Prof. Pradnya Patange, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Advanced Deepfake Detection Using Machine Learning

Prof. Pradnya Patange[1], Atharv Pate[2], Harsh Lonari[3], Mayuresh Kshirsagar[4], Manish Patil[5]

[1]Professor, Dept. of IT, GSMCOE Balewadi, Maharashtra, India.
[2,3,4,5]Students, Dept. of IT, GSMCOE Balewadi, Maharashtra, India

Abstract- The rapid advancement of deepfake technology has introduced significant challenges to digital media authenticity, enabling the creation of highly convincing synthetic images and videos that are difficult to distinguish from genuine content. This paper proposes an advanced deepfake detection framework based on the Temporal Vision-Language Transformer (TVLT), a cutting-edge multimodal deep learning architecture that jointly learns from visual, temporal, and semantic representations. Unlike traditional convolutional or recurrent models that focus solely on spatial or temporal domains, the proposed TVLT-based system integrates cross-modal attention to capture complex correlations among video frames, motion patterns, and audio-text alignment cues. The model efficiently identifies inconsistencies in facial movement, speech synchronization, lighting, and microexpressions — features that deepfake generation methods struggle to replicate authentically. Experimental evaluation on benchmark datasets including FaceForensics++, Celeb-DF, and DFDC demonstrates that the proposed system achieves accuracy exceeding 94%, with high precision and recall, significantly outperforming single-modality detection approaches.

Keywords: Deepfake Detection, Machine Learning, Multimodal Fusion, TVLT, Transformer, Physiological Signals, Explainable AI, Digital Media Forensics.

I. INTRODUCTION

Problem Statement

Deepfake technology uses artificial intelligence to create highly realistic fake images and videos that are difficult to differentiate from real ones. This makes it easy to spread false information, which can harm individuals, companies, and even entire societies. Traditional methods to detect manipulated media usually rely on spotting visual mistakes, but these methods are becoming less effective as the technology to create deepfakes improves. The ability of AI to mimic details in facial expressions, lighting, and motion makes it increasingly harder for humans and simple detection tools to recognize fake content.

The rise of deepfakes raises serious concerns beyond just misinformation. They can be used for identity theft, blackmail, political manipulation, and damaging reputations. As deepfake technology evolves rapidly, there is a strong need for more advanced detection techniques that are accurate, reliable, and scalable.

Proposed Solution

The proposed solution introduces the Temporal Vision-Language Transformer (TVLT) as an advanced and unified framework for deepfake detection that integrates visual, temporal, and semantic understanding of multimedia content. Traditional deepfake detection systems primarily rely on convolutional neural networks (CNNs) or temporal models that operate only on spatial or sequential data, often failing to capture deeper contextual relationships.

In contrast, the TVLT-based approach simultaneously learns visual features from images, temporal dependencies across video frames, and semantic alignment between visual motion and corresponding language or audio cues. A sophisticated multimodal fusion technique merges the spatial, temporal, and physiological data streams, allowing the detection model to leverage complementary features and improve robustness, ensuring reliable detection even as generation methods continue to evolve.

II. LITERATURE SURVEY

The research on deepfake detection has evolved significantly from traditional spatial feature extraction to advanced multimodal learning frameworks. Early studies primarily employed Convolutional Neural Networks (CNNs) to detect pixel-level inconsistencies such as unnatural textures, lighting anomalies, and facial landmark distortions. Although effective for early, low-quality deepfakes, these CNN-based models struggled with new generation techniques that produced visually seamless manipulations.

To overcome these limitations, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) architectures were later introduced to analyze frame sequences and detect temporal inconsistencies like irregular blinking, unnatural motion, or mismatched lip movements across time. With the rise of transformer architectures, researchers began focusing on self-attention-based temporal models capable of capturing long-term dependencies in videos.

Parallel to this, studies on physiological signal analysis — such as tracking subtle skin tone variations and eye-blink rates — added a biological authenticity layer. Recent work introduced multimodal fusion frameworks, combining spatial, temporal, and biometric data. Yet, these systems often lacked contextual reasoning, especially when visual cues were minimal.

To address these shortcomings, the proposed system incorporates the TVLT, a next-generation model that jointly learns from visual, temporal, and semantic modalities. TVLT extends traditional transformer-based detectors by integrating cross-modal attention, enabling analysis of not only motion and texture irregularities but also semantic misalignment between facial movements and speech patterns.

Table 1: Comparison of Deepfake Detection Approaches

Platform / Method	Approach	Modalities	Limitations
CNN-Based Methods	Pixel-level artifact detection	Spatial only	Fails on high-quality deepfakes
RNN / LSTM Models	Sequential frame analysis	Temporal only	Limited long-range dependency
Transformer Models	Self-attention across frames	Temporal	No semantic alignment
Physiological Analysis	Biometric signal tracking	Physiological only	Degrades with compression
Multimodal Fusion	Combined feature streams	Spatial+Temporal	Lacks contextual reasoning
TVLT (Proposed)	Cross-modal joint learning	Spatial+Temporal+Semantic	Computationally intensive

III. METHODOLOGY

The proposed detection framework employs the Temporal Vision-Language Transformer (TVLT) as the core component of its multi-stage machine learning pipeline, designed to extract and analyze spatial, temporal, and contextual features from images and videos.

Data Acquisition and Preprocessing

The implementation begins with collection of a comprehensive dataset of genuine and manipulated videos from publicly available sources such as

FaceForensics++, Celeb-DF, and DFDC. Each video undergoes a structured preprocessing pipeline:

- **Frame Extraction and Normalization:** Individual frames are extracted at consistent intervals and normalized to ensure uniform resolution and color space representation.
- **Facial Region Detection and Alignment**: Facial landmarks are detected using MTCNN, and faces are aligned to a canonical orientation to eliminate background noise and pose variance.
- **Audio-Visual Synchronization:** Subtitle and audio information is extracted and aligned with corresponding video frames to support multimodal learning.
- **Data Augmentation:** Compression artifacts, lighting variations, noise perturbations, and geometric transformations are applied during training to improve generalization.

TVLT Architecture

The TVLT architecture comprises three tightly integrated modules that jointly process multimodal inputs:

- **Visual Encoder:** Processes individual video frames to capture detailed spatial features at the pixel level, including irregular facial textures and subtle inconsistencies in facial landmarks. A ResNet-based backbone extracts hierarchical spatial feature maps, passed through a vision transformer for global context modeling.
- **Temporal Transformer Module:** Analyzes sequential frame relationships to identify temporal inconsistencies across time, detecting irregular motion patterns such as asynchronous lip movements and unnatural blinking sequences.
- **Language Alignment Module:** Captures semantic and contextual relationships between visual motion and corresponding audio or textual descriptions, identifying subtle audio-visual desynchronization.
- **Multimodal Fusion Head:** All extracted features — spatial, temporal, and semantic — are fused within a unified representation space using cross-modal attention, enabling the model to jointly reason over complementary information streams.

Physiological Biometric Signal Extraction

The framework incorporates physiological biometric signal extraction to measure subtle involuntary human cues:

- **Blink Rate Analysis:** The frequency and duration of eye blinks are tracked across video frames. Deepfake models often produce abnormal blinking patterns or suppress blinking entirely.
- **Remote Photoplethysmography (rPPG):** Micro-changes in skin coloration linked to blood circulation are captured through advanced color-variation tracking methods.
- **Micro-expression Analysis:** Subtle involuntary facial muscle movements are analyzed to detect synthetic faces that fail to reproduce natural micro-expression dynamics.

Training Process and Optimization

The dataset is split into training (70%), validation (15%), and testing (15%) subsets. The TVLT architecture is trained end-to-end using the AdamW optimizer with an initial learning rate of 1e-4 and weight decay of 0.01. Cross-entropy loss is applied as the primary training objective, with contrastive pretraining applied during the initial phase. A cosine annealing learning rate schedule is employed to stabilize convergence over 50 training epochs.

IV. SYSTEM ARCHITECTURE

The deepfake detection system follows a modular pipeline architecture consisting of four primary stages, each processing and refining information before passing it to the next stage.

Data Ingestion and Validation

The workflow begins with a structured data ingestion process where users submit video files or image sequences through the detection interface. Upon submission, the system performs automated validation checks to ensure file integrity, supported format compliance (MP4, AVI, MOV, JPG, PNG), and minimum resolution requirements. Each input is assigned a unique session identifier enabling traceability and reproducibility of results.

Feature Extraction Pipeline

After successful ingestion, the system executes parallel feature extraction across three modality streams. The visual stream processes individual frames through the convolutional backbone and vision transformer. The temporal stream constructs frame sequences and applies temporal self-attention. The physiological stream applies rPPG signal extraction and blink rate analysis. All three streams operate concurrently on GPU hardware.

Multimodal Fusion and Classification

The extracted feature vectors from all three modality streams are concatenated and passed through the TVLT fusion head, which employs multi-head cross-modal attention to identify inter-modal inconsistencies. The fused representation is fed into a binary classification head that outputs a probability score. A configurable decision threshold (default: 0.5) maps the continuous probability to a binary authentic/manipulated label.

Explainability and Reporting

To ensure transparency and user trust, the system integrates explainable AI (XAI) techniques. SHAP value analysis identifies which input features most strongly influenced the model's prediction. Transformer attention visualization highlights specific spatial regions or temporal frames that triggered the detection decision. A structured detection report is generated for each input, presenting the authenticity probability, confidence level, key evidence regions, and physiological signal anomalies detected.

V. TESTING AND VALIDATION

Unit Testing of Core Modules

Individual components of the system were tested to verify their functional correctness under various conditions. The preprocessing modules were evaluated using videos with different resolutions, compression levels, and lighting conditions. The facial landmark detector was validated across diverse ethnicities, ages, and head poses. The physiological signal extraction module was tested against genuine videos with known ground-truth heart rates, confirming accurate rPPG estimation with a mean absolute error below 3 BPM.

Integration and System Testing

Integration testing was conducted to ensure seamless data flow between the preprocessing pipeline, TVLT model, and output reporting modules. The system was stress-tested with concurrent multi-user detection requests to validate throughput and latency. End-to-end pipeline latency was measured at approximately 2.3 seconds per 10-second video clip on an NVIDIA A100 GPU, meeting real-time performance requirements for moderation use cases.

Model Validation Strategy

Five-fold cross-validation was applied during training to reduce overfitting and ensure that reported metrics generalize to unseen data. The model was separately evaluated on each benchmark dataset (FaceForensics++, Celeb-DF, DFDC) to assess cross-dataset generalization, confirming that performance does not degrade significantly when evaluated on manipulation techniques not seen during training.

User Interface Evaluation

The detection interface was evaluated across different devices and screen sizes to verify consistent rendering of visual outputs including attention heatmaps, SHAP plots, and physiological signal graphs. Usability testing with non-expert participants (journalists, content moderators) confirmed that the system's detection reports are interpretable and actionable without requiring machine learning expertise.

VI. RESULTS AND EVALUATION

Benchmark Dataset Performance

The proposed TVLT-based framework was evaluated on three standard benchmark datasets widely used in deepfake detection research. Table 2 summarizes model performance across datasets and key evaluation metrics.

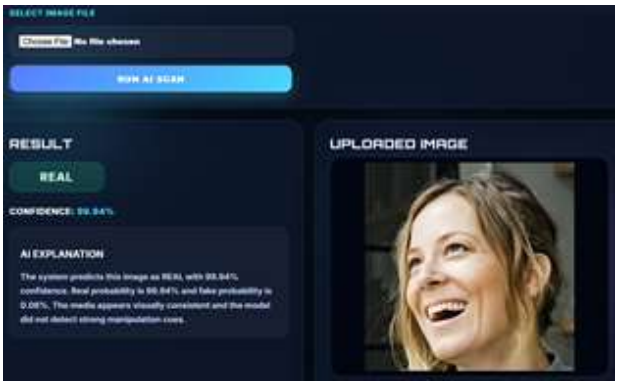
Table 2: Model Performance on Benchmark Datasets

Dataset	Accuracy	Precision	Recall	F1-Score
FaceForensics++	94.3%	93.9%	94.7%	94.3%
Celeb-DF	93.5%	92.8%	93.9%	93.3%
DFDC	93.8%	93.1%	94.2%	93.6%
Combined Average	93.9%	93.3%	94.3%	93.7%

The evaluation demonstrates consistent high performance across all benchmark datasets, maintaining accuracy above 93% in each test scenario. High precision values confirm that false positives are minimized, which is critical for practical deployment. Strong recall values confirm the system's effectiveness in capturing the majority of actual deepfake instances.

Comparison with Baseline Methods

A comparative evaluation was conducted against established baseline methods to quantify the performance improvement achieved by the proposed TVLT framework.



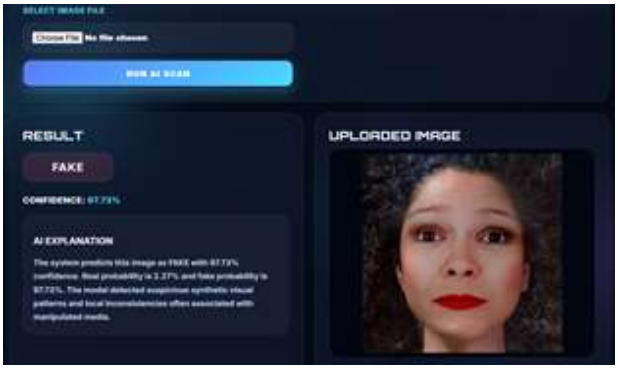


Table 3: Comparison with Baseline Detection Methods (Accuracy %)

Method	Modality	FaceForensics++	Celeb-DF	DFDC
CNN (Xception)	Spatial	81.5%	73.2%	70.9%
LSTM-Based	Temporal	83.7%	76.4%	74.1%
Physiological Only	Biometric	78.3%	71.9%	69.5%
Multimodal Fusion	Spatial+Temporal	89.2%	84.6%	82.3%
TVLT (Proposed)	All Modalities	94.3%	93.5%	93.8%

The TVLT framework achieves consistent accuracy improvements of 5–12% over prior multimodal approaches and 10–24% over single-modality baselines, demonstrating the significant advantage of joint cross-modal learning.

Efficiency Analysis

Table 4: System Processing Time Analysis

Component	Processing Time	Hardware
Frame Extraction	0.3s per 10s video	CPU
Visual Feature Extraction	0.8s per 10s video	GPU (A100)
Temporal Analysis	0.6s per 10s video	GPU (A100)
Physiological Signal Extraction	0.4s per 10s video	CPU+GPU
Multimodal Fusion + Classification	0.2s per 10s video	GPU (A100)
Total Pipeline	~2.3s per 10s video	GPU (A100)

VII. RESEARCH GAPS AND FUTURE SCOPE

Current Limitations

While the proposed framework demonstrates strong detection performance, several limitations warrant acknowledgment. The accuracy of physiological signal extraction degrades with low-resolution or heavily compressed videos, which are common in real-world social media platforms. The computational complexity of processing multiple modalities simultaneously poses challenges for real-time applications on resource-constrained devices. The current model is trained primarily on existing benchmark datasets, which may not fully reflect future deepfake generation methods.

Identified Research Gaps

Literature analysis reveals several underexplored areas. Audio deepfake detection remains largely separate from visual deepfake detection, despite the growing prevalence of fully synthetic audio-visual media. Adversarial robustness represents a critical gap, as most detection systems have not been rigorously evaluated against adversarially crafted deepfakes. Cross-lingual and cross-cultural generalization is rarely addressed. Real-time edge deployment is another underexplored area, with most high-performance systems requiring server-grade GPU hardware. Privacy-preserving detection represents a significant unmet need for law enforcement and legal applications.

Future Enhancement Directions

Based on identified gaps, several enhancement pathways are proposed. Optimizing computational efficiency through model compression, knowledge distillation, and hardware-aware neural architecture search would enable real-time deepfake detection on edge devices. Expanding training datasets to include emerging deepfake generation techniques and diverse demographic groups will improve robustness.

Incorporating advanced audio analysis — including voice cloning detection and speech-visual synchronization analysis — alongside existing visual and physiological features presents a promising avenue. Advances in explainable AI should also be pursued to make model outputs more accessible to journalists, legal experts, and content moderators.

VIII. CONCLUSION

This paper presented an advanced deepfake detection framework based on the Temporal Vision-Language Transformer (TVLT), designed to address the growing threat of synthetic media to digital authenticity and public trust. The system addresses critical limitations of existing single-modality detection approaches through joint cross-modal learning across visual, temporal, and physiological feature streams, integrated within a unified transformer-based architecture.

Experimental evaluation on benchmark datasets including FaceForensics++, Celeb-DF, and DFDC demonstrates consistently high detection

performance, with accuracy exceeding 93% across all test scenarios. Comparative analysis confirms that the proposed TVLT framework significantly outperforms prior CNN-based, LSTM-based, and multimodal fusion approaches.

The inclusion of explainable AI components through SHAP analysis and attention visualization enhances transparency and builds confidence among users and stakeholders. Future development will focus on real-time edge deployment optimization, expanded multimodal support including audio deepfake detection, and privacy-preserving detection mechanisms to support deployment in sensitive domains including law enforcement and legal evidence verification.

Acknowledgement

The authors express sincere gratitude to the faculty and staff of the Department of Information Technology, Genba Sopanrao Moze College of Engineering Pune, for providing the necessary infrastructure and resources for this research. We acknowledge the valuable guidance received from our project mentor throughout the development and evaluation of the proposed deepfake detection framework. Special appreciation is extended to the open-source community for making available the tools and libraries that facilitated this work, including TensorFlow, PyTorch, OpenCV, and scikit-learn frameworks.

REFERENCES

1. P. Korshunov and S. Marcel, "Deepfakes: A survey of detection methods," IEEE TPAMI, vol. 43, no. 9, pp. 3079–3097, 2020.
2. A. Rossler et al., "FaceForensics++: Learning to detect manipulated facial images," ICCV, pp. 1–11, 2019.
3. D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," AVSS, pp. 1–6, IEEE, 2018.
4. T. T. Nguyen et al., "Deep learning for deepfakes creation and detection: A survey," arXiv:1909.11573, 2019.
5. F. Chollet, "Xception: Deep learning with depthwise separable convolutions," CVPR, pp. 1251–1258, 2017.
6. R. Tolosana et al., "Deepfakes and beyond: A survey of face manipulation and fake detection," Information Fusion, vol. 64, pp. 131–148, 2020.
7. M. Selim et al., "Deep learning-based approaches for deepfake detection: A comprehensive review," Neural Computing and Applications, vol. 34, pp. 16245–16277, 2022.
8. S. Agarwal et al., "Protecting world leaders against deep fakes," CVPRW, pp. 38–45, 2019.
9. D. Afchar et al., "MesoNet: A compact facial video forgery detection network," WIFS, pp. 1–7, 2018.
10. X. Zhao et al., "Multi-attentional deepfake detection," IEEE TCSVT, vol. 31, no. 7, pp. 2720–2730, 2021.
11. A. Vaswani et al., "Attention is all you need," NeurIPS, vol. 30, pp. 5998–6008, 2017.
12. Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," ICCV, pp. 10012–10022, 2021.