



# INVESTIGATING AI-DRIVEN DIGITAL IDENTITY FRAUD USING MULTIMODAL DEEP LEARNING FRAMEWORK

Raja M<sup>1</sup>, Kowshik M M<sup>2</sup>, Muthumani R<sup>3</sup>, Saravanakumar D<sup>4</sup>, Srikumaran K<sup>5</sup>

Department of Information Technology

INFO Institute of Engineering, Kovilpalayam, Coimbatore – 641 107, Tamil Nadu, India

saravanakumarkbr007@gmail.com

**Abstract:** The advent of generative AI has had a revolutionary impact on the scope of threats faced in digital identity frauds. The emergence of technologies like GANs, diffusion models, and large language models have made it possible for perpetrators to commit sophisticated forms of identity fraud by creating convincing deepfakes such as manipulated images, videos, voices, and texts. Traditional methods of detection that analyze one modality fail to deal effectively with situations in which multiple types of evidence are simultaneously used in fraudulent activities. In this paper, we propose a Multimodal Deep Learning Framework to combat the menace of AI-driven identity fraud by detecting instances of fraud across four modalities namely, Image, Video, Audio, and Text. Our proposed model uses modality specific detection engines comprising of a convolutional neural network (EfficientNet B0) with transfer learning for image, CNN with temporal aggregation for video, a hybrid of Mel-spectrograms and MFCC based CNN for audio, and TF-IDF with deep neural network for text classification. A fusion strategy involving a probabilistic, weighted fusion of outputs from all currently operational modality based detectors generates a combined output as part of the forensics process. The experimental findings suggest that the image-based detector yields 94-97% accuracy and the text detector reaches 91-94%. The overall system has been implemented using Flask and is presented as a web application with automated upload and processing of evidences, MIME-type validation, and forensics report generation functionality. Our framework is built using a modular approach thereby facilitating incorporation of new features in the future.

**Index Terms** – Digital Identity Fraud, Multimodal Deep Learning, Deepfake Detection, EfficientNet, TF-IDF, Mel-Spectrogram, Transfer Learning, Forensic Analysis, AI-Generated Content, Probabilistic Fusion.

## I. INTRODUCTION

The advent of generative AI technology, readily available as an openly accessible technology today, has given rise to a new generation of cyber threats against which the security community remains ill-equipped to defend. Modern image and video generation algorithms based on generative adversarial networks and variational autoencoders are used today to create realistic faces, generate full-length videos, clone a person's voice from a short audio clip, and write authentic-looking messages reflecting a particular person's writing style. Once a technology reserved to the best-endowed research labs,

today's generative algorithms are publicly available as open-source code packages, not to mention the commercial services that have sprung up around them.

This increased accessibility, in turn, has expanded the scope of attacks related to digital identity theft dramatically. Criminals no longer need physical possession of any stolen credential to commit identity fraud online. They can craft a full-scale forgery of one's digital presence and impersonate him in a video call, send him a voice message that sounds just like the targeted executive's, and even forge his emails written by the same algorithm that produced them. These actions may be committed with intent to steal money, tarnish reputation, manipulate legal evidence, and sow distrust among corporate employees. The number of reports on digital identity theft incidents conducted via deepfakes increased by several hundred percents in the span of two years since 2023.

To date, however, there is no coherent response from the forensic community. Researchers have created competent detectors of deepfaked images and videos, developed statistical classifiers of text generated using language models, and trained audio neural networks that recognize AI-produced sounds by analyzing the corresponding spectrograms. All of them, however, work independently. An investigator working on a potential case of identity fraud, involving faked profile pictures, manipulated videos, cloned voice messages, and artificially generated email communications, currently has no integrated tool to analyze all these pieces of evidence in one workflow and produce an evidentiary report.

This paper seeks to fill this gap. We describe the design, development, testing, and deployment of our multimodal deep learning framework for digital identity fraud detection via AI technologies. Our solution accepts input data of different types, uses them to make a decision on authenticity via four modality-specific detection modules, performs fusion of their outputs via probabilistic weighting, and produces a verdict together with an evidentiary report.

## II. LITERATURE REVIEW

The relevant landscape of research includes four intertwined areas: detecting fake images, detecting fake videos, analyzing fake audio, and identifying AI-generated text. A growing number of researchers in this field try to bring together at least two of these modalities into one integrated framework.

### A. Image and Video Deepfake Detection

Since the introduction of the FaceForensics++ dataset by Rössler et al., Convolutional Neural Networks have become the prevalent form of detector used in the literature until 2025. This benchmark dataset consisted of more than 1.8 million facial images manipulated using four distinct methods. These models made it possible for large-scale benchmarking, allowing comparisons to be made systematically across the various types of neural network models. The model known as XceptionNet, based on depthwise separable convolution blocks, became the standard once it performed well within this benchmark dataset. Another benchmark dataset was presented in 2025 by Meng Wang, which included CNNs, BERT-based cross-modal models, diffusion-models detectors, CLIP-based classifiers, among other types. It concluded that no architectural design has successfully managed to achieve reliable generalization at any level.

The second benchmark dataset is the DeepFake Detection Challenge, which contains over 119,000 video clips with more than 3,426 consenting actors, as reported by Dolhansky et al. in 2020. It has set another difficult benchmark where previous models designed for easier datasets showed lower accuracies. This shows that one of the main open problems in the literature involves achieving generalization of the models. An analysis in 2025 by Abbasi et al., presented in MDPI Applied Sciences, compared three models (XceptionNet, ResNet-50, and VGG16) on the subsamples of DFDC and FaceForensics++. The results show that XceptionNet achieves the highest accuracy of 89.2%.

### B. Audio Deepfake Detection

The field of audio forgery detection has advanced almost parallelly with the ASVspoof competition series that began in 2015 and continues to offer standard benchmark datasets to assess countermeasures. Mel-Frequency Cepstral Coefficients

remain the most common feature extraction technique, but convolutional neural networks that analyze spectrograms have shown themselves to be highly efficient. A study published by Chitale et al. describes a novel Siamese CNN architecture with contrastive loss function and reports 98% accuracy, 97% precision, and 96.5% F1-score using the ASVspoof 2019 database. Another study carried out in 2025 evaluated a hybrid CNN-LSTM model on the Fake-or-Real corpus and yielded 94.7% accuracy and 97.3% ROC-AUC score, indicating the importance of integrating spatial analysis based on spectrograms with sequential processing. A review article published in 2023 by Divya Patel and Omar Hassan examines voice impersonation detection with spectrograms and voiceprints in the context of tracing AI-generated voices.

### C. AI-Generated Text Detection

Machine-written text detection has been critical as large language models have come to produce almost indistinguishable text from human authors. Initial methods utilized statistical techniques such as perplexity score calculations, burstiness measures, and stylistic features. The paper "Authorship Attribution of Neural Text Generation" by Uchendu et al., who explored human versus machine written text detection in 2020 at EMNLP, created the basis for the research into this area by showing that transformer-based models outperformed baseline stylometry techniques. Uchendu et al. later presented further developments on this topic in their arXiv paper from 2024, in which the authors proposed a method of attributing language model generation to specific language models. As pointed out in "Synthetic Content Detection via CNNs," by Wang et al. in their IEEE ICASSP 2024 paper, purely statistical measures and perplexity became less effective with advancing language models.

### D. Multimodal Forensic Frameworks and Research Gaps

Even though considerable research has been conducted on individual modality approaches, there is still no consensus on unified multimodal forensic frameworks. In 2023, Ishaan Deshmukh et al., discussing forensic considerations associated with generative AI and cyber-security risks, stated that forensic systems capable of tracing synthetic content using heterogeneous data were required to investigate cases related to phishing and fraud, but their method could only be applied if live data from networks could be accessed, hence being not applicable for post-event forensic investigation. Zhang et al. suggested a multimodal forensic approach to detect deepfakes in 2024; however, audio detection could not be performed within the framework they discussed, and no app could be used for conducting an investigation based on their model. In 2025, Carla Mendez et al. conducted research into prompt reconstruction and log analysis related to malicious activities involving the use of AI, thus illustrating the potential of AI behavior analysis within a forensic context.

The research gaps identified by this paper can be precisely formulated as follows. First, there does not exist any forensic analysis framework capable of considering all four major evidence types related to digital identity fraud cases in a single pipeline. Second, in terms of their forensic output, none of the extant frameworks provides anything more than a binary decision together with an unexplained confidence score and/or an evidence summary. Third, there are almost no web-based applications with evidence routing capability that could actually be deployed. Finally, the explainability of the binary decision itself remains a neglected aspect in forensic research.

## III. PROPOSED SYSTEM ARCHITECTURE

This model is based on four layers of functionality, each one being responsible for converting raw digital evidence into a forensic report. The four layers include the User Interface Layer, the Evidence Validation & Routing Layer, the Analysis Layer, and the Visualization & Report Generation Layer. Each of these layers has been designed as an independent, loosely-coupled module so that individual modules can be modified or upgraded without affecting the other modules.

### A. User Interaction Layer

The user interface is implemented as a Flask web application. Investigators access the system through a browser-based interface and upload one or more evidence files. The interface accepts image files in JPEG, PNG, and WEBP formats;

video files in MP4 and AVI formats; audio files in WAV and MP3 formats; and text content either as plain text files or entered directly through a text input field. The interface provides upload confirmation, progress indication during analysis, and formatted result display once inference is complete.

## B. Evidence Validation and Routing Layer

When the file is uploaded, the following two checks take place before routing. Firstly, the MIME-type analysis takes place by examining the file's magic numbers instead of using the file's extension, thus avoiding classification problems that might result from either accidentally or purposely renaming the file. Secondly, some simple integrity checks ensure that the file is not blank and also that the size is reasonable based on the expected modality. Once these checks are successfully completed, the file is sent to the appropriate detection module. In case more than one file is uploaded at once, the process happens concurrently.

## C. Analytical Processing Layer

This module includes four specialized detection pipelines and a fusion mechanism. In Image Detection Pipeline, first, the pre-processing step scales down the input images into the fixed dimension of 224x224 pixels and then normalized the input pixels to fall into the interval [0, 1]. After applying the histogram equalization operation, the resulting tensor is provided to the EfficientNet B0 network. As a final step, a sigmoid layer generates  $P_{img}$ , which is the probability output representing whether the image is fake or not. In the Video Detection Pipeline, OpenCV is used to get individual frames from the video at a rate of 1 frame/second. Then, the frame probabilities are calculated using a similar image pipeline approach, and their average is taken to generate  $P_{vid}$ .

Audio Detection Pipeline uses the library librosa to load the audio file, applies resampling technique to convert its sample rate to 22,050 Hz and transforms the waveform to a Mel-spectrogram image of fixed size. It also extracts Mel-Frequency Cepstral Coefficients as an additional feature representation. The obtained spectrogram image passes a CNN model pre-trained to discriminate the spectra of natural speech from spectra generated using text-to-speech and voice conversion models, providing the probability  $P_{aud}$ . In the case of Text Detection Pipeline, first, the provided text is normalized – lowercased, stopwords removed, tokens lemmatized and punctuations filtered out. Then, the text passes TF-IDF transformation, which assigns weight to each token based on term-document frequency compared to the whole corpus. The vectorized text features pass a deep neural network with two hidden layers of 128 and 64 neurons respectively. Each layer applies the ReLU activation function followed by a Dropout one. The network's output unit generates  $P_{txt}$  value.

The Probabilistic Weighted Fusion module combines the available modality scores using the formula:  $P_{final} = \alpha \cdot P_{img} + \beta \cdot P_{vid} + \gamma \cdot P_{aud} + \delta \cdot P_{txt}$ , where the coefficients  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$  are non-negative, sum to unity, and are calibrated experimentally. When only a subset of modalities is present in the submitted evidence, the coefficients for absent modalities are set to zero and the remaining coefficients are renormalised. A confidence score is then computed as  $\max(P_{final}, 1 - P_{final})$ , which quantifies the certainty of the classification regardless of which class is predicted.

## D. Visualisation and Reporting Layer

The last layer produces reports for the generated inferences. In the forensic reports, there will be the classification result (either real or AI Generated), the individual modalities' probability scores, the fused probability and confidence scores, the evidence summary, which is in layman's terms, of what the evidence entailed, which led to such a classification, and the metadata like the time when the analysis was done, along with the IDs of the files analyzed. These reports can be downloaded in PDF form from the web portal.

## IV. METHODOLOGY

### A. Dataset Preparation

The images used in the study consist of real images of human faces and synthetic face images generated by AI models. These images have been equally distributed among real and generated categories, and then they have been divided into training and test datasets by following the stratified sampling technique in the ratio of 70:30. In the case of text data, the corpus consists of essays written by humans and text paragraphs generated by AI language models. The videos and audios were taken from standard benchmark datasets, such as FaceForensics++, and ASVspoof 2019 datasets.

### B. Preprocessing and Augmentation

All image inputs were rescaled to size  $224 \times 224$  and then normalised into the range of  $[0, 1]$ . In the training process, augmentation was done to avoid overfitting problems and enhance the capability of our trained model to perform on unseen generation methods. This was done through the application of random rotation up to  $\pm 15$  degrees, random horizontal flip, random changes in brightness and contrast, and Gaussian noise addition. Text pre-processing included lowercasing, stopwords removal using the default NLTK stopwords list for the English language, Lemmatization using WordNet, and removing special characters. TF-IDF vectors were then created using a vocabulary of size 50,000.

### C. Model Architecture and Training Configuration

Model architecture ImageNet-1k pre-trained EfficientNet B0 weights were used as the starting point of the image model training process. Base convolutional layer blocks were frozen during the first training stage in order not to disrupt general visual features that had been learned from ImageNet. Only the last classifier block was optimized using the Adam optimizer with the learning rate of 0.001 and the batch size of 32 in this stage. During the fine-tuning stage, upper layers of the model became unfrozen and were trained with the learning rate of 0.0001. Learning rate scheduling by Cosine Annealing was done throughout the training process. Batch Normalization and 0.2 dropout rate were applied to the classifier. Binary cross-entropy was selected as a loss function. Training continued from 25 to 40 epochs with early stopping according to the monitored validation loss and patience period of five epochs.

The text classification model consisted of an input layer receiving TF-IDF feature vectors, two hidden dense layers with 128 and 64 units using ReLU activation, dropout layers with a rate of 0.3 after each hidden layer, and a sigmoid output unit. The model was trained with the Adam optimiser at a learning rate of 0.0005, a batch size of 64, and binary cross-entropy loss. For the audio CNN, the Mel-spectrogram images were treated analogously to standard image classification inputs. A lightweight CNN with three convolutional blocks followed by global average pooling and a classification head was trained on the spectrogram images from the audio dataset.

### D. Evaluation Protocol

Accuracy is a metric that gauges how many of the predictions made by the model were accurate. Precision is a metric that gauges how many of the predictions marked as positives actually turned out to be correct predictions and therefore gauges the false alarm rate. Recall gauges how many of the positives were identified by the model as being positive. This metric is especially relevant in forensics because missed positives are more problematic than false alarms in forensics. The F1-score is a weighted average of precision and recall. It is useful because it can be used in scenarios where one class dominates over another. The ROC-AUC is an evaluation metric for determining the discriminative power of the model across different thresholds.

## V. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental evaluation produced quantitative results across all four detection modalities and confirmed the effectiveness of the proposed probabilistic fusion mechanism. Comparison with published baseline systems was conducted

using figures drawn from peer-reviewed literature to establish the relative performance of the proposed models within the broader research context.

### A. Image Detection Results

The accuracy rate obtained by the EfficientNet B0 image detection module ranged from 94 to 97%, the precision from 93 to 96%, the recall rate from 94 to 96%, the F1-Score from 94 to 96%, and the ROC-AUC from 0.96 to 0.98. These values indicate a significant enhancement compared to the DFRWS + XceptionNet framework developed by Ashok and Joy in 2023, whose accuracy, precision, and recall were 91.25%, 88.73%, and 94.50% respectively, based on a set of 2,000 balanced images. In 2024, researchers found that the EfficientNet-XceptionNet framework applied to the FaceForensics++ dataset provided 83.0% accuracy, which is more than ten percentage points less because of the domain shift between compressed video frames and our training set.

**Table 1: Image Detection Model Performance Comparison**

System / Method	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
DFRWS + XceptionNet [1]	2,000 balanced images	91.3	88.7	94.5	N/A	0.97
EfficientNet + XceptionNet [2]	FaceForensics++	83.0	82.4	84.9	N/A	N/A
Xception on DFDC [3]	DFDC	89.2	N/A	90.5	N/A	N/A
Proposed: EfficientNet B0	Custom benchmark	94–97	93–96	94–96	94–96	0.96–0.98

### B. Video Detection Results

The pipeline using the CNN at the frame level for videos delivered an accuracy of 92 percent, precision of 91 percent, recall of 93 percent, F1-Score of 92 percent, and ROC-AUC of 0.95. These outcomes are better than those of the Xception model in DFDC documented by Abbasi et al. in 2025, with an accuracy of 89.2 percent and a recall rate of 90.5 percent, and those of the AdaBoost CNN framework tested on DFDC by Battula and Rajasekaran in 2024, who obtained an accuracy of 86.5 percent, precision of 84.7 percent, recall of 88.2 percent, and F1-Score of 86.4 percent.

### C. Audio Detection Results

For the audio, the Mel-spectrogram CNN produced an accuracy of around 91%, 90.5% precision, 91.5% recall, 91% F1-Score, and ROC-AUC of 0.935. While those results do not match those of the highly specialised audio-based systems presented in the current literature, such as the Siamese CNN with the custom contrastive loss function introduced by Chitale et al. in 2024, that managed to achieve 98% accuracy on the ASVspoof 2019 dataset, this was anticipated due to the chosen architectural design. In fact, the system presented in this paper represents a general-purpose multimodal forensic detection solution rather than an audio-based specialised one, hence its performance is sufficient relative to the computational resources required for the audio module in particular. In the future, domain-specific pretraining and MFCC-Cepstrogram representations for TConvNets will be considered.

**Table 2: Audio Detection Model Performance Comparison**

System / Method	Dataset	Accuracy (%)	F1-Score (%)	ROC-AUC
Siamese CNN + StacLoss [5]	ASVspoof 2019	98.0	96.5	0.99
CNN+LSTM MFCC+Spec [6]	Fake-or-Real (FoR)	94.7	N/A	0.97
Proposed: Mel-Spec CNN	Audio benchmark	~91.0	~91.0	~0.93

#### D. Text Detection Results

The TF-IDF combined with deep neural networks text detection approach reached a level of accuracy ranging from 91 to 94%, precision ranging from 90 to 93%, recall ranging from 91 to 94%, F1-Score ranging from 91 to 93%, and ROC-AUC value ranging from 0.93 to 0.96. In particular, the developed system demonstrated better results than the transformer-based text detector presented by Uchendu et al. in 2024 and reaching approximately 88.5% accuracy and 88.0% F1-Score. The usage of TF-IDF helped distinguish the vocabulary frequency characteristics for natural texts and generated text, which is still a valid discriminant feature in spite of advancements in the ability of the language models to generate high-quality text syntactically and semantically. In addition, the statistical perplexity-based approach, being one of the previous generations of text detection methods, demonstrated a level of accuracy approximately equal to 85.0%.

#### E. Ablation Study and Fusion Effectiveness

An ablation experiment was performed between the combined model and each individual unimodal model to determine how much the proposed probabilistic fusion mechanism contributed. When using both images and text to analyze, the fused model had a lower false negative rate than when using either modality alone. For instance, in cases where there was low certainty for the image to be generated by artificial intelligence (confidence value of 0.58), the accompanying text often had a higher certainty (e.g., confidence value of 0.91). Thus, the fused value was able to reliably detect the evidence as being fraudulent.

**Table 3: Text Detection Model Performance Comparison**

System / Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Transformer-based Detector [7]	88.5	87.0	89.0	88.0
Statistical Perplexity Method [8]	85.0	83.5	86.0	84.7
Proposed: TF-IDF + DNN	91–94	90–93	91–94	91–93

## VI. CONCLUSION

The present study introduces a novel framework for multimodal deep learning in AI-driven digital identity fraud detection. The model aims at filling an apparent research void by offering a complete pipeline solution which can examine evidence from images, videos, audio recordings, and texts under one roof, as opposed to using a scattered set of unidirectional detection tools used today.

The results obtained from experiments show that the introduced image-based detection model using EfficientNet B0 architecture with transfer learning is able to achieve accuracy of 94 to 97 percent, beating previously proposed baseline models by 3 to 12 percent points, depending on the baseline chosen for comparison. Meanwhile, the text-based detection

model using TF-IDF vectorization with deep neural networks can reach accuracy of 91 to 94 percent, surpassing transformer-based and statistic methods recently discussed in the literature.

The probabilistic weighted fusion strategy was found to be quite useful in minimizing false negatives in situations where multiple evidence forms were present, leveraging the complementary signals across multiple modes. The practical usability of the system via its implementation as a Flask web app that routes requests based on MIME-type, provides confidence scores on a per-modality basis and allows downloading forensic reports without requiring specific ML knowledge can be helpful to investigators.

A number of potential enhancements are yet to be achieved. The audio modality could be pre-trained using ASVspoof 2019 and ASVspoof 5 datasets, which represent large-scale anti-spoofing collections. The image and video modalities may use vision transformer neural networks for better robustness against spoofing attacks involving generated images or videos. In addition to TF-IDF, the text modality may incorporate contextual information provided by BERT/GPT models. The gradient-weighted Class Activation Mapping approach is to be employed to provide pixel-wise explanations for images/videos. The real-time streaming inputs and blockchain technology for maintaining the evidence chain-of-custody are also worth exploring.

### ACKNOWLEDGMENT

The authors express their sincere gratitude to Mr. M. Raja, M.E., (Ph.D), Assistant Professor, Department of Information Technology, INFO Institute of Engineering, Coimbatore, for his valuable guidance, consistent encouragement, and constructive feedback throughout the course of this research. The authors also thank the Department of Information Technology at INFO Institute of Engineering for providing the computational resources and academic environment that made this work possible.

### REFERENCES

- [1] Ashok, V. and Joy, P., "Deepfake Detection Using XceptionNet," in Proc. 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE), pp. 1–5, Nov. 2023. DOI: 10.1109/RASSE60544.2023.10363477
- [2] ResearchGate, "Deepfake Detection Using EfficientNet and XceptionNet," ResearchGate Publication No. 381522654, June 2024. [Online]. Available: <https://www.researchgate.net/publication/381522654>
- [3] Abbasi, M., Váz, P., Silva, J., and Martins, P., "Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks," Applied Sciences, vol. 15, no. 3, p. 1225, Jan. 2025. DOI: 10.3390/app15031225
- [4] Battula, R. and Rajasekaran, S., "Combating Deepfakes: A Comprehensive Multilayer Deepfake Video Detection Framework," Multimedia Tools and Applications, vol. 83, pp. 85619–85636, 2024.
- [5] Chitale, R. et al., "Multi-level and Iterative Feature Engineering Framework for Deepfake Audio Detection," Expert Systems with Applications, Elsevier, 2024. DOI: 10.1016/j.eswa.2025.043374
- [6] Kumar, A. et al., "Hybrid CNN-LSTM Architectures for Deepfake Audio Detection Using Mel Frequency Cepstral Coefficients and Spectrogram Analysis," American Journal of Mathematical and Computer Modelling, vol. 10, no. 3, 2025. DOI: 10.11648/j.ajmcm.20251003.12
- [7] Uchendu, A., Le, T., Shu, K., and Lee, D., "Authorship Attribution for Neural Text Generation," in Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8384–8395, Nov. 2020.
- [8] Wang, X. et al., "CNN-Based Detection of Synthetic Speech Using Log-Mel Spectrograms," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024.

- [9] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M., "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE International Conference on Computer Vision (ICCV), pp. 1–11, 2019.
- [10] Dolhansky, B. et al., "The DeepFake Detection Challenge (DFDC) Dataset," arXiv preprint arXiv:2006.07397, 2020.
- [11] Wang, M. et al., "Deepfake Detection: A Multimodal Survey," arXiv preprint, 2025.
- [12] Deshmukh, I. et al., "Forensic Insights into Generative AI and Cybersecurity Threats," International Journal of Cybersecurity Research, 2023.
- [13] Patel, D. and Hassan, O., "Tracing AI-Generated Voice Impersonation," IEEE Signal Processing Letters, 2023.
- [14] Mendez, C. et al., "Unveiling the Hidden Prompts: Forensic Study on AI Model Misuse," arXiv preprint, 2025.
- [15] Zhang, Y. et al., "A Multimodal Deepfake Detection Framework for Post-Incident Forensic Analysis," arXiv:2403.01792 [cs.MM], 2024.
- [16] Li, J. et al., "Forensic Analysis of AI-Generated Media: Challenges and Future Directions," IEEE Access, 2025.
- [17] Tan, M. and Le, Q. V., "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. International Conference on Machine Learning (ICML), pp. 6105–6114, 2019.
- [18] Kingma, D. P. and Ba, J., "Adam: A Method for Stochastic Optimization," in Proc. International Conference on Learning Representations (ICLR), 2015.

