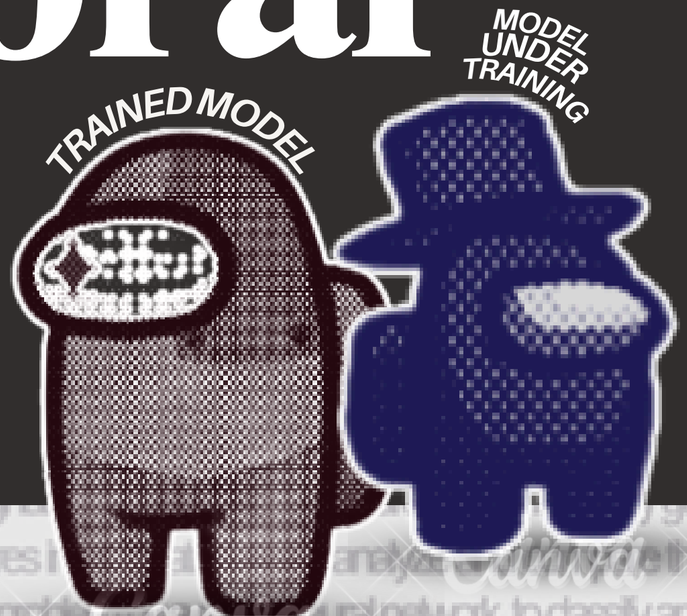
Deepfake detection in human faces using behavioral cues

16 May, 2026

By Manshika, Manaasve, Roshni



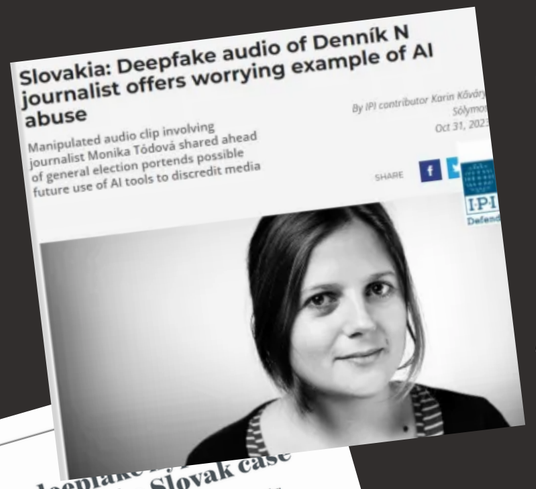
In late 2023, a compromising video allegedly featuring **Rashmika Mandanna** exploded across social media. The clip showed **her entering an elevator in revealing clothing, and within hours**, it spread rapidly across various platforms such as instagram and Whatsapp.

To most viewers, **the video looked completely real.** But the woman in the video was **not Rashmika Mandanna at all.**

It was a deepfake of her, someone had digitally replaced another woman's face with Rashmika's.

The Rashmika Mandanna case exposed how easily AI can manipulate trust, identity, and reality itself. **Deepfakes today, are creating a "crisis of knowing",**

The SMD approach was shown to be an efficient alternative to conventional MMPBSA method, but it can be few orders of magnitude faster [9], which enables screening of a correspondingly larger number compounds. Such gain in performance is possible due to extreme conditions used in SMD simulations, e.g., the pulling speed in simulations is several orders of magnitude higher than that used in single molecu



Political Manipulation

Deepfakes have been weaponized in misinformation campaigns to influence electoral outcomes as seen in the *2023 Slovak parliamentary elections* where a fabricated audio clip targeted party leaders

The current problem?

Cybercrime

Voice cloning enables vishing attacks, identity theft, and financial fraud.

Many **deepfakes involve non-consensu**al pornography, causing severe reputational and psychological harm.

LIAR'S DIVIDEND

The existence of Deepfakes creates doubt around authentic videos**, weakening trust in digital evidence.**

Public awareness of deepfakes enables guilty actors to deny real recordings and avoid accountability.

1. Busch, E., & Ware, J. (2023). The Weaponisation of Deepfakes: Digital Deception by the Far-Right. International Centre for Counter-Terrorism.
2. Singh, S., & Dhumane, A. (2025). Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges. MethodsX, 15, 103632. DOI: 10.1016/j.mex.2025.103632
3. Tolosana, R., Romero-Tapiador, S., Vera-Rodriguez, R., Gonzalez-Sosa, E., & Fierrez, J. (2022). DeepFakes detection across generations: Analysis of facial regions, fusion, and performance evaluation. Engineering Applications of Artificial Intelligence, 110, 104673. DOI: 10.1016/j.engappai.2022.104673

POTENTIAL APPLICATIONS

- **Social Media Filters:** Flagging AI videos before they go viral.
- **Newsroom Verification:** "Truth checker" to verify footage before it is broadcast as news.
- **Court Evidence:** Prove if a video is genuine or a manipulated fake for legal cases.
- **Personal Safety:** identify and block deepfake "revenge porn" and blackmail videos.
- **Mobile Protection:** Because the solution is lightweight, it can run directly on mobile phones to protect everyday users in real-time

The solution:
Detecting "Human Messiness" via Behavioral Patterns

POTENTIAL IMPACT

- **Restoring Digital Trust:** Helps people trust what they see online again.
- **Protecting Elections:** Prevents fake videos of world leaders from tricking voters.
- **Saving Money:** Saves businesses and individuals from losing millions to AI fraud.
- **Safeguarding Privacy:** It provides a way for citizens to fight back against cyberbullying and deepfake-based harassment.
- **Explainable Decisions:** Unlike "black box" AI, this tool can explain why a video is fake by pointing to messy lip-sync or eye-blinking errors.

1. Busch, E., & Ware, J. (2023). The Weaponisation of Deepfakes: Digital Deception by the Far-Right. International Centre for Counter-Terrorism.
2. Singh, S., & Dhumane, A. (2025). Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges. MethodsX, 15, 103632. DOI: 10.1016/j.mex.2025.103632
3. Tolosana, R., Romero-Tapiador, S., Vera-Rodriguez, R., Gonzalez-Sosa, E., & Fierrez, J. (2022). DeepFakes detection across generations: Analysis of facial regions, fusion, and performance evaluation. Engineering Applications of Artificial Intelligence, 110, 104673. DOI: 10.1016/j.engappai.2022.104673

Literature review

METHODOLOGY

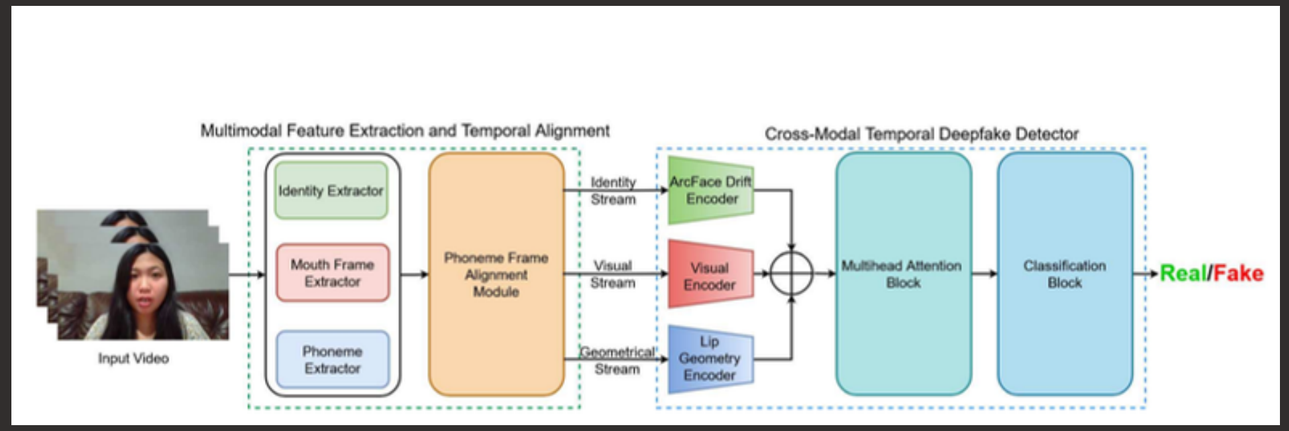
The PIA (Phoneme-Temporal and Identity-Dynamic Analysis) framework utilizes a multimodal detection pipeline that simultaneously analyzes three specific areas:

- **language (phonemes)**
- **lip motion (geometry)**
- **facial identity (embeddings)**

LIMITATIONS

1. **Resolution Dependency:** Works best at the resolution it was trained on.
2. **Language Limitation:** Designed mainly for English speech.
3. **Audio-Only Fakes:** May miss cases where only the voice is cloned.
4. **Noise Sensitivity:** Requires clear audio for accurate lip–speech alignment.

Paper #1 : PIA: Deepfake Detection Using Phoneme-Temporal and Identity-Dynamic Analysis



Achieved a near-perfect 99.8% AUC on the FakeAVCeleb dataset and 98.06% AUC on the high-quality DeepSpeak v2.0

Datta, S. K., Ranga, T., Sun, C., & Lyu, S. (2025). PIA: Deepfake detection using phoneme-temporal and identity-dynamic analysis. In Proceedings of the IEEE PIA: Deepfake detection using phoneme-temporal and identity-dynamic analysis. arXiv preprint, arXiv:2510.14241v1

development of effective protein-peptide docking techniques [4–7] and finding an efficient alternative to binding affinity [8–15] have been a major focus of computational studies in recent years. Recently, steered molecular dynamics (SMD) simulations have become popular to measure mechanical stability which could be used to assess the strength of the molecular interactions

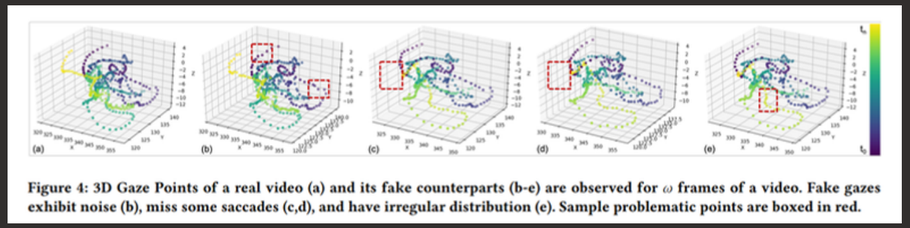


Figure 4: 3D Gaze Points of a real video (a) and its fake counterparts (b-e) are observed for ω frames of a video. Fake gazes exhibit noise (b), miss some saccades (c,d), and have irregular distribution (e). Sample problematic points are boxed in red.

Achieved 92.48% accuracy on FF++ and 99.27% on DeeperForensics, showing robustness to blur and post-processing artifacts.

Paper #2 : Where do deep fakes look? Synthetic face detection via gaze tracking

METHODOLOGY

The method extracts **eye landmarks** from video frames and analyzes **gaze patterns** over time using **visual, geometric, spectral, and symmetry cues**. These features capture inconsistencies in eye behavior and generative noise.

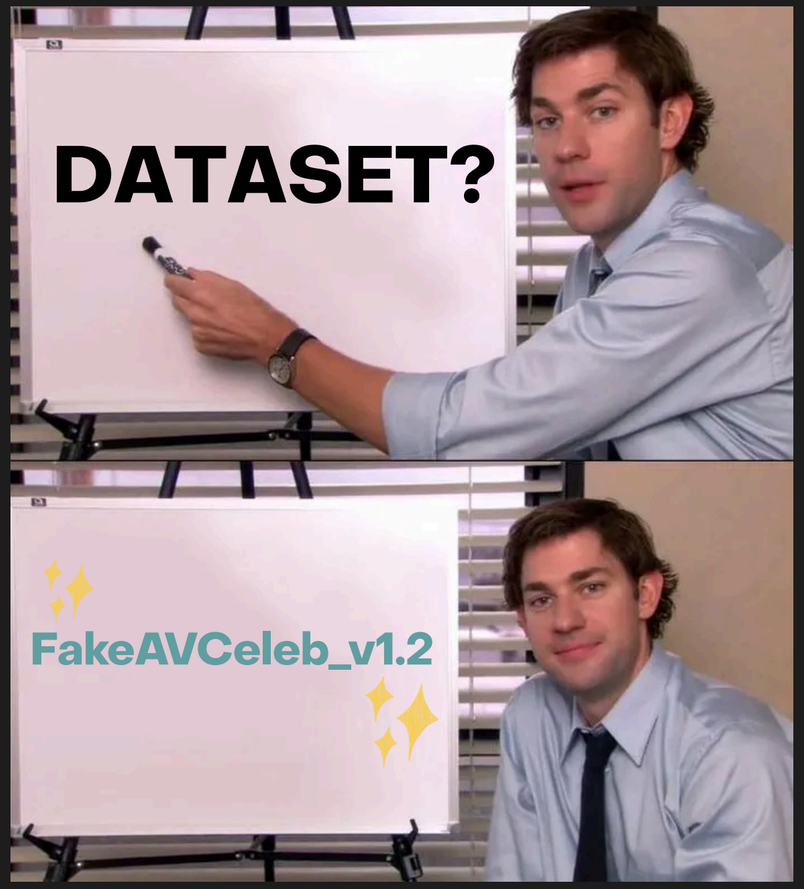
LIMITATIONS

- **Requires clearly visible eyes**; masks, glasses, or low lighting reduce detection reliability.
- **Pose Sensitivity** – Non-frontal head poses can cause gaze inconsistencies that affect detection.
- **Generator Specificity**: Performs worse on GANs like Face2Face or Neural Textures that leave eyes unchanged.

Demir, İ., & Çiftçi, U. A. (2021). Where do deep fakes look? Synthetic face detection via gaze tracking. 2021 Symposium on Eye Tracking Research and Applications (ETRA '21 Full Papers). https://doi.org/10.1145/3448017.3457387

Common Research Gaps

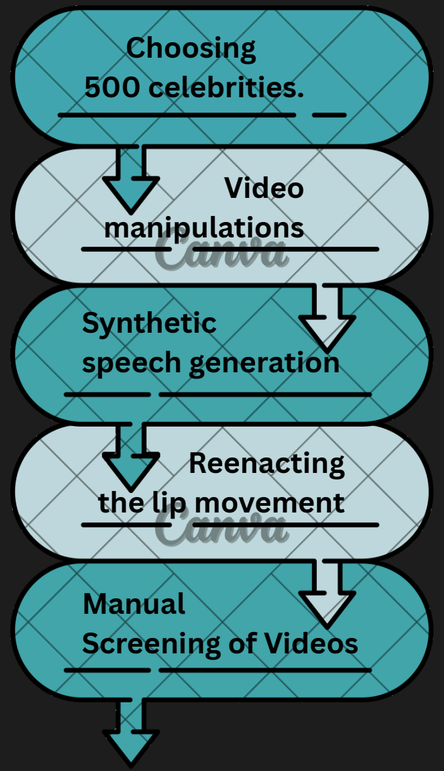
Paper Name (Citation)	Key Insight (The Gap)	Our Motivation (The Solution)	Paper Link
Assessing Deepfake Detection Models (Gamage, 2025)	Poor Generalization: Most models are benchmarked on single datasets and "collapse" when they encounter unseen AI generation tools.	Behavioral Analysis: Our model tracks ALL physical "Human Messiness" (like motion jerkiness ,eye-blinking and lip-syncing) which are harder for any AI tool to simulate perfectly across different platforms.	https://www.utupub.fi/bitstream/handle/10024/182061/LiyanaGamage_Chathura_Thesis.pdf?sequence=1
FaceForensics++: Learning to Detect Manipulated Facial Images (Rössler et al., 2019)	Compression Vulnerability: Standard detectors fail when social media platforms (like WhatsApp) use aggressive compression, which "launders" or erases subtle digital traces.	Neural Fingerprinting: By compressing 1,280 features into a 128-unit "Neural Fingerprint", our model ignores the blurry pixel noise and focuses only on the core behavior that survives compression.	https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9010912
Where Do Deep Fakes Look? (Demir & Çiftçi, 2021)	"Black Box" Problem: Most AI models classify videos without providing an intermediate representation that a human judge or journalist can interpret.	Multi-Expert Branches: We keep the data for Lips, Eyes, and Motion separate, allowing the model to generate a report explaining why it reached a decision (e.g., "Mismatched lip-sync detected").	https://dl.acm.org/doi/epdf/10.1145/3448017.3457387
Comparative Study of MTCNN and YuNet (Adithya et al., 2025)	High Resource Costs: Research often ignores that high-end models require 3–5 GB of RAM, making them too "heavy" to run on normal mobile phones.	Efficiency Focus: We use the YuNet face detector, which is 3x faster than traditional tools (0.008s vs 0.024s), allowing our multi-expert system to run on everyday devices.	https://www.scitepress.org/Papers/2025/138703/138703.pdf



Why FakeAVCeleb?

- Multimodal deepfake dataset (audio + video)
- Contains real and manipulated media
- Includes lip-synced synthetic audio with manipulated video
- Designed for audio-visual deepfake detection

Data Collection and Generation by the authors



VoxCeleb2 dataset (YouTube interviews of 6,000+ celebrities.)

(Using Faceswap and FSGAN)

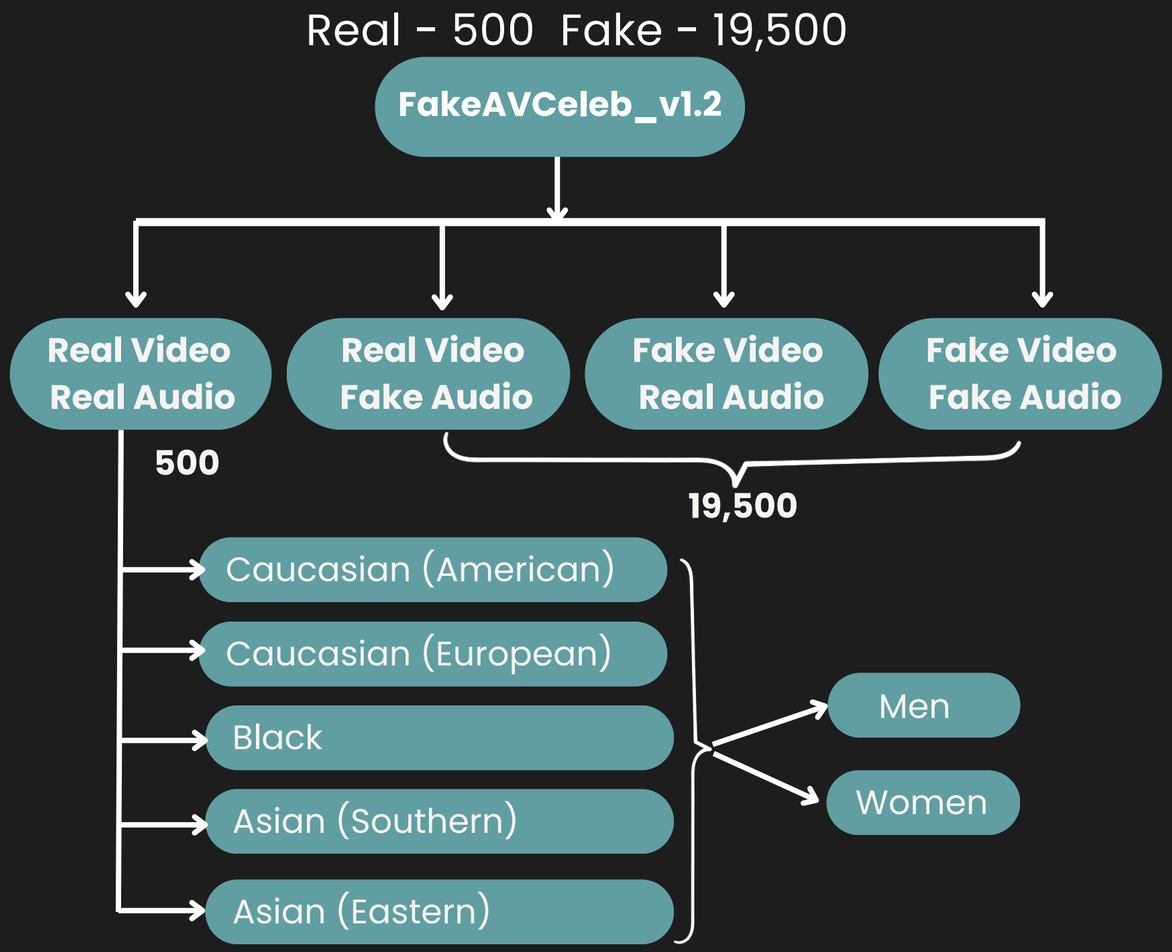
(Using transfer learning based tools SV2TTS)

(Using Wav2Lip)

filtering for high-quality, realistic results

Khalid, H., Tariq, S., Kim, M., & Woo, S. S. (2021). FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. Advances in Neural Information Processing Systems, 34 (NeurIPS 2021) Datasets and Benchmarks Track

Dataset Structure



Quantitative comparison of FakeAVCeleb to existing publicly available Deepfake Dataset

Dataset	Real Videos	Fake Videos	Total Videos	Rights Cleared	Agreeing subjects	Total Subjects	Synthesis Methods	Real Audio	Deepfake Audio
UADFV [14]	49	49	98	No	0	49	1	No	No
DeepfakeTIMIT [41]	640	320	960	No	0	32	2	No	Yes
FF++ [36]	1,000	4,000	5,000	No	0	N/A	4	No	No
Celeb-DF [34]	590	5,639	6,229	No	0	59	1	No	No
Google DFD [36]	363	3,000	3,363	Yes	28	28	5	No	No
DeeperForensics [33]	50,000	10,000	60,000	Yes	100	100	1	No	No
DFDC [35]	23,654	104,500	128,154	Yes	960	960	8	Yes	Yes
KoDF [37]	62,166	175,776	237,942	Yes	403	403	6	Yes	No
FakeAVCeleb	500	19,500	20,000	No	0	500	4	Yes	Yes

"ETHICAL CONCERNS"

The primary ethical concern cited by the authors is racial bias

- Authors attempted to reduce racial bias by selecting 100 celebrities from each of the 5 ethnic groups with a 50/50 male-to-female ratio.
- However, dataset diversity may still be limited, and may not fully represent global populations.

Key Preprocessing Steps

- Dense Frame sampling to extract 30 frames per video. (each video <= 10 seconds)
- Videos with insufficient valid facial frames were skipped to maintain consistent input dimensions.
- A conservative crop with margin scaling (~1.3×) is applied around the face.
- Each cropped face is resized to 224 × 224 pixels (because EfficientNetB0 expects this input size.)
- MediaPipe FaceMesh extracted lip and eye regions using facial landmarks.
- Pixel values are scaled from 0–255 to 0–1 and then passed through EfficientNet preprocessing.
- Audio is extracted and converted to MFCC features by aligning with the 30 video frames so that lip movements and speech can be compared.
- Face/Lip/Eye: 30 × 128
- Audio: 30 × 40
- Joint AV: 30 × 168

Why Oversampling - Handling Class Imbalance in Deepfake Detection

1. Balanced Under-Sampling

- Excess fake samples are discarded to create equal class distribution.
- Prevents model from exploiting class frequency bias.

Limitation:

Large amounts of fake data are never used during training

David M. Montserrat et al. (2020), DFDC dataset.

Strategy -

- Fake videos are under-sampled to a manageable subset.
- Real videos are oversampled to match class parity.
- Final training set:
  - 1250 Real (400 → 1250)
  - 1250 Fake (19,500 → 1250)
- Test Set is locked with 100 real and 400 fake videos.

2. **Oversampling the Minority Class**

- Minority class (real videos) is repeatedly sampled until both classes are equally represented.
- *Wei Ge et al. (2022)* showed that imbalance strongly affects classifier behaviour and explainability.

Limitation:

- Repeated exposure to identical real videos may cause overfitting.

Why this works

- Reduces class bias
- Retains sufficient fake diversity
- Enables stable LightGBM training
- Prevents misleading majority-class predictions

- *Montserrat, D.M., Hao, H., Yarlagadda, S.K., Baireddy, S., Shao, R., Horváth, J., Bartusiak, E., Yang, J., Guera, D., Zhu, F., & Delp, E.J. (2020). Deepfakes Detection with Automatic Face Weighting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 668–669. https://arxiv.org/abs/2004.12027*
- *Ge, W., Todisco, M., & Evans, N. (2022). Explainable Deepfake and Spoofing Detection: An Attack Analysis Using SHapley Additive exPlanations. The Speaker and Language Recognition Workshop (Odyssey 2022), pp. 70–76. https://arxiv.org/abs/2202.13693*

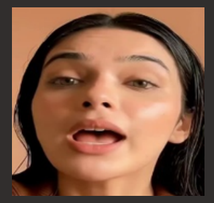
PART I - *Pre Mid Sem*

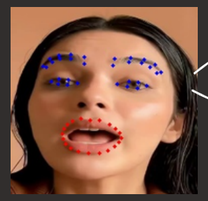
ML Methodology

FEATURE PIPELINE

FACE DETECTION

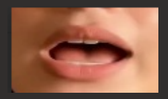
SPATIAL (FULL FACE) FEATURES 30 X 128



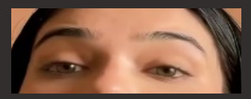


LANDMARK OVERLAY

LIP REGION FEATURES 30 x 128



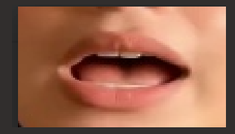
EYE REGION FEATURES 30 x 128



TEMPORAL MOTION FEATURES 1 X 29

AUDIO FEATURES 30 X 40

JOINT AV FEATURES 30 X 168



- Faces detected in each frame using **YuNet**
- Sample 30 frames
- Crop the detected region.
- **Resized to 224 × 224 pixels**

- Pretrained EfficientNetB0 extracted **a 1280D feature vector per frame.**
- A Fully Connected Bottleneck Layer reduced features: **1280D → 128D**
- For 30 frames, the final feature representation became: 30 × 128

- Computed optical flow **(pixel movement between consecutive frames.)**

- Extracted the audio using **MoviePy**
- **Computed MFCC** features using Librosa.
- **Extracted 40 MFCC coefficients (common standard)**
- Audio Features: **30 × 40**

- Concatenate lip features (128) with audio features (40) to create a joint representation
- **Joint AV Feature: 30 × 168**

molecular system is measurable one, however a present applicability of common tools for screening large compounds libraries is limited. As a simple method based on a targeted minimal set biomolecular structural features, which will be able to reveal latent details in quantitative terms about the strength of protein-peptide complex in a consistent and general manner, are still lacking. Consequently, furth development of effective protein-peptide docking techniques [4–7] and finding an efficient alternative to binding affinity [8–15] have been a major focus of computational studies in recent years. Recently, steere molecular dynamics (SMD) simulations have become popular to measure mechanical stability which could be used to assess the strength of the molecular interactions.

PART II- *Post Mid Sem*

ML Methodology

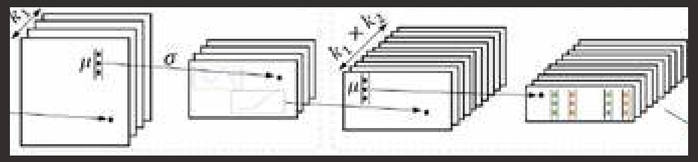
Inputs:

SPATIAL FEATURES (30 X 128)
LIP FEATURES (30 x 128)
EYE FEATURES (30 x 128)
TEMPORAL FEATURES (1X29):

- BLINK RATE
- EAR
- OPTICAL FLOW
- LIP MOTION

TEMPORAL POOLING

mean + std + max pooling

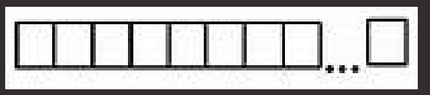


Converts sequential features into fixed-size vectors i.e. Aggregates frame-wise features across time

FIXED-LENGTH VIDEO REPRESENTATIONS

Compact multimodal representation for each video

FEATURE FUSION



- Concatenation of multimodal features
- 3065-D fused representation

ML PREPROCESSING

Scaling + Oversampling

- Standard scaling applied for feature normalization
- Oversampling balanced real and fake training samples
- PCA / UMAP used for feature-space visualization

LightGBM CLASSIFIER

- Builds multiple decision trees sequentially
- Each new tree learns from previous prediction errors
- Combines weak learners into a stronger ensemble
- Learns nonlinear relationships between multimodal features
- Outputs probability scores for REAL / FAKE classification
- SHAP explains which features influenced the prediction

REAL OUTPUT

```
EXTREME_REAL
Probability fake: 0.0001
```

FAKE OUTPUT

```
EXTREME_FAKE
Probability fake: 1.0000
```

SHAP EXPLAINABILITY

```
Predicted label: FAKE
Probability fake: 0.8761
Explanation      : Likely FAKE due to lip movements
not matching audio, irregular audio patterns
```

Leaf-wise tree growth

LightGBM always splits the leaf with the highest error reduction, reaching the same accuracy with fewer splits than XGBoost's level-wise growth.
-Faster training on large feature vectors
Ke et al., NeurIPS 2017

Histogram-based algorithm

Stores binned feature values rather than exact values, critical for our high-dimensional features (EAR, MFCC, optical flow, lip delta).
Ke et al., NeurIPS 2017

Native SHAP support

LightGBM has built-in SHAP integration. Directly enables our XAI layer to explain which features (e.g. EAR spike, MFCC artefact) flag a video as fake.
Ge et al., 2022 — SHAP on spoofing detection

Native SHAP support

Boosting handles class skew
Our dataset: 500 real vs 19,500 fake. Boosting-based methods consistently outperform Random Forest on PR-AUC and G-mean on imbalanced data.
ResearchGate ensemble comparison, 2024

Why *LightGBM* & *SHAP*?

Paper #1 : SHAP for Deepfake/Spoofing Detection

Paper #2 : Lightweight Deepfake Detection with ML Classifiers

Paper #3 : LightGBM vs XGBoost vs RF on Imbalanced Data

1) Ge, W., Todisco, M., & Evans, N. (2022). Explainable Deepfake and Spoofing Detection: An Attack Analysis Using SHapley Additive exPlanations. The Speaker and Language Recognition Workshop (Odyssey 2022), pp. 70–76. DOI: 10.21437/Odyssey.2022-10
2) ResearchGate (2023/2024). Random Forest vs. XGBoost vs. LightGBM: Which Ensemble Learner Performs Best on Imbalanced Data?
3) Yasir, S.M., & Kim, H. (2025). Lightweight Deepfake Detection Based on Multi-Feature Fusion. Applied Sciences, 15(4), 1954. DOI: 10.3390/app15041954

Comparison of Models

Model	Pros	Cons	Final Decision
Logistic Regression	Highest benchmark accuracy and AUC on subset evaluation; extremely fast training and inference; lightweight model size	Linear decision boundary may not capture complex multimodal behavioural interactions; may rely on dataset-specific separability	Not selected due to limited nonlinear modelling capacity for large-scale behavioural deepfake detection
Random Forest	Strong ensemble performance; robust to noise; captures nonlinear relationships better than linear models	Larger ensemble complexity; slower inference; less efficient scaling for high-dimensional fused embeddings	Not selected due to computational overhead and lower scalability
XGBoost	Powerful nonlinear boosting framework; strong feature interaction modelling	Highest training time and computational cost; heavier optimization requirements	Not selected due to significantly higher computational expense compared to LightGBM
LightGBM	Efficient gradient boosting; strong nonlinear learning; scalable for high-dimensional multimodal features; compatible with SHAP explainability; balanced performance-efficiency tradeoff	Slightly lower benchmark accuracy than Logistic Regression on the evaluated subset	**Selected as final model due to best balance between scalability, efficiency, interpretability, and nonlinear multimodal behavioural learning**

MODEL COMPARISON RESULTS WITH SPEED AND SIZE

Model	Acc	BalAcc	F1	AUC	Train(s)	Infer(s)	Size(MB)
Logistic Regression	0.9920	0.9950	0.9949	0.9998	0.1923	0.0055	0.0241
Random Forest	0.9779	0.9637	0.9862	0.9936	3.3431	0.0369	0.9607
XGBoost	0.9517	0.9473	0.9693	0.9905	36.5279	0.0123	0.2782
LightGBM	0.9658	0.9599	0.9784	0.9954	14.6509	0.0228	0.4374

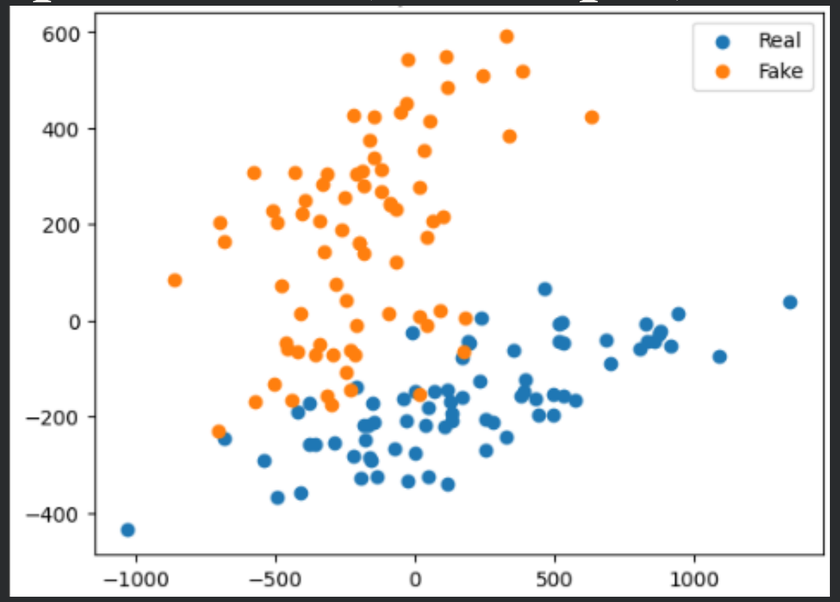
1) Ge, W., Todisco, M., & Evans, N. (2022). Explainable Deepfake and Spoofing Detection: An Attack Analysis Using SHapley Additive exPlanations. The Speaker and Language Recognition Workshop (Odyssey 2022), pp. 70–76. DOI: 10.21437/Odyssey.2022-10

2) ResearchGate (2023/2024). Random Forest vs. XGBoost vs. LightGBM: Which Ensemble Learner Performs Best on Imbalanced Data?

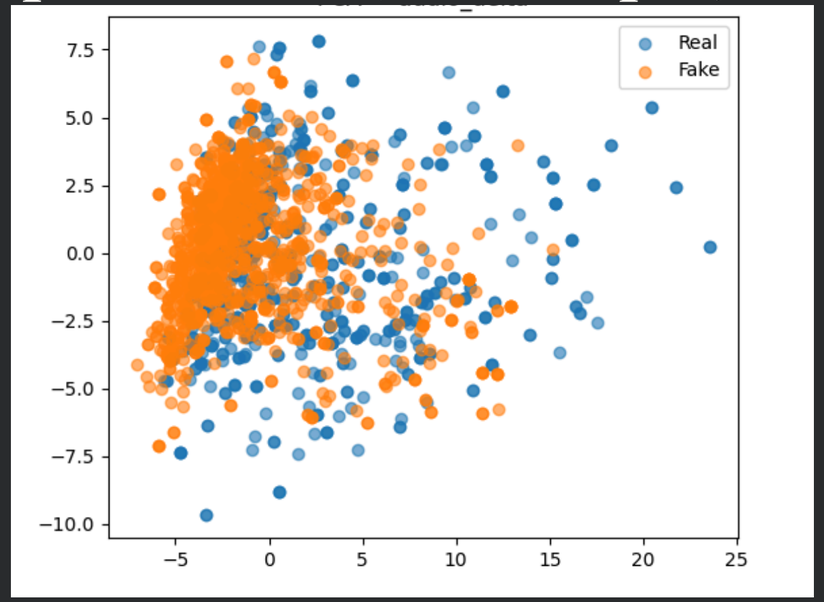
3) Yasir, S.M., & Kim, H. (2025). Lightweight Deepfake Detection Based on Multi-Feature Fusion. Applied Sciences, 15(4), 1954. DOI: 10.3390/app15041954

PCA Analysis

pre mid sem (150 samples)



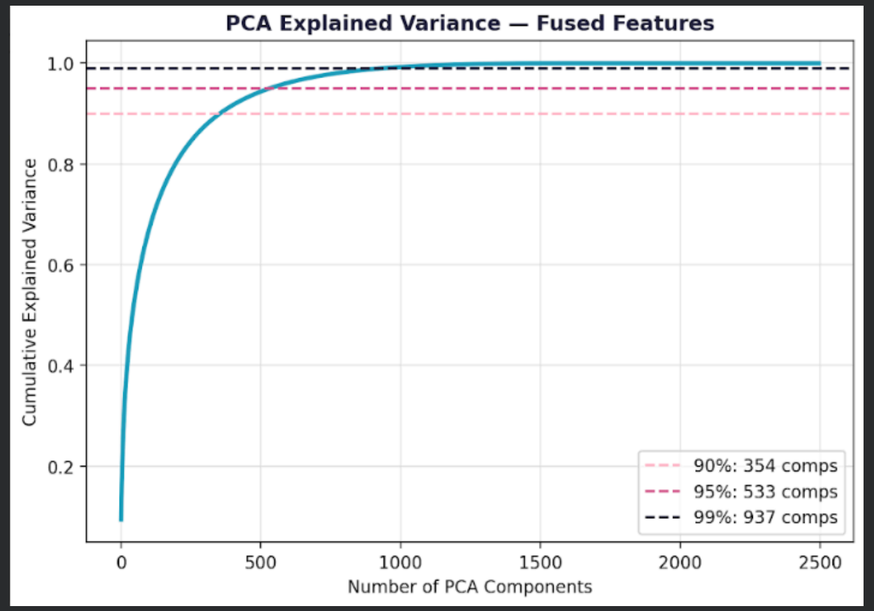
post mid sem (2500 samples)



In the initial mid-sem experiment using 150 samples (75 real and 75 fake), the first two principal components explained approximately 40.5% of the variance (PC1: 30.5%, PC2: 9.99%), showing comparatively clearer clustering and partial separation between real and fake samples.

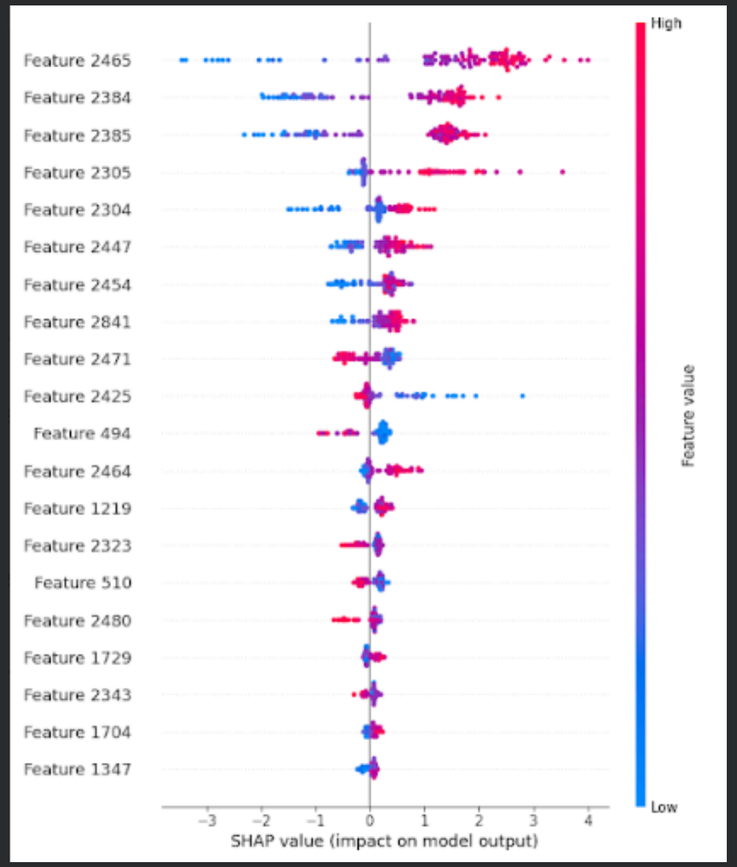
After scaling the pipeline to a larger balanced dataset of approximately 2500 samples (1250 real and 1250 fake), the first two principal components explained 24.2% of the variance (PC1: 17.9%, PC2: 6.3%). The larger dataset introduced significantly greater behavioural variability and feature overlap, producing a more realistic and complex multimodal feature space.

Overall, the increased overlap in the larger dataset suggested that behavioural deepfake detection is not linearly separable at scale, motivating the use of nonlinear ensemble models such as LightGBM for learning complex multimodal interactions.



Moreeee Analysis

- The cumulative explained variance increased gradually across PCA components, indicating that information was distributed across many multimodal behavioural features.
- Approximately 354, 533, and 937 components were required to explain 90%, 95%, and 99% of the variance respectively.
- This suggested that deepfake detection relied on multiple weak behavioural cues rather than a few dominant features.



• SHAP analysis identified the most influential features contributing to REAL / FAKE predictions.

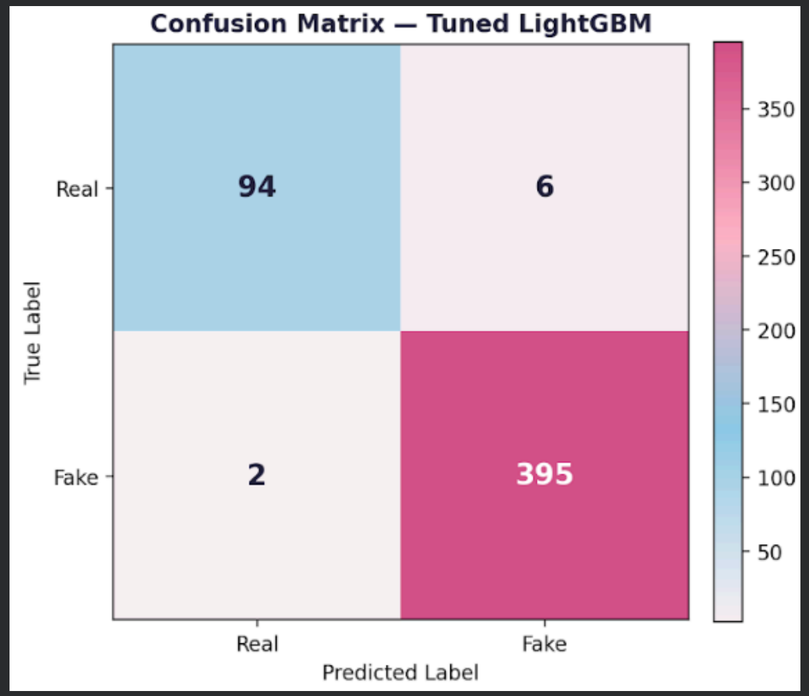
• Audio, delta-MFCC, and joint audio-visual features showed strong impact on model decisions.

• The model relied heavily on behavioural inconsistencies such as lip-sync mismatch, temporal instability, and irregular audio dynamics.


- Should we review the Performance Metrics?
- Sure

Performance Metrics

Accuracy Precision Recall F1-Score AUC-ROC

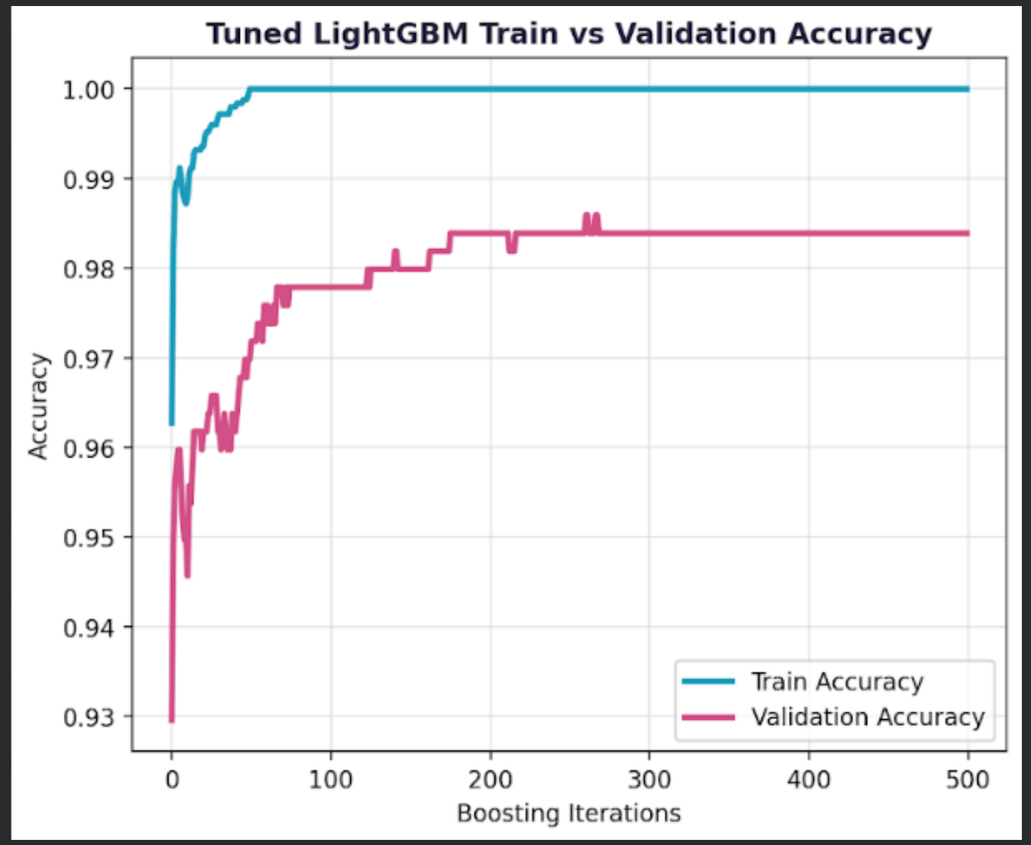



```
FINAL TUNED LIGHTGBM RESULTS
============================================================
              precision    recall  f1-score   support

        Real       0.98      0.94      0.96       100
        Fake       0.99      0.99      0.99       397

    accuracy                           0.98       497
   macro avg       0.98      0.97      0.97       497
weighted avg       0.98      0.98      0.98       497

Accuracy      : 0.9839
AUC-ROC       : 0.9994
Balanced Acc  : 0.9675
F1-score      : 0.9900
```

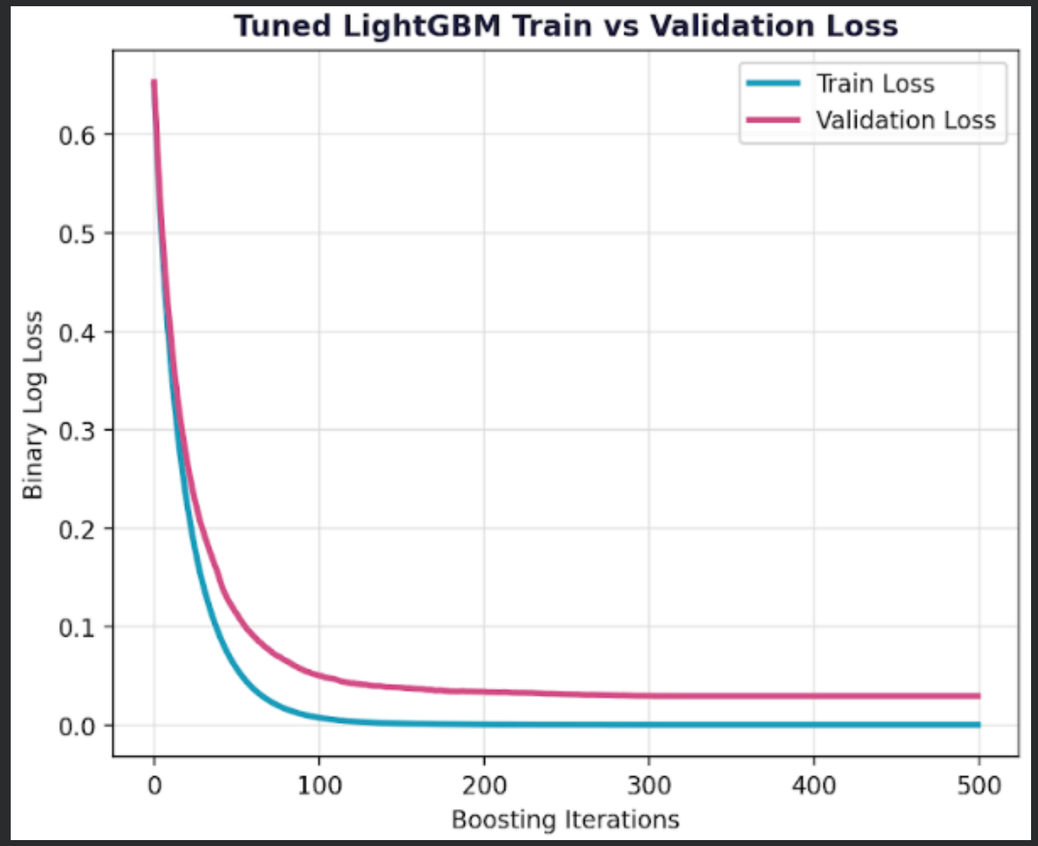
- The tuned LightGBM classifier correctly identified 395 fake videos and 94 real videos.
- Only 8 total misclassifications were observed across 497 test samples.
- The strong diagonal concentration indicated robust classification performance across both classes.

- The final tuned LightGBM model achieved 98.39% accuracy and 0.9994 AUC-ROC on the test set.
- Balanced accuracy of 96.75% demonstrated strong performance despite class imbalance.
- High precision, recall, and F1-score confirmed reliable multimodal behavioural deepfake detection.

Performance Metrics



- Training and validation accuracy increased steadily across boosting iterations before stabilizing near convergence.
- Validation accuracy remained consistently high (~98%), indicating strong generalization performance.
- The small gap between training and validation accuracy suggested limited overfitting after regularization and tuning.



- Both training and validation loss decreased rapidly during early boosting iterations before converging.
- Validation loss stabilized without significant divergence from training loss, indicating stable learning behaviour.
- The smooth convergence pattern demonstrated effective optimization and controlled model complexity.

Challenges faced & Strategies

Software Challenges
- TensorFlow, MediaPipe, and protobuf version conflicts
- Environment instability in Google Colab
- Dependency mismatches across libraries

Solution:
- Version-controlled package installation
- Runtime resets and isolated environments
- Compatibility-based dependency management

Hardware Challenges
- Runtime disconnections during large-scale feature extraction
- High computational cost for multimodal processing
- GPU memory limitations while handling long video sequences

Solution:
- Batch-wise checkpoint saving
- Sequential feature extraction pipeline
- Reduced memory overhead using pooled embeddings

Dataeset Challenges
- Severe class imbalance between real and fake videos
- Failed face detections in low-quality samples
- Variable lighting, resolution, and facial visibility

Solution:
- Oversampling after feature extraction
- Invalid-video filtering
- Conservative face cropping and dense frame sampling

Algorithmic challenges
- High-dimensional multimodal feature fusion
- Partial overlap between real and fake feature spaces
- Risk of overfitting on benchmark datasets

Solution:
- Temporal pooling for compact representations
- LightGBM-based nonlinear ensemble learning
- SHAP explainability and validation-based monitoring

Deployability at Plaksha

Possible Deployment

The system can be deployed as:

- A Cybersecurity Club demonstration tool
- A browser/web-based fake video checker
- A moderation assistant for student media uploads
- A misinformation awareness platform on campus

Example Workflow

Upload Video → Analyze Face + Audio → Detect Fake → Generate SHAP Explanation

Practical Use Cases

- Detect manipulated student media
- Raise awareness about AI-generated misinformation
- Demonstrate explainable AI in cybersecurity applications

Challenges During Scaling

- Video + audio processing is computationally expensive
- EfficientNetB0 and temporal features increase GPU usage
- Large video datasets require high storage and bandwidth
- Real-time detection may need cloud or optimized hardware
- Rapidly evolving deepfakes require frequent retraining
- Scalable deployment may require cloud GPUs, distributed storage, and secure upload pipelines

Future Scope

- Integration with Plaksha Cybersecurity Club initiatives
- Real-time deepfake monitoring dashboard
- Explainable AI reports for non-technical users
- Multi-agent reasoning systems for advanced forensic analysis

Thank You!

Any Questions?