1. Project Objective

The objective of this project is to design, train, and evaluate a multimodal deep learning system for deepfake detection that combines visual and audio information.
The primary research focus is to evaluate whether a supervised multimodal deep learning model can generalize to unseen deepfake manipulation methods using a structured Leave-One-Method-Out (LOMO) evaluation protocol.
Secondary objectives include performance benchmarking, modality contribution analysis, and qualitative failure analysis.

2. Motivation and Research Justification

Deepfake generation techniques evolve rapidly, often rendering supervised detectors ineffective when exposed to manipulation methods not seen during training.
Existing literature shows strong multimodal methods but often relies on self-supervised learning, synthetic audio generation, identity reference data, or inconsistent evaluation protocols.
This project focuses on a simpler, reproducible, supervised multimodal framework with disciplined evaluation, suitable for a college-level journal and major project.

3. Dataset Strategy

Primary Dataset: FaceForensics++
- Contains real videos and visually manipulated fake videos.
- Manipulation methods include DeepFakes, FaceSwap, Face2Face, and NeuralTextures.
- Audio streams remain original and unmanipulated.
- Used to train the multimodal model and perform Leave-One-Method-Out evaluation.

Secondary Dataset (Supplementary):
- FakeAVCeleb OR a small subset of DFDC.
- Contains audio-only, video-only, and audio-video manipulated samples.
- Used only for additional validation to demonstrate robustness to audio manipulation scenarios.
- Not used as the primary training dataset.

4. Data Preparation

Video Processing:
- Extract video frames at a fixed rate (e.g., 10–20 frames per video).
- Detect and crop face regions.
- Resize to fixed spatial resolution (e.g., 224x224).

Audio Processing:
- Extract audio waveform from video.
- Convert audio to Mel-spectrogram representation.
- Normalize and resize spectrograms to fixed dimensions.

Each training sample consists of synchronized video frame sequences and corresponding audio spectrograms.

5. Model Architecture

Video Branch:

- CNN backbone (ResNet-18 or EfficientNet-B0).
- Initialized with ImageNet pretrained weights.
- Fine-tuned on deepfake video frames.

Audio Branch:
- CNN operating on Mel-spectrogram inputs.
- Initialized with pretrained or random weights.
- Trained to learn speech and acoustic cues.

Fusion Module:
- Concatenation of audio and video embeddings.
- Fully connected layers for joint representation learning.

Classifier:
- Binary classification head (Real vs Fake).

6. Training Strategy

Training Type:
- Supervised deep learning.
- End-to-end training of audio branch, video branch, and fusion layers.
- Optionally freeze early layers of pretrained backbones.

Loss Function:
- Binary Cross-Entropy Loss.

Optimizer:
- Adam optimizer with learning rate scheduling.

Training Duration:
- 5 to 10 epochs per LOMO split.
- Early stopping based on validation performance.

7. Evaluation Protocol

Primary Evaluation: Leave-One-Method-Out (LOMO)
- For each manipulation method M:
- Train on real videos and fake videos excluding M.
- Test on fake videos of M and real videos.
- Metrics: AUC, Accuracy, F1-score.

Secondary Evaluations:
- Cross-dataset testing using FakeAVCeleb or DFDC.
- Compression robustness tests.
- Modality ablation:
- Video-only
- Audio-only
- Audio + Video

8. Analysis and Explainability

Analysis includes:
- Comparison of multimodal vs unimodal performance.
- Identification of failure cases.
- Qualitative inspection of misclassified samples.
- Discussion of limitations related to silent clips, background noise, and extreme compression.

Explainability is provided via:
- Attention or activation visualization (optional).

- Frame-level confidence trends.
- Descriptive failure analysis.

9. Expected Contributions

1. A reproducible supervised multimodal deepfake detection framework.
2. A clean and explicit LOMO evaluation protocol.
3. Empirical evidence of generalization behavior on unseen manipulation methods.
4. Modality contribution analysis and practical failure insights.

10. Ethical Considerations

All datasets used are publicly available and intended for research.
No new data is collected.
Results are reported responsibly, avoiding misuse or overclaiming capabilities.

11. Project Deliverables

- Trained multimodal deep learning model.
- Experimental results and tables.
- Major project report.
- College journal paper submission.
- Codebase with documentation.