Deepfake Detection and Attribution

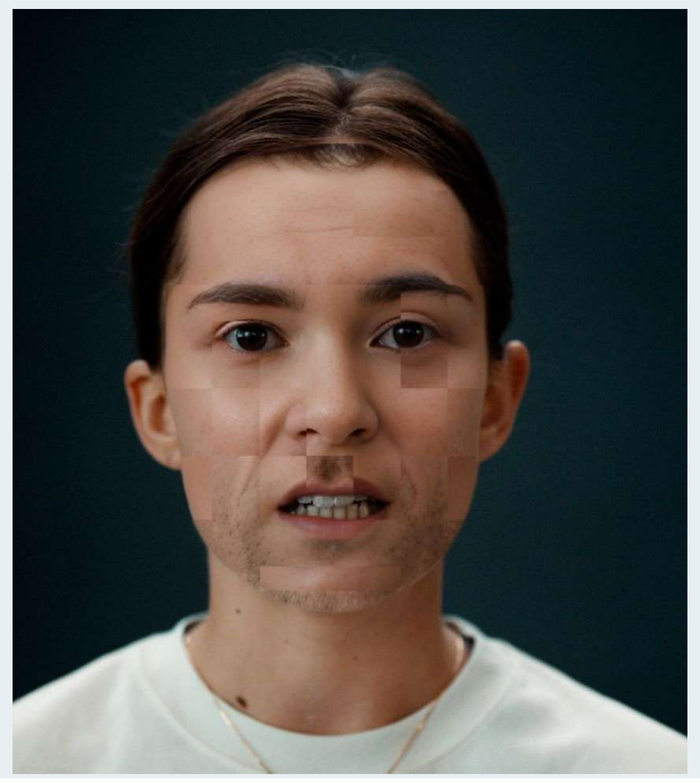
Team:

J. Sri Varsha	2303A51L99	B. Jayanth	2303A51LA7	Guided By:
K. Sindhu	2303A51LA0	G. Mahendra	2303A51LA9	Dr. Ramesh Babu A
		K. Nikhil	2303A51LB0	Assistant Professor

SR UNIVERSITY

INTRODUCTION

Deepfakes are fake videos, images, or voices created using artificial intelligence. They can easily spread false information, damage reputations, and cause confusion. This project, called **Deepfake Detection and Attribution** System, helps identify whether a piece of media is real or fake. It works with three types of data — audio, image, and video — using deep learning models to find small clues that reveal manipulation. The goal is to protect digital truth and make media more trustworthy.

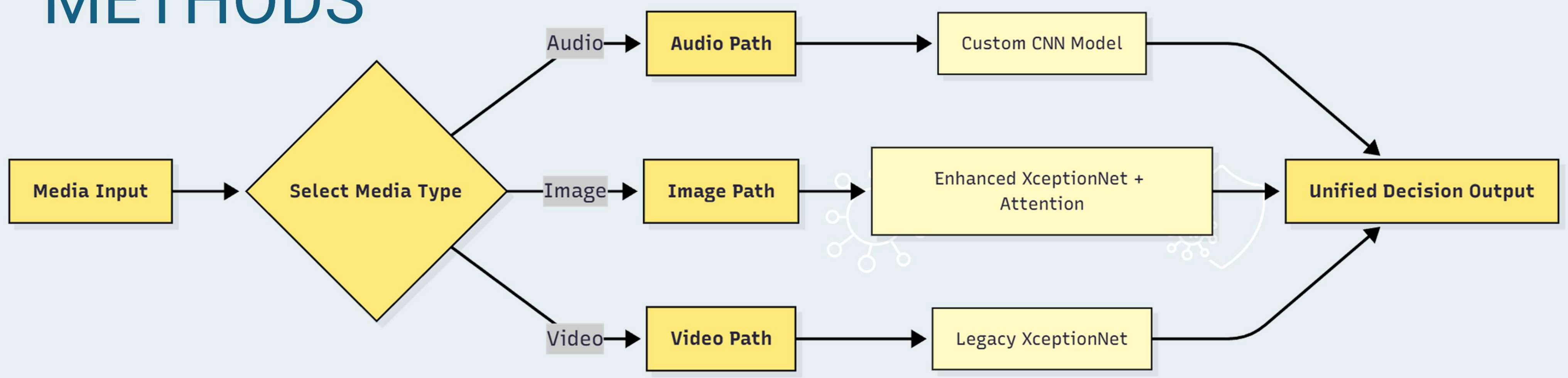


95000+ deepfake videos were identified online in 2023, 550% increase from 2019.

Another source says **43%** of people aged 16+ had seen at least one synthetic (deepfake) media item in the last six months.

Deepfakes are increasing rapidly every year, making online media harder to trust.

METHODS



Media Input → Select Media Type

Audio → Audio Path → Custom CNN Model → Unified Decision Output

Image → Image Path → Enhanced XceptionNet + Attention → Unified Decision Output

Video → Video Path → Legacy XceptionNet → Unified Decision Output

1. Audio Model

The audio model was trained on real and synthetic voice datasets using mel-spectrograms as input. It used a CNN architecture with cross-entropy loss and the Adam optimiser to learn differences in tone, pitch, and frequency patterns.

2. Image Model

The image model was trained with large datasets containing both real and AI-generated images. It used an enhanced XceptionNet with attention layers, batch normalisation, and dropout to improve feature learning and prevent overfitting.

3. Video Model

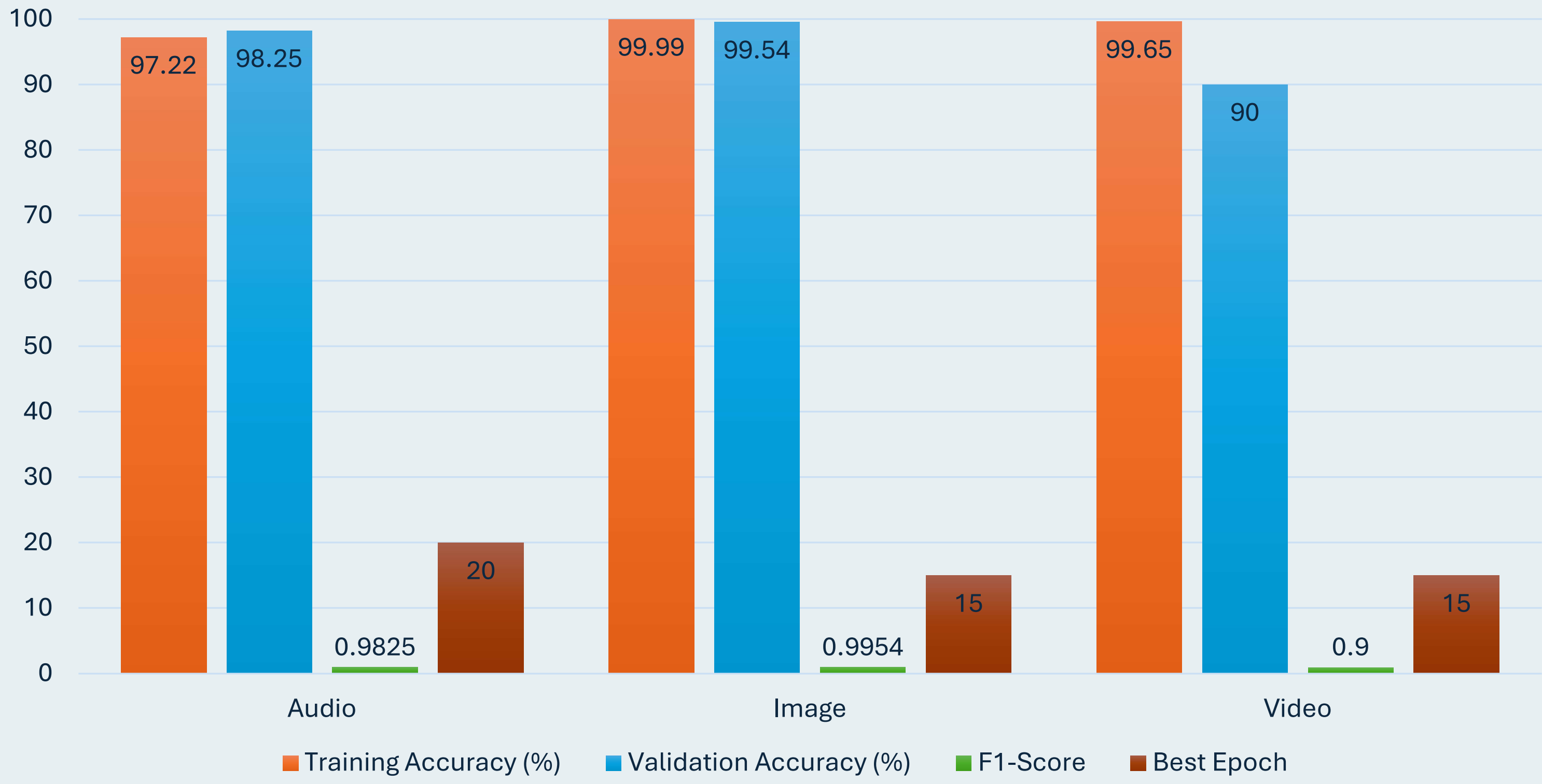
The video model was trained on frame sequences extracted from deepfake and real videos. It employed a legacy Xception-based architecture with temporal frame aggregation to capture motion inconsistencies and classify manipulated clips accurately.

System Integration:

All three trained models were combined into a single multi-modal framework for unified deepfake detection. A Streamlit-based interface was developed to integrate the models, allowing users to upload media, run analysis, and view real-time detection results.

RESULTS

The System delivered highly accurate results across all three media categories. The audio model achieved a validation accuracy of 98.25%, proving effective in detecting cloned and synthesized voices with minimal false positives. The image model produced outstanding performance, reaching a 99.54% accuracy and accurately distinguishing between real and AI-generated images. The video model achieved a 90% validation accuracy, identifying deepfake manipulations such as face swaps and unnatural motion transitions.



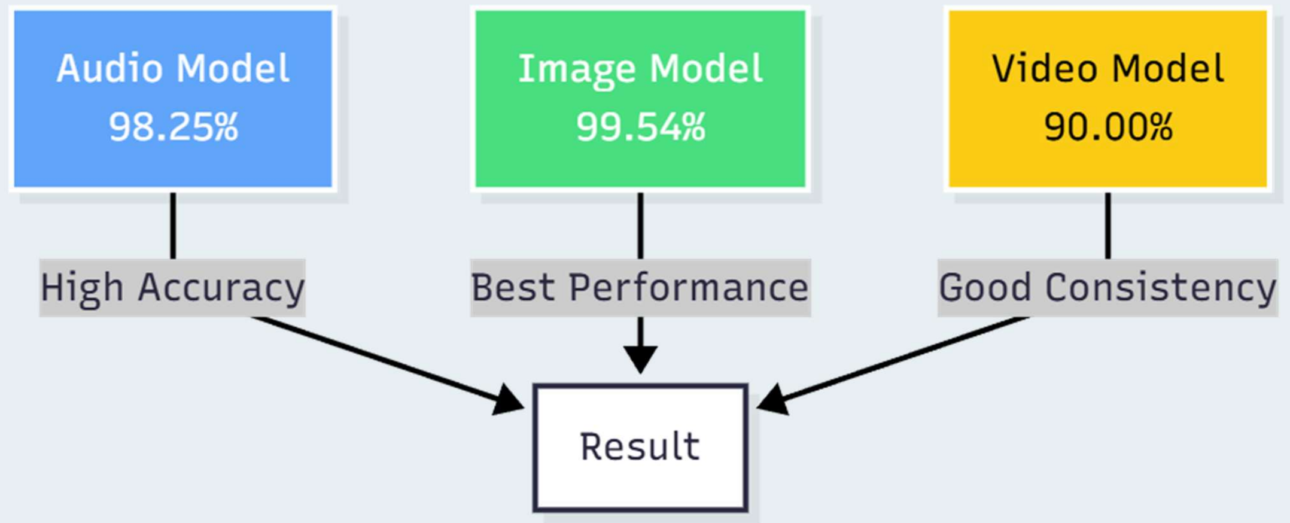
During training, all models showed stable convergence and strong generalization across datasets. **The attention-based XceptionNet** significantly improved the image model's ability to localize manipulation areas, while the temporal frame aggregation enhanced the video model's motion analysis. The confusion matrices and F1-scores confirmed balanced precision and recall across all classes. Overall, the results highlight that the integrated multi-modal design provides a powerful, dependable framework for detecting and attributing deepfakes across different media types.

DISCUSSION

The results of this project highlight the growing need for reliable deepfake detection methods across multiple forms of media. Each model played a key role in addressing different manipulation types, the audio model effectively caught cloned and synthetic voices, while the image model performed exceptionally well in identifying AI-generated visuals. The video model, though slightly lower in accuracy, proved highly useful for detecting unnatural movements and frame inconsistencies. Together, they formed a strong multi-modal framework that covered most common deepfake formats found online.

The project demonstrates that a combination of spectral, spatial, and temporal analysis provides a complete solution for deepfake detection. It not only strengthens digital media verification but also lays the groundwork for future research in automated content authentication and misinformation control.

CONCLUSION



Audio Model 98.25% → High Accuracy → Result

Image Model 99.54% → Best Performance → Result

Video Model 90.00% → Good Consistency → Result

The Deepfake Detection and Attribution System identifies fake audio, image, and video content using advanced deep learning techniques. Each module, including the CNN-based audio model, the Enhanced XceptionNet for images, and the Legacy XceptionNet for videos, showed strong accuracy and consistency, confirming system reliability. The integration of these models into a multi-modal framework improved detection efficiency and reduced misclassification. The Streamlit dashboard provides an interactive interface for real-time analysis and clear visual representation of results. Overall, the project strengthens media authenticity verification and offers a dependable solution to tackle digital misinformation.