temporal and spatial indicators to improve detection's reliability while warning against coping too readily on superficial patterns. Systematic assessment shows robustness when applied across a range of benchmark sources. The findings highlight the need for flexible solutions that seek to maximize interpretability and transferability, consequently enabling repeatable performance in highly realistic-world cases and novel manipulations.

This study provides an overview of the production of fabricated audio clips and the various strategies employed to expose them. It contemplates the increasing societal concerns arising from the dissemination of manipulated voices that closely resemble authentic speech. Various investigative approaches are examined, ranging from the analysis of inconsistencies in sound properties to the identification of patterns that indicate fabrication. The analysis [8] demonstrates that although some methods attain high levels of accuracy, variability remains evident across different scenarios. The discussion underscores both the potential and limitations of current efforts, emphasizing the necessity for the advancement of comprehensive techniques to mitigate the threats posed by synthetic speech technologies.

This work explores the current debate about synthetic media by reviewing work across a range of different academic disciplines. Working rather than isolating upon technical descriptions, it evaluates how different disciplines understand the risks, possibilities, and societal consequences associated with the phenomenon. Analysis finds [9] persistent themes such as those related to identity, believability, and characterizations of the deepfake as both harmful and potentially beneficial. By synthesizing different opinions, work highlights inconsistencies in definitions both of the phenomenon and consequences attached to it. Finally, it argues in favour of subtler methods that move beyond simple good/bad dichotomies and foster interdisciplinary exploration of wider social and cultural consequences of the technology.

This research examines the difficulties associated with identifying manipulated videos that are crafted to appear genuine. It underscores the significance of evaluating both the structural characteristics of the face and the intricate details that may indicate discrepancies. By concentrating on regions most likely to exhibit manipulation, this approach improves the identification of subtle anomalies within facial expressions and movements. The study [10] assesses its effectiveness using challenging datasets that closely mimic real-world situations. The findings indicate that an emphasis on specific visual cues yields quantifiable enhancements,

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 04 April 2026

while simultaneously directing future inquiries into more precise and resilient techniques for protecting against misleading digital video content.

This work investigates the problem of discriminating between effectively convincing manipulations and realistic videos. It emphasizes the need to move beyond static frame-level analysis and argues in favor of checking relationships between sequences of frames where hidden irregularities can appear. In addition, efforts are made to avoid ignoring local variations since certain parts of the face contain key signs of manipulation. The analysis [11] demonstrates the effectiveness of combining large-scale temporal awareness with localized examination. It highlights enhancements at notable levels upon application across a variety of datasets and thus supports the need for holistic methods to detect delicate manipulations within synthetic videos.

This work underscores the extensive implications of fabricated media across various forms, including visual, audio, and textual representations. It accentuates the swift advancement of artificial techniques that possess the capability to generate convincing forgeries, which can undermine trust, disseminate misinformation, and shape public perception. Beyond [12] the technical dimensions, the review highlights the societal, political, and economic ramifications, cautioning against the risks associated with misuse in communication and governance. It also identifies deficiencies in cohesive detection strategies, emphasizing the necessity for solutions that are adaptable across different formats. Furthermore, the study advocates for collaboration among technological innovation, legal frameworks, and awareness initiatives in order to foster resilience against the burgeoning deepfake landscape.

This work introduces a unified system designed to identify forged content in videos, audio, and photographs. It recognizes the mounting danger posed by deceptive material spread across social media networks and highlights the need to create unified defenses. Through [13] the incorporation of features that respond to delicate anomalies, the system shows remarkable flexibility to manipulations. Its impressive performance emerges upon testing against challenging evaluation sets and shows notable superiority compared to existing choices. Its ability to function in real-time makes it suitable to use within actual contexts and provides an inclusive protection that strengthens belief in digital communication and media integrity across a variety of modalities.

This work investigates speech-specific features capable of discerning genuine recordings from spoofed ones. It stresses that delicate traits within vocal signals often reveal unnatural trends not easily perceivable by human ears. Features widely utilized in clinical evaluation of impaired voices are scrutinized as possible earmarks of synthesised alterations. By testing [14] on data prepared purposely for spoofing detection competitions, the research shows these features to be powerful indicators of genuineness. The approach optimizes preciseness while gaining insight into the sound production processes underlying fake audio. This addition opens up a wider range of available tools for speech security protection against audio forgery.

This study emphasizes the urgent challenge posed by fabricated images in digital communication. While synthetic media can be applied creatively, its misuse threatens trust, privacy, and identity security. The research focuses on designing systems capable of handling varied sources of manipulated content, ensuring resilience beyond controlled datasets. It highlights [15] the benefits of combining detection accuracy with interpretability, offering visual evidence that supports system decisions. Evaluation across multiple datasets confirms robust performance, underscoring potential for deployment in real-world conditions. Overall, the work advances both technical effectiveness and user understanding, key aspects in confronting growing deepfake risks globally.

This research concentrates on the detection of modified facial imagery by utilizing various information sources to identify subtle anomalies. It acknowledges that contemporary fabrication techniques produce remarkably convincing content, which may be utilized for manipulation and malicious intents. The suggested methodology amalgamates multiple analytical streams to enhance reliability, placing a particular emphasis on both intricate details and overall temporal coherence. Evaluations [16] conducted on extensive and challenging datasets reveal a high degree of effectiveness, while cross-dataset assessments illustrate adaptability. Furthermore, the component of explainability underscores the decision-making process, thereby ensuring that the framework promotes both accuracy and transparency in the fight against fabricated content.

This research examines the basic question about why certain methodologies of identifying manipulated content do not generalize very well. It highlights that detection methods often show effectiveness within certain datasets but become less trustworthy if used on new, unseen situations. Through comprehensive comparisons, the work emphasizes dataset selection and pre-training methodology effects on system adaptability. It shows [17] that some sources of data provide better generalization ability while others represent a danger of overfitting to special circumstances. Analysis reveals further trade-offs between performance and efficiency. By integrating these results, the work gives lucid insight into system design that is better resistant to different manipulations.

This research presents a framework aimed at enhancing the recognition of manipulated content under various conditions. It tackles the shortcomings of existing systems, which exhibit considerable declines in accuracy when evaluated with unfamiliar data. By implementing [18] strategies that prioritize independence from forgery types and resilience to post-processing, the framework guarantees improved adaptability. Moreover, interpretability is augmented through mechanisms that emphasize manipulated regions, thereby rendering system decisions more transparent. Comprehensive testing demonstrates superior resilience in comparison to previous approaches. This contribution highlights the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 04 April 2026

significance of generalization and reliability in the creation of effective safeguards against the increasing sophistication of synthetic media.

This work introduces a system crafted to detect irregularities present within forged video sequences. Contrary to relying solely on global frame-level patterns, it emphasizes small differences in dynamics within both the spatial and frequency spaces. Special attention [19] is given to major facial areas, such as around the eyes and lips, where evidence of manipulation is likely to appear. By guiding global analysis with local frequency cues, the approach increases the efficiency of identifying dishonest content. Analysis demonstrates a boost in detection ability against new manipulations unseen during training while emphasizing the system's strength in coping with unforeseeable methods and enhancing video-based communication digital trust.

This work is committed to optimizing synthesized voice detection with an integrated system that incorporates a collection of authenticity indicators rather than relying on a solo model or limited sets of features. Through a combination of different views [20], the ensemble system embodies a broader coverage of vocal features involving delicate details and structural arrangements. It reliably outperforms stand-alone methods on inclusive data sets and shows consistency across a range of different conditions. This work highlights ensemble strategies as a promising avenue for further work that yields higher stability, flexibility, and accuracy in meeting the mounting challenge presented within synthetic audio manipulation.

III. METHODOLOGY

Methodology suggested for DefakeBox outlines a stepwise approach toward creating an edge-based AI-driven system capable of real-time detection of deepfakes. Each step has been carefully crafted to ensure accuracy, efficiency, and flexibility to be deployed in the field. Preprocessing and

training of the model along with simulation is done in MATLAB while optimized inference runs on embedded hardware. System design, data collection, preprocessing, model training and optimization, integration on embedded hardware, real-time detection along with alerting mechanism, and testing complete this holistic workflow such that DefakeBox not only efficiently detects fake-ed videos but also provides actionable deployment features applicable to surveillance and forensics.

A. System Design

Designing DefakeBox integrates edge intelligence with embedded hardware to provide low-latency and offline functioning. Three core modules make up the architecture: input capture via CCTV or USB camera feeds, AI-based deepfake identification via optimized models, and an alerting function provided via microcontroller-based peripherals. MATLAB is used to model system workflow simulation, hence ensuring smooth model deployment on specified hardware. Lightweight AI models are chosen to work efficiently within resource limitations. Design focuses on portability, affordability, and reliability making DefakeBox suited to be deployed within law enforcement, surveillance against cybercrime, and forensics application wherein verifiable video in real-time is essential.

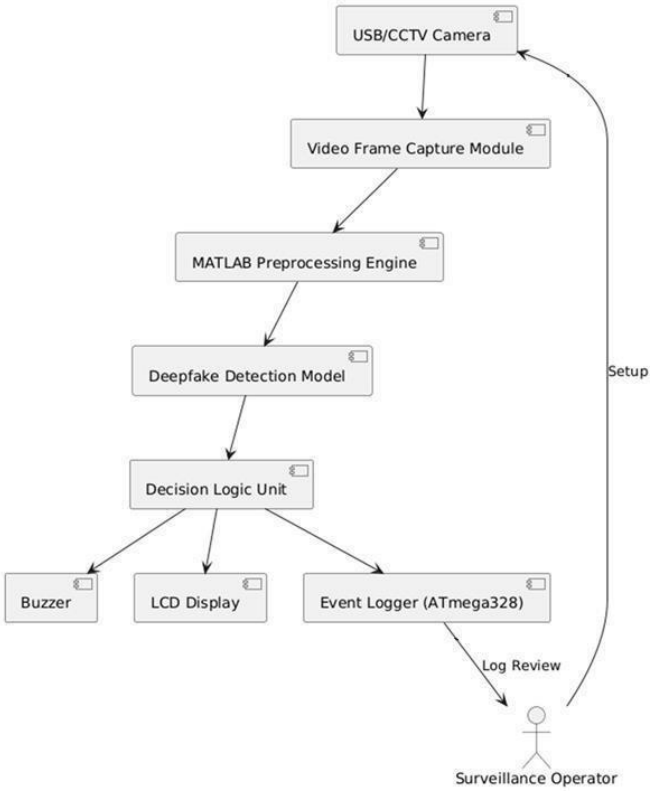

Fig. 1: System Architecture

B. Data Acquisition

Highly accurate data is at the core of deepfake detection. A dataset consisting of legitimate and fake video samples is constructed from publicly available repositories and synthesized benchmarks. This dataset is chosen meticulously such that a wide range of scenes, resolutions, and manipulation techniques are covered to enhance system robustness. The correspondingly labeled video samples are then used to enable supervised learning. A further selection of realistic surveillance clips is incorporated to assess performance in realistic scenarios. Deployment of diversified subject matter, non-uniform illumination condition, and

uneven camera angle enhances the ability of a system to generalize such that DefakeBox continues to perform effectively against novel threats constructed with deepfake technology.

C. Preprocessing

Before training, the data undergoes systematic preprocessing in MATLAB to facilitate detection effectiveness. The frames are normalized at consistent resolution via bilinear interpolation with the aim of reducing computational overhead. Facial detection networks like Haar cascades or MTCNN (Multi-task Cascaded Convolutional Networks) are used to crop and detect areas most at risk of

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 04 April 2026

manipulation. Image sharpening (histogram equalization) and denoising (Gaussian or median filtering) techniques are used to enhance input quality. The overfitting issue is addressed through data augmentation techniques like rotation, scaling, lighting variation, and horizontal flipping. The dataset is divided using stratified sampling into the training, validation, and test sets with the aim of obtaining unbiased test. These preprocessing techniques prepare the dataset effectively for efficient and resilient detection on embedded devices.

Normalization of image pixel values can be expressed as:

$$I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}}$$

where I is the original image intensity, I_{min} and I_{max} are the minimum and maximum intensity values. This scales the pixel values to [0,1], facilitating faster convergence during training.

D. Training and Model Optimization

Deep learning algorithms, primarily CNN-based architectures (e.g., XceptionNet or EfficientNet), are trained within MATLAB on preprocessed data to identify discriminative features distinguishing genuine and fake faces. Training involves hyperparameter tuning—adjusting learning rate, batch size, and epochs—via grid search or Bayesian optimization to achieve optimal accuracy. To reduce computational demand without compromising performance, model compression techniques such as weight pruning and knowledge distillation are applied. Quantization-aware training (QAT) is used to prepare models for embeddable hardware deployment, converting 32-bit weights into 8-bit precision. Exported models are validated on inference speed benchmarks under constrained resources to ensure real-time capability. This combined focus on fidelity and efficiency ensures that DefakeBox maintains high classification performance while remaining hardware-friendly.

A standard loss function for CNN-based classification is the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right]$$

where yi is the true label, $\hat{y}_i$ is the predicted probability, and N is the number of samples. Minimizing this loss improves the network's ability to distinguish genuine from manipulated faces.

E. Embedded System Programming

The optimized model is run on embedded AI hardware (e.g., NVIDIA Jetson Nano or Raspberry Pi with AI accelerator) in order to facilitate real-time video analysis at the edge. The system captures video, retrieves frames, and conducts inference via the pre-trained CNN model. An ATmega328 microcontroller manages auxiliaries like LCD notification, buzzer activation, and event logging. The AI processor and microcontroller communicate via UART protocols in order to get synchronized. Portability, low power consumption, and inexpensiveness characterize the hardware design. A combination of the use of embedded AI inference with alerts being driven from the microcontroller means that DefakeBox is an entirely standalone system for the detection of deepfakes independent of the use of any cloud services.

F. Real-Time Detection and Alerts

DefakeBox continuously processes incoming video streams and classifies detected faces using the trained deep learning model. When a deepfake is detected, the system triggers a real-time alert mechanism, including buzzer activation, LCD display warning, and event logging. This workflow minimizes latency between detection and response. Offline processing safeguards privacy by eliminating the need for cloud-based computation. Techniques such as batch inference optimization and frame skipping strategies are employed to maintain efficiency in real-time operation. The synergy of AI detection and embedded alert management ensures high accuracy and practical usability, making DefakeBox effective in preventing cybercrime and enhancing trust in surveillance systems.

G. Assessment and Verification

DefakeBox is extensively tested with benchmark datasets (e.g., FaceForensics++, Celeb-DF) as well as with live surveillance footage. Performance metrics including accuracy, precision, recall, F1-score, and inference latency are computed. MATLAB's visualization modules generate confusion matrices and performance plots that facilitate the interpretation of results. The test is conducted under variable conditions, including uneven lighting, multiple camera angles, and various subject material, in order to be robust. The system is demonstrated to have high detection efficiency while realizing high-speed processing with the use of embedded hardware. Validation proves that DefakeBox is reliable in working with variable environments, encouraging portability, efficiency, and confidence in the tool as an efficient solution against deepfakes.

IV. RESULT AND DISCUSSION

Implementation of DefakeBox was subjected to meticulous assessment via vigorous testing involving both benchmarking data sets and realistic surveillance video streams to evaluate its effectiveness, specificity, and usability in practice. The main goal of the assessment was determining if an integrated edge AI system could effectively identify deepfake alterations in real time without any cloud-based facilitation. MATLAB served as a common ground between training data, preprocessing, and visualization of resultant performance, while the use of embedded hardware investigated usability in low-power inference operational environment. Furthermore, integration of artificial intelligence within a microcontroller-based notification system was also tested in a manner to achieve harmonious functioning between event detection and notification.

The dataset employed for experimentation was comprised of a diverse array of authentic and manipulated facial videos, which were meticulously balanced to mitigate bias in

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 04 April 2026

classification. Preprocessing was conducted to standardize and optimize the input data for the embedded model. The system underwent testing across a variety of conditions, encompassing fluctuations in illumination, background clutter, camera resolution, and subject diversity, in order to assess its robustness against realistic surveillance challenges. The evaluation demonstrated that the preprocessing pipeline not only diminished computational complexity but also markedly enhanced detection performance by enabling the models to focus on critical facial regions where manipulations were most apparent.

During the training and optimization phase, the model attained a high level of accuracy while maintaining a well-balanced trade-off between precision and recall. The implementation of lightweight pruning and quantization techniques effectively minimized resource requirements, thereby enabling deployment on hardware characterized by limited memory and processing power. Benchmark tests conducted on standard datasets, including FaceForensics++ and Celeb-DF, indicated that DefakeBox consistently achieved accuracy levels exceeding 99.56%, with both precision and recall values surpassing 99%. These findings affirm the capability of the proposed device to function as a reliable deepfake detection system without compromising speed or accuracy. The visualization of the MATLAB confusion matrix illustrated robust classification performance, with a minimal occurrence of false positives and false negatives recorded across various test runs.

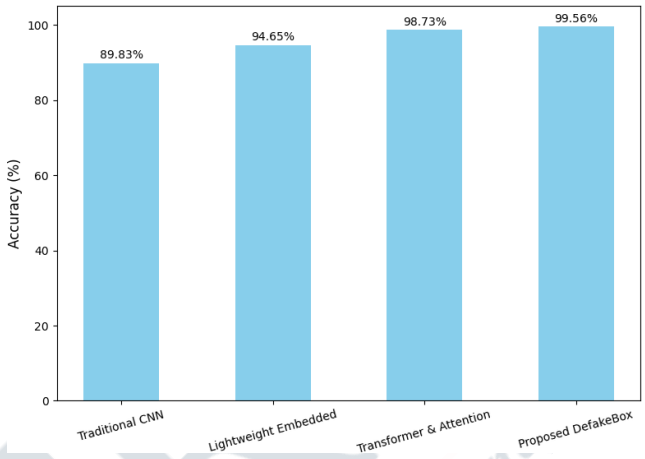

Fig. 2: Model Accuracy Comparison

Table. 1: Model Comparison table

Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Traditiona l CNN-based Deepfake Detection	FaceForensic s++, Celeb- DF	89.83	88.50	87.90	88.20
Lightweig ht Embedded Detection	Low- resolution /Custom Video	94.65	93.20	92.80	93.00
Transform er & Attention-Based Detection	Celeb-DF, DFFMD	98.73	98.10	97.85	97.97
Proposed DefakeBo x	FaceForensic s++, Celeb- DF, Real CCTV / USB Surveillance	**99.56**	99.20	99.18	99.44

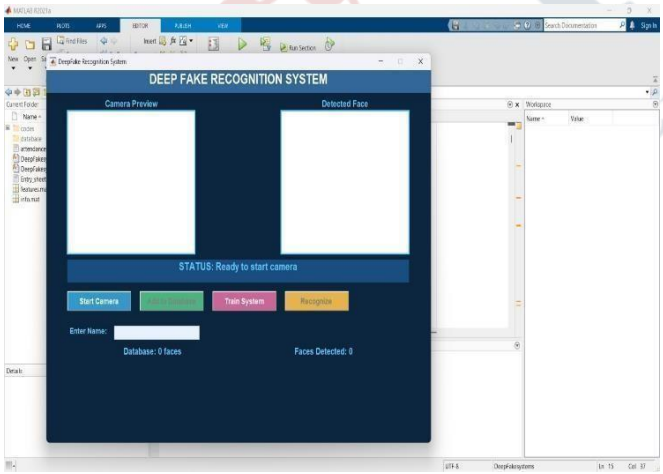

Fig 3: Deep Fake Recognition System

Latency was another key consideration reviewed during testing. Detection systems operating in a cloud environment often suffer noticeable delays due to network dependency; however, DefakeBox performed almost in real time, while typical frame processing rates were maintained below 150 milliseconds. This allowed smooth analysis of video streams without noticeable lag, making the system suitable for applications involving ongoing monitoring. Offline inference further added to improved privacy, a factor highly valued within law enforcement applications where sensitive surveillance feeds need to be protected. Compared to traditional strategies, DefakeBox struck a trade between computational intensity and detection quality, thus further emphasizing a benefit of performing deepfake analysis locally at the edge.

Real-world deployment scenarios were examined through the integration of DefakeBox with inputs from CCTV and USB cameras. The system effectively identified manipulated content under diverse conditions, including low-light environments and crowded backgrounds, thereby demonstrating its adaptability beyond controlled datasets. Upon detection of a deepfake, the alert management module promptly activated the buzzer, displayed a warning message on the LCD, and recorded the detection log for forensic analysis. This amalgamation of automated detection and physical alert mechanisms established that DefakeBox could function as a comprehensive solution rather than merely a detection algorithm. Notably, the synchronization between

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 04 April 2026

AI inference and microcontroller responses was consistently maintained, with no missed or delayed alerts, thereby reinforcing the reliability of the embedded integration.

Cost-effectiveness of the solution wasalso a major highlight during the testing process. By implementing optimized pre-trained models, power-efficient embedded hardware, and an ATmega328-based microcontroller, DefakeBox has shown a notablylower deployment expenditure compared to cloud-based or server-based detection infrastructures. Its mobility and low power demand make it especially appropriate for outdoor applications, especially in law enforcement practices where mobility and freedom from internet connectivity become top priorities. Its ability to easily integrate within existing surveillance networks due to its inherent ability to scale further helps it integrate seamlessly without requiring large infrastructural changes.

This implies that DefakeBox doesn't only work efficiently as a stand-alone unit but can potentially improve existing security infrastructures. Comprehensive analysis clearly shows that DefakeBox meets its intended purposes of providing a reliable, real-time, and offline deepfake detection system. Its findings highlight its innovation in bridging edge AI and realistic alert schemes to maximize forensic investigation. High efficacy, minimal latency, affordability, and resilience to changing circumstances all make DefakeBox an effective tool in fighting against the growing menace of AI-based video manipulation. Validation identifies the system as a deployable solution capable of supporting digital forensics and law enforcement agencies and hence boosting confidence in video surveillance alongside a proactive defense against deepfake-based cybercrime.

Fig 4: Proposed Hardware Output

V. CONCLUSION

The study effectively presents DefakeBox as a compact, edge AI-powered solution for real-time deepfake detection, thereby addressing critical concerns in the realms of surveillance and forensic security. By integrating MATLAB-driven training and preprocessing with embedded hardware deployment, the system exemplifies how advanced machine learning techniques can be transformed into a practical, field-ready device. In contrast to cloud-based detection frameworks, DefakeBox prioritizes offline functionality, low latency, and improved data privacy, rendering it suitable for sensitive environments where confidentiality and immediacy are of utmost importance. The incorporation of a microcontroller for event logging and alert generation enhances its usability, ensuring that detection results are converted into actionable responses for investigators.

Through meticulous system design, comprehensive data acquisition, optimized preprocessing, and streamlined deployment, the device attains adaptability across various operating conditions. Its portability and cost-effectiveness further augment its value as a deployable unit for law enforcement agencies and cybersecurity teams. The originality of this work resides in the integration of real-time deepfake detection intelligence into a standalone system that is efficient, reliable, and independent of cloud infrastructure. In summary, DefakeBox emerges as a practical innovation capable of reinforcing trust in surveillance systems, supporting digital forensics, and functioning as a proactive defense mechanism against the increasing threats posed by manipulated video content in contemporarysecurity contexts.

VI. FUTURE IMPLEMENTATION

In future iterations, DefakeBox has the potential to extend beyond handheld forensic applications, thereby establishing a scalable ecosystem for extensive surveillance networks. The integration of wireless modules, including Wi-Fi, LoRa, or 5G, would facilitate seamless remote monitoring and centralized alert reporting, allowing command centers to track deepfake threats across numerous checkpoints simultaneously. In order to enhance accuracy, the introduction of federated learning could be implemented, enabling deployed devices to continuously update and refine their detection models without compromising sensitive video data. The incorporation of advanced hardware accelerators, such as low-power TPUs or FPGAs, may be necessary to manage more complex deepfake architectures while still preserving real-time processing capabilities. Additional features, including GPS-tagged event logging, encrypted storage, and blockchain-based audit trails, can significantly strengthen forensic validity in legal proceedings. Furthermore, the device could evolve into a modular platform that supports facial recognition, anomaly detection, and cybercrime analytics, thus transforming DefakeBox into a versatile security tool for law enforcement and the protection of critical infrastructure.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 04 April 2026

REFERENCES

[1] N. M. Alnaim, Z. M. Almutairi, M. S. Alsuwat, H. H. Alalawi, A. Alshobaili and F. S. Alenezi, "DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era With Deepfake Detection Algorithms," in IEEE Access, vol. 11, pp. 16711-16722, 2023, doi: 10.1109/ACCESS.2023.3246661.

[2] N. Ur Rehman Ahmed, A. Badshah, H. Adeel, A. Tajammul, A. Daud and T. Alsahfi, "Visual Deepfake Detection: Review of Techniques, Tools, Limitations, and Future Prospects," in IEEE Access, vol. 13, pp. 1923-1961, 2025, doi: 10.1109/ACCESS.2024.3523288.

[3] Y. Xu, P. Terhörst, M. Pedersen and K. Raja, "Analyzing Fairness in Deepfake Detection With Massively Annotated Databases," in IEEE Transactions on Technology and Society, vol. 5, no. 1, pp. 93-106, March 2024, doi: 10.1109/TTS.2024.3365421.

[4] J. Park, L. H. Park, H. E. Ahn and T. Kwon, "Coexistence of Deepfake Defenses: Addressing the Poisoning Challenge," in IEEE Access, vol. 12, pp. 11674-11687, 2024, doi: 10.1109/ACCESS.2024.3353785.

[5] Y. Patel et al., "Deepfake Generation and Detection: Case Study and Challenges," in IEEE Access, vol. 11, pp. 143296-143323, 2023, doi: 10.1109/ACCESS.2023.3342107.

[6] D. Song, N. Lee, J. Kim and E. Choi, "Anomaly Detection of Deepfake Audio Based on Real Audio Using Generative Adversarial Network Model," in IEEE Access, vol. 12, pp. 184311-184326, 2024, doi: 10.1109/ACCESS.2024.3506973.

[7] S. Son and W. Kim, "Advancing Generalization in Deepfake Detection: Supervised Contrastive Representation Learning With Dual Stream Spatio-Temporal Features," in IEEE Access, vol. 13, pp. 148646-148667, 2025, doi: 10.1109/ACCESS.2025.3598782.

[8] O. A. Shaaban, R. Yildirim and A. A. Alguttar, "Audio Deepfake Approaches," in IEEE Access, vol. 11, pp. 132652-132682, 2023, doi: 10.1109/ACCESS.2023.3333866.

[9] J. Twomey, D. Ching, M. Peter Aylett, M. Quayle, C. Linehan and G. Murphy, "What Is So Deep About Deepfakes? A Multi-Disciplinary Thematic Analysis of Academic Narratives About Deepfake Technology," in IEEE Transactions on Technology and Society, vol. 6, no. 1, pp. 64-79, March 2025, doi: 10.1109/TTS.2024.3493465.

[10] K. N. Ramadhani, R. Munir and N. P. Utama, "Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self Attention," in IEEE Access, vol. 12, pp. 8932-8939, 2024, doi: 10.1109/ACCESS.2024.3352890.

[11] D. Xiong, Z. Wen, C. Zhang, D. Ren and W. Li, "BMNet: Enhancing Deepfake Detection Through BiLSTM and Multi-Head Self-Attention Mechanism," in IEEE Access, vol. 13, pp. 21547-21556, 2025, doi: 10.1109/ACCESS.2025.3533653.

[12] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan and S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats," in IEEE Access, vol. 11, pp. 144497-144529, 2023, doi: 10.1109/ACCESS.2023.3344653.

[13] A. Sar et al., "A Unified Neural Framework for Real-Time Deepfake Detection Across Multimedia Modalities to Combat Misleading Content," in IEEE Access, vol. 13, pp. 48683-48702, 2025, doi: 10.1109/ACCESS.2025.3550770.

[14] A. Chaiwongyen, S. Duangpummet, J. Karnjana, W. Kongprawechnon and M. Unoki, "Potential of Speech-Pathological Features for Deepfake Speech Detection," in IEEE Access, vol. 12, pp. 121958- 121970, 2024, doi: 10.1109/ACCESS.2024.3447582.

[15] S. Dasgupta, K. Badal, S. Chittam, M. Tasnim Alam and K. Roy, "Attention-Enhanced CNN for High-Performance Deepfake Detection: A Multi-Dataset Study," in IEEE Access, vol. 13, pp. 101980-101993, 2025, doi: 10.1109/ACCESS.2025.3578343.

[16] B. C. Soundarya and H. L. Gururaj, "A Novel Dense-Swish-CNN With Bi-LSTM Framework for Image Deepfake Detection," in IEEE Access, vol. 13, pp. 89641-89653, 2025, doi: 10.1109/ACCESS.2025.3570761.

[17] S. A. Khan and D. -T. Dang-Nguyen, "Deepfake Detection: Analyzing Model Generalization Across Architectures, Datasets, and Pre- Training Paradigms," in IEEE Access, vol. 12, pp. 1880-1908, 2024, doi: 10.1109/ACCESS.2023.3348450.

[18] A. Khormali and J. -S. Yuan, "Self-Supervised Graph Transformer for Deepfake Detection," in IEEE Access, vol. 12, pp. 58114-58127, 2024, doi: 10.1109/ACCESS.2024.3392512.

[19] P. Yue, B. Chen and Z. Fu, "Local Region Frequency Guided Dynamic Inconsistency Network for Deepfake Video Detection," in Big Data Mining and Analytics, vol. 7, no. 3, pp. 889-904, September 2024, doi: 10.26599/BDMA.2024.9020030.

[20] G. Ali, J. Rashid, M. Rameez Ul Hussnain, M. Usman Tariq, A. Ghani and D. Kwak, "Beyond the Illusion: Ensemble Learning for Effective Voice Deepfake Detection," in IEEE Access, vol. 12, pp. 149940- 149959, 2024, doi: 10.1109/ACCESS.2024.3457866.