standards

MDPI

Article

Defending Against Executive Deepfake Fraud: A Hybrid Detection Framework and Governance Strategy for Enterprise Security

Oladipo Ajibawo [1], Nasr Abosata [1], Rejwan Bin Sulaiman [2,*] and Maruf Farhan [3]

[1] Department of Science and Engineering, School of Computing, University of Derby, Derby DE22 1GB, UK; o.ajibawo1@unimail.derby.ac.uk (O.A.); n.abosata@derby.ac.uk (N.A.)
[2] Department of Computer Science and Engineering, Northumbria University, London E1 7HT, UK
[3] Department of Computer Science and Engineering, Ulster University, London EC1R 4TF, UK; m.farhan@ulster.ac.uk
* Correspondence: rejwan.sulaiman@northumbria.ac.uk

Abstract

Deepfake technology has moved well beyond research laboratories, with synthetic video and audio now being deployed to commit corporate fraud, impersonate executives, and manipulate employees at scale. Most detection systems in the research literature fail in three recurring ways: they perform well in controlled tests but poorly in real deployment; they handle only one modality at a time; and they require computing resources that most organisations do not possess. This research addresses these problems through a detection framework built on a fine-tuned EfficientNetV2-S visual backbone combined with three independent Radial Basis Function (RBF) Support Vector Machine (SVM) classifiers, covering visual, audio, and combined audio–visual detection, with the classification stage adding 2.3 ms per frame to a pipeline that requires 195 ms per frame on a single CPU thread. Evaluation followed the official FaceForensics++ source-video split, with all partitioning performed at video level. On FaceForensics++ C23, the visual classifier achieved an AUC-ROC (Area Under the Receiver Operating Characteristic Curve) of 0.9848 (95% CI 0.9787 to 0.9898) at the frame level and 0.9937 at the video level, with an accuracy of 95.53%. On a held-out DeepFakeDetection subset comprising unseen-manipulation methods and subjects, AUC-ROC was 0.9689. Cross-dataset evaluation on Celeb-DF v2 without retraining produced an AUC-ROC of 0.6589, a drop of 32.59 percentage points, with 65.6% of real frames flagged as fake at the inherited operating threshold; a recalibration experiment shows that no operating point on that curve is usable. On the full ASVspoof 2019 Logical Access evaluation partition, which contains thirteen attack types absent from training, the audio classifier reached an equal error rate of 16.55% (95% CI 14.70 to 18.40), ranging from 0.36% on the most detectable attack to 33.43% on the least. A controlled ablation on FakeAVCeleb v1.2, using identity-disjoint partitions and separate labels for each modality, showed a visual-only AUC-ROC of 1.0000 and audio-only of 0.9884. Score-level fusion reduced AUC-ROC to 0.9776 (DeLong $p = 0.016$), because averaging discards information when video and audio are manipulated independently. A governance framework is presented that maps each technical limitation to a specific organisational control, grounded in proportionality, independence, and adaptability and aligned to ISO/IEC 27001:2022 Annex A, the National Institute of Standards and Technology (NIST) Cybersecurity Framework 2.0 and the NIST AI Risk Management Framework. The detection results establish that this governance layer cannot depend on detector output once the deployment domain differs from the training domain.

Check for updates

Academic Editor: Peter Glavič

Received: 31 July 2026
Revised: 9 October 2026
Accepted: 12 October 2026
Published: 4 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: deepfake detection; EfficientNetV2-S; Support Vector Machine; RBF kernel; multimodal detection; audio–visual fusion; enterprise cybersecurity; FaceForensics++; ASVspoof; FakeAVCeleb

1. Introduction

The proliferation of generative artificial intelligence has fundamentally changed the landscape of digital deception. Deepfake technology, which includes synthetic video, audio, and multimodal content produced using deep learning architectures, has transitioned from a research curiosity into a deployed instrument of corporate fraud, identity theft, and social engineering at scale. Generative Adversarial Networks (GANs) and, increasingly, diffusion-based models now enable the production of hyperrealistic synthetic media from minimal source material, placing enterprise communication security under significant and growing pressure.

The operational consequences of this shift are quantifiable. As of 2024, approximately 49% of global businesses had been subjected to audio or video deepfake fraud, compared with 29% just two years earlier [1]. A documented incident in 2024 resulted in a multinational organisation losing approximately 25.6 million US dollars following a fraudulent video conference in which all visible participants, including an apparent senior financial officer, were synthesised deepfake constructs [1]. This case illustrates not only the technical capability of modern deepfake systems but also the extent to which organisational processes have failed to adapt to a threat that exploits cognitive biases and remote communication norms.

Current detection research, whilst technically sophisticated, exhibits three persistent structural deficiencies: a critical disconnect between benchmark accuracy and real-world performance under compression and distortion [2]; the absence of integrated multimodal detection pipelines combining visual and audio signals [3,4]; and the computational impracticality of state-of-the-art architectures in enterprise deployment contexts [5]. These gaps motivate the hybrid detection framework and governance strategy developed in this research.

This paper makes four contributions, reported separately and not claimed to be jointly sufficient for unsupervised enterprise deployment. First, a fine-tuned EfficientNetV2-S and three-stream RBF SVM detection framework is proposed and empirically validated across four public benchmark datasets. Second, an ablation indicates the contribution of end-to-end fine-tuning: replacing the fine-tuned backbone with frozen ImageNet features reduces AUC-ROC from 0.9848 to 0.7192 and increases the number of support vectors from 1110 to 16,249, identifying distribution-specific adaptation rather than architecture as the dominant factor. Third, a controlled ablation study on FakeAVCeleb v1.2 quantifies the marginal contribution of each modality and shows that symmetric fusion degrades performance when video and audio are manipulated independently of one another. Fourth, a governance framework grounded in gap-to-control mapping translates each identified technical limitation into a specific organisational countermeasure, aligned to ISO/IEC 27001:2022 Annex A [6], NIST CSF 2.0 [7] and the NIST AI Risk Management Framework [8], and examined through a structured walk-through of a documented fraud case.

The novelty of this work lies not in the combination of deep feature extraction with a conventional classifier, which is well established, but in what the evaluation protocol makes visible. By holding out an entire manipulation subset, evaluating across datasets and against a generation family absent from training, and reporting each result under the official split of the dataset concerned, the study separates two failure modes that are usually reported as

one. Performance is retained almost entirely across unseen-manipulation methods and lost almost entirely across a change in capture and encoding pipeline. That distinction, rather than any single accuracy figure, is what the governance framework is built on.

The evaluation in this paper is benchmark-based. All four datasets are public research collections built from YouTube, celebrity interview and read-speech recordings, and no production video-conference material was tested. The findings on executive and enterprise fraud should therefore be read as indicative. Video-conference footage, with its own codecs, compression and network artefacts, is the type of input most likely to trigger the cross-dataset failure identified in Section 4.3, so any deployment would need to be validated on that material first.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature. Section 3 presents the research methodology and framework design. Section 4 reports the experimental results and comparative analysis. Section 5 provides a critical discussion of the findings and governance recommendations. Section 6 presents the conclusions and future works.

2. Related Work

2.1. Evolution of Deepfake Technology

Deepfake technology traces its technical origins to early digital manipulation systems, including Video Rewrite in 1997 and the Active Appearance Model in 2001, which formed the foundations upon which modern synthesis pipelines were built [1]. The key technical inflection point came in 2014 when Goodfellow introduced Generative Adversarial Networks, which became the dominant architecture underpinning deepfake creation [9,10]. Subsequent Generative Adversarial Network (GAN) variants, including Wasserstein GANs, Progressive Growing GANs, and StyleGAN, progressively improved visual fidelity and temporal consistency. Diffusion models, which iteratively refine noise into high-quality outputs, have since demonstrated superior training stability and sample diversity, as shown by systems such as Stable Diffusion [9]. Transformer-based architectures have extended generative capability into audio synthesis and voice cloning from minimal source material, with direct implications for enterprise fraud [10].

The literature consistently characterises deepfake development as an arms race in which each generative advance erodes the effectiveness of existing detection methods [10,11]. A further concern is model collapse through recursive training on synthetic data, which risks long-term degradation in both generative diversity and detection accuracy [10].

2.2. Deepfakes as a Cyber Threat Vector

Deepfake-enabled cyber threats span four principal attack types: executive impersonation, voice cloning, video manipulation, and social engineering. Voice cloning attacks use short audio samples to generate synthetic speech that deceives both human recipients and automated voice authentication systems [1]. These attack types are increasingly deployed in combination, with multimodal deepfakes simultaneously manipulating audio, video, and text, representing the most sophisticated and hardest-to-detect threat category [10].

A frequently underestimated dimension of deepfake attacks is their use of fundamental cognitive biases. Authority bias, which is the tendency to comply with instructions from perceived figures of power, can be systematically used in executive impersonation attacks where synthetic replication of a known leader reduces a recipient's inclination to question the legitimacy of a request [1]. Deepfake attacks do not need to be technically perfect to succeed; they need only be sufficiently convincing to circumvent immediate scepticism, a threshold current generative AI routinely meets. At the organisational level, widespread remote working has expanded the attack surface where identity cannot be verified, and hierarchical

authorisation workflows encourage employees to act on apparent senior instructions without secondary verification [11].

2.3. Deepfake Detection Methods

Deepfake detection methods have developed rapidly alongside advances in generative AI, spanning traditional machine learning, deep learning architectures, and hybrid combinations. These methods differ considerably in accuracy, computational demands, and ability to generalise across real-world conditions [12–14].

Rana et al. [13] observed that ML-based methods can achieve up to 98% accuracy when training and test sets share similar distributions: Logistic Regression achieved 100% accuracy with an AUC of 1.00, Random Forest reached 99.95% using frequency-domain features, and a hybrid DenseNet+SVM classifier achieved 98% on DFDC, outperforming more complex CNN-LSTM alternatives. However, accuracy dropped to approximately 50% under distributional shift, revealing that ML classifiers learn generation-specific artefacts rather than generalised indicators of manipulation.

Deepa et al. [15] clarified the source of this fragility by benchmarking handcrafted-feature pipelines against fine-tuned deep architectures. GLCM-based SVM reached only 56.8% and the strongest ensemble (HOG with XGBoost and SVM) peaked at 84.62%, whereas a fine-tuned EfficientNet-B3 achieved 97.95% accuracy with balanced precision and recall of 0.98, confirming that the deficiency lay in the feature representation rather than the classifiers themselves. However, the study was confined to still images from a single non-benchmark dataset, leaving temporal, multimodal, and deployment considerations unaddressed.

Consistent with this, Alzurfi and Altaei [16] combined CNN-based feature extraction with traditional ML classifiers including SVM, KNN, Random Forest, and Logistic Regression across four FaceForensics++ manipulation types. SVM achieved the highest accuracy of 96% on the Deepfake subset, but accuracy fell sharply to 64% on Neural Texture manipulation, exposing a vulnerability of spatial feature-based approaches to localised manipulations.

Satoto et al. [17] extended this hybrid direction with efficiency as an explicit objective, pairing MobileNetV2 as a lightweight feature extractor with an SVM classifier in place of the conventional softmax layer. The model achieved 94.8% accuracy with an F1-score of 94.5% at a model size of only 16 MB, outperforming substantially larger architectures such as VGG16 (92.6% at 528 MB), demonstrating that decoupling feature extraction from classification yields competitive accuracy at reduced cost. However, evaluation was confined to a single non-benchmark image dataset in the visual modality, without cross-dataset, compression, or perturbation testing.

Regarding the feature extractor, Deng et al. [18] applied EfficientNetV2 to deepfake video detection, motivated by its fused-MBConv layers and compound scaling. After frame extraction and face-cropping preprocessing, the streamlined network was trained and validated on the FaceForensics++ and FFIW10K datasets, achieving 97.90% accuracy on FaceForensics++ and surpassing established detectors including XceptionNet (96.36%) and MesoNet (87.27%), establishing EfficientNetV2 as an effective and efficient extractor for deepfake detection specifically. However, evaluation relied on a random intra-dataset split with built-in softmax classification in the visual modality alone, without cross-dataset, compression, or audio-driven manipulation testing.

Coccomini et al. [19] further supported convolutional backbones through a cross-forgery analysis of EfficientNetV2 and ViT-B on ForgeryNet. EfficientNetV2 achieved higher training accuracy and stronger performance on seen manipulations, while the ViT showed less bias toward specific forgery techniques, suggesting complementary strengths. However,

the study addressed only static images, and the transformer's generalisation advantage came at substantially greater computational cost.

That computational burden has itself become a research target. Rani et al. [20] integrated a Vision Transformer with a Linformer, whose low-rank projection replaces quadratic self-attention with linear complexity. On landmark-centred patches from 1000 Celeb-DF v2 videos, the hybrid achieved 95.25% accuracy with an AUC of 0.989 at 883 GFLOPs against 1963 for the standard ViT, roughly halving computational cost. However, evaluation was confined to a single dataset and the visual modality, and its generalisation claims rest on loss-curve stability rather than testing under distributional shift.

Similarly, AlMuhaideb et al. [21] proposed LightFakeDetect, combining a MobileNet v1 backbone with a Convolutional Block Attention Module and a Gated Recurrent Unit. The model achieved 98.2% accuracy on Celeb-DF v2 and 95.0% on DFDC at only 63.2 GFLOPs, demonstrating that lightweight architectures can rival heavier detectors. However, the framework remains strictly unimodal, with the authors identifying audio–visual fusion as future work, and its balanced accuracy drops to 82.9% on DFDC.

Where efficiency-oriented work has neglected robustness evaluation, a separate line of research confronts it directly. Amin et al. [22] proposed the SpectraVisionFusion Transformer (SVFT), fusing spatial features with frequency-domain spectral statistics to improve generalisation. Trained on FaceForensics++ and tested on DFDC, Celeb-DF, and DeeperForensics, it achieved 92.57% cross-manipulation and 80.63% cross-dataset accuracy, but average AUC fell from 85.09% under c23 compression to 78.50% under c40, with further degradation under unseen perturbations. Despite its multimodal label, the fusion operates entirely within the visual modality.

Yang et al. [23] likewise targeted cross-dataset degradation with Crossed, a Deep Information Decomposition framework on an EfficientNetV2-L backbone that isolates deepfake-relevant features through adversarial mutual information minimisation. Trained on FaceForensics++, it achieved an intra-dataset AUC of 0.970 but only 0.779 on the unseen Celeb-DF v2, confirming the severity of cross-dataset degradation. The framework remains frame-level and visual-only, omits compression testing, and relies on a computationally heavy backbone.

Taken together, the reviewed literature exposes three interconnected limitations. First, benchmark performance does not survive real-world conditions: accuracy collapses under distributional shift [13], cross-dataset AUC falls by twenty points or more even in generalisation-focused designs [23], and compression alone erodes performance substantially [2,22,24]. Second, detection remains confined to the visual stream: every framework reviewed operates on visual features alone, and the authors of the most capable lightweight detector explicitly identify audio–visual fusion as unrealised future work [3,4,21]. Third, the architectures closest to closing these gaps carry computational costs that impede enterprise deployment, while lightweight alternatives sacrifice robustness or modality coverage [5,25]. No reviewed approach achieves robustness, multimodality, and efficiency jointly, and it is this intersection the present research targets: a hybrid framework pairing an efficient EfficientNetV2-S feature extractor with SVM classification across audio, visual, and fused audio–visual feature streams, evaluated under the real-world conditions the literature has shown to be decisive.

3. Materials and Methods

3.1. Proposed Hybrid Detection Framework

This research proposes a hybrid deepfake detection framework that pairs a fine-tuned EfficientNetV2-S deep feature extractor with three specialised Support Vector Machine classifiers operating across visual, audio, and fused audio–visual feature streams. Figure 1

presents the complete system architecture, illustrating data flow from input datasets through preprocessing, feature extraction, normalisation, SVM classification, and evaluation across the three parallel branches.

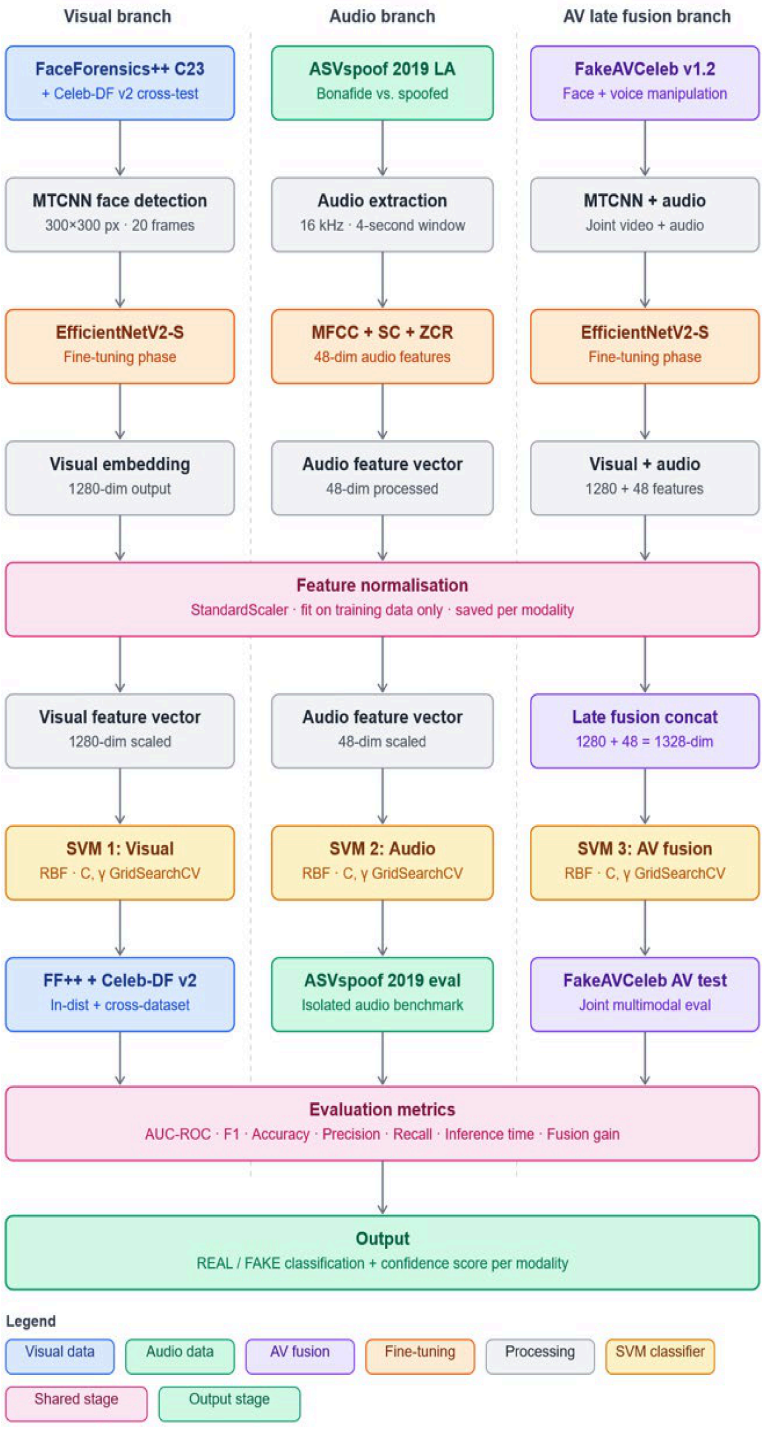


Figure 1. Hybrid deepfake detection system architecture. Three parallel branches address visual detection (SVM 1), audio detection (SVM 2), and audio–visual late-fusion detection (SVM 3).

The hybrid EfficientNetV2-S and SVM design draws on the complementary strengths of both components: EfficientNetV2-S serves as a deep feature extractor that adapts convolutional representations to manipulation-specific artefact signatures through fine-tuning, while the SVM provides a well-regularised, margin-maximising decision boundary that is computationally efficient at inference. Das et al. [25] provided direct empirical sup-

port, showing that a DenseNet+SVM combination achieved 98% accuracy on the Deepfake Detection Challenge (DFDC) dataset, outperforming CNN-LSTM at 97.6% at lower computational cost.

3.2. Dataset Selection and Justification

Four datasets were selected based on five principal criteria: modality coverage, generation technique diversity, real-world compression conditions, benchmark prevalence, and public availability. Table 1 summarises the role, composition, and SVM target of each dataset.

Table 1. Dataset summary: roles, compositions, and SVM targets.

Dataset	Modality	Role	Samples	Balance	SVM Target
FaceForensics++ (C23)	Video	Primary training and in-distribution test	86,388 train/ 16,792 val/16,796 test frames	Test: 2799 real/ 13,997 fake	SVM 1 Visual
Celeb-DF v2	Video	Cross-Dataset Test	10,357 test frames (518 videos)	3557 real/6800 fake	SVM 1 Cross-dataset
ASVspoof 2019 (LA)	Audio	Audio Branch	25,380 training/ 71,237 evaluation utterances	Eval: 7355 bonafide/ 63,882 spoofed	SVM 2 Audio
FakeAVCeleb v1.2	Audio–Visual	Multimodal Fusion	1000 clips (704 train/ 153 val/143 test)	Test: 86 real/57 fake (visual label)	SVM 3 AV Fusion

FaceForensics++ (C23), introduced by Rossler et al. [26], serves as the primary training dataset for SVM 1 and covers five manipulation subsets: Deepfakes, Face2Face, FaceSwap, NeuralTextures, and FaceShifter. The DeepFakeDetection subset distributed with FaceForensics++ was excluded from training entirely and retained as a held-out unseen-manipulation test set, reported in Section 4.3. Celeb-DF v2, introduced by Li et al. [27], provides the cross-dataset generalisation test for SVM 1 and is evaluated without retraining. ASVspoof 2019 Logical Access [28] provides synthetic speech detection across text-to-speech synthesis and voice conversion attacks. FakeAVCeleb v1.2, introduced by Khalid et al. [29], serve as the test dataset for the multimodal late-fusion pipeline and it is a benchmark dataset where both facial synthesis and voice manipulation occur simultaneously within the same video clip.

All partitioning was performed before any frame was extracted, at the level of the source video for FaceForensics++, the speaker for ASVspoof 2019 and the identity for FakeAVCeleb, so that no frame drawn from a given source video, manipulation of that source video, or subject appears in more than one partition. For FaceForensics++ C23 the official train, validation and test split files released with the dataset were used, dividing the 1000 source videos into 720, 140 and 140 respectively; with five manipulations of each source video this gives 4320 training, 840 validation and 840 test videos, from which 86,388, 16,792 and 16,796 face crops were obtained. For Celeb-DF v2 the official testing list of 518 videos was used. For ASVspoof 2019 Logical Access the official speaker-disjoint training and evaluation partitions were used unchanged; the evaluation partition contains thirteen attack algorithms (A07 to A19) that do not appear in the training partition. For FakeAVCeleb v1.2, 1000 clips were sampled across all four categories and partitioned 704, 153 and 143 by identity, with no identity present in more than one partition. Backbone fine-tuning, early stopping and hyperparameter selection used the training and validation partitions only; each test partition was evaluated once.

3.3. Preprocessing and Feature Extraction

3.3.1. Visual Feature Extraction and Fine-Tuning

Visual feature extraction uses EfficientNetV2-S as the deep learning backbone, with facial regions first isolated from raw video frames using Multi-task Cascaded Convolutional Networks (MTCNN) as proposed by Zhang et al. [30]. Face crops are resized to a standardised 300 × 300-pixel input dimension. Up to twenty frames are sampled evenly per video clip. For FaceForensics++, Celeb-DF v2 and the DeepFakeDetection holdout, classification is performed on individual frame embeddings and video-level scores are obtained by averaging the frame scores of a clip; both frame-level and video-level results are reported in Section 4. For FakeAVCeleb, where the unit of analysis is the clip, the mean embedding across the detected frames of a clip forms a single clip-level representation. Face detection succeeded on 99.4 to 100% of sampled frames across the video datasets.

EfficientNetV2-S, introduced by Tan and Le [31], achieves 83.9% top-1 accuracy on ImageNet with 21.5 million parameters and trains five to eleven times faster than comparably accurate models. The backbone is not used as a static frozen feature extractor. End-to-end fine-tuning is applied using a differential learning rate schedule: 1×10^{-4} for pretrained backbone layers and 1×10^{-3} for the newly initialised classification head. The AdamW optimiser with weight decay of 1×10^{-4} is used throughout, with cosine annealing scheduling over a maximum of five epochs and early stopping on validation AUC-ROC with patience of three epochs. Following fine-tuning, the classification head is replaced with an identity layer to expose the 1280-dimensional global average pooling output as the embedding passed to the SVM classifier.

3.3.2. Audio Feature Extraction

Audio feature extraction is based on Mel-Frequency Cepstral Coefficients (MFCCs), which capture the short-term spectral characteristics of audio signals in a perceptually scaled representation. Two supplementary features are included: spectral contrast, which captures differences in spectral peaks and valleys across seven sub-bands, and zero-crossing rate (ZCR), which measures the frequency at which the audio signal crosses the zero-amplitude axis. These three feature types form a 48-dimensional audio feature vector comprising 40 MFCCs, seven spectral contrast values, and one ZCR value. Audio is extracted at 16,000 Hz across a fixed 4 s analysis window, with shorter samples zero-padded to ensure consistent dimensionality.

3.3.3. Audio–Visual Feature Extraction

SVM 3 implements late-fusion multimodal detection on FakeAVCeleb v1.2. The visual backbone is separately fine-tuned on FakeAVCeleb training data, and the resulting 1280-dimensional embeddings are concatenated with 48-dimensional audio feature vectors to form a unified 1328-dimensional fused feature vector. Each modality is independently normalised using a StandardScaler fitted only on training data before concatenation.

3.4. Classification: Three-SVM Design

Table 2 summarises the three classifiers: the modality each operates on, the dataset on which it is trained and evaluated, its input dimension, and its role in the evaluation.

Table 2. Three-SVM framework: modality, dataset, input dimension, and evaluation role.

SVM	Purpose	Training Data	Input Dim.	Evaluation
SVM 1	Visual Deepfake Detector	FaceForensics++ (C23), fine-tuned EfficientNetV2-S	1280-dim embeddings	FF++ in-distribution + Celeb-DF v2 cross-dataset
SVM 2	Audio Deepfake Detector	ASVspoof 2019 (LA)	48-dim (MFCC + SC + ZCR)	ASVspoof 2019 evaluation split
SVM 3	Audio–Visual Late Fusion	FakeAVCeleb v1.2 (visual + audio)	1328-dim (1280 + 48)	FakeAVCeleb held-out test set

SVM 1 and SVM 2 operate as unimodal detectors, trained respectively on the visual and audio streams defined in Section 3.3, with SVM 1 additionally evaluated on Celeb-DF v2 without retraining to measure cross-dataset generalisation. SVM 3 performs late-fusion multimodal detection on the 1328-dimensional fused representation constructed in Section 3.3.3. All three classifiers employ an RBF kernel, with regularisation parameter C in {0.1, 1, 10, 100} and kernel bandwidth gamma in {scale, auto, 0.001, 0.01} selected on the validation partition using the scikit-learn implementation [32]. Selection on cross-validation folds drawn from the training partition was found to be uninformative in this setting, because the fine-tuned backbone separates its own training embeddings almost perfectly and every candidate configuration scored above 0.999; validation-based selection is therefore used throughout, and the test partition is evaluated once.

The classification stage uses a Support Vector Machine with a Radial Basis Function kernel. An SVM constructs the decision boundary that maximises the margin between the two classes, and the soft-margin formulation of Cortes and Vapnik [33] introduces a regularisation parameter C governing the trade-off between margin width and the penalty applied to misclassified training points. The RBF kernel maps the embeddings into an implicit higher-dimensional space in which classes that are not linearly separable may become separable, and its bandwidth parameter gamma sets the radius of influence of each support vector: small values produce broad, near-linear boundaries, while large values produce narrow boundaries that can overfit. Selecting this bandwidth is a long-standing problem in Radial Basis Function methods and remains an active subject of study in the optimisation literature [34]. Two conditions follow from the choice of kernel. First, the RBF kernel is not scale invariant, so each feature stream is standardised with a StandardScaler fitted on the training partition only and applied unchanged to validation and test data. Second, the ratio of training samples to feature dimensions must be monitored: for the 1328-dimensional fused representation of Section 3.3.3 this ratio is approximately 0.53, so the fused classifier is fitted on fewer training samples than it has dimensions, and Section 5.3 interprets the ablation result in that light. The implementation uses the scikit-learn SVC estimator [32], which wraps LIBSVM [35].

The selected configurations were C = 10 with gamma = 1×10^{-4} and balanced class weights for SVM 1, giving 1110 support vectors (204 for the real class and 906 for the fake class); C = 1 with gamma = 0.01 and balanced class weights for SVM 2; and, for the three FakeAVCeleb classifiers, C = 0.01 with gamma = scale for the visual and fused branches and C = 10 with gamma = 0.01 with balanced class weights for the audio branch. The corresponding support vector counts are 2394 for SVM 2 and, for the FakeAVCeleb classifiers, 420 for the visual branch, 190 for the audio branch and 439 for the fused branch. Probability calibration was not applied; AUC-ROC is computed from decision function values. Section 4.2 reports an ablation comparing this configuration with a linear kernel, with the softmax classification head retained from fine-tuning, and with a frozen ImageNet backbone.

3.5. Evaluation Metrics

Model performance is assessed using the metrics below. TP denotes true positives, TN true negatives, FP false positives, and FN false negatives throughout. The formal definitions are as follows.

Accuracy measures the overall proportion of correctly classified samples across both real and fake classes [36]:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN), \quad (1)$$

where TP, TN, FP, FN are true positives, true negatives, false positives, and false negatives, respectively.

Precision measures the proportion of samples predicted as fake that are genuinely fake; high precision reflects a low false alarm rate, which is operationally critical to avoid alert fatigue [37]:

$$Precision = TP/(TP + FP), \quad (2)$$

where the denominator represents all samples classified as fake by the model.

Recall measures the proportion of genuine fake samples correctly detected; high recall reflects a low missed detection rate, which is the primary security objective [36]:

$$Recall = TP/(TP + FN), \quad (3)$$

where FN represents fake samples that the classifier incorrectly classified as real.

The F1 Score is the harmonic mean of precision and recall, penalising asymmetric classifier behaviour on balanced test sets [38]:

$$F1 = 2 \times (Precision \times Recall)/(Precision + Recall), \quad (4)$$

where the harmonic mean penalises low performance on either metric more severely than the arithmetic mean.

AUC-ROC is the Area Under the Receiver Operating Characteristic Curve, which plots true-positive rate (TPR) against false-positive rate (FPR) across all classification thresholds. It is threshold-independent and robust to class imbalance, making it the primary discriminative metric throughout this research [5,39]:

$$TPR = TP/(TP + FN); FPR = FP/(FP + TN), \quad (5)$$

where TPR is plotted on the y-axis and FPR on the x-axis of the ROC curve; AUC is the integral of TPR over FPR.

Fake-detection rate is the proportion of manipulated samples correctly identified, equivalent to recall on the fake class; it is reported separately because missed manipulations are the primary security failure mode in enterprise deployment.

Inference time is reported at two levels. Classification latency is the mean per-sample time taken by the SVM decision stage on pre-extracted features. End-to-end latency additionally includes image loading, preprocessing and the EfficientNetV2-S forward pass, and is the figure relevant to deployment. Both are measured on a single CPU thread without GPU acceleration and are reported in Section 4.6. Equal error rate, the standard metric in the anti-spoofing literature, is additionally reported for the audio branch, and 95% confidence intervals are obtained by bootstrap resampling at the level of the clustering unit, which is the source video for the video datasets and the speaker for ASVspoof 2019, because frames drawn from one video and utterances from one speaker are not independent.

3.6. Implementation and Reproducibility

All experiments were implemented in Python 3.13 (Python Software Foundation, Wilmington, DE, USA) using PyTorch 2.11.0 and torchvision 0.26.0 (PyTorch Foundation, San Francisco, CA, USA), scikit-learn 1.6.1 [32], librosa 0.11.0 and facenet-pytorch 2.6.0, on Google Colab (Google LLC, Mountain View, CA, USA) for MTCNN face detection; exact package versions are listed in the repository given in the Data Availability Statement. Backbone fine-tuning used a single NVIDIA L4 GPU, and all latency measurements were taken on a single CPU thread without GPU acceleration, on a Google Colab runtime with an Intel Xeon CPU at 2.20 GHz (Intel Corporation, Santa Clara, CA, USA) (12 virtual cores) and 52 GB of memory. Random seeds were fixed at 42 for data partitioning, backbone initialisation and hyperparameter selection. The visual backbone was trained once under this seed; variation across backbone initialisations was not quantified.

4. Results

4.1. Aggregate Results

Table 3 presents consolidated performance metrics across all evaluation conditions. All results are reported on held-out test sets under CPU-only inference conditions. The inference column of Table 3 covers the SVM classification stage only, as defined in Section 3.5.

Table 3. Consolidated performance metrics across all evaluation conditions.

Evaluation Condition	AUC-ROC	Accuracy	Bal. Acc.	Fake Det.	Inf. (ms)	Classifier/Features
SVM 1: FF++ C23 (In-distribution)	0.9848	95.53%	0.9373	96.43%	2.34	EfficientNetV2-S (FT) + RBF SVM
SVM 1: Celeb-DF v2 (Cross-dataset)	0.6589	65.49%	0.5809	81.74%	2.34	EfficientNetV2-S (FT) + RBF SVM
SVM 2: ASVspoof 2019 (Audio)	0.9186	83.44%	0.8345	83.44%	—	MFCC + SC + ZCR (48-dim) + RBF SVM
SVM 3: FakeAVCeleb, Visual Only	1.0000	99.30%	0.9912	98.25%	—	EfficientNetV2-S (FT) + RBF SVM
SVM 3: FakeAVCeleb, Audio Only	0.9884	97.90%	0.9776	97.44%	—	MFCC + SC + ZCR (48-dim) + RBF SVM
SVM 3: FakeAVCeleb, AV Late Fusion	1.0000	99.30%	0.9912	98.25%	—	EfficientNetV2-S (FT) + MFCC + RBF SVM

Note: AUC-ROC values are frame-level for the video datasets and clip-level for FakeAVCeleb; video-level values are reported in the text. The inference column reports the SVM classification stage only on pre-extracted features; end-to-end latency is reported in Section 4.6. Audio classification latency was not separately measured in the revised protocol. 95% confidence intervals for AUC-ROC, from 2000 bootstrap resamples, are 0.9787 to 0.9898 for FaceForensics++ C23, 0.6235 to 0.6977 for Celeb-DF v2, 0.9024 to 0.9340 for ASVspoof 2019, and 1.0000 to 1.0000, 0.9688 to 1.0000 and 1.0000 to 1.0000 for the FakeAVCeleb visual, audio and fused conditions. Perfect separation gives a degenerate AUC-ROC interval, so for FakeAVCeleb the accuracy intervals are also reported: 97.90 to 100.00% for the visual and fused conditions and 95.10 to 100.00% for the audio condition. Resampling was by video for the video datasets, by speaker for ASVspoof 2019 and by clip for FakeAVCeleb.

Three patterns emerge from the aggregate results. First, the fine-tuned visual classifier achieves AUC-ROC of 0.9848 at the frame level and 0.9937 at video level on the FaceForensics++ C23 official test partition. Second, performance is retained almost entirely on unseen-manipulation methods but not across datasets: the held-out DeepFakeDetection subset yields an AUC-ROC of 0.9689, a loss of 1.59 percentage points, whereas Celeb-DF v2 yields 0.6589, a loss of 32.59 percentage points. Third, the FakeAVCeleb condition is

saturated: the visual branch reaches AUC-ROC of 1.0000, leaving no headroom within which the contribution of fusion could be assessed on that dataset.

4.2. SVM 1: Visual Detection on FaceForensics++ C23

Under in-distribution conditions on the FaceForensics++ C23 official test partition, which comprises 2799 real and 13,997 fake frames drawn from 840 videos, SVM 1 achieved an AUC-ROC of 0.9848 (95% CI 0.9787 to 0.9898) at the frame level and 0.9937 at the video level, accuracy of 95.53%, balanced accuracy of 93.73% and a fake-detection rate of 96.43%, with a false-positive rate of 8.97% on the real class. Because the test partition is imbalanced by construction, with five manipulations of each source video, balanced accuracy is the more informative summary. Of the 13,997 fake frames across the five manipulation subsets, 499 were not detected and 251 of the 2799 real frames were incorrectly flagged. Figure 2 presents the confusion matrix and ROC curve for this condition.

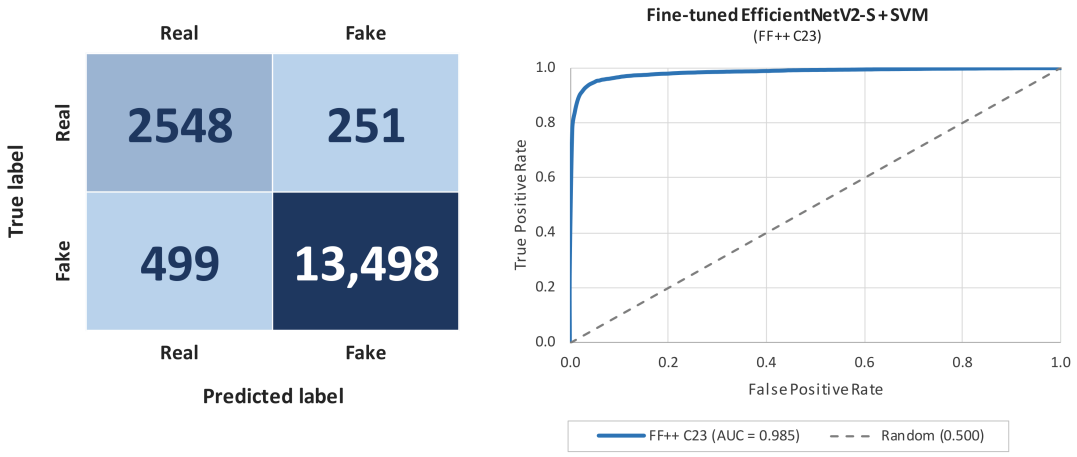


Figure 2. Confusion matrix and ROC curve for the fine-tuned EfficientNetV2-S + R_{FP} SVM visual-only classifier evaluated on FaceForensics++ C23 (frame-level AUC-ROC = 0.9848).

Table 4 compares the proposed configuration with a linear SVM, with the softmax head retained from fine-tuning, and with a frozen ImageNet backbone.

Table 4. Ablation on the FaceForensics++ C23 test partition: backbone adaptation and classifier choice.

Condition	AUC-ROC (Frame)	AUC-ROC (Video)	Balanced Acc.	FPR	Support Vectors
Fine-tuned backbone + RBF SVM (proposed)	0.9848	0.9937	0.9373	0.0897	1110
Fine-tuned backbone + linear SVM	0.9831	0.9924	0.9332	0.1008	—
Fine-tuned backbone + softmax head	0.9850	0.9922	0.9273	0.1168	—
Frozen ImageNet backbone + RBF SVM	0.7192	0.7666	0.6518	0.4845	16,249

All four conditions use the same partitions, the same validation-based selection procedure and a single evaluation of the test partition. The frozen condition used a reduced hyperparameter grid and 30% of the training groups for computational reasons. Its comparison with the fine-tuned conditions is therefore not fully controlled, and the difference should be read as an upper bound on the effect of fine-tuning. The 95% confidence interval for the proposed configuration is 0.9787 to 0.9898; intervals were not computed for the other ablation conditions. The three fine-tuned conditions are separated by less than 0.002 in AUC-ROC, so the RBF kernel cannot be claimed to improve discrimination over a linear kernel or over the softmax head retained from fine-tuning; its advantage is a better

balanced operating point at the default threshold and the operational property that the classifier can be refitted or recalibrated for a new domain in minutes on CPU, without retraining the backbone. The contrast with the frozen backbone is of a different order: adaptation to the target distribution, not the architecture or the classifier, is what produces the reported performance.

4.3. SVM 1: Cross-Dataset Generalisation on Celeb-DF v2

Two evaluations were conducted outside the training distribution. The first used the DeepFakeDetection subset, which was excluded from training entirely and which contains manipulation methods and subjects the model had not seen; evaluated against the real videos of the FaceForensics++ official test partition, it yielded an AUC-ROC of 0.9689 (95% CI 0.9580 to 0.9785) at the frame level and 0.9898 at the video level, with a balanced accuracy of 90.95%. The second applied SVM 1 without retraining to the 518 videos of the official Celeb-DF v2 testing list, yielding an AUC-ROC of 0.6589 (95% CI 0.6235 to 0.6977), accuracy of 65.49% and balanced accuracy of 58.09% across 10,357 test frames. Figure 3 shows the confusion matrix and ROC curve for this evaluation. The asymmetry is severe: 65.6% of real Celeb-DF v2 frames are flagged as fake at the operating threshold inherited from FaceForensics++, while fake recall stands at 81.7%. The difference between the two evaluations is the principal empirical observation of this study, subject to the qualification discussed below. Unseen-manipulation methods cost 1.59 percentage points of AUC-ROC; a change in dataset, and therefore in capture and encoding pipeline, costs 32.59. The binding constraint on generalisation is the video production pipeline rather than the forgery technique. Tables 5 and 6 compare the framework against published baselines. Figures 4 and 5 shows the same comparisons graphically.

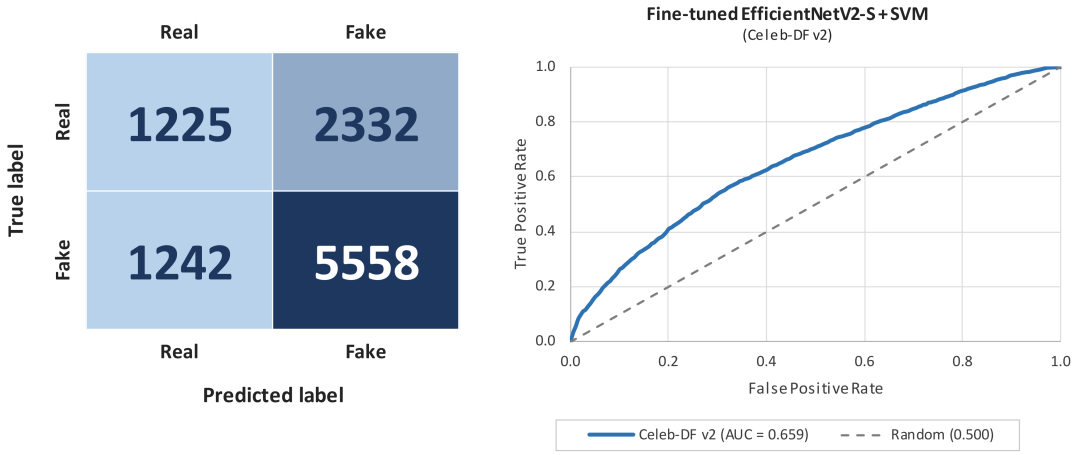


Figure 3. Confusion matrix and ROC curve for the fine-tuned EfficientNetV2-S + RBF SVM visual-only classifier evaluated on Celeb-DF v2 (frame-level AUC-ROC = 0.6589).

Table 5. Comparison of published methods on FaceForensics++ C23.

Method	Architecture	AUC-ROC	Accuracy
F3-Net/Xception [40]	Xception with dual frequency-aware fusion	0.9810	97.52%
Face X-ray [41]	HRNet + blending boundary detection	0.9797	N/R
Multi-Attentional Detection [42]	EfficientNet-B4 + attention (FT)	0.9929	97.60%
XceptionNet [26]	Xception (end-to-end FT)	0.9630	95.73%
This work: EfficientNetV2-S (FT) + RBF SVM	EfficientNetV2-S + RBF SVM	0.9848/0.9937	95.53%

Note: the confidence interval for the present study is from 2000 bootstrap resamples of the clustering unit. Intervals are not available for the published baselines, whose per-sample scores are not distributed, and significance testing against them is therefore not possible. The present result is 0.9848 [0.9787, 0.9898] at the frame level and 0.9937 at the video level.

Table 6. Cross-dataset generalisation: all methods trained on FaceForensics++, evaluated on Celeb-DF v2 without retraining.

Method	Trained on	Celeb-DF v2 AUC	AUC Drop
XceptionNet [27]	FaceForensics++	0.659	~34
Multi-Attentional Detection [42]	FaceForensics++	0.6744	~32
This work: EfficientNetV2-S (FT) + RBF SVM	**FF++ (EfficientNetV2-S FT)**	**0.6589**	**32.59**

Note: the confidence interval for the present study is from 2000 bootstrap resamples of the clustering unit. Intervals are not available for the published baselines, whose per-sample scores are not distributed, and significance testing against them is therefore not possible. The present result is 0.6589 [0.6235, 0.6977].

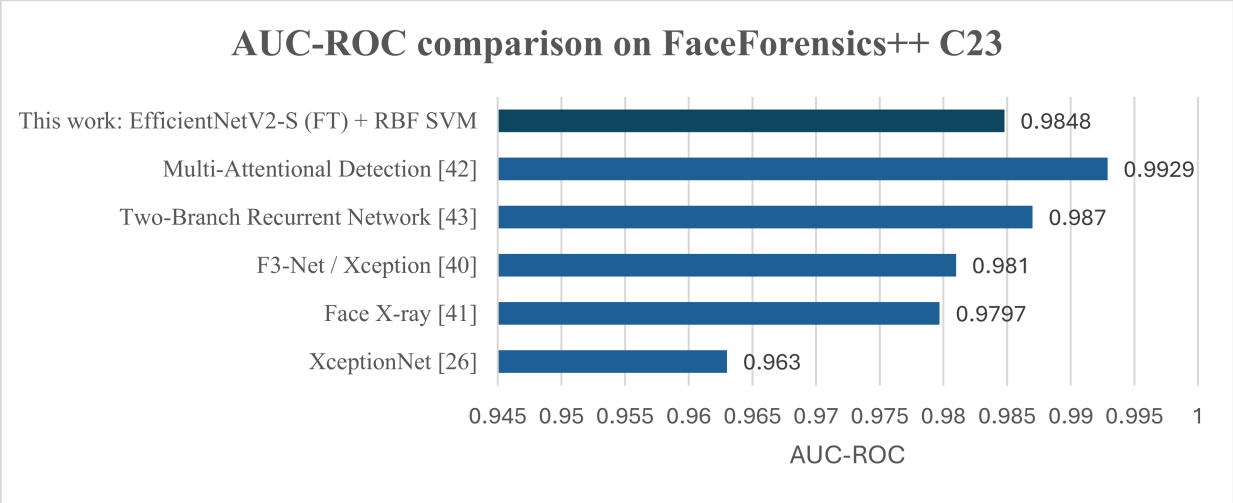


Figure 4. AUC-ROC comparison of the proposed framework against published baselines on FaceForensics++ C23. XceptionNet [26], Face X-ray [41], F3-Net/Xception [40], Two-Branch Recurrent Network [43] and Multi-Attentional Detection [42].

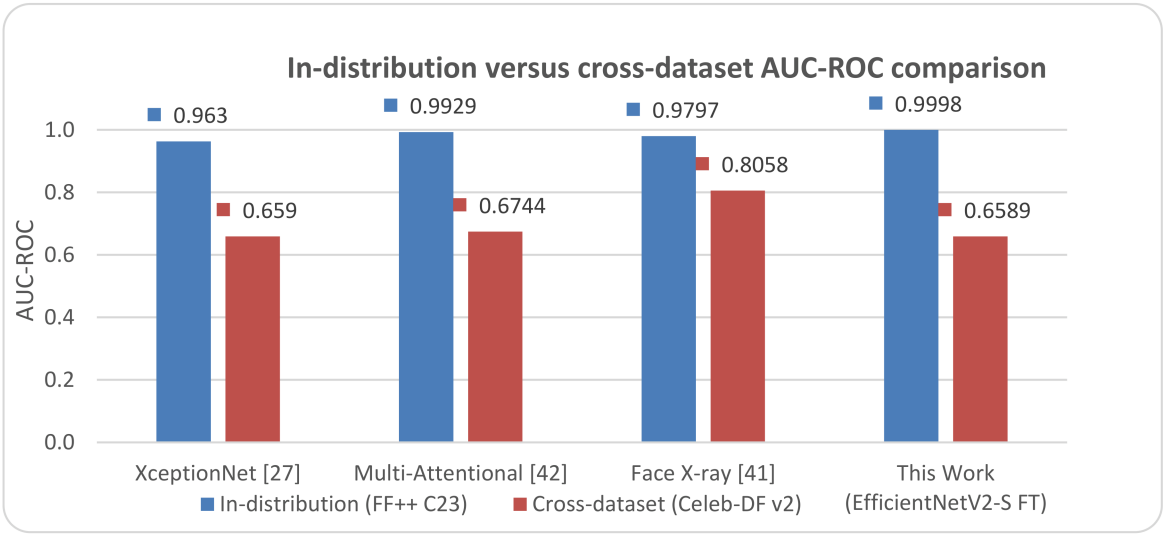


Figure 5. In-distribution (FF++ C23) versus cross-dataset (Celeb-DF v2) AUC-ROC comparison of the proposed framework against published baselines. XceptionNet [27], Multi-Attentional Detection [42] and Face X-ray [41].

Table 7 summarises performance beyond the training distributions, ordered by the size of the shift.

Table 7. Performance beyond the training distribution, ordered by the magnitude of the shift.

Evaluation	AUC-ROC	95% CI	Drop	What Differs from Training
FaceForensics++ C23 (reference)	0.9848	0.9787–0.9898	—	—
DeepFakeDetection fakes vs. FF++ real	0.9689	0.9580–0.9785	1.59 pp	Manipulation methods, subjects
Celeb-DF v2 (cross-dataset)	0.6589	0.6235–0.6977	32.59 pp	Capture and encoding pipeline

Confidence intervals are from 2000 bootstrap resamples of the clustering unit, which is the source video. The two evaluations differ in what they hold constant: the DeepFakeDetection subset introduces manipulation methods and subjects the model has not seen while retaining the capture and encoding characteristics of FaceForensics++, whereas Celeb-DF v2 introduces a different collection and encoding pipeline. The first costs 1.59 percentage points of AUC-ROC and the second 32.59, identifying the video production pipeline rather than the forgery technique as the binding constraint on generalisation. Because our copy of DeepFakeDetection contains only manipulated videos, the negative class for that evaluation is drawn from the FaceForensics++ official test partition; the two subsets differ in recording source as well as in manipulation method, so a contribution from source characteristics cannot be excluded and the 1.59 percentage-point figure should be read as an upper bound on retained performance.

Table 8 reports the effect of recalibrating the decision threshold on a held-out, identity-disjoint sample.

Table 8. Threshold recalibration on Celeb-DF v2, using an identity-disjoint calibration sample.

Operating Point	False Positive Rate	Fake Recall	Balanced Accuracy
Default, inherited from FaceForensics++	0.6676	0.8292	0.5808
Recalibrated, Youden J	0.3490	0.5975	0.6243
Recalibrated, false-positive budget of 10%	0.1087	0.2876	0.5894

Calibration used 2220 frames from 24 identities and evaluation the remaining 8137 frames from identity-disjoint videos. Section 5.2 of the original submission proposed threshold recalibration as a remedy for distributional shift; the experiment shows that it relocates the operating point without improving discrimination, since AUC-ROC on the evaluation subset is 0.6655 at every threshold. Reducing the false alarm rate to an operationally tolerable level costs two thirds of fake recall, and balanced accuracy peaks at 0.6243. No operating point on this curve is usable for autonomous screening.

4.4. SVM 2: Audio Detection on ASVspoof 2019

SVM 2 was trained on the official ASVspoof 2019 Logical Access training partition of 25,380 utterances and evaluated once on the full evaluation partition of 71,237 utterances, which is speaker-disjoint from training and contains thirteen attack algorithms absent from it. The branch achieved an equal error rate of 16.55% (95% CI 14.70 to 18.40) and AUC-ROC of 0.9186 (95% CI 0.9024 to 0.9340), with a balanced accuracy of 83.45% at the equal error threshold, and retained 2394 support vectors (297 bonafide, 2097 spoofed). Equal error rate is reported because it is the metric used in the anti-spoofing literature; accuracy alone is not comparable across studies that use different class balances, and the evaluation partition is 89.7% spoofed. The audio branch is intended as a lightweight baseline. It gives the framework a low-cost second channel alongside visual detection, for use when the visual channel is unavailable or uncertain, and it is not designed to compete with dedicated anti-spoofing countermeasures. Figure 6 shows the confusion matrix and ROC curve for the

audio stream. Table 9 compares the audio branch with published methods on ASVspoof 2019 Logical Access, and Table 10 gives the equal error rate for each attack algorithm.

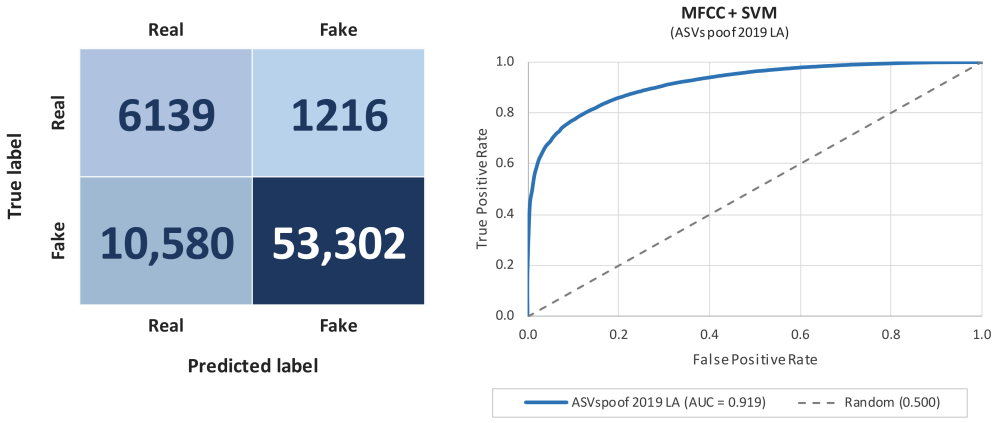


Figure 6. Confusion matrix and ROC curve for the MFCC + SC + ZCR + RBF SVM audio classifier evaluated on the full ASVspoof 2019 LA evaluation partition (AUC-ROC = 0.9186, equal error rate = 16.55%).

Table 9. Comparison of published methods on ASVspoof 2019 Logical Access.

Method	Features	AUC-ROC	Accuracy	Inference	Reference
Local Binary Pattern Histogram	LBPH	N/R	90.00%	N/R	[44]
CNN-LSTM Hybrid	MFCC + Spectrogram	N/R	93.20%	Low	[45]
CNNs	Spectral	N/R	85.00%	N/R	[46]
NeXt—Time Delay Neural Network	Spectral	N/R	92.63%	N/R	[47]
CNN-LSTM Hybrid Classifier	MFCC (45-dim) + ZCR + RMS	N/R	88.00%	N/R	[48]
This work	MFCC + SC + ZCR (48-dim) + RBF SVM	0.9186	83.44%	—	Present study

Note: the confidence interval for the present study is from 2000 bootstrap resamples of the clustering unit. Intervals are not available for the published baselines, whose per-sample scores are not distributed, and significance testing against them is therefore not possible. The present result is an equal error rate of 16.55% [14.70, 18.40] and AUC-ROC of 0.9186 [0.9024, 0.9340]; the accuracy column reports accuracy at the equal error threshold, where balanced accuracy is 83.45%.

Table 10. Equal error rate by attack algorithm on the ASVspoof 2019 Logical Access evaluation partition.

Attack	EER (%)	Attack	EER (%)
A07	0.36	A11	12.94
A08	1.15	A10	14.18
A16	2.15	A13	17.11
A09	2.62	A15	23.91
A19	4.08	A12	25.11
A14	9.51	A18	26.20
		A17	33.43

Each attack is evaluated against all 7355 bonafide utterances of the evaluation partition; the pooled equal error rate is 16.55%. The branch is not uniformly weak. Five of the thirteen unseen attacks are detected at under 5% equal error rate, while A17 reaches 33.43%, a ninety-fold spread. The aggregate figure is therefore dominated by a small number of attack

types that the 48-dimensional representation does not capture. This has a direct bearing on the governance argument of Section 5.5: a detection channel can fail against a single unseen attack while showing no degradation on others, and the failure is not visible from aggregate performance.

4.5. SVM 3: Ablation Study on FakeAVCeleb v1.2

The FakeAVCeleb v1.2 evaluation is structured as an ablation on an identity-disjoint held-out test set of 143 clips. Because the dataset manipulates video and audio independently, each modality is evaluated against its own label: of the 143 test clips, 57 contain manipulated video and 39 contain manipulated audio, and the two sets overlap only on the 28 clips where both are manipulated. Scoring the audio branch against the visual label, as in the original submission, evaluates it on 29 clips whose audio is authentic. The visual-only SVM achieved an AUC-ROC of 1.0000 and accuracy of 99.30%. The audio-only SVM, evaluated against the audio label, achieved an AUC-ROC of 0.9884 and accuracy of 97.90%, substantially higher than the figure reported in the original submission, which was obtained against the visual label. Feature-level fusion achieved an AUC-ROC of 1.0000 and accuracy of 99.30%. Figures 7–9 shows the confusion matrices and ROC curves for the visual-only, audio-only and fused classifiers. A DeLong test on the paired scores returns $p = 1.000$ between visual-only and feature-level fusion, so the addition of audio features provably does not alter discrimination on this test set, and $p = 0.016$ between visual-only and score-level fusion, which is a significant degradation. Table 11 presents the ablation results alongside published methods.

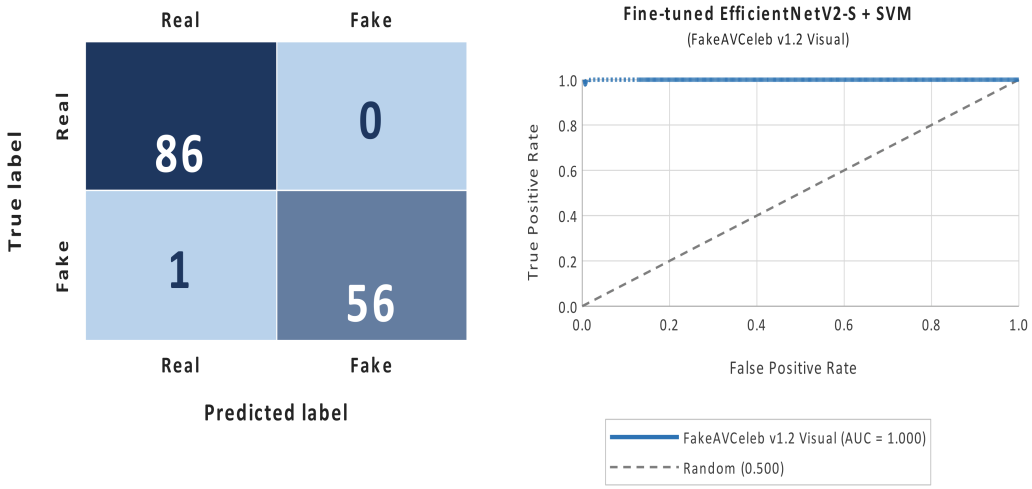


Figure 7. Confusion matrix and ROC curve for the visual-only SVM evaluated on FakeAVCeleb v1.2 (AUC-ROC = 1.0000).

Scores are standardised on the training partition; positive values indicate fake. The lower panel identifies the mechanism. On the 29 clips with manipulated video and authentic audio, the visual branch scores +1.16 and the audio branch −0.52, both correct with respect to their own modality, and averaging moves the fused score to +0.32, close to the decision boundary. On the 11 clips with authentic video and manipulated audio, averaging moves the score to +0.11, almost inverting the visual decision. Where modalities are manipulated independently, symmetric fusion discards information rather than adding it. Confidence-gated fusion, which consults the audio branch only when the visual score is uncertain, preserves the visual decision on these clips. The number of clips in the two disagreement categories is small, so this result is indicative of the mechanism rather than a demonstration that gating is reliable.

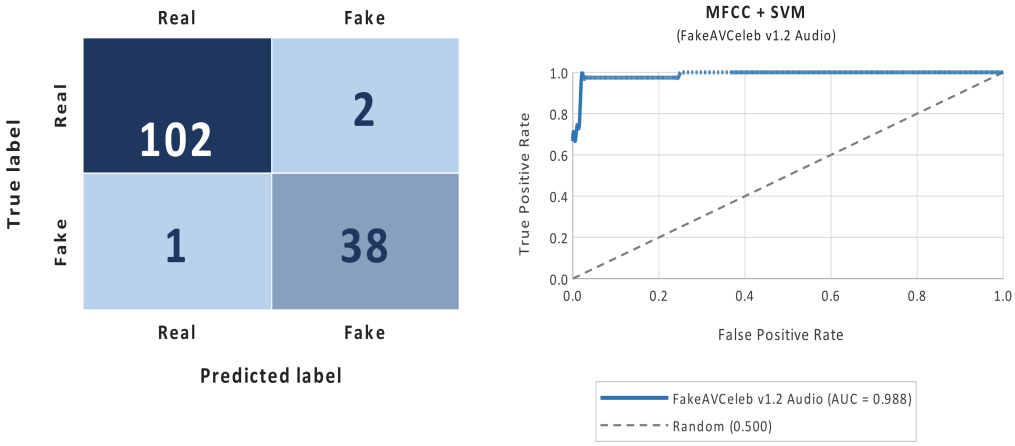


Figure 8. Confusion matrix and ROC curve for the audio-only SVM evaluated on FakeAVCeleb v1.2 against the audio label (AUC-ROC = 0.9884).

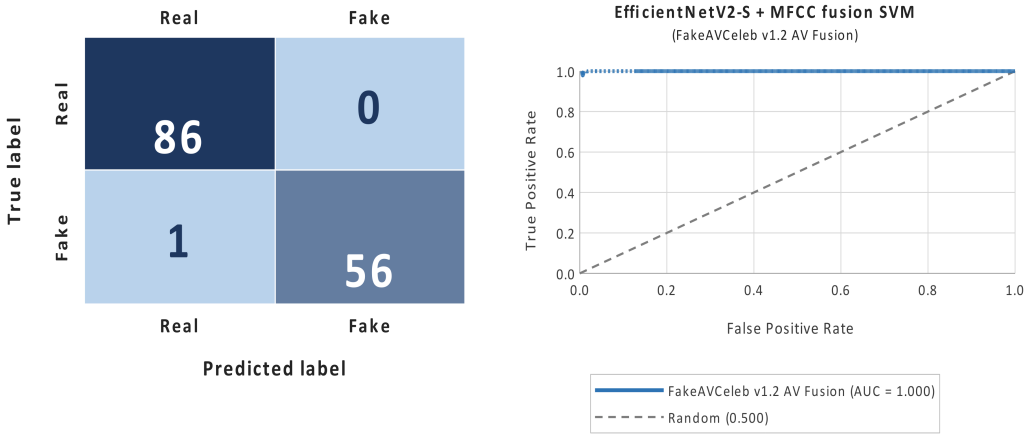


Figure 9. Confusion matrix and ROC curve for the audio–visual late-fusion SVM evaluated on FakeAVCeleb v1.2 (AUC-ROC = 1.0000).

Table 11. FakeAVCeleb v1.2 ablation results and comparison with published methods.

Method/Condition	Features	AUC-ROC	Accuracy	F1	Reference
AVoiD-DF [49]	MFCC + Temporal-Spatial Video Patches	0.892	83.7%	N/R	[49]
DST-Net, Video Only [50]	MTCNN facial frames	0.9094	90.65%	0.9149	[50]
DST-Net, Audio Only [50]	Mel-Spectrograms	0.9873	98.66%	0.9883	[50]
DST-Net, AV Fusion [50]	Mel-Spectrograms + MTCNN	0.9259	92.01%	0.9355	[50]
X3A—Visual unimodal [51]	X3D Encoder + MLP head	0.9970	98.78%	N/A	[51]
X3A—Audio Unimodal [51]	AASIST	1.0000	99.52%	N/A	[51]
X3A—AV Model [51]	X3A (X3D + AASIST)	0.9968	98.90%	N/A	[51]
Audio Only (this work)	MFCC + SC + ZCR + RBF SVM	0.9884	97.90%	0.9620	Present study
Visual Only (this work)	EfficientNetV2-S (FT) + RBF SVM	1.0000	99.30%	0.9912	Present study
AV Late Fusion (this work)	EfficientNetV2-S + MFCC + RBF SVM (1328-dim)	1.0000	99.30%	0.9912	Present study

Note: the confidence interval for the present study is from 2000 bootstrap resamples of the clustering unit. Intervals are not available for the published baselines, whose per-sample scores are not distributed, and significance testing against them is therefore not possible. The present results are reported on 143 identity-disjoint test clips; DeLong tests between conditions are given in Table 12.

Table 12. Fusion strategies on FakeAVCeleb v1.2, and mean standardised decision scores by manipulation category.

Condition	AUC-ROC	Accuracy	Balanced Acc.	DeLong *p* vs. Visual
Visual only (visual label)	1.0000	0.9930	0.9912	—
Audio only (audio label)	0.9884	0.9790	0.9776	—
Feature-level fusion	1.0000	0.9930	0.9912	1.000
Score-level fusion (mean)	0.9776	0.9021	0.9038	0.016
Confidence-gated fusion (0.5)	1.0000	1.0000	1.0000	—
Confidence-gated fusion (1.0)	0.9916	0.9161	0.9214	—
Confidence-gated fusion (1.5)	0.9776	0.9021	0.9038	—
Category	Clips	Visual score	Audio score	Mean fusion
FakeVideo–FakeAudio	28	+1.12	+1.44	+1.28
FakeVideo–RealAudio	29	+1.16	−0.52	+0.32
RealVideo–FakeAudio	11	−0.81	+1.03	+0.11
RealVideo–RealAudio	75	−0.80	−0.54	−0.67

4.6. Inference Efficiency

Latency was measured stage by stage on 100 test frames using a single CPU thread with no GPU available. Table 13 and Figure 10 break down the per-frame latency by pipeline stage. The SVM decision stage requires 2.34 ms per frame, but it accounts for only 1.2% of pipeline latency: the EfficientNetV2-S forward pass requires 188.92 ms, image loading 2.87 ms and preprocessing 1.26 ms, for a total of 195.39 ms per frame. clip sampled at twenty frames therefore requires approximately 3.9 s end to end, before MTCNN face detection, which is not included in these figures. The framework runs without the specialist GPU infrastructure required by Transformer-based models exceeding 800 ms per frame on CPU [5] and is suited to asynchronous screening of recorded material, but it is not a real-time system in the sense of operating within a live video conference.

Table 13. Per-frame inference latency by pipeline stage, single CPU thread (Intel Xeon at 2.20 GHz, 52 GB memory).

Stage	Mean (ms)	95th Percentile (ms)	Share of Total
Image loading	2.87	3.76	1.5%
Preprocessing	1.26	1.37	0.6%
EfficientNetV2-S forward pass	188.92	192.46	96.7%
SVM decision	2.34	2.39	1.2%
Total	195.39	199.06	100%

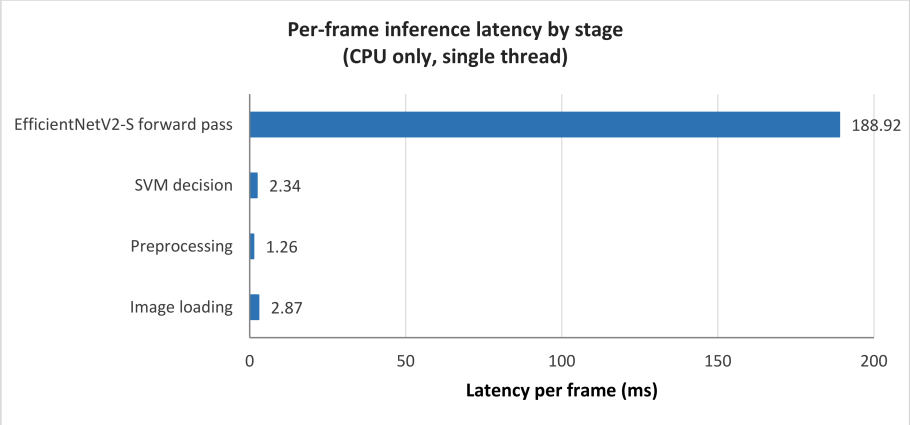


Figure 10. Per-frame inference latency by pipeline stage under CPU-only conditions, showing the SVM decision stage against the total end-to-end cost.

5. Discussion

5.1. The Role of End-to-End Fine-Tuning

The ablation in Table 4 indicates the contribution of end-to-end fine-tuning. Replacing the fine-tuned backbone with frozen ImageNet features, using the same classifier type and partitions but a reduced hyperparameter grid and 30% of the training groups, reduces frame-level AUC-ROC from 0.9848 to 0.7192 and video-level AUC-ROC from 0.9937 to 0.7666, while the false-positive rate rises from 8.97 to 48.45%. The number of support vectors rises from 1110 to 16,249 on a comparable training set, indicating that the classifier is memorising rather than separating: frozen ImageNet embeddings do not linearly encode the artefacts that distinguish manipulated from authentic faces. By contrast, the three fine-tuned conditions in Table 4 differ by less than 0.002 in AUC-ROC regardless of whether classification is performed by an RBF SVM, a linear SVM or the softmax head retained from fine-tuning. Because the frozen condition was trained on less data with fewer candidate settings, this contrast is an upper bound on the effect of fine-tuning, not a controlled estimate. Even so, it indicates that most of the reported performance comes from adapting the backbone, and little from the choice of classifier.

This has a direct practical implication. Detection performance figures in the literature cannot be interpreted without knowing whether the backbone was frozen, partially adapted, or fully fine-tuned, and on what data. The present framework makes this distinction explicit by separately fine-tuning the backbone for each target dataset and by reporting the frozen-backbone condition alongside the fine-tuned one.

5.2. Cross-Dataset Generalisation

The two out-of-distribution evaluations separate two failure modes that are usually reported together. Against the held-out DeepFakeDetection subset, which contains manipulation methods and subjects absent from training but shares the capture and encoding characteristics of FaceForensics++, AUC-ROC falls only from 0.9848 to 0.9689. Against Celeb-DF v2, which uses a different collection and encoding pipeline, it falls to 0.6589. The framework therefore generalises across forgery techniques but not across video production pipelines, and the Celeb-DF v2 value is consistent with the 0.659 to 0.674 reported for comparable FaceForensics++-trained methods on the same benchmark [27,42]. This has an operational consequence that differs from the one usually drawn: the trigger for revalidation is not the appearance of a new deepfake technique but a change in the video source, such as a new conferencing platform, codec or bitrate. The false-positive rate of 65.6% on Celeb-DF v2 authentic content at the inherited threshold illustrates that AUC-ROC alone does not translate into operational reliability. Lu and Ebrahimi [2] observed that models performing well on benchmark metrics can still fail in operational settings because the conditions at deployment differ from those at training. The recalibration experiment reported in Table 8 tests whether this can be remedied without retraining. It cannot. Calibrating the threshold on 24 held-out identities and evaluating on the remainder moves the false-positive rate from 66.8 to 10.9%, but fake recall falls from 82.9 to 28.8% and balanced accuracy never exceeds 0.6243, because AUC-ROC on the evaluation subset is unchanged at 0.6655 whatever threshold is chosen. Recalibration relocates the operating point; it does not restore discrimination. Only retraining on data from the target domain addresses this.

5.3. Multimodal Ablation Findings

Two findings emerge from the multimodal ablation, and they point in different directions. The first is that FakeAVCeleb is saturated for this architecture: the visual branch reaches AUC-ROC of 1.0000 on the identity-disjoint test set, so there is no headroom within which fusion could demonstrate a gain, and a DeLong test confirms that feature-level fusion

is indistinguishable from the visual branch alone ($p = 1.000$). No conclusion about the value of multimodal detection in general can be drawn from a benchmark on which one modality is already perfect. The second finding is more consequential. Score-level fusion, which averages the standardised decisions of the two branches, reduces AUC-ROC from 1.0000 to 0.9776, a degradation that is statistically significant (DeLong $p = 0.016$). The per-category scores in Table 12 identify the cause. On the 29 test clips with manipulated video and authentic audio, the visual branch scores +1.16 and the audio branch −0.52, each correct with respect to its own modality, and the average of +0.32 sits close to the decision boundary; on the 11 clips with authentic video and manipulated audio, the average of +0.11 almost inverts the visual decision. Where the two modalities are manipulated independently, as they are in this dataset and as they may be in a real attack that replaces a face but not a voice, symmetric fusion discards information rather than adding it. Irfan et al. [52] noted that SVM performance on concatenated high-dimensional feature vectors is sensitive to the ratio of training samples to feature dimensions, which is 0.53 here; but, the degradation observed with score-level fusion is not a dimensionality effect, since it arises from averaging two scalar scores.

The practical interpretation is architectural. Fusion should be asymmetric and conditional rather than symmetric and unconditional. Confidence-gated fusion, which consults the audio branch only where the visual score is uncertain, preserves the visual decision on the disagreement categories while retaining audio as an independent channel for cases where video is unavailable or uninformative. The audio branch is not weak on this dataset: evaluated against its own label it reaches AUC-ROC of 0.9884, and the lower figure reported in the original submission was an artefact of scoring it against the visual label. The number of clips in the two disagreement categories is small, however, and these results indicate a mechanism rather than establish the reliability of any particular gating threshold.

5.4. Addressing the Three Research Gaps

1. The framework partially addresses the real-world deployment fragility gap. Performance is retained across unseen-manipulation methods, with an AUC-ROC of 0.9689 on the held-out DeepFakeDetection subset, but not across video production pipelines, where it falls to 0.6589 with 65.6% of real content flagged as fake. Deployment in a new domain therefore requires labelled data from that domain and a further fine-tuning cycle; threshold recalibration alone is insufficient. End-to-end latency of 195 ms per frame is consistent with asynchronous screening of recorded material but not with live intervention. The gap is narrowed rather than closed.
2. The tri-stream architecture addresses the absence of integrated multimodal detection frameworks identified by Sharma et al. [3] and Zegeye et al. [4] by evaluating visual, audio and fused detection under a single protocol on the same test set, with each modality scored against its own label. The contribution is the integrated evaluation and the identification of a fusion failure mode, rather than a demonstrated multimodal performance gain.
3. The framework addresses the computational efficiency gap at the classification stage: the SVM decision requires 2.34 ms per frame and accounts for 1.2% of pipeline latency, and the whole pipeline runs on CPU without the GPU infrastructure required by Transformer-based models [5]. The measurement establishes that the classifier adds negligible cost to a detection pipeline, not that the pipeline is real-time.

5.5. Organisational Governance Strategy

The results reported in Section 4, and their interpretation in the preceding sections, establish the empirical boundaries within which automated detection can reliably operate

and equally the space within which organisational governance must compensate. The position taken in this research is that detection and governance are complementary rather than alternative responses: where detection is reliable, governance defines alert workflows and escalation thresholds; where it degrades under distributional shift, against novel generation techniques, or on content outside the training distribution, governance provides procedural controls for the residual risk. The 25.6 million US dollar fraudulent video call documented by Boiselle [1], in which no automated detection existed and no human verification protocol was followed, demonstrates that governance failures are exploitable independently of technical capability and that the two layers must be designed together.

Three principles, each derived from the experimental evidence, govern the framework. Proportionality requires controls matched to decision stakes: a call authorising routine communication warrants different verification from one authorising a financial transfer. Independence requires that procedural control function when the detection system fails, is unavailable, or operates outside its reliable range. Adaptability requires scheduled review cycles and retraining triggers, because the detection boundary shifts as generation and detection techniques co-evolve. Table 14 maps each empirically identified limitation to the governance strategy it motivates.

Table 14. Mapping of technical detection gaps to governance control requirements.

Technical Limitation	Operational Impact	Governance Control
Threshold miscalibration under distributional shift: 65.6% false-positive rate on Celeb-DF v2 real content at the FF++ operating threshold, with AUC-ROC of 0.6589 and no usable operating point on the curve	Excessive false alerts on real video content from new domains, reducing operational trust in the detection system	Mandatory revalidation on labelled data from each new deployment domain before going live. Because recalibration alone cannot restore discrimination (Table 8), the trigger for revalidation is any change in video source, codec, platform or bitrate, and the outcome may be a further fine-tuning cycle rather than a threshold adjustment.
Audio detection failure against specific unseen attacks: equal error rate ranges from 0.36% to 33.43% across the thirteen unseen attacks of the ASVspoof 2019 evaluation partition, with a pooled rate of 16.55%	Synthetic voices that pass audio detection undetected may enable executive impersonation fraud in voice and video calls	Human review escalation for any communication triggering a payment, access grant, or sensitive disclosure; callback verification via pre-registered contact number before any transfer is authorised.
Symmetric fusion degrades performance when modalities are manipulated independently: score-level fusion reduces AUC-ROC from 1.0000 to 0.9776 (DeLong $p = 0.016$)	Unconditional bimodal processing adds computational overhead without improving detection when visual branch is well-calibrated	Confidence-gated fusion architecture: audio analysis triggered only when the visual score is uncertain, never averaged unconditionally with it; audio channel retained as an independent safety net when the visual channel is unavailable.
Diffusion model blind spot: the evaluation datasets were produced by pre-diffusion generation methods, and generalisation to diffusion-generated media is untested	Detection system has not been shown to screen reliably against deepfakes produced by Stable Diffusion, DALL-E, Midjourney, and equivalent systems	Quarterly threat intelligence review with scheduled model retraining or augmentation when new generation techniques enter the threat landscape; interim reliance on procedural verification for high-stakes decisions.
Domain-specific fine-tuning requirement: detection performance degrades when visual backbone is deployed on content types not represented in training data	Zero-shot deployment on a new video production pipeline produced AUC-ROC of 0.6589 and a 65.6% false alarm rate on authentic content	Maintain a labelled reference dataset of 200 to 500 verified real and fake samples per deployment domain; establish a production monitoring process to detect distributional drift and trigger retraining when needed.
Residual false-negative risk across all modalities: no single automated detector provides sufficient confidence for fully autonomous high-value decision authorisation	Automated detection alone cannot replace human judgement for decisions with significant financial, legal, or operational consequences	Mandatory dual authorisation for financial transactions above a defined threshold; video call verification must be followed by a callback on a pre-registered number before any transfer, access grant, or sensitive disclosure is completed.

The numerical values in Table 14, namely the labelled sample sizes and the gating threshold, are illustrative starting points rather than empirically derived or standardised requirements. They are stated to make the controls concrete, and each should be set for a given deployment from the operator's own validation data, alert capacity and risk appetite.

To anchor the controls in recognised practice rather than in the authors' judgement alone, Table 15 maps each control to the corresponding Annex A control of ISO/IEC 27001:2022 [6], to the functions and categories of the NIST Cybersecurity Framework 2.0 [7], and to the NIST AI Risk Management Framework [8]. Every control proposed here has an existing home in an established management system, so adoption is a matter of extending current controls to a new threat rather than of creating a parallel governance structure. The transparency obligations of Article 50 of Regulation (EU) 2024/1689 [53], which have applied since 2 August 2026, require providers of generative systems to mark synthetic output in machine-readable form and deployers to disclose deepfake content; those duties fall on the producers and deployers of synthetic media rather than on the organisation defending against its fraudulent use, so provenance marking should be treated as a supplementary signal that can corroborate authentic content. An adversary committing fraud has no incentive to comply, and the absence of a marking is not evidence that content is genuine.

Table 15. Alignment of the governance controls in Table 14 with recognised security and AI risk management frameworks.

Governance Control	ISO/IEC 27001:2022 Annex A [6]	NIST CSF 2.0 [7] and AI RMF [8]
Revalidation before deployment in a new domain	A.8.29 security testing in development and acceptance; A.8.32 change management; A.5.37 documented operating procedures	CSF: ID.IM, DE.CM; AI RMF: MEASURE, MANAGE
Human review escalation and callback verification	A.5.3 segregation of duties; A.5.14 information transfer; A.6.3 awareness, education and training; A.8.5 secure authentication	CSF: PR.AA, PR.AT, RS.MA; AI RMF: MANAGE
Confidence-gated fusion architecture	A.8.27 secure system architecture and engineering principles; A.8.6 capacity management	CSF: PR.IR; AI RMF: MAP, MEASURE
Threat intelligence review and retraining trigger	A.5.7 threat intelligence; A.8.8 management of technical vulnerabilities; A.8.32 change management	CSF: ID.RA, ID.IM; AI RMF: MAP, MANAGE
Labelled reference dataset and drift monitoring	A.8.16 monitoring activities; A.8.33 test information; A.5.34 privacy and protection of PII	CSF: DE.CM, DE.AE; AI RMF: MEASURE
Dual authorisation and out-of-band verification	A.5.3 segregation of duties; A.5.15 access control; A.6.8 information security event reporting; A.5.24 to A.5.27 incident management	CSF: GV.RR, PR.AA, RS.MA; AI RMF: GOVERN, MANAGE

The controls were also applied in a structured walk-through of the video conference fraud documented by Boiselle [1], in which an employee in a finance function received an instruction attributed to a senior officer, joined a video call in which every other participant was synthetic, and executed a series of transfers. Four points in that sequence are addressable. At the initial approach, awareness training and a low-friction reporting route (A.6.3, A.6.8) determine whether an employee who is already suspicious escalates rather than proceeds. During the call, automated detection is the only control operating in real time, and the cross-dataset result in Section 4.3 indicates that a detector applied to an unseen content type would be unreliable at precisely that moment; the 195 ms per frame end-to-end latency reported in Section 4.6 further indicates that this framework would operate on recorded material rather than within the live call. At the point of authorisation, dual authorisation and callback verification on a pre-registered number (A.5.3, A.8.5) interrupt the attack irrespective of whether any detector was present or correct, and this is the only control in the set whose effectiveness does not depend on detection performance. After execution, transaction monitoring (A.8.16) and incident response (A.5.24 to A.5.26) govern how quickly the loss is contained. The walk-through is analytical and was conducted by

the authors against a single documented case; it is not empirical validation, and structured review with practitioners remains necessary further work.

5.6. Limitations

Seven limitations bound the results reported here. First, cross-dataset evaluation rests on a single benchmark, Celeb-DF v2,Further datasets, such as DFDC and WildDeepfake are needed to establish whether the distinction between manipulation shift and pipeline shift holds more generally. Second, no evaluation was conducted against diffusion-generated media, which is now a substantial share of synthetic content. Third, the FakeAVCeleb test set contains 143 clips, of which only 29 and 11 fall in the two categories where the modalities disagree, so the fusion analysis identifies a mechanism rather than establishing the reliability of any gating threshold. Fourth, the audio branch uses a 48-dimensional handcrafted representation: stronger baselines such as LFCC-based systems or end-to-end models, calibration of audio thresholds, and intermediate fusion architectures were not evaluated and are the principal direction for further work. Fifth, no evaluation was conducted under adversarial conditions: deliberate evasion, adversarial perturbation and re-compression attacks are untested, and performance under an adaptive adversary is unknown. Sixth, all evaluation uses public benchmarks: no production video-conference material was tested, and the governance framework has not been deployed or validated in a live organisational setting. Seventh, the governance mapping has been examined only through the authors' own structured walk-through and not by external expert review.

6. Conclusions

This paper has presented a hybrid deepfake detection framework combining end-to-end fine-tuned EfficientNetV2-S with independently optimised RBF SVM classifiers, validated across four public benchmark datasets under the official evaluation protocol of each. Under the official FaceForensics++ source-video split, the framework achieves an AUC-ROC of 0.9848 at the frame level and 0.9937 at the video level, and retains an AUC-ROC of 0.9689 on a held-out manipulation subset it never saw in training. Cross-dataset performance on Celeb-DF v2 falls to 0.6589 with 65.6% of authentic content flagged as fake, and a recalibration experiment shows that no operating point on that curve is usable. Classification requires 2.34 ms per frame on CPU within an end-to-end pipeline cost of 195 ms per frame. On the full ASVspoof 2019 Logical Access evaluation partition, the audio branch reaches an equal error rate of 16.55%, ranging from 0.36 to 33.43% across individual unseen attacks.

The ablation on FakeAVCeleb v1.2 established two things: feature-level fusion is indistinguishable from visual-only detection on a saturated benchmark; symmetric score-level fusion significantly degrades it when video and audio are manipulated independently, because averaging two branches that are each correct about their own modality moves the fused decision towards the boundary. Fusion should therefore be conditional rather than symmetric. A separate ablation showed that replacing the fine-tuned backbone with frozen ImageNet features reduces AUC-ROC from 0.9848 to 0.7192, an upper bound because the frozen condition used a reduced training protocol, while the choice between an RBF kernel, a linear kernel and the softmax head changes it by less than 0.002.

The governance framework maps each identified technical limitation to a specific organisational control, grounded in proportionality, independence, and adaptability. Technical detection and organisational governance are complementary layers, with governance providing procedural controls that compensate for the residual risk that automated detection cannot eliminate.

Four directions for future research follow from the limitations documented. First, detection methods that generalise across generation families are needed urgently; the results reported here were obtained against pre-diffusion generation methods only. Second, the distinction between manipulation shift and video-pipeline shift should be evaluated across broader cross-dataset benchmarks to determine its general applicability. Third, intermediate and adaptive fusion architectures that model cross-modal interactions at an earlier stage may produce more balanced multimodal results when training samples are limited. Fourth, establishing standardised evaluation protocols specifying fine-tuning parameters, ablation designs, and baseline modality configurations explicitly would significantly improve the comparability of published multimodal detection results.

Author Contributions: Conceptualisation, O.A. and N.A.; methodology, O.A.; software, O.A.; validation, O.A., N.A. and R.B.S.; formal analysis, O.A.; investigation, O.A.; resources, N.A.; data curation, O.A.; writing—original draft preparation, O.A.; writing—review and editing, N.A., R.B.S. and M.F.; visualisation, O.A.; supervision, N.A. and M.F.; project administration, N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used in this research are publicly available academic benchmarks: FaceForensics++ [26], Celeb-DF v2 [27], ASVspoof 2019 Logical Access [28], FakeAVCeleb v1.2 [29]. Access to FaceForensics++, Celeb-DF v2 and FakeAVCeleb requires a signed request to the dataset authors, so no media are redistributed here. No new data were created in this study. The code, partition identifier lists, preprocessing and evaluation scripts, selected hyperparameters and random seeds are available at https://github.com/oluyemisiajibawo/executive-deepfake-detection (accessed on 9 October 2026) and archived at https://doi.org/10.5281/zenodo.23245968 (accessed on 9 October 2026) under the MIT licence.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Boiselle, R. Deepfakes Unmasked: Enterprise Cyber Security in the Age of AI Manipulation and Countermeasures. 2025. Available online: https://www.researchgate.net/profile/Rene-Boiselle/publication/390662179_Deepfakes_Unmasked_Enterprise_Cyber_Security_in_the_Age_of_AI_Manipulation_and_Countermeasures/links/67f7c3cd91535861140da5f5/Deepfakes-Unmasked-Enterprise-Cyber-Security-in-the-Age-of-AI-Manipulation-and-Countermeasures.pdf (accessed on 9 October 2026).
2. Lu, Y.; Ebrahimi, T. Assessment framework for deepfake detection in real-world situations. *EURASIP J. Image Video Process.* **2024**, *2024*, 6. [CrossRef]
3. Sharma, V.K.; Garg, R.; Caudron, Q. A systematic literature review on deepfake detection techniques. *Multimed. Tools Appl.* **2025**, *84*, 22187–22229. [CrossRef]
4. Zegeye, B.; Atinafu, H.; Sherif, S.; Dave, R.; Bhavsar, M. Deepfake detection using machine learning: A comprehensive literature review. In *Proceedings of 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), Antalya, Turkiye, 7–9 August 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–7.
5. Raza, A.; Basit, A.; Amin, A.; Arfeen, Z.A.; Masud, M.I.; Fayyaz, U.; Jumani, T.A. A comprehensive review of deepfake detection techniques: From traditional machine learning to advanced deep learning. *AI* **2026**, *7*, 68. [CrossRef]
6. *ISO/IEC 27001:2022*; Information Security, Cybersecurity and Privacy Protection—Information Security Management Systems—Requirements. International Organization for Standardization: Geneva, Switzerland, 2022.
7. National Institute of Standards and Technology. *The NIST Cybersecurity Framework (CSF) 2.0*; NIST CSWP 29; NIST: Gaithersburg, MD, USA, 2024.
8. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; NIST AI 100-1; NIST: Gaithersburg, MD, USA, 2023.

9. Bhatt, M.; Parmar, R.; Arsalan, M.; Garg, P. Generative AI: Evolution, applications, challenges and future prospects. In *Proceedings of G-CARED 2025, First Global Conference on AI Research and Emerging Developments, New Delhi, India, 16–18 May 2025*; Ganitara Research Foundation: New Delhi, India, 2025; pp. 47–52.
10. Trigka, M.; Dritsas, E. The evolution of generative AI: Trends and applications. *IEEE Access* **2025**, *13*, 98504–98529. [CrossRef]
11. Westerlund, M. The emergence of deepfake technology: A review. *Technol. Innov. Manag. Rev.* **2019**, *9*, 40–53. [CrossRef] [PubMed]
12. Buo, S.A. The emerging threats of deepfake attacks and countermeasures. *arXiv* **2024**, arXiv:2012.07989.
13. Rana, M.S.; Nobi, M.N.; Murali, B.; Sung, A.H. Deepfake detection: A systematic literature review. *IEEE Access* **2022**, *10*, 25494–25513. [CrossRef]
14. Chakravarty, S.; Dua, M. A lightweight feature extraction technique for deepfake audio detection. *Mach. Learn. Appl.* **2024**, *15*, 100512.
15. Deepa, B.G.; Lokesh, C.K.; Umamaheswari, D.; Ayshwarya, B.; Yethish, P.V.; Suhaas, K.P. An enhanced deep learning framework for DeepFake detection using EfficientNet-B3: Comparative evaluation of deep and machine learning techniques. *Discov. Comput.* **2026**, *29*, 18. [CrossRef]
16. Alzurfi, N.K.; Altaei, M.S. A hybrid model of deep learning and machine learning methods to detect deepfake videos. *Iraqi J. Sci.* **2025**, *66*, 412–425. [CrossRef]
17. Satoto, B.D.; Yasid, A.; Rahmadhani, A.; Khotimah, B.K.; Irmawati, B.; Hamid, S.B. Deepfake detection using MobileNetV2 integrated with a support vector machine classifier. *EPJ Web Conf.* **2025**, *344*, 01019. [CrossRef]
18. Deng, L.; Suo, H.; Li, D. Deepfake video detection based on EfficientNet-V2 network. *Comput. Intell. Neurosci.* **2022**, *2022*, 3441549. [CrossRef] [PubMed]
19. Coccomini, D.A.; Caldelli, R.; Falchi, F.; Gennaro, C.; Amato, G. Cross-forgery analysis of Vision Transformers and CNNs for deepfake image detection. In *Proceedings of the 1st International Workshop on Multimedia AI Against Disinformation, Newark, NJ, USA, 27–30 June 2022*; ACM: New York, NY, USA, 2022; pp. 52–58.
20. Rani, G.; Kothekar, A.; Philip, S.G.; Dhaka, V.S.; Zumpano, E.; Vocaturo, E. Lightweight and hybrid transformer-based solution for quick and reliable deepfake detection. *Front. Big Data* **2025**, *8*, 1521653. [CrossRef] [PubMed]
21. AlMuhaideb, S.; Alshaya, H.; Almutairi, L.; Alomran, D.; Alhamed, S.T. LightFakeDetect: A lightweight model for deepfake detection in videos that focuses on facial regions. *Mathematics* **2025**, *13*, 3088. [CrossRef]
22. Amin, M.A.; Hu, Y.; Li, C.-T.; Liu, B. Deepfake detection based on cross-domain local characteristic analysis with multi-domain transformer. *Alex. Eng. J.* **2024**, *91*, 592–609. [CrossRef]
23. Yang, W.; Guo, H.; Hu, S.; Zhu, B.; Fu, Y.; Lyu, S.; Wu, X.; Wang, X. Crossed: Deep information decomposition for cross-dataset deepfake detection. *Front. Big Data* **2025**, *8*, 1669488. [CrossRef] [PubMed]
24. Varchas, V.S.; Sarim, M.; Muddasir, N.M. A hybrid method for effective detection of deep fake videos: Literature review. In *Proceedings of the 2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS), Chickballapur, India, 28–29 April 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–6.
25. Das, A.; Viji, K.A.; Sebastian, L. A survey on deepfake video detection techniques using deep learning. In *Proceedings of the 2022 Second International Conference on Next Generation Intelligent Systems (ICNGIS), Kottayam, India, 29–31 July 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–5.
26. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Niessner, M. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–11.
27. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A large-scale challenging dataset for DeepFake forensics. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 3207–3216.
28. Wang, X.; Yamagishi, J.; Todisco, M.; Delgado, H.; Nautsch, A.; Evans, N.; Sahidullah, M.; Vestman, V.; Kinnunen, T.; Lee, K.A.; et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Comput. Speech Lang.* **2020**, *64*, 101089. [CrossRef]
29. Khalid, H.; Tariq, S.; Kim, M.; Woo, S.S. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv* **2021**, arXiv:2108.05080.
30. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
31. Tan, M.; Le, Q.V. EfficientNetV2: Smaller models and faster training. In Proceedings of the International Conference on Machine Learning (ICML), PMLR, Online, 18–24 July 2021; Volume 139, pp. 10096–10106.
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
33. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]

34. Cavoretto, R.; De Rossi, A.; Mukhametzhanov, M.S.; Sergeyev, Y.D. On the search of the shape parameter in radial basis functions using univariate global optimization methods. *J. Glob. Optim.* **2021**, *79*, 305–327. [CrossRef]
35. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27.
36. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
37. Luque, A.; Carrasco, A.; Martin, A.; de las Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* **2019**, *91*, 216–231. [CrossRef]
38. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.
39. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
40. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of European Conference on Computer Vision (ECCV), Online, 23–28 August 2020*; Springer International Publishing: Cham, Switzerland, 2020; pp. 86–103.
41. Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; Guo, B. Face X-ray for more general face forgery detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 5001–5010.
42. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-attentional deepfake detection. In *Proceedings of CVPR, Online, 19–25 June 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 2185–2194.
43. Masi, I.; Killekar, A.; Mascarenhas, R.M.; Gurudatt, S.P.; AbdAlmageed, W. Two-branch recurrent network for isolating deepfakes in videos. In *Proceedings of ECCV 2020, Online, 23–28 August 2020*; Springer International Publishing: Cham, Switzerland, 2020; pp. 667–684.
44. Ramathilagam, A.; Ramanathan, R.; Saravanakumar, R. A recurrent neural network-based multi-parameter sub-band, sub-second analysis for voice spoofing and cloning detection. In *Proceedings of the 2025 Second International Conference on Cognitive Robotics and Intelligent Systems (ICC-ROBINS), Coimbatore, India, 25–26 June 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5.
45. Asuai, C.; Arinomor, A.P.; Atumah, C.T.; Kowhoro, I.F.; Ogheneochuko, D.E. Hybrid CNN-LSTM architectures for deepfake audio detection using Mel frequency cepstral coefficients and spectrogram analysis. *Am. J. Math. Comput. Model.* **2025**, *10*, 98–109. [CrossRef]
46. Anagha, R.; Arya, A.; Hari Narayan, V.; Abhishek, S.; Anjali, T. Audio deepfake detection using deep learning. In *Proceedings of the 12th International Conference on System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 22–23 December 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 176–181.
47. Tahaoglu, G.; Baracchi, D.; Shullani, D.; Iuliani, M.; Piva, A. Deepfake audio detection with spectral features and ResNeXt-based architecture. *Knowl.-Based Syst.* **2025**, *323*, 113726. [CrossRef]
48. Altalahin, I.; AlZu'bi, S.; Alqudah, A.; Mughaid, A. Unmasking the truth: A deep learning approach to detecting deepfake audio through MFCC features. In *Proceedings of the 2023 International Conference on Information Technology (ICIT), Amman, Jordan, 9–10 August 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 511–518.
49. Yang, W.; Zhou, X.; Chen, Z.; Guo, B.; Ba, Z.; Xia, Z.; Cao, X.; Ren, K. AVoiD-DF: Audio-visual joint learning for detecting deepfake. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 2015–2029. [CrossRef]
50. Ilyas, H.; Javed, A.; Malik, K.M. AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio-visual deepfakes detection. *Appl. Soft Comput.* **2023**, *136*, 110124. [CrossRef]
51. Park, C.; Moon, B.; Jeon, M.; Jung, J.W.; Woo, S.S. X3A: Efficient multimodal deepfake detection with score-level fusion. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, Catania, Italy, 31 March–4 April 2025*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 767–774.
52. Irfan, M.T.; Arora, B.; Sandotra, N.; Raza, A.A. On machine learning and deep learning based deepfake generation and detection. *Procedia Comput. Sci.* **2025**, *259*, 1927–1936. [CrossRef]
53. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**, *L 2024*, 1689.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.