Journal of Soft Computing Paradigm (ISSN: 2582-2640)
www.irojournals.com/jscp/

Multi-Modal Deepfake Detection via Spatial, Temporal, and Audio-Visual Fusion with Vision Transformers

Merlin Gethsy D.[1], Sharmila V[2].

[1]Assistant Professor, [2]PG Student, Department of Computer Science and Engineering, V V College of Engineering, Tisaiyanvilai, Thoothukudi, India.

E-mail: [1]merlin@vvcoe.org, [2]sharmilasharmila9498@gmail.com

Abstract

The rapid advancement of the deepfake generation technologies has intensified concerns related to digital misinformation, identity impersonation, and media manipulation. Although numerous deepfake detection methods have been developed by mitigate these threats, most rely on a single modality and exhibit limited robustness when confronted with diverse manipulation techniques and cross-dataset scenarios. To overcome these deficiencies, we propose VeriSphere, a multimodal deepfake detection framework that combines spatial, temporal, and audiovisual forensics in one system. It uses a Vision Transformer for detecting spatial artifacts, an X-CLIP-based module for capturing temporality, and an AV synchronization module to examine whether speech aligns with lip movements. The outputs are then fused using a weighted strategy to produce a single trust score for prediction. Results show that VeriSphere achieves a high accuracy of 92.1%, an AUC of 0.963, and an F1-score of 0.924 across three benchmark datasets: FaceForensics++, Celeb-DF, and DFDC.

Keywords: Deepfake Detection, Vision Transformer, Multi-Modal Fusion, Temporal Consistency Analysis, Audio-Visual Synchronization, Digital Media Forensics.

DOI: https://doi.org/10.36548/jscp.2026.3.003
Extended version of article received from **International Conference on Next Generation Engineering Innovations, Al and Sustainable Automation - 2026**
Received: 23.04.2026, received in revised form: 29.05.2026, accepted: 18.06.2026, published: 30.06.2026

1. Introduction

Recent advances in deepfake generation, driven by generative adversarial networks (GANs), diffusion models and neural face synthesis, have enabled the creation of highly realistic synthetic videos that are increasingly difficult to distinguish from authentic content. This threatens the veracity of online media with grave consequences including misinformation, identity theft, deep fakes, and the erosion of public trust in digital communication. Furthermore, the rapid spread of fake content over social and other digital platforms makes it crucial that we have tools that can quickly separate fake from real in order to inform the public and prevent the spread of false news. Despite some progress in recent years, the constantly shifting nature of deepfake generation continues to be the bane of researchers working on a one size fits all detection system.

Most of the current deepfake detection strategies are based on a single modality, such as spatial artifacts, temporal consistency and audio-visual sync [7,8]. Spatial methods focus on visual inconsistencies in single frames, while temporal ones-based continuity motion through video sequences. The audio-visual techniques also assess the association between speech signals and signals corresponding to facial movements. While these approaches have shown good performance under certain conditions, their effectiveness usually deteriorates in the presence of unseen manipulation techniques, compressed videos or cross-domain datasets. Therefore, a holistic detection framework that simultaneously leverage complementary information of multi-modality is needed to enhance the robustness, generalization ability and detection reliability under different deepfake scenarios.

To address these limitations, a multi-modal deepfake detection framework are integrating spatial, temporal, and audio-visual analysis is proposed. The framework uses a Vision Transformer (ViT) to detect spatial manipulation artifacts, combined with an X-CLIP encoder and temporal consistency analysis to obtain frame-level anomalies and a Wav2Vec2–MediaPipe synchronization module to assess the correspondence between speech and lip movements. The outputs of these independent detection branches are fused through a weighted fusion strategy to yield one single trust score and final classification decision. The proposed framework utilizes complementary cues from visual appearance, temporal behaviour and audio-visual correspondence to enable accurate detection, broaden cross-dataset generalization capabilities, and provide a credible yet interpretable solution for on-line deepfake video verification.

2. Related Works

The rapid advancement of deep learning and generative artificial intelligence led to a significant increase in the quality of synthetic media synthesis. The emergence of Generative Adversarial Networks (GANs) allowed synthesizing photorealistic images and video sequences, creating the basis for deepfake technology [13]. With deepfake generation techniques becoming more advanced, the issues related to disinformation, fraud, identity theft, and media manipulation become a cause for concern. A number of comprehensive studies noted that it has become harder to identify manipulated media and called for the development of new forensics systems able to deal with advancing methods of media manipulation [12], [14]. In order to develop such systems, several benchmark datasets were created, including FaceForensics++, Celeb-DF, and DeepFake Detection Challenge (DFDC) [1], [7], [8].

The early methods used for deepfake detection focused mainly on the detection of spatial artifacts in individual image frames. CNN-based architectures proved to be highly successful by detecting discrepancies in face manipulations and image generation processes. Depthwise separable convolution-based architectures showed remarkable results in face manipulation detection by learning discriminative features of visual artifacts [10]. Attention-based techniques also contributed towards achieving good results for deepfake detection through better feature localization and recognition [15], [11]. Recently, transformer architectures have proven to be highly capable of learning long-range dependencies and global context information. ViT architectures presented a novel patch-based learning technique which enabled highly efficient image understanding and image classification tasks [9]. By applying the principles of transformers in multi-modal and multi-scale transformer architectures, more advanced capabilities for deepfake detection have been achieved [2].

Spatial analysis can produce useful forensic information, frame-based analysis falls short of uncovering inconsistencies in video sequences because inconsistencies are detected at the frame level and not on video sequences. Temporal modeling methods have proven to be a great remedy to detecting motion inconsistencies and synthetic patterns in time. Methods using biological signals showed that in manipulated videos, there are typically abnormal blinkings and inconsistencies in physiological features that cannot be easily replicated [3]. Advances in transformer-based video representation learning have enabled temporal models to effectively analyze motion consistency and cross-frame relationships [4]. Even though such methods can

detect complex manipulations, temporal methods alone may fall short where visual anomalies are absent or realistic motions are produced.

Recent research has also explored audio-visual synchronization as another forensic clue in the detection of deepfakes. Lip synchronization analyzes the correlation between speech and facial motion and hence detects any inconsistencies caused by video editing [5]. The self-supervised speech representation models have also improved the audio feature extraction process by learning strong contextual representations from the speech signals [6]. Although there is an excellent performance of the spatial, temporal and audio-visual approaches independently, most current works concentrate on a single modality, which restricts their capacity for generalization and adaptation to new unseen data. In addition, more advanced deepfake generation processes continue to remove artifacts in each individual modality, making the unimodal detectors susceptible to failure. As a result, there is a need for a multimodal framework that makes use of both the spatial artifacts, temporal consistency and audio-visual synchronization for improved deepfake detection. In view of this challenge, the proposed VeriSphere framework uses vision transformers for spatial analysis, X-CLIP for temporal modeling and wav2vec2 for audio-visual synchronization in deepfake detection.

3. Proposed Methodology

The proposed VeriSphere framework offers a robust and interpretable approach to deepfake video detection by integrating spatial, temporal, and audio-visual forensic evidence. Most of the current deepfake detection approaches are modality-specific thus limited to isolated domains that can be easily masked using advanced techniques. In order to overcome this limitation, the proposed framework can jointly analyze visual appearance, motion consistency and speech–lip synchronization to extract complementary forensic cues from multiple modalities at once. Specifically, our framework can be illustrated as three side-by-side detection branches with a weighted fusion module which aggregates confidence scores generated in each branch into an overall trust score and final classification.

3.1 Multi-Modal Feature Extraction

Stage one of VeriSphere is focused on extracting discriminative features from an input video with a combination of visual and audio elements. The uploaded video is considered to be initially decomposed into video frame and audio segments. Video frames are used for both

spatial and temporal analysis, and video signals are analyzed separately to determine synchronization. Our model consists of three independent detection branches: spatial artifact, temporal consistency errors, and audio-visual synchronization, as depicted in Figure 1.

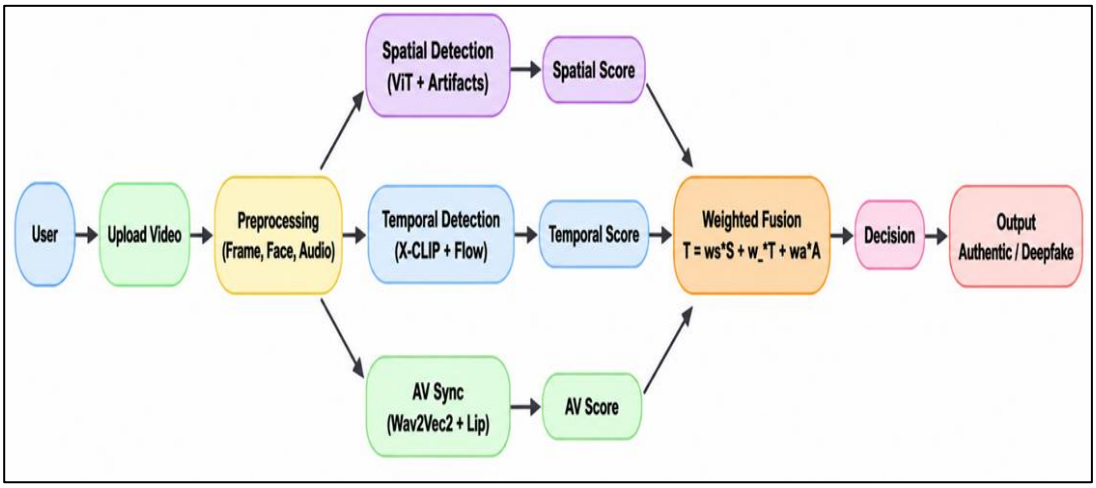

Figure 1. Proposed Verisphere Multi-Modal Deepfake Detection Framework

3.1.1 Spatial Artifact Analysis

Spatial analysis is created to detect the manipulation artifacts which are inserted in he synthesis and swapping face facial features and image generation process. To this end, facial regions are detected and normalized to be fed through a Vision Transformer ViT-Base/16 burada. The transformer architecture differs radically from conventional convolutional neural networks by using self-attention to detect long-range dependencies and consequently enabling detection of fine discrepancies across different facial regions.mechanisms, enabling the identification of subtle inconsistencies across different facial regions.

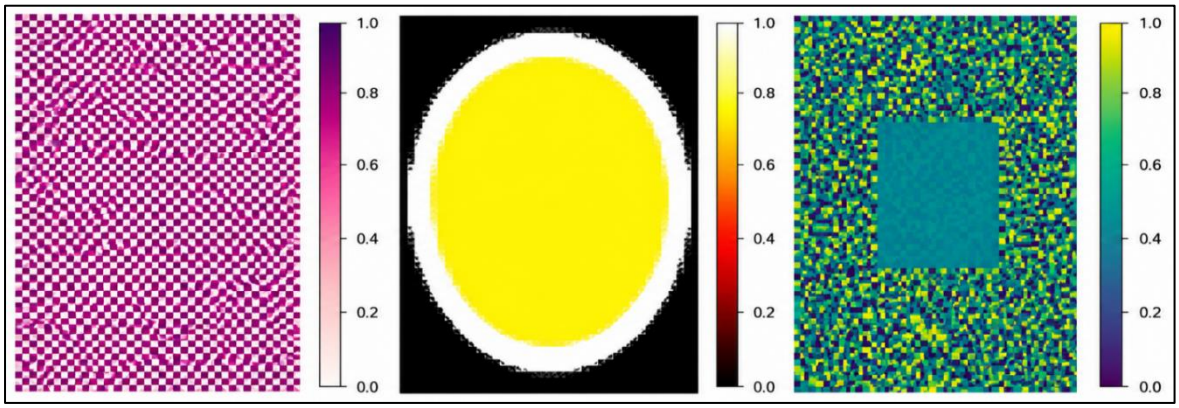

Figure 2. Spatial Forensic Feature Maps for Deepfake Artifact Detection

In figure 2, the left panel illustrates the DCT fingerprint response, highlighting high-frequency spectral patterns associated with synthetic image generation. The middle panel indicates the blend boundary response, where edge discontinuities can signify possible inconsistencies between manipulated facial regions and nearby genuine content. Right Panel: The response of skin to noise this shows examples of textures (from our paper) that we use in machine learning to tell apart the natural pattern you see with human skin from artificial surfaces. Together, these forensic signals constitute complementary evidence to detect visual artefacts and can enhance Vision Transformer-based deepfake detection methods.

The spatial confidence score is computed as

$$S = 0.55V_s + 0.20D_s + 0.15B_s + 0.10N_s \quad (1)$$

Where,

- S = final spatial confidence score
- V_s = Vision Transformer confidence score
- D_s = DCT fingerprint confidence score
- B_s = blend boundary confidence score
- N_s = skin noise confidence score

The spatial confidence score is obtained from a fusion of transformer-based visual representations with their complementary forensic indicators. Larger values of imply more compelling evidence for visual genuineness, while small predictions signal distortions being present in the facial areas. The weighting coefficients assign greater importance to the Vision Transformer output because it captures global facial inconsistencies more effectively than individual handcrafted features.

3.1.2 Temporal Consistency Analysis

Although spatial artifacts provide a valuable forensic evidence, advanced deepfake generation techniques often reduce visible frame-level inconsistencies. Consequently, temporal modeling is incorporated to analyze motion continuity and frame-to-frame coherence throughout the video sequence.

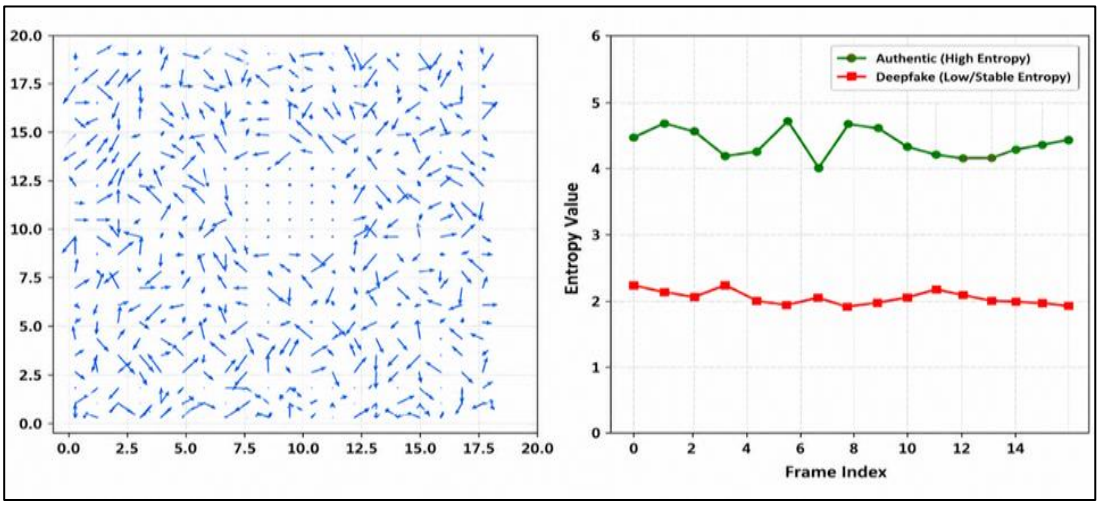

Figure 3. Temporal Forensic Analysis Using Optical Flow and Temporal Entropy

The temporal branch employs an X-CLIP encoder to extract temporal representations from consecutive video frames. The encoder captures contextual dependencies across multiple frames and facilitates the detection of anomalous motion patterns introduced by video manipulation. In order to improve temporal reliability, the analysis process is performed on third party motion fields, comprised of dense optical flow and temporal entropy measurements. Dense optical flow measures the consistency of motion in regions of the face and temporal entropy is used to measure frame-level variation over the full sequence. In general, authentic videos contain unobstructed motion transitions and dynamic entropy patterns, while manipulated videos often possess abnormal motion discontinuities and low temporal variations. Examples of temporal forensic signals shown in Figure 3

The temporal confidence score is expressed as

$$T = \alpha X_t + \beta O_t + \gamma E_t \tag{2}$$

subject to

$$\alpha + \beta + \gamma = 1 \tag{3}$$

Where,

- T = temporal confidence score
- X_t = X-CLIP confidence score

- O_t = optical flow consistency score
- E_t = temporal entropy score
- α, β, γ = weighting coefficients

The temporal confidence score measures the consistency of facial motion across sparsely sampled video frames. The X-CLIP encoder is designed to be able to achieve long-range temporal dependencies; optical flow and temporal entropy yield further estimates of motion coherence and dynamic variation. Videos having comparably smooth temporal transitions, generally have higher temporal confidence values.

3.1.3 Audio-Visual Synchronization Analysis

Deepfake videos are frequently exhibit inconsistencies between spoken audio and visible lip movements because visual synthesis and audio generation are often performed independently. To exploit this characteristic, audio-visual synchronization branch is incorporated into the framework. Figure 4 demonstrates the synchronization of speech energy and lip movement in real videos.

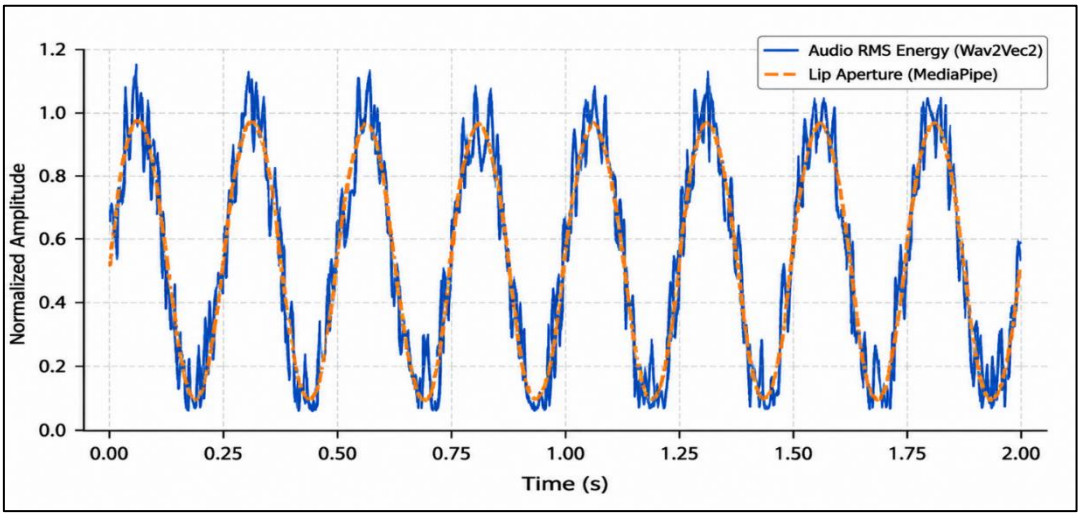

Figure 4. Audio-Visual Synchronization Analysis Based on Speech and Lip-Motion Alignment

Wav2Vec2 framework processes the audio stream by providing speech embeddings through the context of the raw audio. At the same time, MediaPipe Face Mesh is used to extract facial landmarks to provide trajectories of lip movement. Features like lip aperture, lip width,

and mouth curvature are calculated based on the extracted landmarks and then compared to the respective speech embeddings.

The synchronization score is determined through normalized cross-correlation:

$$A = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (4)$$

Where,

- A = audio-visual synchronization score
- x_i = audio feature sequence
- y_i = lip-motion feature sequence
- $\bar{x}$ = mean audio feature value
- $\bar{y}$ = mean lip-motion feature value
- n = number of synchronized observations

The normalized cross-correlation measures the degree of the correspondence between speech characteristics and lip movements. A higher synchronization score indicates the strong temporal alignment between audio and visual modalities, whereas lower values suggest potential manipulation or compositing artifacts.

3.2 Multi-Modal Fusion and Deepfake Classification

The outputs generated by spatial, temporal, and audio-visual branches provide complementary forensic evidence regarding video authenticity. To exploit the strengths of each modality, a weighted fusion mechanism is employed to aggregate their confidence scores into unified trust score.

The overall fusion score is calculated as

$$F = w_s S + w_t T + w_{av} A \quad (5)$$

Where,

- F = overall fusion confidence score

- S = spatial confidence score
- T = temporal confidence score
- A = audio-visual confidence score
- w_s = spatial weight (0.30)
- w_t = temporal weight (0.40)
- w_{av} = audio-visual weight (0.30)

The fusion weights are determined empirically through validation experiments using the training datasets. Multiple weight combinations were evaluated by the systematically varying the contributions of the spatial, temporal, and audio-visual branches while monitoring the detection accuracy and AUC on validation set. The optimal configuration assigns weights of 0.30, 0.40, and 0.30 to the spatial, temporal, and audio-visual branches, respectively. A slightly higher weight was assigned to the temporal branch because temporal inconsistencies remain more difficult to reproduce consistently in manipulated videos and therefore provide stronger discriminative information. The spatial and audio-visual branches were assigned equal weights since both contribute complementary forensic evidence that improves the robustness of the final decision.

The final classification decision is obtained through threshold-based evaluation:

$$\hat{y} = \begin{cases} 1, & F \geq \tau \\ 0, & F < \tau \end{cases} \tag{6}$$

where:

- $\hat{y}$ = predicted class label
- F = fusion confidence score
- τ = classification threshold

The final classification is obtained through threshold-based decision making. A prediction value of the $\hat{y}= 1$ denotes an authentic video, whereas $\hat{y}= 0$ indicates the deepfake video. The threshold value is selected through validation experiments to maximize classification performance while maintaining balanced precision and recall.

Although attention-based fusion has demonstrated promising performance in several multimodal learning applications, it generally introduces additional trainable parameters and increases computational complexity during optimization. The objective of the proposed VeriSphere framework is to achieve an effective balance between detection accuracy, interpretability, and computational efficiency. Therefore, a weighted fusion strategy was adopted, as it enables explicit control over the contribution of each modality while maintaining a lightweight architecture suitable for practical deployment. Moreover, the experimentally optimized weights provide a transparent interpretation of the relative importance of spatial, temporal, and audio-visual forensic evidence, making the final decision process easier to analyze and reproduce.

3.3 Dataset Description and Preprocessing

The proposed VeriSphere framework was tested on three publicly available benchmark datasets, namely FaceForensics++ [1], Celeb-DF [7] and DeepFake Detection Challenge (DFDC) dataset [8] which together ensure a variety of manipulations, compression rates, and realistic conditions for testing. FaceForensics++ dataset contains several face forgery techniques and is used for visual artifacts analysis, Celeb-DF provides high-quality deepfakes with fewer visual artifacts that allows assessing robustness of the model, DFDC dataset being one of the largest benchmarks enables large scale testing with various manipulations. Prior to training, the videos were preprocessed by extracting frames at 15 fps using OpenCV, face detection and alignment with MTCNN as well as resizing to 224 × 224 for Vision Transformer input. The audio was extracted at 16 kHz and converted to mono format for synchronization analysis. Moreover, data augmentation techniques such as JPEG recompression, changes of brightness and contrast, addition of Gaussian noise, and horizontal flip were performed.

Table 1. Dataset Statistics Used for VeriSphere Framework

Dataset	Real Videos	Fake Videos	Manipulation Types	Training Split	Validation Split	Test Split
Face Forensics++ (HQ) [1]	1,000	5,000	DeepFakes, Face2Face, FaceSwap, NeuralTextures, FaceShifter	80%	10%	10%

Face Forensics++ (LQ) [1]	1,000	5,000	DeepFakes, Face2Face, FaceSwap, NeuralTextures, FaceShifter	80%	10%	10%
Celeb-DF [7]	590	5,639	GAN-Based Face Manipulations	70%	15%	15%
DFDC Subset [8]	10,000	20,000	Multiple Deepfake Generation Techniques	75%	12.5%	12.5%

Table 1, Statistical description of the benchmark data sets used for training and testing the proposed system. The data sets contain an array of genuine and forged videos that can be used to comprehensively test the ability to detect deepfakes in space-time and audio-visual domains.

4. Results and Discussion

4.1 Model Performance Analysis

The effectiveness of the proposed VeriSphere framework was first evaluated by analyzing the training behavior of the spatial detection branch.

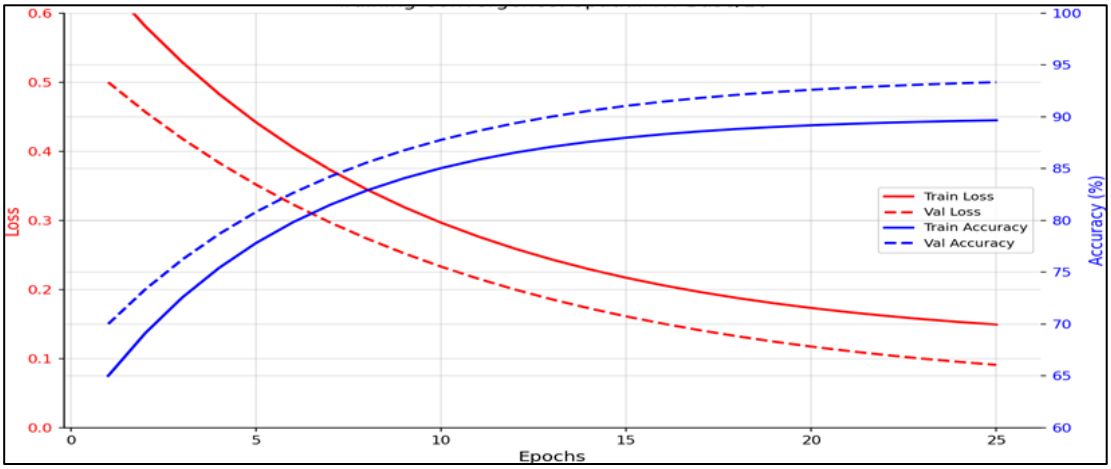

Figure 5. Training and Validation Performance of the Vision Transformer Model

Figure 5 presents the training and validation accuracy and loss curves obtained during the fine-tuning of the Vision Transformer model. it is a good indication of convergence in the process of learning. Moreover, compared to training and validation trends that are well aligned

with each other indicates the optimization strategy that minimises overfitting while retaining generalisation on unseen samples. These observations support the appropriateness of manipulating facial content using the Vision Transformer architecture, to extract spatially-discriminative representations.

Following convergence analysis, a contribution of each detection branch was independently examined. The results presented in Table 2 the spatial branch alone achieved the best standalone performance since it is able to capture global facial inconsistencies and synthesis artifacts spread across multiple facial regions. The temporal and audio-visual branches achieved lower performance in comparison, but were able to detect motion irregularities and synchronization inconsistencies that are not typically detected using pure image-based methods. The fact that these branches are each capturing a unique class of forensic evidence demonstrates the complimentary nature of the multi-modal framework proposed.

Table 2. Performance Comparison of Individual Detection Branches on the FaceForensics++ Test Set

Detection Layer	Accuracy (ACC %)	AUC	F1-Score
Spatial (ViT-Base/16)	86.4	0.921	0.871
Temporal (X-CLIP + Flow)	79.2	0.863	0.803
AV-Sync (Wav2Vec2 + MediaPipe)	74.5	0.812	0.758

Multiple fusion configurations were assessed to further investigate the effectiveness of feature integration. As indicated in Table 3, these results show that adding more modalities to the framework generally leads to an increased ability to detect. The model leverage inherent complementarity between temporal and audio-visual information which significantly improved classification reliability by overcoming the shortcomings of individual detection branches. The superior performance of the full fusion architecture confirms that deepfake manipulations simultaneously exert effects on visual appearance, temporal continuity and speech–lip correspondence. Thus, the adjacency of those heterogeneous forensic clues allows the framework to build a better indication on video authenticity and enhance its detection robustness overall.

Table 3. Performance Detection of Single-Modal and Multi-Modal Fusion Configurations

Configuration	Accuracy (ACC %)	AUC	F1-Score	EER (%)
Spatial Only	86.4	0.921	0.871	12.3
Temporal Only	79.2	0.863	0.803	18.7
AV-Sync Only	74.5	0.812	0.758	22.1
Spatial + Temporal	89.7	0.941	0.901	9.8
Spatial + AV-Sync	87.9	0.929	0.884	11.2
Temporal + AV-Sync	83.6	0.904	0.847	14.5
Full Fusion	92.1	0.963	0.924	7.4

4.2 Robustness and Generalization Analysis

The proposed framework was also analyzed in terms of its robustness to different types of manipulation used in deepfakes. Figure 6 shows the results of detection achieved using manipulation categories included in the FaceForensics++ dataset. The ability to achieve high-quality results regardless of the manipulation type proves that the framework does not rely on any artifacts created by one particular generation method. This observation implies that the suggested fusion approach is capable of extracting general patterns useful for forensic analysis regardless of significant changes in the characteristics of manipulations.

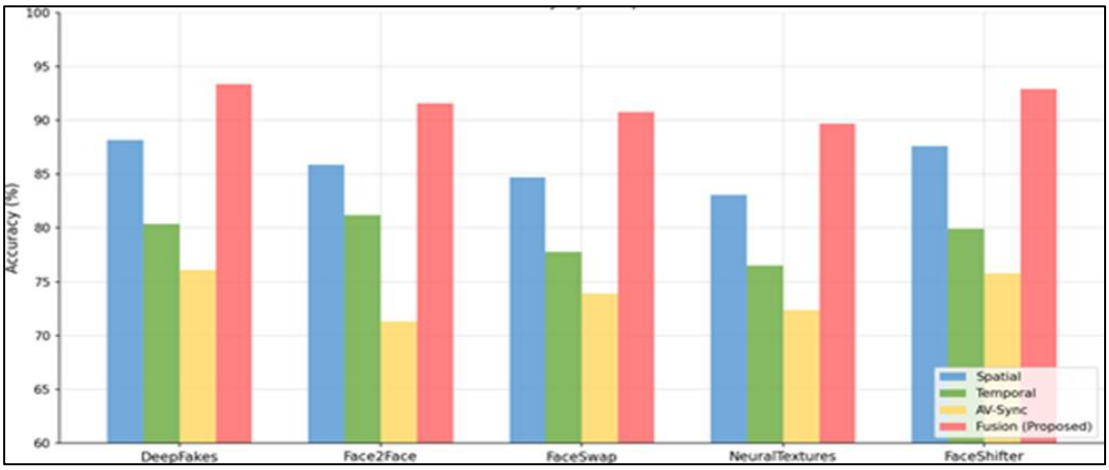

Figure 6. Detection Accuracy Across Different Deepfake Manipulation Methods.

Generalization capability was subsequently assessed through cross-dataset evaluation using FaceForensics++, Celeb-DF, and DFDC. The results presented in Table 4 highlight the challenges due to domain shift. Models trained on individual datasets performed poorly when

evaluated on other unseen datasets, suggesting that dataset-specific artifacts are not enough for generalizable detection. On the other hand, by including samples from multiple benchmark datasets, cross-domain detection reliability was significantly enhanced. The performance gain indicates that different manipulation styles can guide the framework to learn more generalized forensic representations, which makes it more scalable in real-world deepfake cases.

Table 4. Cross-Dataset Generalization Performance

Training Dataset(s)	Evaluation Dataset	Purpose of Evaluation	Accuracy (ACC %)	AUC
FaceForensics++ (HQ only)	FaceForensics++ (LQ)	Evaluate robustness across different compression levels	83.2	0.891
FaceForensics++ (HQ only)	Celeb-DF	Establish baseline cross-dataset performance on high-quality deepfake videos	78.6	0.841
FaceForensics++ (HQ only)	DFDC	Establish baseline performance on the most challenging benchmark	71.3	0.793
FaceForensics++ + Celeb-DF	DFDC	Assess the influence of increased training diversity on generalization	76.8	0.832
FaceForensics++ + Celeb-DF + DFDC	DFDC	Evaluate the final generalized model after progressive multi-dataset training	81.4	0.874

The cross-dataset experiment was carried out to understand how increased training diversity impacts model generalization and not to cover all possible training/testing combinations. In the first step, FaceForensics++, Celeb-DF, and DFDC benchmarks were considered individually to measure their cross-domain generalization performance under various domains. Once the baseline for these individual datasets was set, Celeb-DF and DFDC datasets were incrementally added during the training while DFDC remained the evaluation dataset. DFDC is chosen for its high variability of subjects, recording environment, compression artifacts, and face manipulation methods in comparison with other benchmark datasets. Hence, it is capable of providing the toughest test of cross-domain generalization. It is clear from the results obtained in incremental addition of datasets that the VeriSphere model

is able to learn more generalized forensics representations when exposed to different patterns of face manipulation.

4.3 Ablation and Computational Analysis

To determine the relative significance of each module in terms of its contribution towards the fusion process, an ablation analysis was conducted on the basis of contribution from the spatial, temporal, and audio-visual modules. As can be seen from Table 5, combining the three branches proved to be the optimal combination for accurate detection. Any other approach focusing on only one branch was found to lower the level of accuracy, suggesting that any individual forensics clue will not be able to cover all the aspects of deepfake manipulation.

Table 5. Ablation Analysis of Fusion Weight Configurations

Spatial Weight (w_s)	Temporal Weight (w_t)	AV-Sync Weight (w_{av})	Accuracy (ACC %)	AUC
0.50	0.25	0.25	88.3	0.934
0.33	0.33	0.33	90.7	0.951
0.25	0.50	0.25	91.2	0.957
0.30	0.40	0.30	92.1	0.963
0.20	0.60	0.20	89.5	0.944

The influence of the classification threshold on model performance was furtherly investigated using the results presented in the Figure 7.

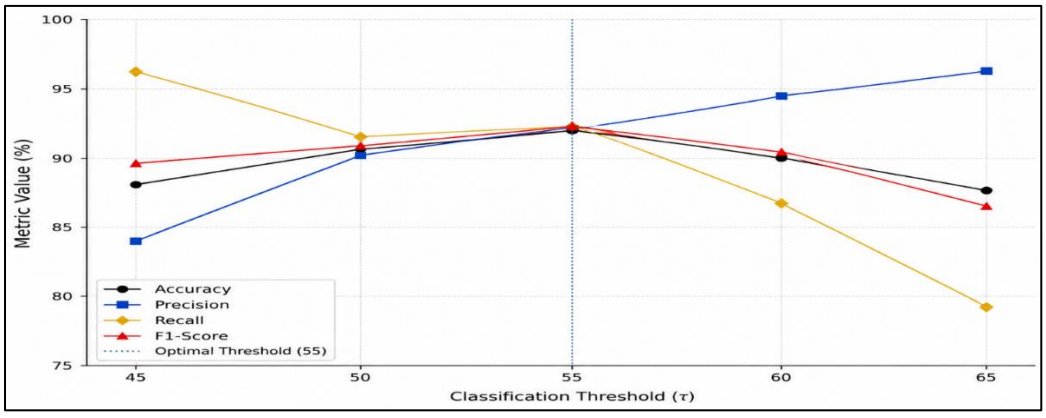

Figure 7. Threshold Sensitivity Analysis of the Proposed VeriSphere Framework

The threshold controls the trade-off between precision and recall, thereby affecting the sensitivity of the detection system. The experimental analysis was demonstrates the existence of an optimal operating region where false positives and false negatives are simultaneously

minimized. Threshold values below this region increase detection sensitivity but introduce additional false alarms, whereas higher thresholds improve confidence at the expense of missed detections. The selected operating threshold therefore provides a balanced compromise between detection reliability and classification stability.

Finally, the computational efficiency of the framework proposed was tested through the use of the standard CPU deployment platform. From the latency values presented in Table 6, temporal analysis is the most computationally intensive step because of the sequential processing of videos. However, despite this limitation, the total computation time can be considered reasonable for real-time forensic analysis. Processing of videos independent of any GPU hardware proves the viability of using this framework for content moderation platforms and other forensic analysis systems.

Table 6. Computational Efficiency and Inference Latency Analysis

Component	Average Latency (ms)
Video Loading + Frame Extraction	124
Face Detection + Alignment (MTCNN)	38
Spatial ViT Inference	612
Temporal X-CLIP Encoding	743
AV-Sync (Wav2Vec2 + MediaPipe)	394
Score Fusion + Output	< 1
Total (10-Second Clip)	∼1.91 s

Overall, the experimental results show that the proposed VeriSphere framework can effectively exploit the complementary forensic information from the spatial, temporal and audio-visual domains. The proposed fusion-based architecture improves the detection accuracy, as well as the robustness to different manipulation techniques and unseen datasets. The results verify the effectiveness of multi-modal deepfake detection and show its potential for reliable use in real-world applications of digital media authentication.

5. Conclusion and Future Work

This project presented the VeriSphere, a multi-modal deepfake detection framework that integrates spatial, temporal, and audio-visual forensic analysis within a unified architecture. The proposed approach combines Vision Transformer-based spatial artifact detection, X-CLIP-

driven temporal consistency analysis and Wav2Vec2–MediaPipe synchronization assessment to capture a complementary manipulation cues across multiple modalities, with a weighted fusion strategy used to aggregate outputs and generate a unified trust score for classification. Experimental evaluation on the FaceForensics++, Celeb-DF, and DFDC benchmark datasets demonstrated the effectiveness of the framework, achieving a detection accuracy of 92.1%, an AUC of 0.963, and an F1-score of 0.924, while significantly outperforming individual detection branches and improving robustness against diverse manipulation techniques. Overall, the findings indicate that integrating spatial artifacts, temporal inconsistencies, and audio-visual synchronization cues provides a more comprehensive and dependable assessment of video authenticity. Future work will focus on enhancing robustness against the emerging diffusion-based deepfake generation methods and adversarial manipulation strategies, exploring adaptive fusion mechanisms, self-supervised representation learning, and lightweight model optimization for real-time deployment on resource-constrained platforms, as well as incorporating physiological and behavioral biometric cues and leveraging larger, more diverse datasets to further improve generalization and practical applicability in real-world digital media forensics and content verification scenarios.

References

[1] Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” In Proceedings of the IEEE/CVF International Conference of Computer Vision (ICCV), Seoul, South Korea, 2019: 1–11.

[2] Wang, Junke, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. "M2tr: Multi-Modal Multi-Scale Transformers for Deepfake Detection." In Proceedings of the International Conference on Multimedia Retrieval, 2022: 615-623.

[3] Li, Yuezun, Ming-Ching Chang, and Siwei Lyu. "In Ictu Oculi: Exposing Ai Created Fake Videos by Detecting Eye Blinking." International workshop on information forensics and security (WIFS), IEEE, 2018: 1-7.

[4] Ni, Bolin, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. "Expanding Language-Image Pretrained Models for

General Video Recognition." In European conference on computer vision, Cham: Springer Nature Switzerland, 2022: 1-18.

[5] Chung, Joon Son, and Andrew Zisserman. "Out of Time: Automated Lip Sync In The Wild." In Asian conference on computer vision, Cham: Springer International Publishing, 2016: 251-263.

[6] Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." Advances in neural information processing systems 2020, vol. 33: 12449-12460.

[7] Li, Yuezun, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. "Celeb-Df: A Large-Scale Challenging Dataset for Deepfake Forensics." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2020: 3207-3216.

[8] Dolhansky, Brian, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. "The Deepfake Detection Challenge (DFDC) Dataset." arXiv preprint 2020, arXiv:2006.07397.

[9] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv preprint 2020, arXiv:2010.11929.

[10] Chollet, François. "Xception: Deep Learning with Depthwise Separable Convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition 2017: 1251-1258.

[11] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry "Learning Transferable Visual Models from Natural Language Supervision." In International conference on machine learning, Proceedings of Machine Learning Research, 2021, vol. 139: 8748-8763.

[12] Tolosana, Ruben, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection." Information fusion 2020, vol. 64: 131-148.

[13] Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets." Advances in neural information processing systems 2014, vol. 27: 1-9.

[14] Nguyen, Thanh Thi, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. "Deep Learning for Deepfakes Creation and Detection: A Survey." Computer Vision and Image Understanding 2022, vol. 223: 103525.

[15] Zhao, Hanqing, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. "Multi-Attentional Deepfake Detection." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2021: 2185-2194.