FAKE AUDIO DETECTION AND CLASSIFICATION PPLATFORM

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING LABORATORY

Vaani

Authenticity in Every Waveform

Presented By

Nittya Kanse (UCE2023545)
Purva Battawar (UEC2023206)
Rishika Bora (UIN2023747)

MOTIVATION

Fraud Escalation

According to a McAfee report, 83% of Indian victims of AI voice scams reported monetary loss.

Trust Preservation

40% of respondents believe their own voice might have been cloned, according to a survey

Regulatory Pressure

Increasing legal and policy demand for AI-generated content labeling and verification.

Security Vulnerability

Voice-based systems and 2FA are now easily spoofed, increasing risk.

PROBLEM STATEMENT

Deepfake audio poses a serious threat to digital communication, enabling voice fraud and misinformation. Existing detection methods are either slow, unreliable, or lack transparency.To address this, a system is needed that can accurately detect fake audio in real time and ensure the authenticity of results through secure verification.

OBJECTIVES

To design an AI model using MFCC feature extraction and a Random Forest Classifier for efficient real vs. fake voice detection.

To integrate machine learning predictions with blockchain for secure and verifiable result storage.

To provide a Flask-based user interface for real-time detection and QR-based authenticity certificate generation.

To ensure model explainability through feature visualization and blockchain traceability.

Vaani

Authenticity in Every Waveform

OUR SOLUTION

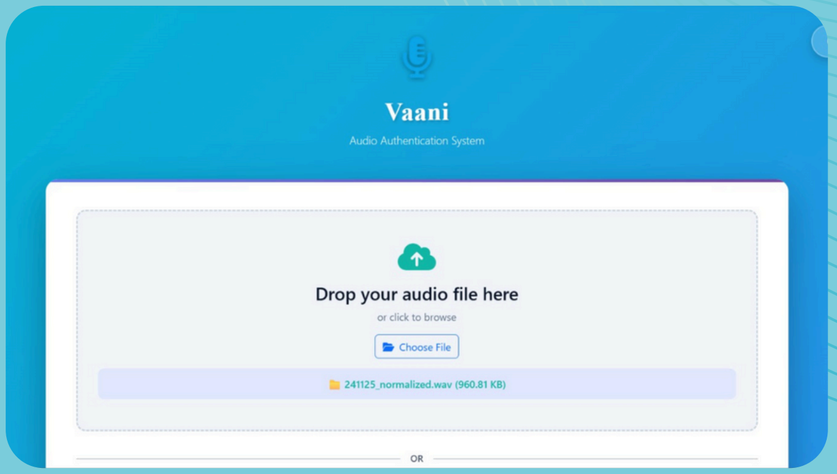
Vaani, a lightweight audio-forensics system detecting AI-generated speech using MFCC features and a Random Forest model, storing results immutably on blockchain while generating verification certificates and optional transcript-based categorization of fake audio.

Platform Features:

1. **Real-time detection of AI-generated vs. real audio** using MFCC and spectral features.
2. **Lightweight, privacy-preserving processing** with no audio storage.
3. **Immutable blockchain logging** for secure forensic audit trails.
4. **Automatic authenticity certificate generation** with QR-based verification.
5. **Transcript generation** to analyze linguistic patterns in fake audio.
6. **Categorization of audio** (Legitimate, Neutral, Slightly Suspicious, Suspicious, Highly Suspicious, Scam, Potential Scam)
7. **Fast, deployable system** with clear detection history and high reliability.

Vaani

Authenticity in Every Waveform



How we are different from the existing solution?

- Ewe classify intent via transcripts.
- Heavy deep-learning models dominate; our system is lightweight and real-time.
- Most platforms store audio; we never store inputs (privacy-preserving).
- Others lack secure provenance; we use blockchain logs + QR-based verification.
- Competing solutions give raw scores; we give clear, interpretable results with verifiable certificates.

EXISTING AI/ML METHODS

Method / Model Type	Method / Model Type	Method / Model Type	Method / Model Type
CNN-based Deepfake Audio Detector	2D/1D CNN on Mel-Spectrograms	Needs GPU, slow, fails on noisy or short clips	Fast MFCC-based **Random Forest** works in real time (<0.01s), CPU-friendly
RNN/LSTM Voice Spoofing Detector	LSTM/GRU Sequence Models	Overfits easily, weak against unseen deepfake styles	Robust **feature-driven ML model**, generalizes better to unknown TTS
Classical MFCC + SVM/RF Systems	SVM / Logistic Regression	Only real vs fake classification, limited functionality	Adds **transcription + scam classification + blockchain logging**
Cloud API Deepfake Services	Black-box DL Models (Google, Microsoft)	Costly, internet required, privacy concerns	Fully **local**, offline, transparent—no data leaves user's device
Keyword-Based Scam Classifiers	Rule-based NLP / Keyword Matching	Fails when scammers paraphrase or use soft language	Uses **BETTER30-trained ML text classifier** for context-aware behavioral tagging

METHODOLOGY

The system follows a lightweight, real-time deepfake audio detection pipeline composed of five key stages:

01

Dataset Preparation

- DEEP-VOICE dataset with real + AI voice-conversion samples
- 150 training and 13 testing clips
- All audio standardized to 16-bit WAV, 22,050 Hz, 10-second length

02

Feature Extraction

- 15-dimensional vector using Librosa:
  - 12 MFCCs (mean-pooled)
  - Spectral Centroid, Spectral Rolloff (85%), Zero-Crossing Rate

03

Model Training

- Random Forest (100 estimators)
- Stratified 80–20 split
- Scale-invariant, robust to small datasets
- Outputs Real / Fake classification

04

Real-Time Detection Pipeline

- Validate WAV → preprocess → extract features → load model → classify
- Generates a signed certificate containing prediction, timestamp, and feature hash
- Entire process <0.01 sec with no audio storage

03

Blockchain-Based Forensics

- Custom linked-list blockchain
- Each block stores: prediction, SHA-256 feature hash, timestamp
- Ensures immutability and verifiable audit trails

04

Transcript-Based Scam Analysis

- Converts speech → text using lightweight ASR
- TF-IDF + Logistic Regression for scam-pattern detection
- Identifies urgency cues, impersonation, payment pressure
- Provides a dual-layer forensic output: real/fake audio + scam/normal transcript

IMPACT AND BENEFITS OF THE SOLUTION

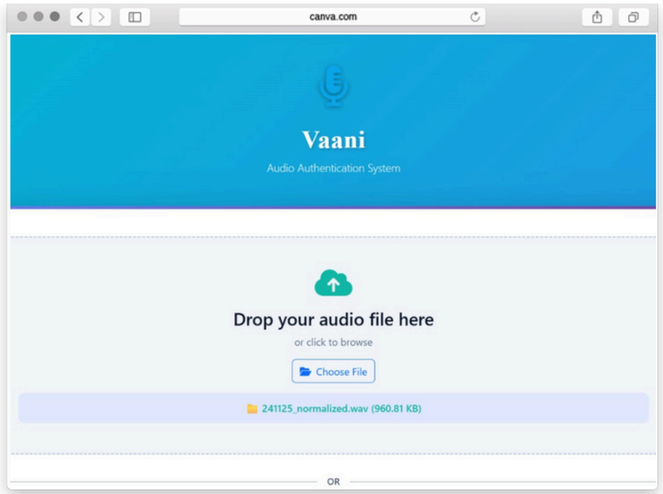
Impact:

1. **Builds trust** by verifying real vs. fake audio.
2. Helps **prevent deepfake** misuse and fraud.
3. Supports digital forensics with verifiable timestamps and hashes.
4. Enhances **security** in voice-based authentication systems.
5. **Reduces misinformation** spread through manipulated audio.

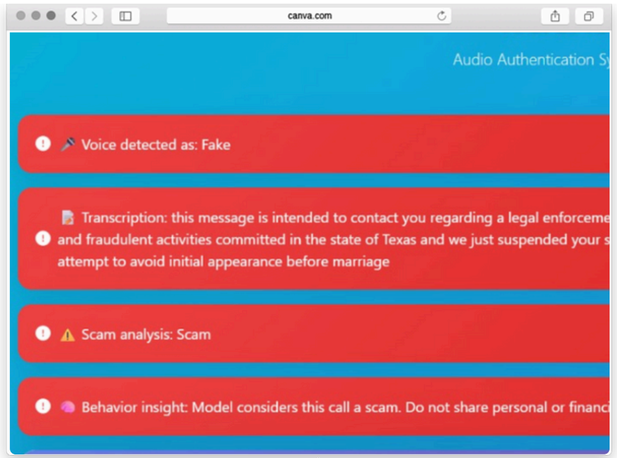
Benefits:

- Fast and accurate fake-audio detection.
- Lightweight and easy to deploy.
- User-friendly interface with certificate + QR verification.
- Tamper-proof validation using block hash.
- Scalable for real-world applications (banking, media, education).

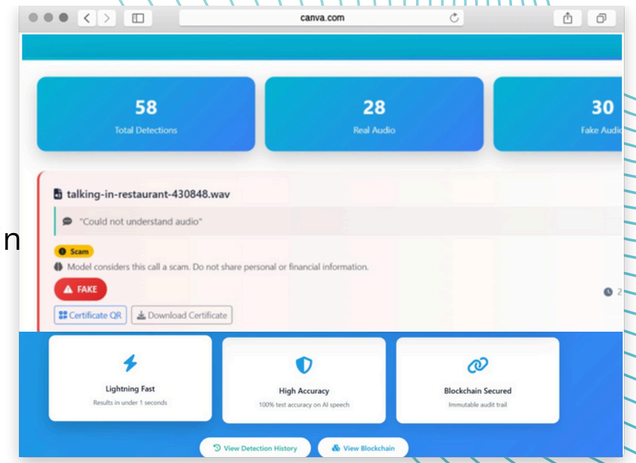
USER FLOW

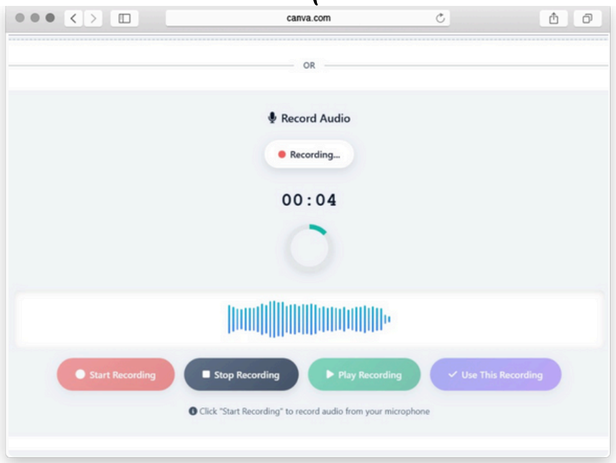


Audio Detection



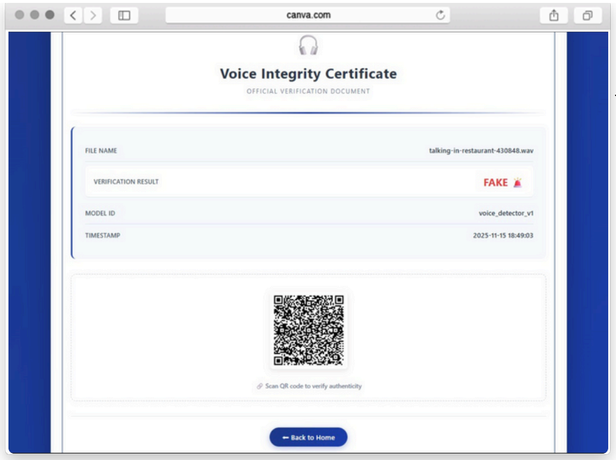
Dashboard and blockchain overview





Record audio

Certificate generation



Vaani
AUTHENTICITY IN EVERY WAVEFORM

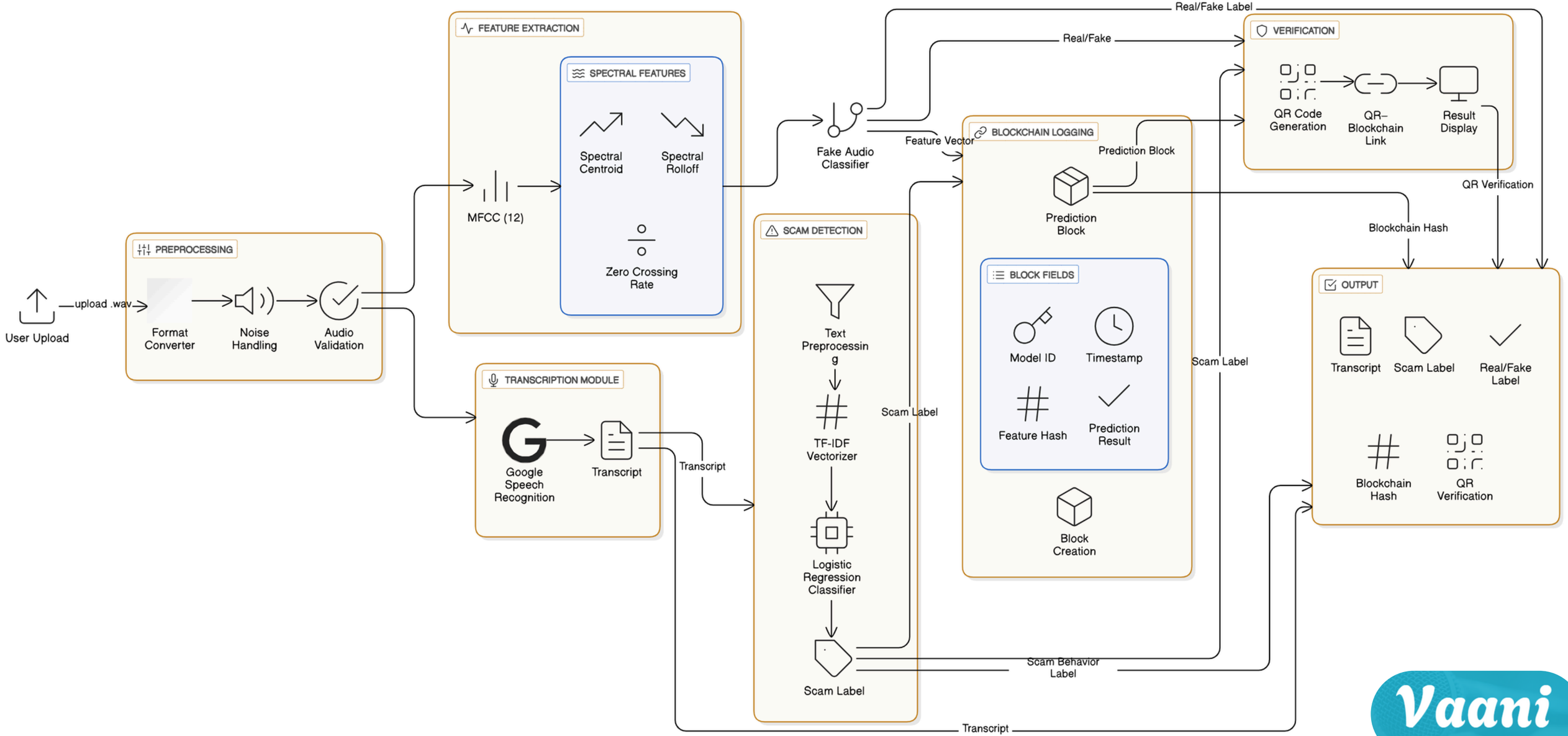
MODULE DESCRIPTION

Module	Module	Module
Frontend UI	templates/index.html, logs.html, blockchain.html	Provides upload interface, displays deepfake detection results, transcript, scam classification, QR verification link, history logs, and blockchain visualization.
Main Application	app.py	Central controller handling file uploads, audio preprocessing, MFCC & spectral feature extraction, deepfake prediction, transcription, scam classification, blockchain updates, and QR generation.
Audio Preprocessing & Feature Extraction	Inside app.py	Converts input audio to WAV if needed; extracts 12 MFCCs + spectral centroid, rolloff, zero-crossing rate → generates 15-feature vector for the ML classifier.
Deepfake Detection Model	model/voice_detector.pkl	Random Forest classifier trained on MFCC + spectral features. Predicts REAL vs FAKE audio for uploaded speech samples.
Transcription Module	Uses speech_recognition in app.py	Converts audio to text using Google Speech Recognition API (after ensuring WAV format). Provides transcript for scam analysis.
Scam / Behavior Text Classifier	models/better30_scam_text_model.pkl	TF-IDF + Logistic Regression model trained on BETTER30 dataset. Classifies transcript into 7 levels (e.g., Legitimate, Suspicious, Scam, etc.) and generates behavior comments.
Blockchain Module	blockchain.py, blockchain.db	Stores immutable logs containing prediction results, audio hash, transcript, scam label, timestamp, and QR verification reference. Ensures tamper-proof evidence.
QR Certificate Generator	Inside app.py	Generates QR codes for each blockchain entry, enabling quick verification of authenticity records.
User Logs Module	user_actions.py	Maintains structured JSON logs of predictions, transcripts, scam labels, and timestamps for UI visualization.
Datasets	BETTER30.csv, uploaded audio files	BETTER30 dataset for text classification; temporary uploaded audio files stored during analysis.
Storage	data/sample_alerts.json	Stores detection summaries including deepfake prediction, scam behavior label, comments, and timestamps.
Static Assets	static/	Contains CSS, JavaScript, and QR images for UI styling and rendering.

TECHSTACK USED

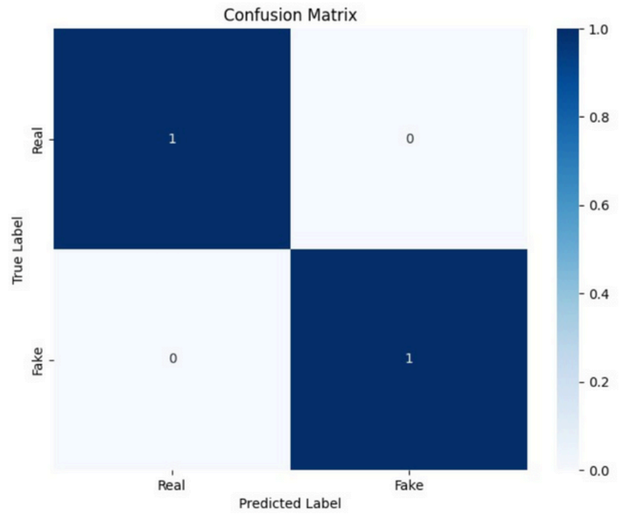
Technology	Technology	Technology
Flask (Python Web Framework)	Backend server + API + UI integration	Routing, file upload, rendering templates, connecting audio/text models, blockchain & QR modules
Librosa (Audio Analysis Library)	Audio preprocessing & feature extraction	librosa.load(), MFCCs, spectral centroid, rolloff, zero-crossing rate → 15-feature vector
Scikit-learn (ML Library)	Model training and prediction	Random Forest classifier (Real vs Fake), train–test split, classification report
Joblib	Model serialization	joblib.dump() & joblib.load() for audio and text classifier models
speech_recognition (Google Speech API client)	Audio transcription (primary)	Recognizer, AudioFile, recognize_google() — converts WAV → text (requires internet)
pydub	Audio format conversion & preprocessing for ASR	Convert MP3/MP4 → WAV, normalize, export WAV compatible with speech_recognition
(Optional) OpenAI Whisper / Local ASR	Offline/local transcription alternative	Better accuracy in noisy/multilingual scenarios; avoids external API calls
Custom Blockchain (Python)	Immutable tamper-proof detection history	Block creation, hashing, chain validation, stored in blockchain.db
QR Generation & Verification (qrcode, pyzbar)	Generate & verify certificate QR codes	QR encoding of block hash / verification URL; scan to validate record
JSON Storage	Lightweight logging format	Stores detection history, timestamps, predictions (data/sample_alerts.json)
HTML/CSS + Jinja Templates	Frontend interface	Upload page, results display, logs viewer, blockchain explorer

ARCHITECTURE DIAGRAM



Vaani
AUTHENTICITY IN EVERY WAVEFORM

ACCURACY SCORE



	precision	recall	f1-score	support
0	0.85	1.00	0.92	11
1	0.00	0.00	0.00	2
accuracy			0.85	13
macro avg	0.42	0.50	0.46	13
weighted avg	0.72	0.85	0.78	13

Classification report:

	precision	recall	f1-score	support
Scam	0.00	0.00	0.00	1
citing urgency"	0.00	0.00	0.00	2
emphasizing security and compliance"	0.00	0.00	0.00	0
highly_suspicious	0.00	0.00	0.00	2
legitimate	0.44	1.00	0.61	11
neutral	0.93	0.61	0.74	46
polite_ending	0.00	0.00	0.00	1
potential_scam	0.00	0.00	0.00	0
scam	0.86	0.84	0.85	37
scam_response	0.79	0.86	0.83	22
slightly_suspicious	1.00	0.50	0.67	2
standard_opening, identification_request	0.00	0.00	0.00	1
suggesting a dangerous situation"	0.00	0.00	0.00	0
suspicious	0.40	0.80	0.53	5
accuracy			0.72	130
macro avg	0.32	0.33	0.30	130
weighted avg	0.78	0.72	0.72	130

Vaani
Authenticity in Every Waveform

RESULTS

- **Feature Extraction:**
  The system successfully extracted Mel-Frequency Cepstral Coefficients (MFCC) and other spectral features using the Librosa library.These features effectively represented the key characteristics of the input audio signals, enabling accurate classification.

- **Model Performance:**
  The Random Forest Classifier achieved an accuracy of 87%, demonstrating strong performance in distinguishing between authentic andmanipulated or spam audio samples.The model's robustness was validated across multiple test samples, maintaining consistent prediction reliability.

- **Blockchain Integration:**
  The authenticity verification results were immutably stored in a Blockchain Ledger, ensuring transparency, traceability, and prevention of tampering.Each verification result was synchronized with SQLite for historical tracking and fast query access.

- **Certificate Generation:**
  Upon successful verification, the system generated a QR-based digital authenticity certificate in JPEG format, which can be downloaded via the Flask web interface.

FUTURE SCOPE

1. Integrate advanced deep-learning models like wav2vec-X and Whisper-v3 for **improved detection accuracy**.
2. Expand dataset with **multilingual** and **multi-speaker** deepfake samples for better generalization.
3. Develop a **real-time browser** or **mobile-based** fake audio detection tool.
4. Add **fuzzy-logic** based confidence scoring for more human-interpretable decisions.
5. Enable **cross-media deepfake detection** (audio + video + text) for multi-modal security systems.
6. **Deploy** lightweight on-device models for edge AI applications (IoT, smartphones).

CONCLUSION

The integration of audio analysis, machine learning, and blockchain creates a secure and reliable system for audio authenticity verification. With an achieved accuracy of 87%, the MFCC-based features combined with a Random Forest classifier effectively distinguish genuine from fake or spam audio. Immutable blockchain storage further strengthens trust by preserving tamper-proof verification records. This makes the system suitable for sensitive applications such as spam detection, voice-based authentication, and providing reliable, verifiable evidence for legal or forensic audio analysis.

REFERENCES

[1] M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, “The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation,” Procedia Computer Science, vol. 219, pp. 211–219, 2023.

[2] A. Hamza, M. D. Hassan, and M. Sajjad, “Deepfake Audio Detection via MFCC Features using Machine Learning,” ResearchGate, 2022.

[3] M. Ganavi, S. Narayan, and M. A., “AudioVeritas: A Machine Learning Model to Detect Deepfake Audio,” Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET), vol. 11, no. 1, 2023.

[4] J. Yi, R. Fu, and S. Watanabe, “A Comprehensive Survey on Spoofed Speech Detection,” arXiv preprint arXiv:2308.14970, 2023.

[5] J. Jung, H. Kim, and H. Shin, “Adversarial Attacks on Speaker Verifi cation Systems Using Neural TTS,” in Proc. Interspeech, 2021.

[6] A. Patel and P. Sharma, “Detection of Manipulated Speech Using MFCCs and Classical Machine Learning,” Int. J. Speech Technol., 2021.

[7] M. Todisco, H. Delgado, and N. Evans, “Constant Q Cepstral Coeffi cients for Spoofing Detection,” in Proc. Odyssey: Speaker and Language Recognition Workshop, 2016.

Thank You

Authenticity in Every Waveform