The Digital Trust Dilemma

In an age of synthetic reality, how can we determine what is real?

The rapid advancement of artificial intelligence has created a new frontier of media manipulation. Tools can now generate human faces that have never existed, clone voices with unnerving accuracy, and alter videos so seamlessly that the human eye can no longer be the final arbiter of truth. This reality poses a fundamental threat to security, public discourse, and the very concept of shared reality. This presentation introduces a system designed to restore clarity.

Understanding the Spectrum of Synthetic Media

“Deepfakes”—a portmanteau of “deep learning” and “fake”—are artificially generated or manipulated media created by advanced AI. As generative models like GANs and diffusion models become more powerful and accessible, the threat landscape has expanded.

Face Replacement & Reenactment

Replacing a person's face with another or manipulating their expressions and lip movements.

› **Primary Threat:** Political manipulation, creation of false evidence.

AI-Generated Imagery

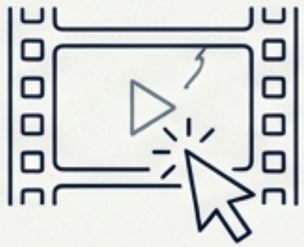
Creating photorealistic images of people and scenes that do not exist.

› **Primary Threat:** Propaganda, creating false identities, synthetic explicit content.

Voice Cloning & Synthesis

Generating speech that is indistinguishable from a real person's voice, often from just a few seconds of sample audio.

› **Primary Threat:** Impersonation for financial fraud, spreading misinformation.



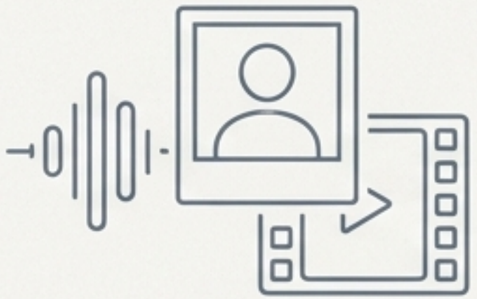
Video Manipulation

Altering entire video sequences, including backgrounds and actions, to create entirely false narratives.

› **Primary Threat:** Disinformation campaigns, eroding trust in video evidence.

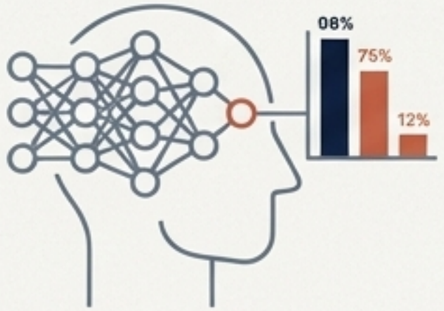
A Purpose-Built Response: The Deepfake Detection and Attribution (DFDA) System

DFDA is a multi-modal artificial intelligence system designed not only to identify synthetic media but also to empower users with clear, understandable insights.



Accurate Multi-Modal Detection

Analyse audio, images, and video using dedicated, specialised AI models to cover the full spectrum of deepfake threats.



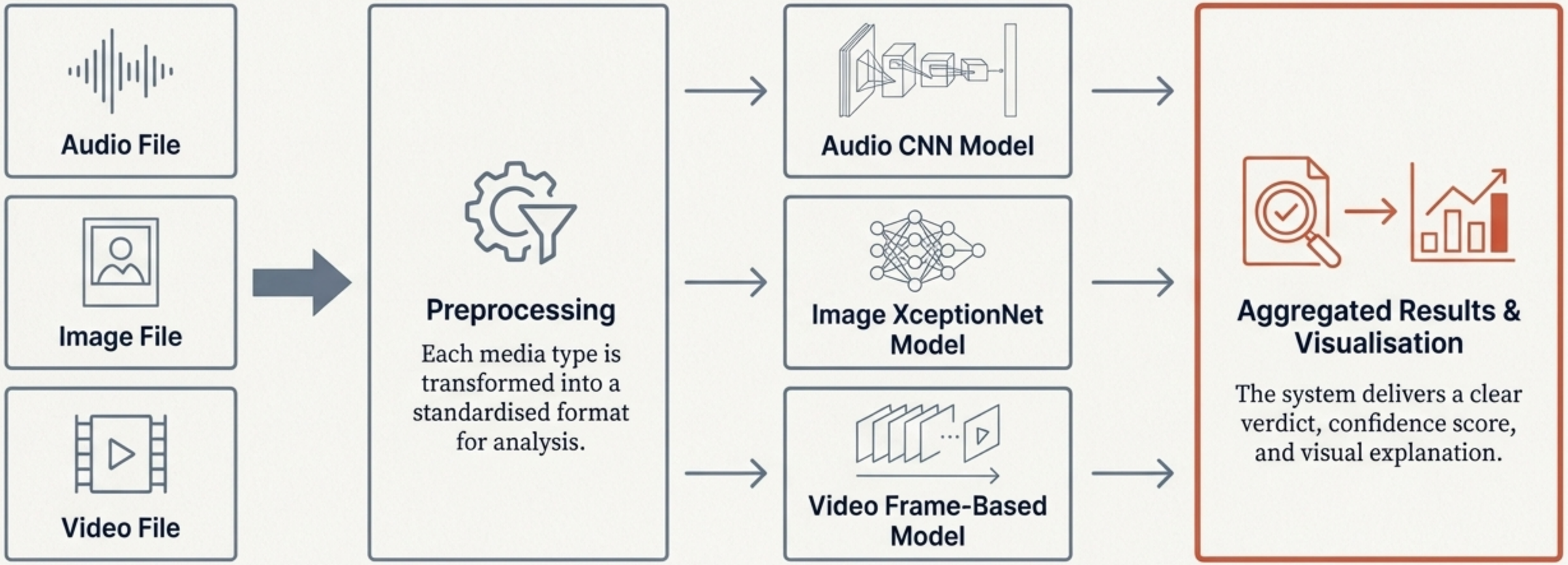
Explainability and Accessibility

Present results in simple language with intuitive visuals, making advanced detection accessible to non-experts like journalists, researchers, and forensic teams.

Cross-Model Attribution

Move beyond a simple ‘real vs fake’ verdict to identify the specific AI tool or method used to generate the synthetic media, aiding in digital forensics.

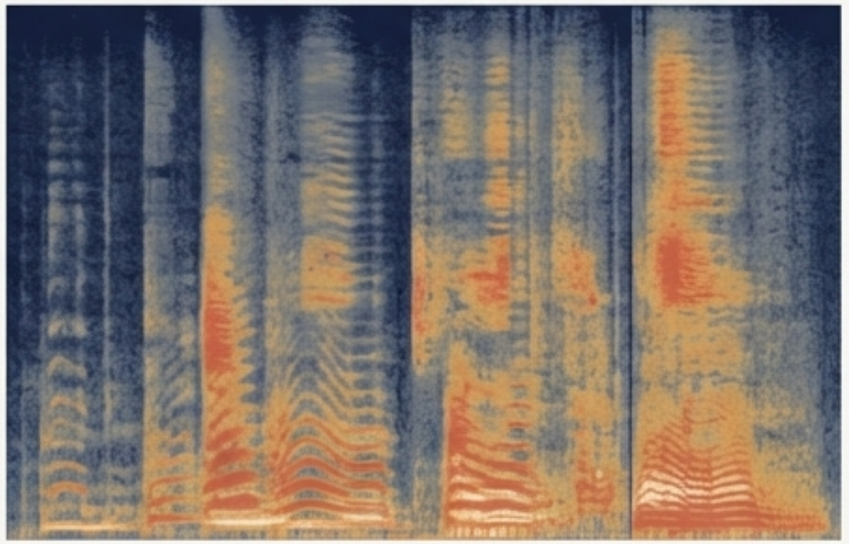
The DFDA Architecture: A Unified Multi-Branch Pipeline



This modular design ensures flexibility, high accuracy, and the ability to integrate new detection capabilities in the future.

The Audio Model: Learning to See the ‘Picture of Sound’

The model does not ‘listen’ to audio; it analyses a visual representation called a mel-spectrogram—a colourful heatmap of sound frequencies over time. This approach makes subtle, machine-generated artefacts visible.



What the CNN Detects

- Overly smooth or unnaturally perfect harmonics from voice cloning.
- Repetitive, uniform frequency bands characteristic of text-to-speech systems.
- Abrupt or unnatural transitions between sounds.

Training Foundation

Trained on thousands of samples from sources including

Original human recordings, ElevenLabs, Tacotron, and Voice Conversion models.

The Image Model: An Attention-Driven Hunt for Imperfections

The model is built on an enhanced XceptionNet, a powerful network for feature extraction. Its key innovation is an **Attention Block** (in terracotta HEX #D95F43), which allows the model to focus computational resources on the most informative regions of an image, similar to how a human expert would examine a photograph.

Micro-Artefacts Detected

- Inconsistent lighting and shadowing, especially on the face.
- Distorted or asymmetrical details in teeth, pupils, and ears.
- Irregular or strangely blurred backgrounds.
- Unnatural skin texture (often overly smooth).
- Incorrect reflections in the eyes that do not match the environment.

The Video Model: Exposing Flaws Through Temporal Analysis

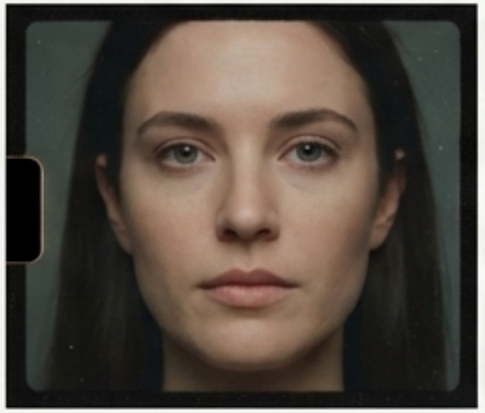
Detecting video deepfakes requires analysing motion and consistency over time, as forgeries that look perfect in a single frame can reveal themselves through unnatural movement. The DFDA model breaks videos into individual frames for analysis.

The Power of Aggregation: The system classifies dozens or hundreds of frames from a single video. The final verdict is based on probability averaging across all frames. This method provides stability and is highly resistant to occasional perfect-looking frames.

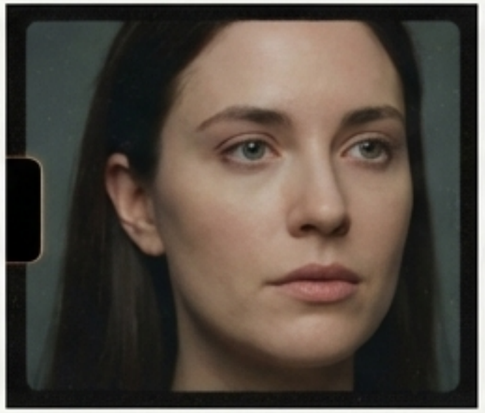
Temporal Inconsistencies Detected

- Jitter or warping around the hairline and jaw as the head turns.
- Asynchronous or unnatural blinking patterns.
- Inconsistent skin tones that shift under changing light.
- Sudden distortions around the mouth during speech.

Frame 132



Frame 142



Frame 152

The Foundation of Accuracy: A Diverse, Large-Scale Dataset

The performance of any AI system is determined by the quality and breadth of its training data. DFDA is built on a comprehensive, **multi-modal dataset** covering a wide range of real-world and synthetic media.

Image Dataset

19,000+ images

across 10 distinct classes

Sources include:

- Real photos, DALL-E, Midjourney, Stable Diffusion, FaceSwap, DeepFaceLab, NeuralTextures.

Audio Dataset

Thousands

of 3-5 second clips

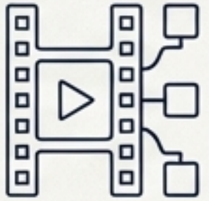
Sources include:

Real human speech, ElevenLabs, Tacotron, Voice Conversion models, Generic TTS.

Video Dataset

Tens of thousands

of frames from hundreds of videos



Sources include:

Real videos, Face2Face, FaceSwap, NeuralTextures manipulation datasets.

Validated Performance: High Accuracy Across All Modalities

Rigorous testing on validation datasets confirms the DFDA system's effectiveness in accurately distinguishing authentic content from sophisticated AI-generated fakes.

Audio Detection Accuracy

~98%

Demonstrates high reliability in identifying synthetic voices and cloned audio.

Image Detection Accuracy

99%+

Achieves near-perfect accuracy in classifying images from ten different real and synthetic sources.

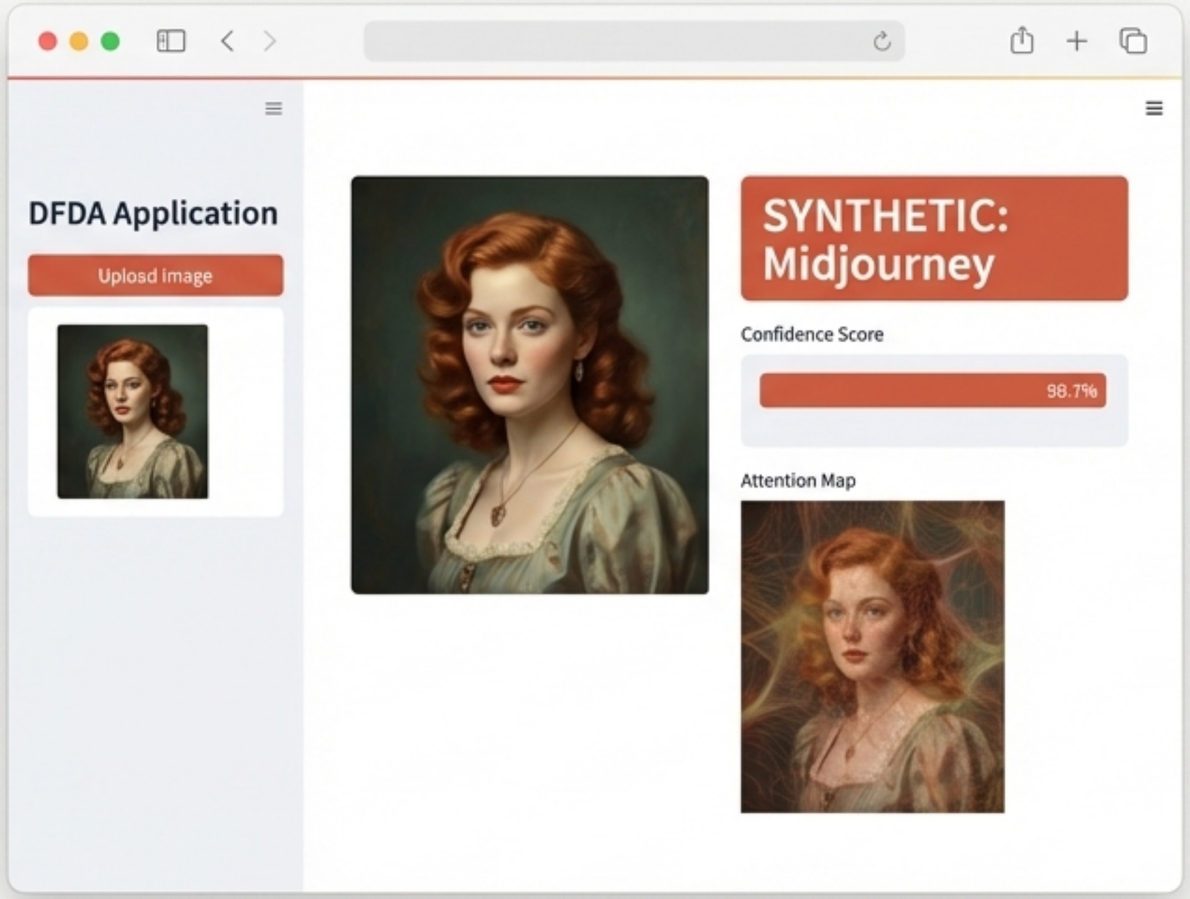
Video Detection Accuracy

~90%

Reflects strong performance on the complex task of frame-based video analysis, successfully identifying manipulated footage.

From Complex System to Accessible Tool: The DFDA Application

To bridge the gap between advanced AI and practical use, the entire DFDA system is accessible through an intuitive Streamlit application. This empowers users without a background in machine learning to leverage its full capabilities.



Key Features

- Simple drag-and-drop interface for uploading audio, image, or video files.
- Clear presentation of the final prediction (e.g., ‘REAL’ or ‘SYNTHETIC: Midjourney’).
- A confidence score bar chart showing the model’s certainty.
- Supporting visualisations, such as the mel-spectrogram for an audio file, to provide context for the decision.

Empowering Industries: Key Applications for DFDA

By providing reliable and accessible deepfake detection, DFDA serves as a critical tool across multiple domains dedicated to upholding authenticity and security.

Digital Forensics

Aiding investigators in verifying the authenticity of digital evidence and attributing synthetic media to specific generation tools.

Journalism & Media

Enabling newsrooms to authenticate video and audio sources, combating the spread of misinformation.

Cybersecurity

Protecting against fraud schemes that use cloned voices or synthetic identities for social engineering attacks.

Academic Research

Providing a robust platform for researchers studying the behaviour of generative models and detection techniques.

Public Awareness

Serving as an educational tool to demonstrate the capabilities and risks of deepfake technology.

Defending Digital Authenticity

The challenge posed by deepfakes is not a single event, but an ongoing technological race. As generative models evolve, so too must our methods for detection and verification. Systems like DFDA are not just defensive tools; they are essential infrastructure for maintaining trust in a digital world. By combining advanced AI with a focus on accessibility and clear explanation, we can equip society to navigate the complexities of synthetic reality with confidence and clarity.