iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81335

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

AI Generated Media Classification Using Deep Learning

Nizampatnam Sai Siri[1], Meruga Asha[2], Uppalapati Nithisha[3], Puli Sathwika[4]
[1, 3, 4]*Dept. of CSE, Bapatla Women's Engineering College, Bapatla, India*
[2]*Assistant Professor, Dept. of CSE, Bapatla Women's Engineering College, Bapatla, India*

Abstract: ***The rapid evolution of generative artificial intelligence has significantly accelerated the creation of highly realistic deepfake content across multiple media formats, posing serious threats to digital trust, cybersecurity, and information authenticity. This paper presents a unified multi-modal deepfake detection framework capable of identifying AI-generated or manipulated content across four modalities: image, video, audio, and text. The proposed system integrates lightweight deep learning architectures and classical machine learning models within a scalable Django-based web application.For image and video detection, a MobileNetV3-based convolutional neural network is employed to enable efficient spatial feature extraction and real-time inference. Video classification is performed through frame-wise analysis with prediction aggregation to ensure computational efficiency without sacrificing reliability. Audio deepfake detection is implemented using Mel-Frequency Cepstral Coefficient (MFCC) feature extraction combined with a Support Vector Machine classifier. Text-based AI content detection utilizes TF-IDF vectorization with Logistic Regression to distinguish human-written content from machine-generated text. Experimental evaluation demonstrates strong performance in image classification with an overall accuracy of 89.29%, while maintaining low computational overhead suitable for real-time deployment. The proposed architecture emphasizes modularity, scalability, and cross-modal adaptability, making it suitable for practical cybersecurity applications and web-based deployment environments. The framework demonstrates the feasibility of integrating heterogeneous detection models into a unified, production-ready deepfake detection system.***

Keywords: ***Deepfake Detection, Multi-Modal Learning, MobileNetV3, Convolutional Neural Networks, Support Vector Machine, Logistic Regression, MFCC, TF-IDF, Video Frame Analysis, Real-Time Detection, Cybersecurity, Artificial Intelligence, Django Deployment.***

I. INTRODUCTION

The rapid advancement of generative artificial intelligence, particularly Generative Adversarial Networks (GANs) and diffusion-based models, has significantly transformed the landscape of synthetic media generation. Modern AI systems are now capable of producing hyper-realistic images, videos, audio clips, and textual content that are often indistinguishable from authentic data. While such advancements have enabled innovation in creative industries, virtual environments, and digital media production, they have simultaneously introduced serious concerns related to misinformation, identity manipulation, and digital security.

The widespread availability of AI-powered content generation tools has accelerated the proliferation of deepfake images and videos, synthetic voice cloning, and AI-generated textual misinformation. These technologies pose significant threats to public trust, cybersecurity, and media authenticity. Deepfake videos can manipulate public figures, AI-generated audio can impersonate individuals, and machine-generated text can be used to distribute misleading information at scale. Consequently, the development of robust and scalable detection mechanisms has become an urgent research priority. Traditional image verification techniques, which rely primarily on pixel-level statistical analysis or handcrafted features, are increasingly ineffective against modern generative models. Contemporary synthetic content generation leverages complex adversarial training mechanisms and high-capacity neural architectures capable of capturing fine-grained visual and temporal patterns. As a result, advanced detection strategies based on deep learning have emerged as the dominant approach for identifying manipulated or AI-generated content. Convolutional Neural Networks (CNNs) have demonstrated remarkable effectiveness in visual recognition tasks such as image classification, object detection, and feature extraction. Lightweight architectures such as Mobile-Net have further enabled efficient deployment of deep learning systems in real-time and resource-constrained environments. Motivated by these developments, this work proposes a unified multi-modal deepfake detection framework that integrates deep learning and classical machine learning techniques within a web-deployable architecture.

©IJRASET: All Rights are Reserved

Unlike approaches that focus solely on image-level analysis, the proposed system addresses four complementary modalities: image, video, audio, and text. For image and video detection, a Mobile-Net-based convolutional neural network is employed to perform spatial feature extraction. Video analysis is conducted through frame-wise classification with prediction aggregation to ensure computational efficiency while maintaining detection reliability. Audio deepfake detection utilizes Mel-Frequency Cepstral Coefficient (MFCC) feature extraction combined with a Support Vector Machine (SVM) classifier, while AI-generated text detection is implemented using TF-IDF vectorization with Logistic Regression.

The proposed system is implemented as a Django-based web application, enabling real-time and offline detection across multiple media formats. This deployment-oriented design demonstrates the practical feasibility of integrating heterogeneous detection models into a unified, production-ready cybersecurity framework.

Furthermore, the architecture emphasizes computational efficiency, modular scalability, and cross-modal adaptability. By leveraging lightweight neural networks and classical machine learning techniques, the system balances performance and inference speed, making it suitable for real-world deployment scenarios.

In summary, this research contributes a comprehensive multi-modal deepfake detection framework capable of identifying AI-generated content across diverse media types. The results demonstrate the effectiveness of Mobile-Net-based convolutional architectures for visual deepfake detection while highlighting the importance of integrating multiple modalities to strengthen digital media authentication systems.

The remainder of this paper is organized as follows. Section II presents the related work. Section III describes the proposed methodology and system architecture. Section IV discusses experimental results and performance evaluation. Finally, Section V concludes the paper and outlines future research directions.

II. LITERATURE SURVEY

The rapid growth of generative models such as Generative Adversarial Networks (GANs) and diffusion-based architectures has significantly intensified research in synthetic media detection. As AI-generated images, videos, audio, and textual misinformation become increasingly realistic, distinguishing fabricated content from authentic media has become a critical challenge in digital forensics and cybersecurity. Recent literature focuses on deep learning-based detection mechanisms, multimodal fusion strategies, spectral artifact analysis, and explainability-driven forensic techniques.

Several studies have explored feature fusion and ensemble strategies to enhance detection robustness. Alrowais *et al.* [1] proposed a deep feature fusion-based model for detecting GAN-generated fake faces, demonstrating that combining complementary deep representations improves classification accuracy. Similarly, Raza *et al.* [8] introduced MMGANGuard, a multi-model framework combining ResNet and DenseNet architectures to achieve high-accuracy GAN image detection. These approaches highlight the importance of leveraging diverse feature extractors to capture subtle inconsistencies introduced during synthetic generation.

Diffusion model detection has also gained attention due to the rise of text-to-image generative systems. Bammey [9] proposed "Synthbuster," which identifies diffusion-generated images through spectral residual artifacts present in the frequency domain. Kang *et al.* [10] further demonstrated that residual noise patterns and manipulation traces can be effectively exploited for detecting various deepfake types. These methods emphasize that generative models often leave detectable statistical footprints, particularly in high-frequency components.

Explainability in deepfake detection has emerged as a critical research direction. Bird and Lotfi [7] introduced CIFAKE, which integrates explainable AI mechanisms to identify regions that contribute most significantly to classification decisions. Cartella *et al.* [13] analyzed human gaze patterns when observing fake images, contributing insights into interpretability and cognitive response modeling. Such works underline the importance of transparency in detection systems to avoid black-box decision-making.

Beyond unimodal image detection, multimodal fake news and misinformation detection frameworks have shown superior performance by integrating text and visual features. Hamed *et al.* [2], Luqman *et al.* [3], Li *et al.* [4], and Liu *et al.* [5] proposed hybrid and contrastive learning-based multimodal fusion mechanisms that combine textual and visual embeddings to improve detection reliability. These studies demonstrate that inconsistencies across modalities often reveal deceptive content more effectively than single-modality analysis.

Digital watermarking has also been investigated as a preventive and detection mechanism. Malanowska *et al.* [6] provided a meta-survey on watermarking techniques for fake news detection, highlighting both its potential and limitations in large-scale misinformation environments. While watermarking provides proactive verification, it is not always applicable to already-circulating synthetic media.

Hybrid deep learning and machine learning classifiers have further enhanced detection systems. Al-Dulaimi and Kurnaz [11] combined CNN architectures with Random Forest and SVM classifiers for multi-dataset deepfake detection, demonstrating improved generalization. Bhargava *et al.* [12] proposed a CNN-based neural network for forgery identification through pixel-level manipulation analysis. These works show that combining deep feature extraction with classical classifiers can improve robustness and adaptability. Image registration and cross-domain feature alignment techniques have also contributed indirectly to synthetic content detection. Wang *et al.* [14] proposed improved image registration algorithms for complex scenes, which are relevant for identifying spatial inconsistencies in manipulated media. Additionally, research in real-time face recognition and surveillance systems [15] has provided foundational advances in CNN efficiency and deployment strategies, supporting real-time deepfake detection implementations. Despite these advancements, many existing works focus primarily on a single modality—typically images—without addressing cross-modal threats such as AI-generated audio and machine-generated text. Moreover, several high-accuracy models rely on computationally intensive architectures that may not be suitable for real-time web deployment.

In contrast to prior studies, the proposed system emphasizes a unified multi-modal detection framework integrating lightweight convolutional neural networks (MobileNet) for image and video frame analysis, MFCC-based SVM classification for audio deepfake detection, and TF-IDF with Logistic Regression for AI-generated text detection. The framework prioritizes computational efficiency, modular scalability, and practical deployment within a Django-based web environment. Thus, while existing literature provides strong foundations in GAN detection, spectral analysis, multimodal fusion, and explainability, there remains a need for integrated, deployable, and computationally efficient multi-modal systems. The proposed work addresses this gap by combining deep learning and classical machine learning models into a unified architecture capable of detecting synthetic media across multiple formats in real-time scenarios.

III. METHODOLOGY

This work proposes a unified multi-modal deepfake detection framework designed to identify AI-generated content across images, videos, audio, and text. Unlike approaches that focus on a single modality, the proposed system integrates lightweight deep learning and classical machine learning models into a scalable Django-based web application. The overall methodology, illustrated in Fig. 1, consists of data acquisition, preprocessing, model training, evaluation, and deployment.

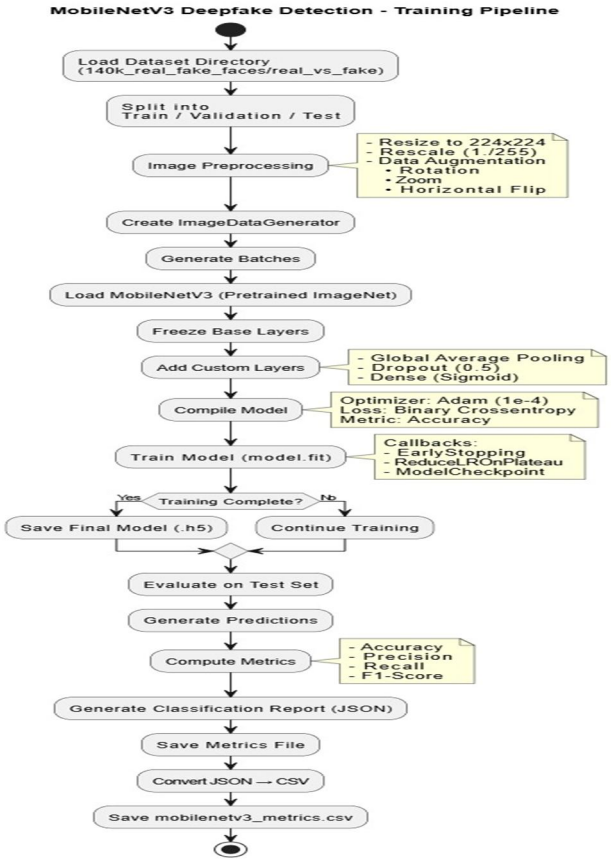


Fig. 1.1 : Methodology Flowchart

A. Data Acquisition and Preprocessing

For image-based detection, the system utilizes the 140K Real and Fake Faces dataset, containing approximately 140,000 images categorized into real and AI-generated classes. The dataset is organized into binary class directories ("real" and "fake"), enabling supervised learning.

All images are pre-processed before model training:

- Resizing to 224 × 224 resolution
- Pixel value normalization (rescale = 1/255)
- Data augmentation techniques:
    - Rotation
    - Horizontal flipping
    - Zoom transformation

These augmentation strategies increase dataset diversity and improve model generalization.

For video detection, frames are extracted using OpenCV. A fixed number of frames (e.g., 30–60 per video) are sampled uniformly. Each frame undergoes the same preprocessing steps as image inputs.

For audio detection, Mel-Frequency Cepstral Coefficients (MFCC) features are extracted using Librosa. Thirteen MFCC features are computed and standardized using feature scaling before classification.

For text detection, input text is cleaned and transformed using TF-IDF vectorization with unigrams and bigrams (maximum 10,000 features) prior to classification.

B. Model Implementation

1) Image and Video Detection (Mobile-Net-Based CNN)

A MobileNetV3 architecture pre-trained on ImageNet is employed as the backbone model for visual deepfake detection. Transfer learning is applied by freezing the base convolutional layers and adding custom classification layers:

- Global Average Pooling
- Dropout (0.5)
- Dense layer with sigmoid activation

Mobile Net is selected due to:

- Lightweight architecture
- Reduced computational cost
- Faster inference time
- Suitability for real-time deployment

For video detection, frame-wise classification is performed using the same Mobile Net model. Final video classification is obtained by aggregating frame-level predictions using average confidence scoring.

2) Audio Deepfake Detection

Audio deepfake detection is implemented using a Support Vector Machine (SVM) classifier trained on MFCC features. The extracted MFCC vectors are scaled using Standard Scaler before classification. This approach provides efficient and reliable detection with low computational overhead.

3) Text AI-Generated Content Detection

Text classification is performed using Logistic Regression combined with TF-IDF vectorization. This model distinguishes between human-written and AI-generated text by capturing statistical differences in lexical patterns and phrase structures.

C. Model Training

The Mobile-Net model is trained using the binary cross-entropy loss function and Adam optimizer with a learning rate of 1e-4. Training is conducted using mini-batch gradient descent with a batch size of 32 and image size of 224 × 224.

To prevent overfitting and improve training stability, the following callbacks are applied:

- Early Stopping
- ReduceLROnPlateau
- Model Checkpoint

The dataset is divided into training, validation, and testing subsets to ensure unbiased performance evaluation.

D. Model Evaluation

Model performance is evaluated using standard classification metrics:

- Accuracy
- Precision
- Recall
- F1-Score

Evaluation is performed on an independent test dataset. For video detection, evaluation is conducted using aggregated frame predictions. For audio and text detection, evaluation is based on classification accuracy and F1-score.

The trained Mobile-Net-based image detection model achieved an overall accuracy of 89.29%, demonstrating strong capability in distinguishing real and AI-generated images while maintaining computational efficiency.

E. System Integration and Deployment

The trained models are integrated into a Django-based web application. The backend handles:

- Media upload
- Preprocessing
- Model inference
- Result visualization

Live detection is implemented through continuous frame capture and sliding-window averaging for stable predictions. All trained models and evaluation metrics are stored in the media directory for deployment and monitoring.

IV. RESULTS AND DISCUSSION

In this study, the proposed multi-modal deepfake detection system was experimentally evaluated using the 140K Real and Fake Faces dataset for image classification, along with separate datasets for audio and text detection. The primary visual detection model was based on MobileNetV3, selected for its lightweight architecture and suitability for real-time deployment.

For image classification, approximately 100,000 images were used for training and 20,000 images for testing. All images were resized to 224 × 224 pixels and normalized using a rescaling factor of 1/255. A batch size of 32 was used during training to balance computational efficiency and learning stability. The model was trained using the Adam optimizer with a learning rate of 1e-4 and binary cross-entropy loss function.

To prevent overfitting and improve convergence, the training process incorporated Early Stopping, ReduceLROnPlateau, and Model Checkpoint callbacks. A dropout layer with a rate of 0.5 was included in the classification head to enhance regularization.

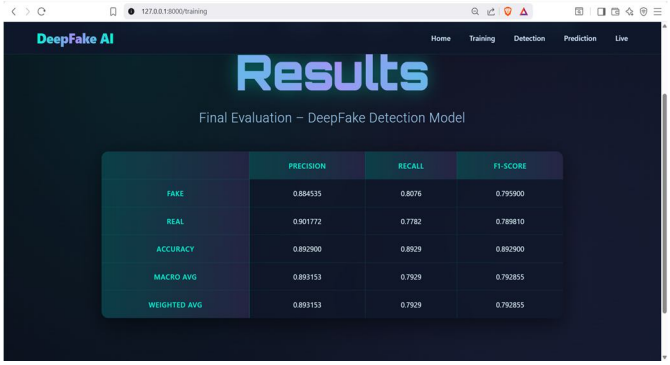


Fig. 1.2 : Model Performance Results

A. Image Detection Performance

The MobileNet-based model achieved an overall accuracy of 89.29% on the independent test dataset. The detailed performance metrics are summarized below:

The results demonstrate strong discriminative capability in distinguishing real from AI-generated images. The relatively balanced precision and recall values indicate that the model maintains stability across both classes, although minor misclassifications occur in challenging or borderline cases.

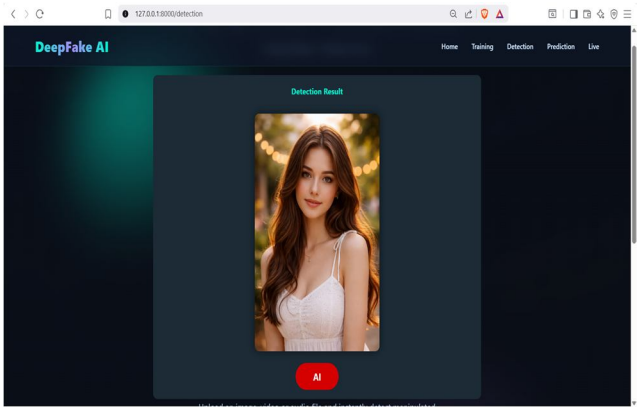


Fig. 1.3 : Image Detection

B. Video Detection Performance

Video deepfake detection was implemented using frame-wise MobileNet classification with confidence aggregation. By sampling 30–60 frames per video and computing average prediction scores, stable classification results were achieved. The frame aggregation strategy reduced noise from individual frame mispredictions and improved overall robustness.

The MobileNet-based approach significantly reduced computational complexity compared to heavy CNN-LSTM architectures, enabling near real-time inference suitable for web deployment.

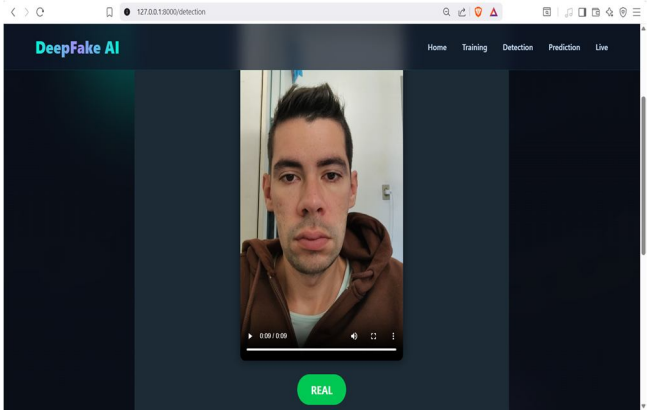


Fig. 1.4.1 : Video Detection

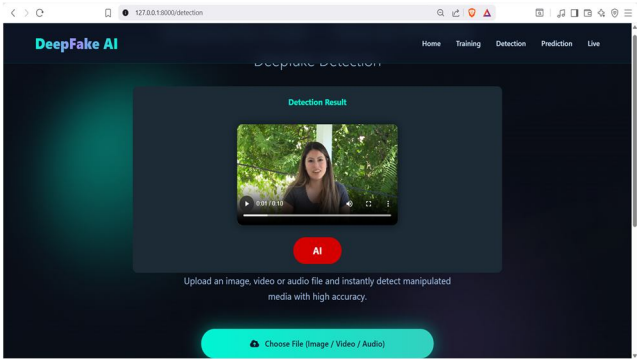


Fig. 1.4.2 : Image Detection

C. Audio Detection Performance

Audio deepfake detection was performed using MFCC feature extraction combined with an SVM classifier. The MFCC representation effectively captured spectral and temporal characteristics of speech signals. The SVM classifier demonstrated efficient separation between genuine and synthetic audio samples while maintaining low computational overhead.

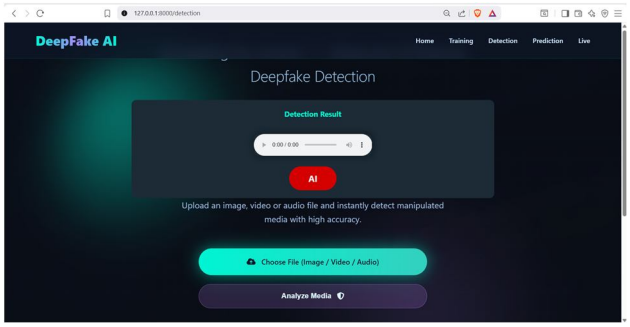


Fig. 1.5.1 : Audio Detection

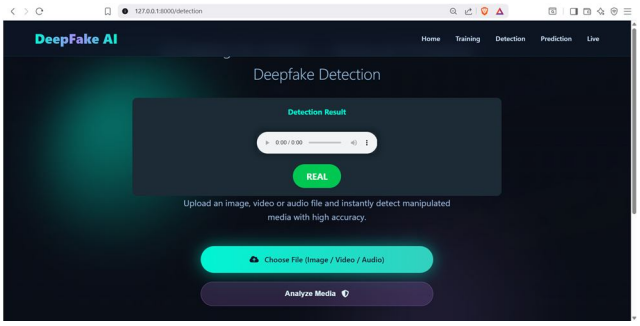


Fig. 1.5.2 : Audio Detection

D. Text Detection Performance

For AI-generated text detection, TF-IDF vectorization with Logistic Regression was used. The model successfully captured lexical and structural differences between human-written and machine-generated content. The lightweight architecture ensured fast inference, making it suitable for integration within the web-based system.

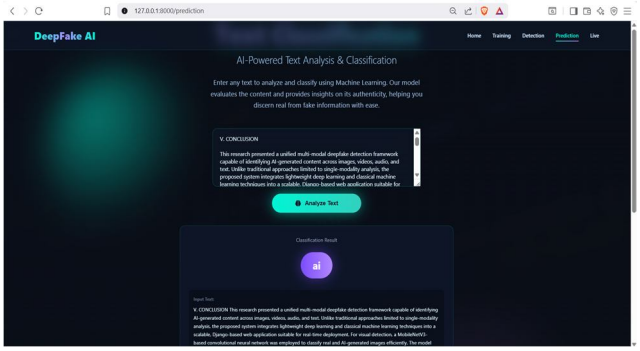


Fig. 1.6.1 : Text Detection

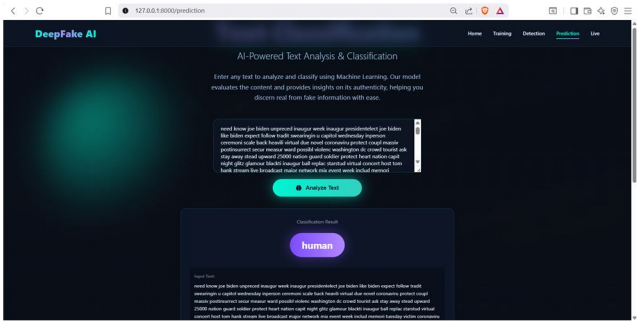


Fig. 1.6.2 : Text Detection

E. Discussion

The experimental results indicate that MobileNetV3 provides an effective balance between accuracy and computational efficiency. While deeper architectures such as VGG or ResNet or LSTM may achieve slightly higher accuracy in controlled environments, they require significantly higher computational resources. The proposed MobileNet-based framework prioritizes deploy ability and real-time performance, making it practical for web-based deepfake detection systems.

The use of callbacks such as Early Stopping and learning rate scheduling contributed to stable convergence and reduced overfitting. The dropout layer further enhanced generalization performance, particularly when handling large-scale datasets such as the 140K real/fake image collection.

Although the achieved accuracy of 89.29% demonstrates reliable classification performance, there remains scope for improvement in handling adversarial or highly refined synthetic content. Future enhancements may include transformer-based vision models, cross-modal consistency analysis, and ensemble learning approaches.

Overall, the results confirm that the proposed multi-modal detection framework is capable of effectively identifying AI-generated content across images, videos, audio, and text, while maintaining computational efficiency suitable for real-world deployment.

V. CONCLUSION

This research presented a unified multi-modal deepfake detection framework capable of identifying AI-generated content across images, videos, audio, and text. Unlike traditional approaches limited to single-modality analysis, the proposed system integrates lightweight deep learning and classical machine learning techniques into a scalable, Django-based web application suitable for real-time deployment.

For visual detection, a MobileNetV3-based convolutional neural network was employed to classify real and AI-generated images efficiently. The model achieved an overall accuracy of 89.29% on the 140K Real and Fake Faces dataset, demonstrating strong discriminative capability while maintaining computational efficiency. Video deepfake detection was implemented through frame-wise MobileNet classification with prediction aggregation, enabling stable and near real-time inference. Audio detection leveraged MFCC feature extraction with a Support Vector Machine classifier, while AI-generated text detection utilized TF-IDF vectorization with Logistic Regression to distinguish between human-written and machine-generated content.

The results confirm that lightweight architectures such as MobileNet can effectively balance accuracy and deployment efficiency, making them suitable for real-world digital media authentication systems. The integration of multiple modalities strengthens the robustness of the framework against diverse synthetic manipulation techniques. Furthermore, the web-based deployment demonstrates the practical feasibility of implementing deepfake detection systems for content moderation, cybersecurity, and digital forensics applications.

REFERENCES

[1] F. Alrowais, A. A. Hassan, W. S. Almukadi, M. H. Alanazi, R. Marzouk and A. Mahmud, "Boosting Deep Feature Fusion-Based Detection Model for Fake Faces Generated by Generative Adversarial Networks for Consumer Space Environment," IEEE Access, vol. 12, pp. 147680-147693, 2024, doi: 10.1109/ACCESS.2024.3470128.

[2] S. K. Hamed, M. J. A. Aziz and M. R. Yaakub, "Improving Data Fusion for Fake News Detection: A Hybrid Fusion Approach for Unimodal and Multimodal Data," IEEE Access, vol. 12, pp. 112412-112425, 2024, doi: 10.1109/ACCESS.2024.3443092.

[3] M. Luqman, M. Faheem, W. Y. Ramay, M. K. Saeed and M. B. Ahmad, "Utilizing Ensemble Learning for Detecting Multi-Modal Fake News," IEEE Access, vol. 12, pp. 15037-15049, 2024, doi: 10.1109/ACCESS.2024.3357661.

[4] Y. Li, K. Jia and Q. Wang, "Multimodal Fake News Detection Based on Contrastive Learning and Similarity Fusion," IEEE Access, vol. 12, pp. 155351-155364, 2024, doi: 10.1109/ACCESS.2024.3481311.

[5] Y. Liu, W. Bing, S. Ren and H. Ma, "BC-FND: An Approach Based on Hierarchical Bilinear Fusion and Multimodal Consistency for Fake News Detection," IEEE Access, vol. 12, pp. 62738-62749, 2024, doi: 10.1109/ACCESS.2024.3392409.

[6] A. Malanowska, W. Mazurczyk, T. K. Araghi, D. Megías and M. Kuribayashi, "Digital Watermarking—A Meta-Survey and Techniques for Fake News Detection," IEEE Access, vol. 12, pp. 36311-36345, 2024, doi: 10.1109/ACCESS.2024.3374201.

[7] J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," IEEE Access, vol. 12, pp. 15642-15650, 2024, doi: 10.1109/ACCESS.2024.3356122.

[8] S. A. Raza, U. Habib, M. Usman, A. A. Cheema and M. S. Khan, "MMGANGuard: A Robust Approach for Detecting Fake Images Generated by GANs Using Multi-Model Techniques," IEEE Access, vol. 12, pp. 104153-104164, 2024, doi: 10.1109/ACCESS.2024.3393842.

[9] Q. Bammey, "Synthbuster: Towards Detection of Diffusion Model Generated Images," IEEE Open Journal of Signal Processing, vol. 5, pp. 1-9, 2024, doi: 10.1109/OJSP.2023.3337714.

[10] J. Kang, S.-K. Ji, S. Lee, D. Jang and J.-U. Hou, "Detection Enhancement for Various Deepfake Types Based on Residual Noise and Manipulation Traces," IEEE Access, vol. 10, pp. 69031-69040, 2022, doi: 10.1109/ACCESS.2022.3185121.

[11] O. A. H. H. Al-Dulaimi and S. Kurnaz, "Deep Fake Image Detection Based on Deep Learning Using a Hybrid CNN-LSTM with Machine Learning Architectures as Classifier," in 2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Istanbul, Turkiye, 2024, pp. 1-7, doi: 10.1109/HORA61326.2024.10550728.

[12] D. Bhargava, S. Rani, M. Singh, N. Tripathi, A. Bhargava and G. Panwar, "Deep-Fake Finder: Uncovering Forgery Image Through Neural Network Analysis," in 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), Gautam Buddha Nagar, India, 2024, pp. 1-5, doi: 10.1109/IC3SE62002.2024.10592889.

[13] G. Cartella, V. Cuculo, M. Cornia and R. Cucchiara, "Unveiling the Truth: Exploring Human Gaze Patterns in Fake Images," IEEE Signal Processing Letters, vol. 31, pp. 820-824, 2024, doi: 10.1109/LSP.2024.3375288.

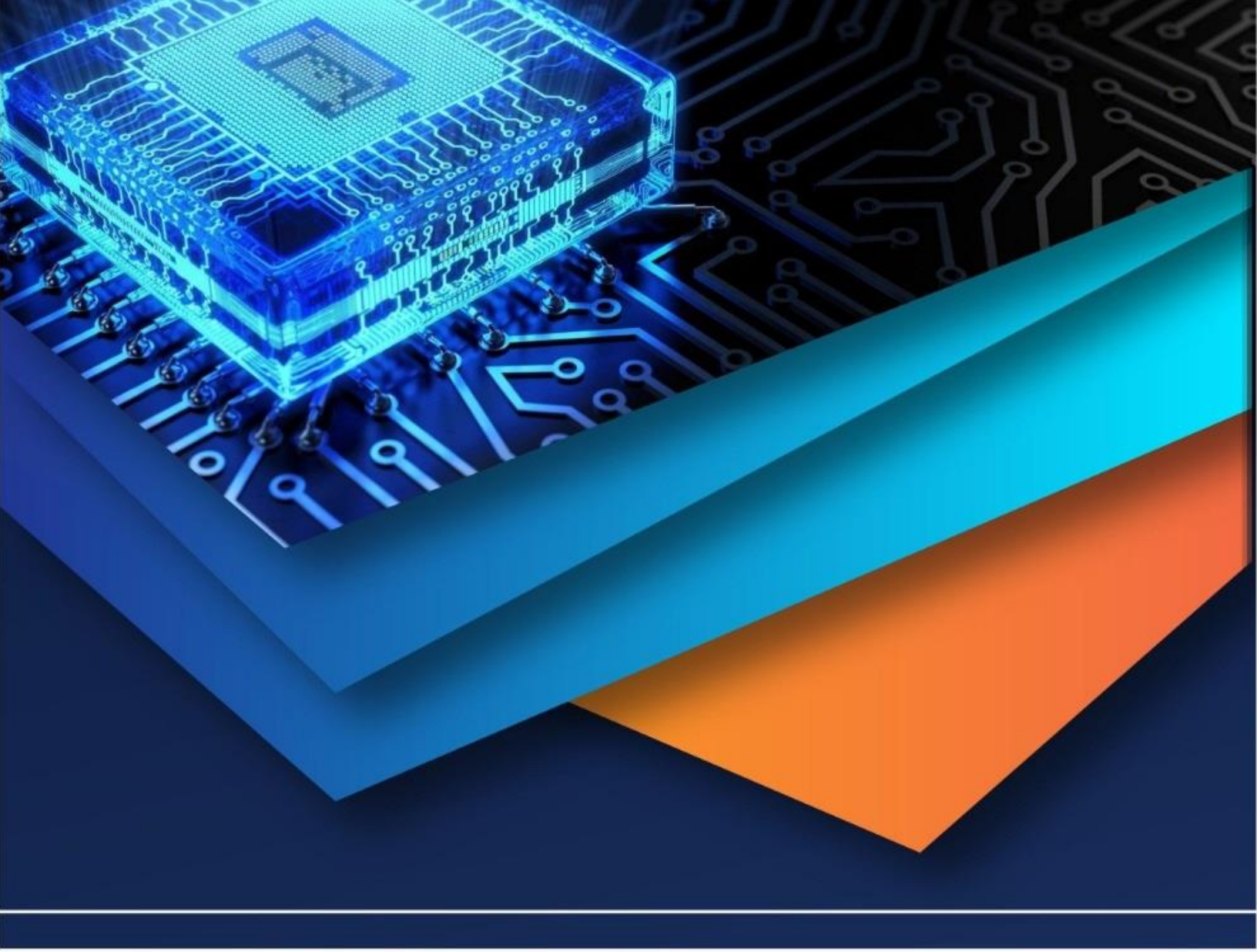
[14] Y. Wang, X. Liang and L. Chen, "Research on Infrared and Visible Image Registration Algorithm for Complex Road Scenes," IEEE Access, vol. 11, pp. 78511-78521, 2023, doi: 10.1109/ACCESS.2023.3299266.

[15] Kanagamalliga, S., Abishek, R., Krishna, B. B. S., & Vinayagam, P., "Advancements in Real-Time Face Recognition Algorithms for Enhanced Smart Video Surveillance," Procedia Computer Science, vol. 230, pp. 486-492, 2023.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)