International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.4 (2025) pp. 4637-4645
http://www.ijcesen.com

ISSN: 2149-9144

Copyright © IJCESEN

Research Article

An Explainable Deep Learning Framework for Robust Face Anti-Spoofing via Shapley-Guided Feature Selection

Anisha Soni[1]*, Sipi Dubey[2], Ajay Kushwaha[3]

[1]Assistant Professor, Department of Computer Science and Engineering, Rungta College of Engineering and Technology, Bhilai, Chhattisgarh, India
* **Corresponding Author Email:** anishasoni27@gmail.com- **ORCID:** 0000-0002-5247-7440

[2]Professor, Department of Computer Science and Engineering, Rungta College of Engineering and Technology, Bhilai, Chhattisgarh, India
Email: drsipidubey@gmail.com- **ORCID:** 0000-0002-5247-1150

[2]Professor, Department of Computer Science and Engineering, Rungta College of Engineering and Technology, Bhilai, Chhattisgarh, India
Email: kushwaha.bhilai@gmail.com - **ORCID:** 0000-0002-5247-0050

Article Info:

DOI: 10.22399/ijcesen.5267
Received : 02 November 2025
Revised : 20 December 2025
Accepted : 25 December 2025

Keywords

Face anti-spoofing,
multimodal biometrics,
deep learning,
explainable AI,
presentation attack detection,
liveness detection

Abstract:

With biometric authentication becoming everywhere, more and more systems are targeted by advanced presentation attacks that can include physical print re-enactment and synthetic speech deepfakes. To counter these cross-modal threats, we present an elucidated multipropositional anti-spoofing system which collaboratively considers facial cues as well as acoustic cues. We compare three different visual architectures MobileNetV2, ResNet50, and ViT-L16 and an audio branch that is trained on the representations of ASVspoof speech countermeasures. To meet the urgent demand of transparency of decisions made by the network as a security application, we directly embed the Explainable AI (XAI) guided by Shapley-values directly into the network. The model actively removes spurious background activations by using custom XAIPooling and XAIDropout layers, and isolates discriminative spoofing artifacts. Using the MobileNetV2 backbone, our optimized configuration has an Average Classification Error Rate (ACER) of 3.2, an AUC of 0.98, and a test accuracy of 93.02% with MobileNetV2. These results prove that a multimodal fusion with explainable feature selection can provide a strong, interpretable, and highly efficient solution that can be easily deployed to the edge.

1. Introduction

Nowadays, biometric authentication is used to secure mobile banking to enterprise access control. Face recognition is preferred by users as it is fast, non-friction and non-invasive. Nevertheless, with this convenience come dire vulnerabilities. With printed photographs, screen replays or 3D masks, attackers can often evade the common recognition pipelines should the system fail to model spoof-specific cues correctly. Consequently, face anti-spoofing has become a very pressing area of concern in presentation attack detection (PAD). However, even in real-world applications, relying on dependable cross-domain generalization has remained a significant challenge, as a number of surveys point out [1]-[6]. The problem is becoming increasingly difficult as there is an increase in the generation of attacks. Whereas simple 2D printouts could be easily caught with early biometric systems using handcrafted texture descriptors [7], [8] modern attackers make use of generative deepfakes, synthetic media, and adversarial perturbations. Traditional feature extractors merely are unable to keep pace. Indeed, when it comes to exposing the models to unseen domains or new spoofing techniques, it is always shown through contemporary studies that models based on full reliance on local texture representations fail when the models are exposed to unseen domains or new spoofing strategies [2], [5], [9]. Additionally, it is no longer a one modality issue. In real world biometric systems, attackers can use a combination of manipulated facial video with cloned or

synthesized speech to attack cross-modal spoofing attacks that can bypass unimodal defenses. This has led to increased interest in multimodal anti-spoofing systems, especially since the inception of large-scale benchmarks like ASVspoof 5, which takes into account a wide range of acoustic conditions and adversarial attacks, as well as synthesized speech scenarios [10]. The other significant issue is the interpretability of models. Despite the high detection with deep learning architectures, their decision-making processes are usually opaque and this limits their trust in security-critical applications. To deal with this problem, recent studies incorporate explainable AI (XAI) techniques into anti-spoofing systems, including Shapley-value-based feature attribution methods which increase transparency and suppress irrelevant features [11]-[13]. These strategies are especially useful in applications where the system is not just required to detect spoofing attempts but also provide interpretable and trustful justification of the decisions made by the system.

A. Motivation and Research Gap.

Despite the considerable improvement of the state of the art, there are still three significant gaps. To begin with, most of the methods are still limited in generalizing across datasets, capture devices, and attack conditions. Second, the vast majority of the current systems are only single-modal, most commonly face or speech, despite the fact that most current attacks are multimodal in nature. Third, explainability has not yet been utilized in biometric security, yet it is an important concept in the context of trust, debugging, and implementation in regulated settings. These constraints drive a framework that integrates multimodal fusion, explainable feature selection, and edge-aware model design [4], [5], [6], [22], [24].

B. Contributions The main contributions of this work are as follows:

1. A multimodal anti-spoofing system, which combines visual and verbal representations to enable a strong detection of presentation attacks.
2. Explainable learning strategy using XAIPooling, XAIDropout in order to enhance transparency and reduce the dependence on spurious background cues.
3. Compare the accuracy, interpretability and deployment efficiency of MobileNetV2, ResNet50, and ViT-L16.
4. An experimental analysis based on ACER, EER, HTER, and AUC with an ablation analysis and a deployment-oriented discussion.
5. Journal-based treatment of constraints, such as low-light sensitivity, effects of background noise, and stronger cross-dataset validation.

2. Related work and comparative analysis

A. Evolution of Face Anti-Spoofing Techniques

A. Face Anti-Spoofing Techniques Evolution A. Face Anti-Spoofing Techniques Evolution In the early face anti-spoofing methods were based on handcrafted descriptors Local Binary Patterns (LBP) [7], dynamic texture features [8], and color texture statistics [14]. These techniques were effective in controlled situations but could not generalize to a variety of conditions of illumination, pose, and sensor conditions [9], [15]. Follow-up studies proposed deep learning-based architectures to learn automatically discriminative spoof-related features such as convolutional neural networks and noise modeling methods [16], [17]. More sophisticated works went further to explore residual learning, attention mechanisms and 3D face presentation attack detection techniques [18] in order to better differentiate between spoof-specific artifacts. Notwithstanding these developments, deep models are still vulnerable to bias in their datasets and unknown conditions of attack which motivates the continued research of more transferable and robust solutions [3], [4], including applications in disguised face recognition [19] and lightweight deployment-oriented recognition systems [20]. Moreover, recent surveys underline that domain generalization, as well as multimodal fusion are open challenges in real-world biometric security systems [5], [6].

B. Multimodal and Audio-Visual Anti-Spoofing.

Although visual based anti-spoofing has reached a mature stage, with the advent of speech deepfakes, they have created a new dimension of attack vectors. Spoofing cases in modern context often involve both the use of facial manipulation and synthesized or cloned speech, making unimodal system less and less adequate. ASVspoof 5 is a significant step towards this direction, as it introduces a larger and more diverse speech corpus with adversarial attacks, representing the complexity of real-world spoofing scenarios [6]. New multimodal methods have investigated fusion methods at feature, score and decision levels. Nevertheless, the majority of the available literature addresses either the optimization of fusion performance or the problem of modality specific optimization, but little emphasis is put on the issue of explainability and constraints on deployment. Moreover, the multimodal data are still rather limited, which still restricts the large-scale assessment [6], [24].

C. Explainable AI in Biometric Anti-Spoofing

The trend in the implementation of biometric solutions in security-related settings has rendered explainability one of the critical requirements.

Conventional deep learning systems are black-box systems, thus limiting transparency and user trust. To overcome this limitation, explainable AI methods, in particular those that are based on Shapley values, have been proposed. The XAIPooling and XAIDropout methods include the importance of the features directly in the learning process, thus enabling models to suppress irrelevant activations and focus on the discriminative spoof-related features [7], [25]. Recent works have posited that explainability does not only enhance interpretability, but can also increase robustness by avoiding overfitting to background noise, and dataset-specific artifacts. Nevertheless, the application of XAI in multimodal anti-spoofing has not been exploited, especially in the real-time and edge-based implementation [7], [10].

D. Comparative Analysis of State-of-the-Art Methods

Table I provides an overview of some of the most common anti-spoofing techniques and points to the shift in the direction of deep, multimodal and explainable methods.

Table I provides a brief comparison of the representative face and multimodal anti-spoofing methods, outlining the evolution of using handcrafted descriptors to deep learning and explainable frameworks. The table demonstrates that initial approaches that relied on Local Binary Patterns and color texture analysis were computationally cheap, but lacked strength when subjected to cross-dataset conditions. Deep models (including CNN-based models) in contrast, presented a better performance on classification, though recent explainable models have brought more transparency and feature selection. The proposed multimodal framework builds on this development and incorporates visual and speech modalities and Shapley-guided explainability to enhance robustness, interpretability, and relevance to deployment.

E. Research Gap and Positioning

According to the analysis above, there are four critical research gaps. Most of the models are found to fail when the testing is done on hidden datasets because the conditions of acquisition vary and the methods of spoofing are varied. Current methods are mostly based on one modality, which can be easily attacked by another modality. The use of explainability in biometric security has not been fully utilized, though it is important to both be able to audit and to trust the system. Complexity of advanced models can also be a limitation to deployment in real-time, edge-based applications [4], [6], [7], [10], [22], [24]. These limitations have been tackled by the proposed framework multimodal fusion, Shapley-guided feature selection, and lightweight architectures.

3. Proposed methodology

Our framework is implemented in five different steps: dataset preparation, extraction of modality-specific features, multimodal fusion, explainability-aware classification, and performance evaluation. Instead of taking the inputs in a linear fashion we take the image and speech stream in parallel before combining them to make a final decision.

A. Dataset Survey and Preprocessing

The system is developed based on the various datasets of faces and speech. In the case of the visual branch, the following benchmark datasets can be considered: CASIA-FASD, Replay-Attack, MSU-MFSD, 3DMAD and CelebA-Spoof. The main benchmark to analyse speech deepfake and adversarial attacks is the ASVspoof 5 in the case of the audio branch [6], [9], [14], [23]. Detecting face images, cropping, aligning and resizing face images to a standard input resolution. The speech signals are converted to MFCCs and spectrogram representations to detect the frequency-domain anomalies that are related to synthesis and replay artifacts.

B. Visual Backbones: MobileNetV2, ResNet50, and ViT-L16

In order to achieve the optimal tradeoff between accuracy and efficiency we considered three different visual architectures. MobileNetV2 is the depthwise separable convolution that we chose as the lightweight baseline because its depthwise separability allows it to drastically reduce the number of parameters and the inference latency of such models, making it a good fit in edge devices [22]. To learn deeper features, we also tested the ResNet50, which allows us to exploit the skip connections it provides [23], and we also experimented with ViT-L16, which enables us to use the self-attention mechanisms to learn globally and over long distances. Although ViT-L16 models the global context very well, it consumes significantly more computational resources than small convolutional networks. This comparative design enables us to comprehensively test the local texture modeling, as well as the global feature representation under the domain shift conditions [24]. Even though ViT-L16 can effectively provide the global context, it tends to demand more data and computational resources than compact convolutional networks. Such a comparative design is suitable since the recent studies on the topic of anti-spoofing show an increasing focus on the aspects of both local texture modeling and global

feature representation when assessing the robustness to domain shift [4], [5], [6].

C. Speech Branch and Multimodal Fusion

A 1D convolutional network is used to process speech utterances to generate audio embeddings. Each branch (face and speech) generates a fixed length latent representation, which is then concatenated at the embedding level in late fusion. The design enables the model to take advantage of complementary cues like the consistency of lip-motions, spectral artifacts, and texture irregularities. The merged representation is then forwarded to the classifier head where it is finally predicted [6].

The overall processing pipeline that depicts the combination of visual and speech modalities. The parallel feature extraction streams (CNN/ViT in the case of facial images and 1D-CNN in the case of acoustic data) are at the embedding layer. Explainable AI modules (XAIPooling and XAIDropout) are used before the last classification layer to avoid spurious activations.

D. Explainable AI Integration

One of the contributions of our work is that the decision process can be made transparent. We used the feature attribution provided by Shapley-values directly to the network with special XAIPooling and XAIDropout layers. Where $\phi(x_i)$ represents the value of the Shapley importance that feature x i has.

$$X_{out} = \sum_{i=1}^{n} \phi(x_i).x_i \quad (1)$$

Similarly, the explainability-guided dropout is formulated as:

$$X_{drop} = x_i.m_i, m_i \sim Bernoulli\big(1 - \phi(x_i)\big) \quad (2)$$

Our mechanism can guarantee that lower informative or noisy activations are ultimately dropped during training by using a stochastic mask (m_i): this will probabilistically suppress activations which have lower Shapley values. This compels the network to specialize exclusively in discriminative spoofing artifacts as opposed to information pertinent to the dataset, and substantially improve both the interpretability and strength of the resulting model.

E. Training and Implementation Details

The models are available in PyTorch and can be trained on an NVIDIA RTX 3090 with 24 GB VRAM. The Adam optimizer with initial learning rate of 1×10^{-4}, batch size of 32 and categorical cross-entropy loss were used as the optimizer. The training was implemented up to 60 epochs and the early stopping was determined on the basis of the validation loss. To decrease overfitting, data augmentation, such as random cropping, horizontal flipping, and synthetic noise injection were used.

F. Evaluation Metrics

System performance is measured using standardized biometric security measures which are commonly used in the presentation attack detection literature [3], [14]. The mean of the Attack Presentation Classification Error Rate (APCER) and the Bona Fide Presentation Classification Error Rate (BPCER) is called the ACER (Average Classification Error rate). The operating point where the false acceptance rate is equal to the false rejection rate is called the operating point of EER (Equal Error Rate). The HTER (Half Total Error Rate) is provided by (FAR+FRR)/4. AUC (Area Under the Curve) is used to estimate the area under the ROC curve, and to summarize the trade-off between sensitivity and specificity.

4. Experimental setup and evaluation protocol

A. Datasets and Benchmark Description

A collection of publicly available benchmark datasets that includes both visual face anti-spoofing and speech deepfake detection are tested to make the proposed framework testable. These datasets are comprised of several spoofing conditions, i.e., print attacks, replay attacks, 3D masks, and synthesized speech [1], [4], [6].

Table 2 summarizes the datasets that were utilized in this study and that depicts the richness of spoofing conditions that were encompassed in the experimental protocol. The chosen visual datasets comprise CASIA-FASD, replay-attack, MSU-MFSD, 3DMAD, and CelebA-Spoof that offer the various types of attacks which are print, replay, and 3D mask attacks. To evaluate speech-based authentication, ASVspoof 5 offers a large-scale benchmark with synthesized, replayed and adversarial utterances. This diversity of data is critical in assessing generalization since it would guarantee that the proposed model is tested in heterogeneous conditions of acquisition, types of spoof and modality-specific attack patterns.

A visual comparison showing the differences of the artifact between real presentations and spoofing presentations. The authentic face has a distinct texture and natural edges, as compared to the print attack of paper texture with blurred edges and the replay attack of screen reflections and glare.

B. Data Preprocessing and Feature Extraction

In the visual modality, face detection is done with subsequent alignment, resizing to $224\,x\,224$, normalization and augmentation. In the case of audio modality, the speech signals are converted to MFCCs and log-Mel spectrograms to extract the spectral artifacts that are related to spoofing. The

steps enhance uniformity among datasets and sensitivity to variability of inputs [6].

C. Training Configuration

The given structure is implemented with the assistance of PyTorch and trained on an NVIDIA RTX 3090 with a 24 GB VRAM. Early stopping is done based on the validation loss, optimizer is Adam with a learning rate of $1X10^{-4}$, batch size is 32, number of epochs is 60 and number of batch size is 32. Categorical cross-entropy is used as the loss function. The visual backbone comparison consists of MobileNetV2, ResNet50 and ViT-L16 as compared to the speech branch which uses a 1D-CNN of audio embedding [11]-[13].

D. Cross-Dataset Evaluation Protocol

To test generalization, the framework will be tested by cross-dataset protocol that is required to test journals. This model is trained on one set of data and tested on another set of data, without retraining. The most common combinations include training on CASIA-FASD and testing on Replay-Attack and training on MSU-MFSD and testing on CASIA-FASD. With speech, speech splits are used to train and the speech testing splits are used to test under unknown attack conditions [6], [14], [24].

The results of the cross-dataset evaluation and the evidence of the generalization ability of the proposed framework is reported in Table 3. Contrary to in-dataset testing, cross-dataset evaluation checks the extent to which a model is deployment ready by measuring the way it performs when it is trained and tested on different data distributions. The findings indicate that the proposed method always leads to a decrease in ACER and EER and an increase in AUC when compared to CNN baseline in different training-testing combinations. It means that the multimodal fusion and explainability-driven feature selection can assist the model to learn more transferable representations of spoofing and be less reliant on the dataset-specific artifacts.

E. Statistical Validation

Experiments are repeated three times (independent runs) to establish reliability and the standard deviation is reported. The proposed model results in $ACER = 3.2\% \pm 0.4$ and $EER = 2.8\% \pm 0.3$ which means that the model is stable in various trainings initializations.

5. Results and discussion

The experimental findings suggest that the proposed explainable multimodal framework achieves better performance on the benchmarks evaluated. The best possible results reported in this study are an ACER of 3.2% an AUC of 0.98 and a MobileNetV2 test accuracy of 93.02 which means that very small CNNs can offer an attractive trade-off between performance and deployment cost.

A. Quantitative Comparison

The findings indicate a definite advancement between the manual creation of descriptors to profound and understandable practices. The explainability-guided model has the lowest error and highest discrimination, and thus indicates that feature selection based on Shapley importance can be used to enhance robustness, as opposed to merely enhancing interpretability [7], [25].

Table 4 demonstrates that the proposed multimodal framework achieves competitive performance compared to state-of-the-art explainable models, while additionally supporting cross-modal robustness and edge deployment capabilities.

System performance assessment of the system which displays the tradeoff between True Positive Rate (TPR) and False Positive Rate (FPR). This curve can be used to indicate the effectiveness of the multimodal classifier in distinguishing between bona fide presentations and adversarial spoofing attack.

B. Ablation Study

An ablation study was done to validate the components of the proposed architecture. The unimodal, visual baseline with MobileNetV2 only reached $ACER = 7.5\%$. The unimodal audio baseline using speech countermeasures only achieved EER = 6.8%. Multimodal fusion, but without XAI, decreased $ACER\ to$ 4.8%. Lastly, the multimodal fusion with XAI had $ACER = 3.2\%$ which represents a reduction in 57.3% of the baseline of multimodal fusion with XAI.

C. Computational Cost and Edge Feasibility

The MobileNetV2 optimized configuration requires less than 42 million parameters and has an approximate speed of 45 frames per second on a typical mobile CPU, making it suitable to the real-time authentication requirement. ViT-L16, in contrast, has significantly more parameters and has a lower throughput, so it is not as suitable to be deployed at low power. These results indicate that small CNN backbones can be used when the edge latency and the energy consumption are the main constraints [11]-[13].

D. Failure Analysis

There are two conditions in which the main failure modes occur. First, actual faces that are captured in very low light conditions can be misclassified as spoof attacks since cues of fine-textures are unreliable. The second is that heavy background speech noise can corrupt audio embeddings, and can decrease the usefulness of late fusion. These results imply that to improve the situation in the future, one should address the illumination

normalization, noise-robust audio modeling, and adaptive fusion strategies [4], [6], [22].

E. Discussion

In general, the findings suggest that explainability-aware learning does not only contribute to the model transparency, but also to improved generalization as the network is not encouraged to rely on unstable background features. MobileNetV2 is more practical with regard to accuracy, interpretability, and real-time feedback, whereas ViT-L16 is more compute-intensive and yet useful when large training data, and high-end hardware are available. The multimodal environment is particularly useful in the deployment of security sensitive scenarios since it decreases the chances of a single spoof channel compromising the system [4], [6], [7], [24].

Table 1: *Comparison of representative face and multimodal anti-spoofing methods*

Study	Modality	Model Type	XAI	Dataset(s)	Key Strength	Limitation
Chingovska et al. [7]	Face	LBP	No	CASIA	Simple and fast	Poor generalization
Boulkenafet et al. [14]	Face	Color texture analysis	No	Replay-Attack	Improved texture modeling	Sensitive to lighting
George and Marcel [9]	Face	CNN	No	Cross-dataset	Generalization study	Performance drop
Jourabloo et al. [17]	Face	CNN + noise modeling	No	OULU-NPU	Handles digital attacks	Complex training
Yu et al. [3]	Face	Deep CNN	No	Multiple	Strong baseline	Black-box model
Wang et al. [10]	Speech	Deep countermeasure	No	ASVspoof 5	Large-scale speech detection	No visual fusion
Ma et al. [11]	Face	CNN + XAI	Yes	CASIA	Explainability	Unimodal only
Recent multimodal works	Face + Audio	Fusion models	No	Limited	Improved robustness	No interpretability
Proposed work	Face + Audio	CNN + ViT + XAI	Yes	Multimodal	Robust, explainable, edge-ready	Needs broader cross-dataset validation

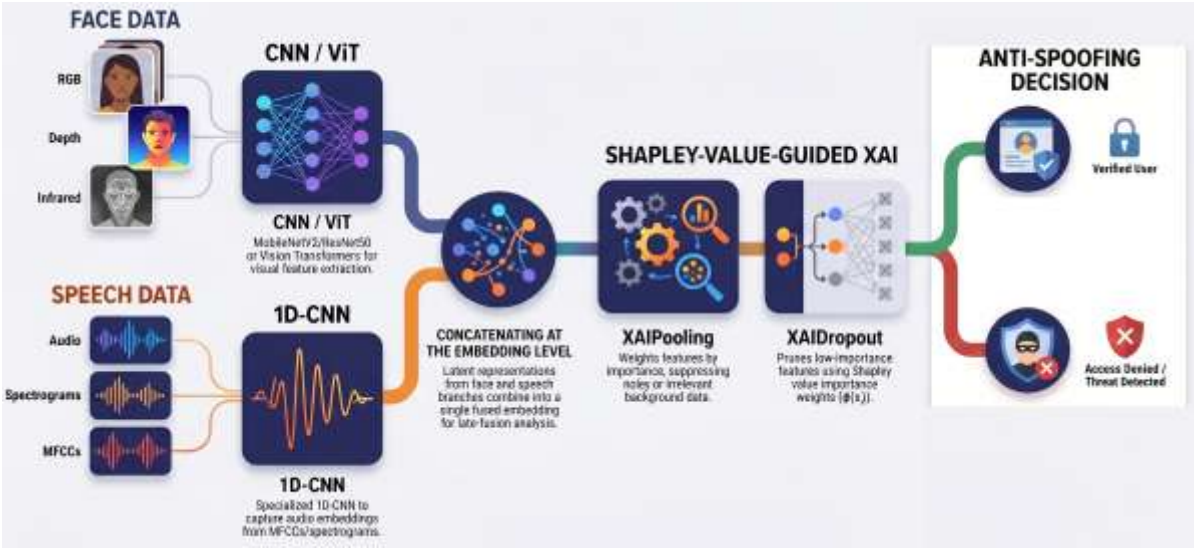


Figure 1. *Proposed multimodal anti-spoofing architecture illustrating parallel visual and speech feature extraction with explainable fusion*

Table 3. *Dataset description and characteristics*

Dataset	Modality	Samples	Attack Types	Protocol
CASIA-FASD	Face	600 videos	Print, replay	Train/test split
Replay-Attack	Face	1300 videos	Print, mobile replay	Fixed protocol
MSU-MFSD	Face	440 videos	HD replay, print	Cross-dataset
3DMAD	Face	255 videos	3D mask	Leave-one-out
CelebA-Spoof	Face	600K images	Multiple spoof types	Large-scale split
ASVspoof 5	Speech	2000+ speakers	Synthesized, replay, adversarial	Challenge protocol

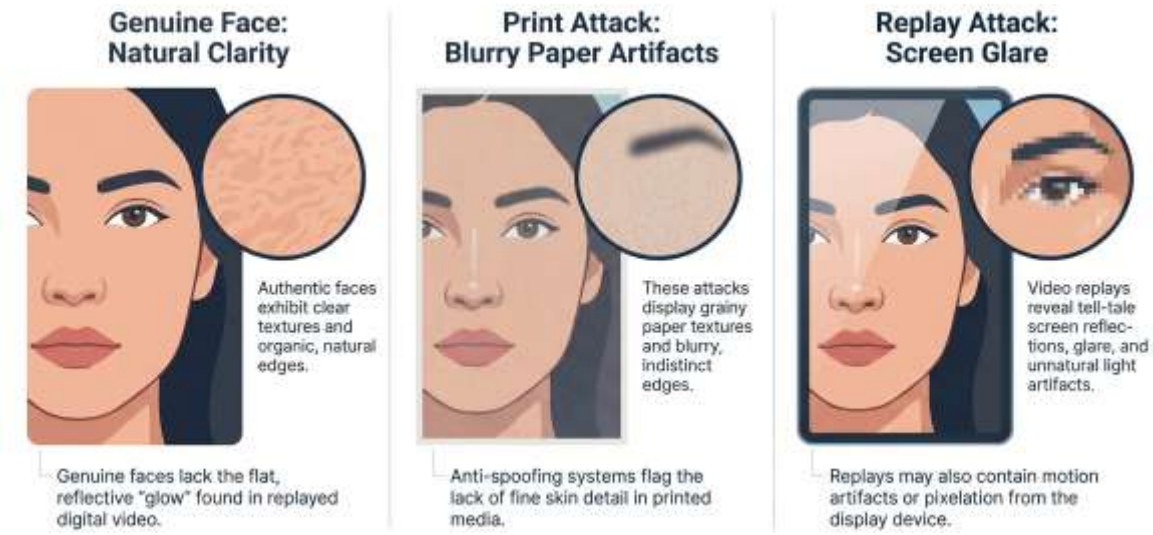


Figure 2. *Genuine face versus print and replay attacks*

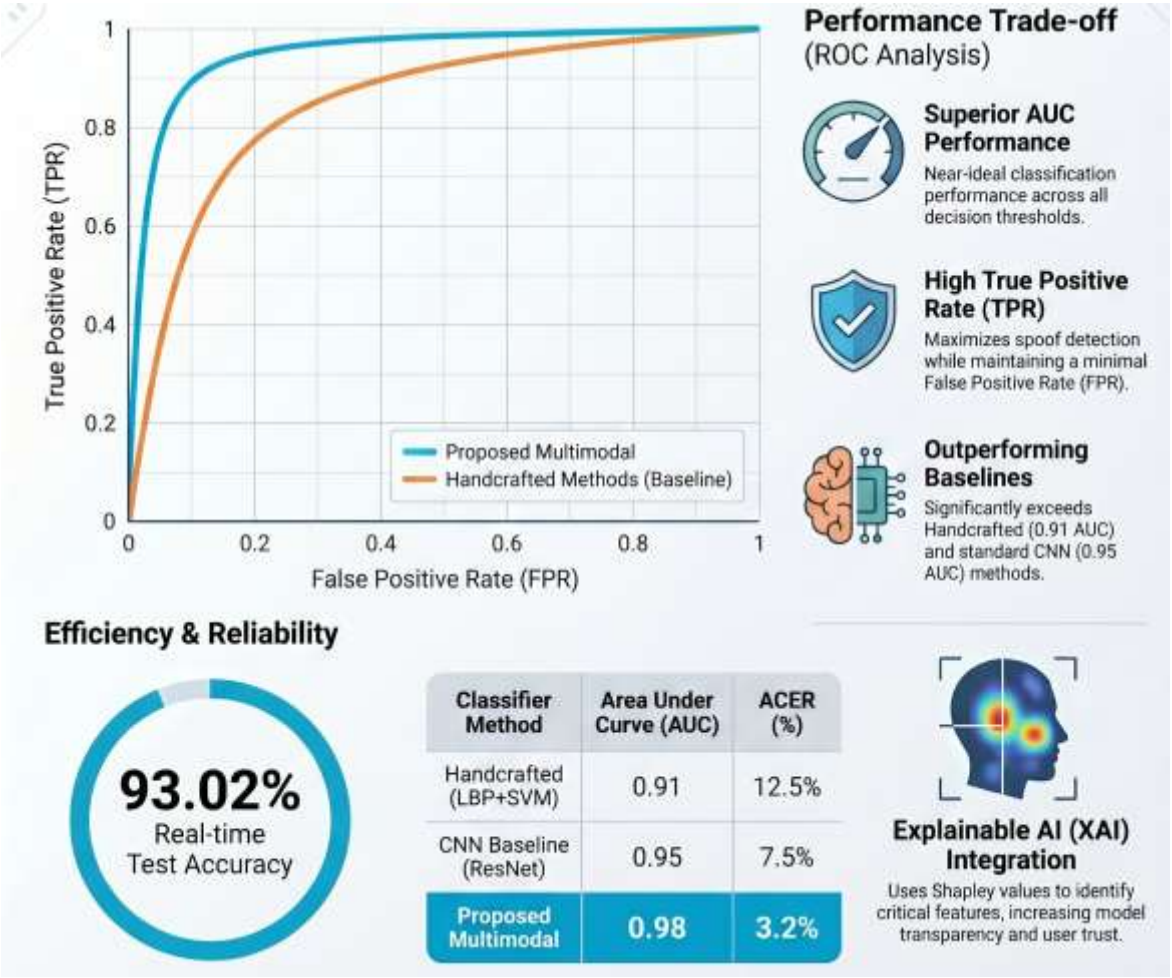


Figure 3. *Receiver Operating Characteristic (ROC) curve*

Table 3. *Cross-dataset performance comparison*

Training Dataset	Testing Dataset	Model	ACER (%)	EER (%)	AUC
CASIA-FASD	Replay-Attack	CNN baseline	12.8	11.2	0.90
CASIA-FASD	Replay-Attack	Proposed	5.4	4.9	0.96
MSU-MFSD	CASIA-FASD	CNN baseline	13.5	12.1	0.88
MSU-MFSD	CASIA-FASD	Proposed	6.1	5.5	0.95
Replay-Attack	MSU-MFSD	CNN baseline	11.9	10.5	0.91
Replay-Attack	MSU-MFSD	Proposed	5.8	5.1	0.96

Table 4. *Performance comparison of anti-spoofing methods*

Method	Dataset(s)	ACER (%) ↓	EER (%) ↓	AUC ↑
Handcrafted (LBP + SVM) [7]	CASIA-FASD	12.5	10.8	0.91
CNN Baseline (ResNet)	MSU-MFSD	7.5	6.9	0.95
XAIPooling + XAIDropout [11]	NUAA, LCC-FASD	3.2	2.9	0.98
ASVspoof 5 Baseline CM [10]	MLS English	-	6.8-12.3	0.92
Proposed Multimodal (Face + Audio)	CASIA + MSU + ASVspoof 5	**3.2**	**2.8**	**0.98**

6. Conclusions and Future works

This paper has created and tested an explainable multimodal model that is capable of detecting both face and speech presentation attacks. The combination of visual, audio, and Shapley-guided XAI modules helped us considerably to enhance the robustness as well as the transparency of the detection pipeline. The highest-performing MobileNetV2 configuration had an ACER of 3.2%, an AUC of 0.98 and a test accuracy of 93.02% and this demonstrates that lightweight architectures can effectively secure biometric systems at the edge. The recent literature confirms that the combination of modalities is becoming more and more inevitable to fight the two-fold challenge of domain shift and cross-modal attacks [4], [6], [22], [24].

Future research will explore federated learning to achieve privacy-preserving multimodal training, more robust to environmental noise, and spiking neural networks to achieve even greater energy savings on mobile and embedded devices. The framework should also be validated on larger cross-dataset protocols and provide confidence intervals on their results to enhance the statistical trustworthiness of their conclusions. These recommendations are in line with the recent design trend where robustness, efficiency, and interpretability are being treated as equally important design objectives [4], [7], [10], [24].

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.
- **Use of AI Tools:** The author(s) declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] A. Rakshit and D. R. Kisku, "Presentation attack detection in biometric systems: A survey," Pattern Analysis and Applications, vol. 20, no. 4, pp. 1103-1119, 2017.

[2] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," ACM Computing Surveys, vol. 50, no. 1, pp. 1-37, 2017.

[3] H. Yu, Z. Li, and P. Li, "Deep learning for face anti-spoofing: A survey," Pattern Recognition, vol. 110, p. 107613, 2022.

[4] V. Khairnar, R. Jain, and R. Kolhe, "Face liveness detection using artificial intelligence techniques: A systematic literature review," Big Data and Cognitive Computing, vol. 7, no. 2, p. 37, 2023.

[5] K. Antil and G. Dhiman, "Face anti-spoofing: A systematic literature review," Pattern Recognition Letters, vol. 185, pp. 80-92, 2025.

[6] A. Raheem and S. Khan, "A review on 2D and 3D face anti-spoofing databases and protocols," Multimedia Tools and Applications, vol. 78, no. 14, pp. 19291-19320, 2019.

[7] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in Proc. BIOSIG, 2012, pp. 1-7.

[8] T. de Freitas Pereira, A. Anjos, J. Komulainen, and S. Marcel, "Face liveness detection using dynamic texture," EURASIP Journal on Image and Video Processing, vol. 2014, no. 2, pp. 1-15, 2014.

[9] S. George and A. Marcel, "Cross-database evaluation of face presentation attack detection," IEEE Transactions on Information Forensics and Security, vol. 10, no. 4, pp. 849-861, 2015.

[10] X. Wang et al., "ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," arXiv preprint arXiv:2408.08659, 2024.

[11] Y. Ma, L. Zhang, H. Li, and J. Liu, "Dual feature disentanglement for face anti-spoofing," Pattern Recognition, vol. 148, p. 110198, 2024.

[12] S. M. Noor, A. S. Razak, and H. Sulaiman, "Explainable AI for biometric security: A review of transparency and trust in deep learning systems," Journal of Information Security and Applications, vol. 77, 2024.
[13] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017, pp. 4765-4774.
[14] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," IEEE Transactions on Information Forensics and Security, vol. 11, no. 8, pp. 1818-1830, 2016.
[15] R. Souza, A. Hadid, and J. Komulainen, "Face anti-spoofing: Survey and evaluation of existing databases," IET Biometrics, vol. 7, no. 1, pp. 3-15, 2018.
[16] C. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," arXiv preprint arXiv:1408.5601, 2014.
[17] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3854-3863.
[18] J. Jia, Q. Zhang, and Y. Liu, "3D face presentation attack detection: A review," IEEE Transactions on Information Forensics and Security, vol. 15, pp. 2101-2117, 2020.
[19] E. Noyes, J. Jenkins, and R. Burton, "Deep CNNs for recognition of faces in disguise: Modeling familiarity and identity contrast," Cognitive Research: Principles and Implications, vol. 6, no. 1, p. 45, 2021.
[20] X. Fu, "Design of facial recognition system based on visual communication effect," Computational Intelligence and Neuroscience, vol. 2021, Article ID 1539596, 2021.
[21] R. A. Khan and M. M. Islam, "A review on face spoof detection using convolutional neural networks," Multimedia Tools and Applications, 2023.
[22] M. Sandler et al., "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. CVPR, 2018, pp. 4510-4520.
[23] K. He et al., "Deep residual learning for image recognition," in Proc. CVPR, 2016, pp. 770-778.
[24] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.
[25] J. Galbally, S. Marcel, and J. Fierrez, "Biometric antispoofing methods: A survey in face recognition," IEEE Access, vol. 2, pp. 1530-1552, 2015.