International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 1282-1316
https://doi.org/10.70917/ijcisim-2026-5393
Article

Beyond Single-Corpus Leaderboards: Controlled Cross-Corpus Evaluation of Frequency-Gated and Multi-Scale Cross-Attention Networks for Image Deepfake Detection

Abhay Mundra, Dr.Mamta Meena
Research Scholar, India. er.abhayjit@gmail.com ,mamtameena@orientaluniversity.in

Abstract: Image deepfake detectors are routinely selected on single-corpus accuracy, yet deployment requires a decision on media whose manipulation family is unknown. This paper asks whether single-corpus ranking predicts ranking under unknown provenance, and answers in the negative. Six architectures are trained from a common initialisation under one identical data, preprocessing, augmentation and optimisation protocol: four re-implement the established hybrid convolutional–recurrent family, and two are proposed here. FA-ViT-BiLSTMcouples vision-transformer patch tokens with log-magnitude Fourier tokens through a learned element-wise modality gate, allowing the spectral stream to be attenuated under compression. MSCA-CTNet exchanges information between three-scale convolutional tokens and global transformer tokens through bidirectional multi-head cross-attention. All six are evaluated on Celeb-DF v2, DFDC, UADFV, FaceForensics++ C23 and DeepDetect-2025, on a mixed-provenance Combined set, and under a leave-one-corpus-out protocol. Dual-backbone hybrids give the strongest single-corpus results, reaching 0.953 accuracy on DFDC, while MSCA-CTNet attains the highest ROC-AUC at 0.987. The central finding is a dissociation: the architecture that wins four of five corpora loses 0.203 accuracy under mixed provenance and ranks fourth on held-out corpora, where the two proposed architectures lead with mean AUCs of 0.849 and 0.830. Class-wise accuracy and the Matthews correlation coefficient are shown to be necessary rather than optional.
Keywords: Deepfake detection, media forensics, bidirectional LSTM, vision transformer, cross-attention, frequency-domain analysis, cross-dataset generalisation, Matthews correlation coefficient

1. Introduction

Synthetic facial media has moved from a research curiosity to an operational threat. Generative adversarial networks and, more recently, diffusion models can synthesise or alter a human face with a fidelity that defeats untrained human inspection, and the resulting artefacts circulate through social platforms within minutes of creation. The documented consequences span non-consensual imagery, financial fraud, identity theft and organised disinformation, and corporate security analyses now treat synthetic media as a first-class social-engineering vector rather than a hypothetical one [11], [19]. Automated detection is therefore not an academic exercise but an operational requirement.

The research response has been substantial. Convolutional detectors that learn low-level generation residuals achieve strong results when the generator family is known [1], [9], [23], [25], and hybrid architectures that append a recurrent stage to a convolutional encoder capture spatio-temporal or spatial-sequential dependencies that a purely convolutional model discards [6], [18], [22]. More recent work has moved toward transformer backbones, frequency-domain representations and adapter-based transfer, motivated by the observation that pixel-domain artefacts are the first casualty of compression and re-encoding [24], [30], [33], [36].

A recurring difficulty undermines the comparison of these results. Published accuracies are typically obtained under study-specific data partitions, preprocessing pipelines, input resolutions and training budgets, and are reported for the corpora on which each method performs best. Two detectors reported at 96% on nominally the same benchmark may have been trained on different partition sizes with different augmentation and different early-stopping criteria, so the difference between them is not attributable to architecture. Benchmark efforts such as DeepfakeBench [27] and Deepfake-Eval-2024 [16] have made this point directly, and cross-generation analyses have shown that detector rankings are unstable when the generator changes [26].

A second and less discussed difficulty concerns the metric itself. On a balanced binary task, a classifier that predicts a single class for almost every input attains exactly 50% accuracy, which is easy to recognise; but the same classifier can simultaneously produce an above-chance ROC-AUC, because the ranking information in its logits is intact even though its decisions at the operating threshold are degenerate. A benchmark that reports accuracy and AUC without a class-wise breakdown or a correlation-based metric will describe such a model as merely weak rather than as broken.

This paper addresses both difficulties. We hold the data, the preprocessing, the augmentation, the optimiser, the schedule, the early-stopping criterion and the model-selection rule fixed, and vary only the architecture. Six architectures are evaluated on five public corpora and on a sixth mixed-provenance corpus assembled from all five, so that generalisation to an unknown manipulation family is measured directly rather than inferred. Every pairing is reported with seven metrics including the Matthews correlation coefficient, which alone among the common metrics approaches zero for a collapsed classifier.

A. Objectives of the Study

The work is organised around four objectives, each of which is realised in an identified part of the paper so that the correspondence between aim and evidence is explicit.

O1. To develop preprocessing that improves feature quality in deepfake images through resolution normalisation and contrast conditioning. This objective is realised in Section IV-C, where the front end that precedes every architecture is specified as a resampling, photometric-conditioning and standardisation chain in (5)–(8), and in Section III-C, which fixes its parameters identically for all six models so that preprocessing cannot confound the architectural comparison.

O2. To design a hybrid deep learning framework that achieves high accuracy in deepfake image detection. This objective is realised in Sections IV-E and IV-F, which propose FA-ViT-BiLSTM and MSCA-CTNet, and is evaluated in Sections VI-A, VI-E and VI-F, where MSCA-CTNet records the highest ROC-AUC of the study and both proposed architectures are ablated component by component.

O3. To integrate interpretability mechanisms that provide visual and readable insight into detection decisions. This objective is realised in Section IV-H, which identifies four quantities the framework exposes without additional computation the modality gate, the cross-attention correspondence, the class-wise decomposition and the component-wise attribution and is exercised in Sections VI-D, VI-F and VI-J, the last of which presents confidence-annotated per-sample decisions for every evaluation set.

O4. To evaluate robustness and generalisability across multiple corpora, including cross-dataset testing. This objective is realised in Sections V-B and V-C and evaluated in Sections VI-B, VI-G and VI-H through a mixed-provenance evaluation, a leave-one-corpus-out protocol in which the test corpus contributes no image to training, and a three-seed dispersion analysis.

B. Contributions

The contributions of this work are as follows.

• A controlled six-architecture, six-corpus benchmark in which every model is retrained from a common initialisation under one fixed protocol, so that observed differences are attributable to architecture rather than to data handling, and reported with six metrics for every model–corpus pairing.

• A preprocessing front end, specified in closed form and applied identically to all six architectures, that normalises spatial sampling rate and global tone across corpora acquired under incompatible capture and compression regimes.

• FA-ViT-BiLSTM, a proposed architecture that couples ViT patch tokens with log-magnitude Fourier tokens through a learned element-wise modality gate, allowing the network to attenuate the frequency stream on corpora where compression has destroyed spectral evidence.

• MSCA-CTNet, a proposed architecture that exchanges information between three-scale convolutional tokens and global transformer tokens through bidirectional multi-head cross-attention, achieving the highest ROC-AUC recorded in this study.

• A mixed-provenance evaluation and a leave-one-corpus-out protocol demonstrating that within-corpus ranking does not predict generalisation: the strongest single-corpus architecture degrades most, and a consistently second-place architecture generalises best.

• Component-wise ablations of both proposed architectures, isolating the contribution of the frequency branch, the modality gate, the multi-scale token construction and the bidirectional cross-attention.

• An analysis of degenerate classifier behaviour showing that above-chance ROC-AUC can coexist with chance-level accuracy and zero MCC, with the methodological consequence that class-wise reporting is necessary rather than optional.

The remainder of this paper is organised as follows. Section II reviews related work. Section III describes the corpora, preprocessing and experimental setup. Section IV presents the proposed methodology, including the two proposed architectures, their governing equations, the preprocessing front end and the interpretability mechanisms. Section V specifies the validation protocols. Section VI reports and discusses the results. Section VII concludes.

2. Related Work

A. Convolutional Detectors for Image Deepfakes

The dominant line of work treats forgery detection as binary classification over learned convolutional features. Soundarya et al. combined a convolutional network with hand-designed deep features [1], and subsequent work refined the feature extractor: Guardian-AI targets image-level artefacts directly [9], while Sharma and Selwal address the specific difficulty of low-quality inputs by optimising the preprocessing stage rather than the classifier [23]. Attention has been introduced into the convolutional stack to weight discriminative facial regions, with multi-dataset validation [25]. Explainability has become a parallel concern; an ensemble of convolutional detectors with attribution analysis has been proposed to make the decision auditable rather than opaque [31].

The baseline for the present study is the Dense-Swish-CNN with Bi-LSTM framework of Soundarya and Gururaj [22], which replaces ReLU with the Swish activation inside densely connected blocks so that negative activations propagate during feature computation, and passes the resulting descriptor to a bidirectional recurrent stage. That work reports accuracies between 96.5% and 97.76% across four corpora. The present study re-implements this architecture, together with the three hybrid baselines evaluated alongside it, so that all comparisons are made under identical conditions.

B. Hybrid and Spatio-Temporal Architectures

Because manipulation artefacts are distributed across a face rather than concentrated at one location, several authors append a sequential stage to the convolutional encoder. Zafar et al. fuse temporal and spatial features in a

hybrid framework [6], and Surapaneni et al. apply long short-term memory networks to video authentication [18]. Anand et al. take a more targeted approach, guiding video representations by facial action units so that localised manipulations, which are easily averaged away by global pooling, remain visible [21]. Parameter-efficient transfer has also been explored: the dual-level adapter of Shao et al. adapts a frozen backbone at both feature and prompt level, avoiding full fine-tuning [24].

C. Frequency-Domain and Generalisation-Oriented Methods

A distinct body of work argues that the pixel domain is the wrong place to look. Generative upsampling leaves periodic traces that are compact in the Fourier domain, and Liu et al. show that the frequency bias of a detector materially determines whether it generalises beyond its training generator [30]. Related efforts pursue generalisation by other routes: language-guided contrastive learning aligns visual evidence with semantic descriptions to transfer across synthesis methods [33], a quality-guided adapter conditions detection on perceived image quality for AI-generated content [36], and semantic-distribution alignment exploits the discrepancy between authentic and generated image statistics [37]. Hafezi et al. approach the problem from the generator side, using StyleGAN3 imagery to harden the detector [38]. A complementary proactive strategy embeds orthogonal moment watermarks before distribution, so that detection becomes verification rather than inference [32].

These results motivate the first architecture proposed here: rather than choosing between the pixel and frequency domains, FA-ViT-BiLSTM encodes both and learns a per-dimension gate that determines their relative contribution, so that the frequency stream can be attenuated where compression has destroyed spectral evidence.

D. Multimodal, Multi-Domain and LLM-Based Detection

Deepfakes are not confined to images. Audio detection has attracted spectral-fusion models [5], mixture-of-experts architectures [17], quantum-trained convolutional networks [2] and dedicated critical surveys [12]; text-level detection has been formulated as a separate paradigm [4]. At the multimodal level, Huang et al. use a large multimodal model to detect, localise and explain social-media image manipulation [10], and Ren et al. evaluate whether multi-modal reasoning language models can serve as detectors directly [15]. Recent transactions-level work extends this to news and rumour verification: two-stage knowledge distillation with multi-teacher self-training addresses the scarcity of labelled multimodal data [28], alternating modality optimisation balances the contribution of each modality during training [29], emotion cues are unified with view-specific learning to model annotation ambiguity [34], and a neuro-symbolic framework with large vision-language models exposes fine-grained deceptive patterns in an interpretable form [35]. The present work is confined to the image modality, but the cross-attention mechanism of MSCA-CTNet is architecturally close to the cross-modal exchange used in these systems.

E. Positioning of the Proposed Architectures

Both proposed architectures build on established ideas, and their novelty must therefore be stated precisely rather than claimed by assembly. Frequency-domain forgery detection is not new: F3-Net mines frequency-aware clues through frequency-component decomposition and local frequency statistics [43], and Liu et al. show that the frequency bias of a detector governs its generalisation [30]. What distinguishes FA-ViT-BiLSTM is not the use of spectral input but the fusion rule. F3-Net and its successors combine frequency and spatial evidence with a fixed architectural weighting learned once at training time and applied uniformly at inference. The gate of (17) is instead computed per input and per embedding dimension from the joint pooled context of both streams, so a single trained model can rely on spectral evidence for an uncompressed image and discount it for a heavily compressed one. Section VI-L reports that FA-ViT-BiLSTM is nonetheless the most compression-sensitive model in the study, which bounds how far this mechanism compensates and is reported as a negative result rather than omitted.

Similarly, hybrid CNN–transformer detectors already exist: ViXNet couples a vision transformer with an Xception network for forgery detection [44], and adapter-based designs attach transformer capacity to a frozen convolutional backbone [24]. These architectures fuse a single convolutional representation with a single transformer

representation, typically by concatenation or sequential stacking. MSCA-CTNet differs in two respects: the convolutional side contributes three distinct spatial scales, addressing the fact that blending seams and texture inconsistencies occupy different scales, and the exchange is bidirectional, so that transformer tokens are refined by convolutional evidence as well as the reverse. The ablations of Section VI-F isolate both choices.

The benchmark itself is positioned against DeepfakeBench [27] and Deepfake-Eval-2024 [16]. Those efforts standardise evaluation of released checkpoints across a large method set; this study retrains a smaller set from scratch under one protocol, which permits attribution of differences to architecture at the cost of breadth.

F. Surveys, Benchmarks and the Evaluation Gap

Several reviews map the field and its datasets [7], [13], [14], [20], and comparative studies assess detectors specifically in the context of misinformation prevention [8]. Two contributions bear directly on the motivation for this paper. Tolosana et al. analyse detection across generator generations and show that performance does not transfer cleanly between them [26]. Yan et al. introduceDeepfakeBench, arguing explicitly that inconsistent preprocessing and evaluation protocols make published numbers incomparable [27], a concern reinforced by the in-the-wild benchmark of Chandra et al. [16]. Broader analyses of robustness [19] and of artificial-intelligence-assisted cybercrime investigation [3] situate the problem in its operational context.

Table I summarises the surveyed literature. The gap this paper addresses is narrow and specific: the existing benchmarks compare published numbers or re-evaluate released checkpoints, whereas this study retrains every architecture from a common initialisation under one protocol and adds a mixed-provenance corpus that measures degradation directly.

TABLE I. Summary of Surveyed Literature

Ref.	Year	Modality / task	Approach and focus
[1]	2025	Image	CNN with deep-feature fusion for face forgery detection
[2]	2025	Audio	Quantum-trained CNN for audio deepfake detection
[3]	2025	General	AI-assisted investigation and prevention of cybercrime
[4]	2025	Text	Text-level paradigm for deepfake content detection
[5]	2025	Audio	Spectral-feature fusion model for audio deepfakes
[6]	2025	Video	Hybrid framework combining temporal and spatial features
[7]	2025	Review	Survey of AI algorithms and future prospects
[8]	2025	Comparative	Assessment of detectors for misinformation prevention

[9]	2025	Image	Guardian-AI deep learning detection model
[10]	2025	Image / multimodal	Large multimodal model for detection, localisation and explanation
[11]	2025	Security	Deepfake-driven social engineering in corporate environments
[12]	2025	Speech	Critical survey of speech deepfake detection
[13]	2025	Video	Review of video detection methods and challenges
[14]	2025	Review	Survey of autonomous detection techniques and evaluation
[15]	2025	Multimodal	Multi-modal reasoning LLMs evaluated as detectors
[16]	2025	Benchmark	In-the-wild multimodal benchmark of 2024 deepfakes
[17]	2025	Speech	Mixture-of-experts architecture for speech deepfakes
[18]	2025	Video	LSTM networks for video authentication
[19]	2025	Robustness	Analysis of detection techniques and their robustness
[20]	2025	Review	Review of datasets, tools and detection features
[21]	2025	Video	Action-unit-guided representations for localised manipulation
[22]	2025	Image	Dense-Swish-CNN with Bi-LSTM (baseline of the present study)
[23]	2025	Image	Optimised preprocessing for low-quality images
[24]	2025	Image / video	Dual-level adapter for parameter-efficient detection
[25]	2025	Image	Attention-enhanced CNN, multi-dataset evaluation

[26]	2022	Video	Facial-region analysis and fusion across generator generations
[27]	2023	Benchmark	DeepfakeBench unified evaluation protocol
[28]	2026	Multimodal news	Two-stage knowledge distillation with multi-teacher self-training
[29]	2026	Multimodal rumour	Alternating modality optimisation network
[30]	2026	Image forgery	Frequency bias for robust and generalised forgery detection
[31]	2026	Image	Explainable AI with an ensemble of CNNs
[32]	2026	Proactive	Orthogonal moment watermarking for proactive detection
[33]	2026	Synthetic image	Language-guided contrastive learning for generalisation
[34]	2026	Multimodal	View-specific learning with emotion-aware ambiguity modelling
[35]	2026	Multimodal news	Explainable neuro-symbolic framework with LVLMs
[36]	2026	AIGC image	Quality-guided forgery adapter
[37]	2026	AI-generated image	Semantic distribution and authenticity discrepancy alignment
[38]	2026	Image	StyleGAN3-based robust forgery identification

3. Datasets and Experimental Setup

A. Corpora

Six evaluation sets are used: five public single-source corpora and one mixed-provenance set assembled from them. They were selected to span three distinct generation regimes identity swapping, expression reenactment and fully synthetic face generation and two distinct quality regimes, uncompressed and compressed. Table II lists their public sources.

1) DFDC

The DeepFake Detection Challenge corpus [39] is the largest publicly released face-swap collection and was produced with paid consenting subjects across varied lighting, pose and background conditions. The facial-crop distribution used here preserves the generator diversity of the original release while removing the video decoding stage

from the pipeline. DFDC forgeries retain comparatively strong low-level generation residuals, which makes the corpus tractable for convolutional encoders and useful as an upper reference point.

2) Celeb-DF v2

Celeb-DF v2 [40] was constructed explicitly to remove the visual artefacts that made earlier corpora easy: colour mismatch at the blending boundary, low-resolution synthesised regions and temporal flicker are substantially reduced relative to first-generation datasets. It is therefore a harder identity-swap benchmark than DFDC, and the accuracy gap between the two corpora observed in Section VI is consistent with that design intent.

3) UADFV

UADFV [41] is a small early corpus of face-swapped videos, historically associated with head-pose inconsistency cues. Its manipulations are less refined than those of the later corpora, and in this study it functions as a low-difficulty control: an architecture that cannot separate UADFV reliably is not converging rather than encountering a hard problem. The cropped facial-image distribution is used.

4) FaceForensics++ (C23)

FaceForensics++ [42] provides four manipulation families Deepfakes, Face2Face, FaceSwap and NeuralTextures spanning both identity swapping and expression reenactment, over a shared set of pristine sequences. The C23 variant is compressed with H.264 at a constant rate factor of 23, approximating the quality of video redistributed through a social platform. This compression attenuates precisely the high-frequency residuals on which artefact-based detectors depend, and the corpus is accordingly the most difficult in this study. Labels are assigned by directory: original sequences and the YouTube source directory map to the Real class, and the four manipulation directories to the Fake class.

5) DeepDetect-2025

DeepDetect-2025 is a recent public compilation of authentic photographs paired with synthetically generated faces, and therefore probes the fully-synthetic regime rather than the face-swap regime addressed by the other four corpora. Because it is a recent community release, its exact generator composition should be verified against the source documentation when reporting.

6) Combined

The Combined set draws samples across all five corpora into a single class-balanced evaluation, so that a detector is confronted with identity swaps, reenactments and fully synthetic faces at several compression levels without any indication of provenance. This is the condition under which a deployed forensic detector operates, and it is the evaluation on which the ranking of Section VI changes.

TABLE II. Evaluation Corpora

Corpus	Generation regime	Public source (Kaggle slug)	Test set
DFDC	Identity swap	ashifurrahman34/dfdc-dataset	150 / 150
Celeb-DF v2	Identity swap (refined)	pranabr0y/celebdf-v2image-dataset	150 / 150
UADFV	Identity swap (early)	dhanvinan/uadfv-cropped-face-dataset	150 / 150
FaceForensics++ C23	Swap + reenactment, compressed	gradientvoyager/faceforensics-c23-extracted-faces-100k	150 / 150
DeepDetect-2025	Fully synthetic	ayushmandatta1/deepdetect-2025	150 / 150

Combined	Mixed provenance	Union of the five corpora above	150 / 150

Test set column gives Real / Fake sample counts. An additional DFDC facial-image source (itamargr/dfdc-faces-of-the-train-sample) was consulted during dataset preparation.

All corpora are used for research purposes only under the terms of their respective releases. Celeb-DF v2 and FaceForensics++ are distributed under end-user licence agreements that restrict redistribution and require acknowledgement of the originating authors; the derived facial-crop distributions used here are cited to their original publications [39]–[42], and no identifiable subject is presented outside the context of forgery analysis. The qualitative panels in Section VI reproduce frames from these public research corpora solely to illustrate detector behaviour.

B. Label Inference and Partitioning

The corpora use incompatible directory conventions, so labels are inferred by matching lowercase path tokens against a real-token set (real, original, authentic, genuine, pristine, youtube) and a fake-token set (fake, deepfake, manipulated, synthetic, forged, altered, faceswap, face2face, neuraltextures), with corpus-specific overrides applied first for FaceForensics++. Each corpus is then split in a stratified manner into 70% training, 10% validation and 20% test partitions, preserving the class ratio in every partition. Because splitting precedes all training, no image appears in more than one partition.

C. Preprocessing Configuration

The preprocessing front end is specified in closed form in Section IV-C; this subsection fixes its parameters. Facial crops arrive at heterogeneous native resolutions, so every image is resampled to a common 160×160 lattice by bicubic interpolation, a resolution that matches the regime of socially redistributed imagery while bounding computational cost, and is then standardised with ImageNet channel statistics as in (8). Training samples additionally receive random horizontal flipping, vertical flipping with probability 0.10, rotation within $\pm 15°$, affine translation of up to 10%, scaling between 0.90 and 1.10, shear of up to 8°, and photometric conditioning in brightness, contrast, saturation and hue. Validation and test samples receive resampling and standardisation only, so that reported metrics are not inflated by test-time augmentation. The entire chain is identical for all six architectures and for all six evaluation sets, which is what permits differences in the results to be attributed to architecture.

D. Implementation and Training Protocol

All models are implemented in PyTorch with pretrained backbones obtained from the timmlibrary, and trained on a single GPU with automatic mixed precision. Every architecture receives an identical optimiser, schedule and stopping rule; only the network differs. Backbone parameters are frozen for the first epoch so that the randomly initialised projection layers do not corrupt pretrained features. Checkpoint selection uses validation ROC-AUC rather than validation accuracy, because AUC is threshold-independent and therefore does not reward a checkpoint that has collapsed toward a single class. All random number generators are seeded identically. Table III lists the complete configuration.

TABLE III. Training Configuration

Parameter	Value	Remark
Framework	PyTorch + timm	Pretrained ImageNet backbones
Input resolution	$160 \times 160 \times 3$	Matches low-quality web imagery
Resampling kernel	Bicubic	`Cubic convolution, a = −0.5`
Batch size	32	Identical for all six models

Optimiser	AdamW	Decoupled weight decay
Learning rate	1×10^{-4}	Initial value
Weight decay	1×10^{-4}	—
Loss	Cross-entropy, smoothing 0.05	Improves calibration
Scheduler	ReduceLROnPlateau	Factor 0.5, patience 2, on val. AUC
Maximum epochs	15	Early stopping patience 4
Backbone freezing	First epoch	Stabilises projection layers
Mixed precision	Enabled	Automatic mixed precision on GPU
Model selection	Best validation ROC-AUC	Threshold-independent criterion
Random seed	42	Python, NumPy and PyTorch
Split ratio	70 / 10 / 20	Stratified train / validation / test

4. Proposed Methodology

A. Overview

Six architectures are evaluated. Four re-implement the established hybrid familyCNN+Bi-LSTM, ResNet34+EfficientNet+Bi-LSTM, MobileNet+EfficientNet+Bi-LSTM and Dense-Swish-CNN+Bi-LSTM and are included so that the comparison is made under identical conditions rather than against numbers reported under differing regimes [22]. Two are proposed in this work: FA-ViT-BiLSTM and MSCA-CTNet. The complete pipeline is shown in Fig. 1: images are indexed from the five corpora, labelled, split, passed through the preprocessing front end of Section IV-C and delivered to each architecture, which is trained separately on each corpus and once on the Combined corpus, giving thirty-six independent runs.

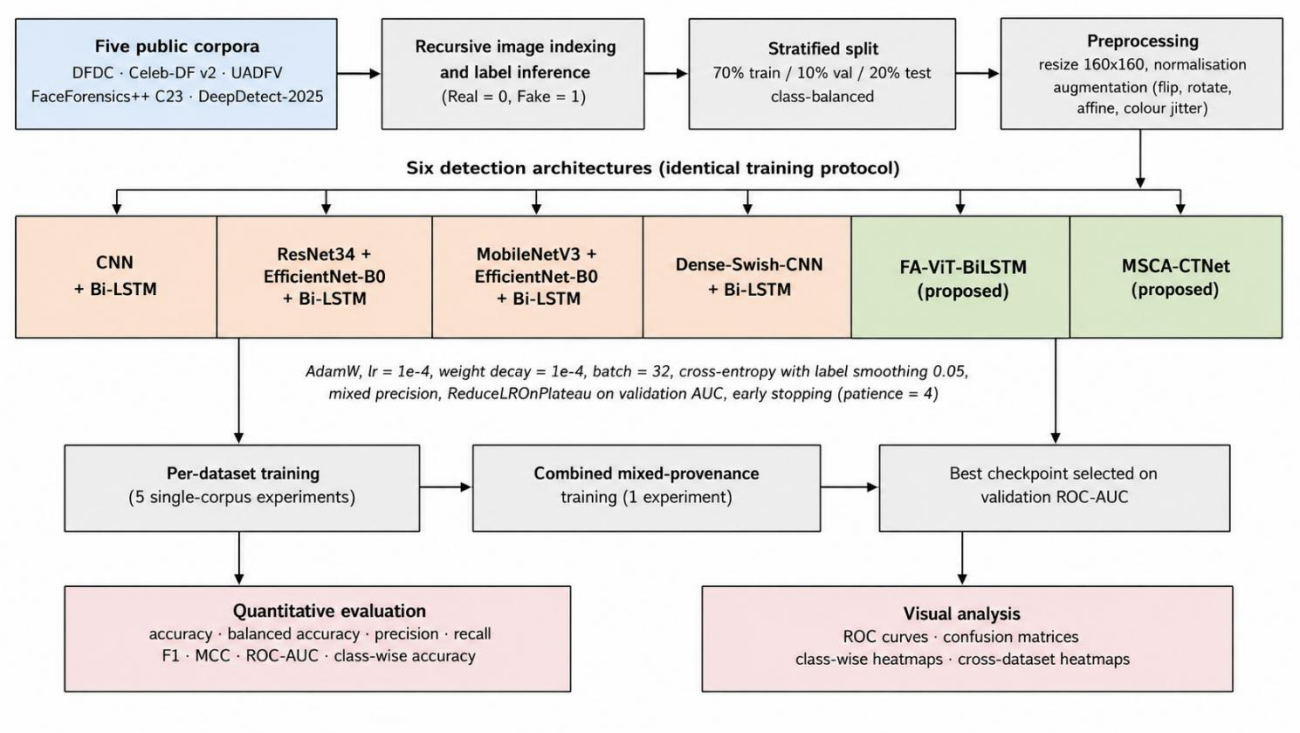

Fig. 1. End-to-end methodology of the proposed benchmarking framework, from corpus indexing through stratified splitting, preprocessing and six-architecture training to quantitative and visual evaluation.

B. Problem Formulation

Let the complete image collection be denoted as in (1), where each sample is a cropped facial image and each label indicates authenticity.

$$I = \{I_1, I_2, \ldots, I_n\},\ Y = \{y_1, y_2, \ldots, y_n\},\ y_i \in \{0, 1\} \qquad (1)$$

The label convention is given in (2): the pristine class is indexed zero and the manipulated class one, so the Fake class is positive for all threshold-dependent metrics.

$$y_i = 0 \text{ if } I_i \in I\_real;\quad y_i = 1 \text{ if } I_i \in I\_fake \qquad (2)$$

With five source corpora contributing real and manipulated subsets, the union is expressed in (3), and detection is framed as learning a mapping $f\theta$ that estimates the posterior probability of manipulation, as in (4).

$$I\_total = \cup_k I\_real^{(k)} \cup \cup_j I\_fake^{(j)},\quad k = 1 \ldots p,\ j = 1 \ldots q \qquad (3)$$

$$f\theta : \mathbb{R}^{(H\times W\times 3)} \rightarrow [0, 1]^2,\quad \hat{p}(y = 1 \mid I) = \text{softmax}(f\theta(I))_1 \qquad (4)$$

C. Preprocessing Front End for Feature-Quality Conditioning

The five corpora were acquired, cropped and compressed under mutually incompatible regimes, so the raw inputs differ in spatial sampling rate, dynamic range and global tone before any manipulation artefact is considered. If those differences are left uncontrolled they enter the model as corpus identity rather than as forensic evidence, and the mixed-provenance evaluation of Section VI-B becomes uninterpretable. The front end specified here therefore has two purposes: to raise the quality of the features presented to every architecture, and to hold the presentation constant so that architecture remains the only free variable.

Resolution is normalised first. Each crop is resampled onto the common 160 × 160 lattice by bicubic convolution, as in (5), where k is the cubic kernel with a = −0.5 and p denotes the target sample position expressed in source coordinates. The choice of kernel is not cosmetic: nearest−neighbour resampling injects a broadband step response into precisely the high−frequency band that carries generative upsamplingtraces, and would therefore contaminate the spectral stream of (16) with resampling artefacts indistinguishable from generator artefacts. Bicubic interpolation is C^1−continuous and confines its own spectral leakage to a narrow band around the Nyquist limit of the target lattice.

$$x_r(p) = \Sigma_m \Sigma_n x(m, n) \cdot k(p_x - m) \cdot k(p\_y - n),\quad k(\cdot) \text{ cubic}, a = -0.5 \qquad (5)$$

Contrast and brightness are conditioned next, as in (6). The contrast operation blends each channel toward the mean luminance $\bar{g}$ of the crop with coefficient λ_c, which expands the dynamic range when $\lambda\_c > 1$ and compresses it when $\lambda\_c < 1$; the brightness operation scales the result by λ_b. Both coefficients are drawn per sample from a bounded interval during training and fixed at unity at evaluation. Because the transform is referenced to the mean luminance of the individual crop rather than to a global constant, it removes the corpus-level tone offsets introduced by different capture and re-encoding chains while preserving the local contrast relations in which blending seams are visible.

$$\tilde{x}\_c = \text{clip}_0^1(\ \lambda\_b\ [\ \lambda\_c x\_c + (1 - \lambda\_c)\ \bar{g}\ ]),\quad \bar{g} = \text{mean}(\text{gray}(x)),\ \lambda\_c, \lambda\_b \sim U[1-\delta, 1+\delta] \qquad (6)$$

Geometric conditioning follows, as in (7), where the sampled affine operator composes a rotation, an anisotropic scaling, a shear and a translation. Applying the geometry after the photometric stage keeps the interpolation

of (5) and (7) on the same kernel and avoids a second resampling of already contrast-stretched values. Finally the tensor is standardised per channel with ImageNet statistics, as in (8), which is the form in which pretrained backbones expect their input and which places all six architectures on the same input scale.

$$x_a(p) = \tilde{x}(A\,p + t), \quad A = R(\theta)\,S(s)\,Sh(\psi), \quad \theta \in [-15°, 15°],\ s \in [0.90, 1.10],\ \psi \le 8° \qquad (7)$$

$$\hat{x}\_c = (x_a\_c - \mu\_c) / \sigma\_c, \quad \mu = (0.485, 0.456, 0.406), \quad \sigma = (0.229, 0.224, 0.225) \qquad (8)$$

Two properties of this front end matter for the claims made later. It is deterministic at evaluation time, so no reported metric benefits from test-time augmentation; and it contains no learned parameters, so it cannot absorb corpus-specific information during training and then release it at test time. The stronger enhancement operators that the literature has proposed for degraded inputs — adaptive histogram equalisation and learned super-resolution among them [23] — are deliberately excluded for this reason: each introduces a second variable whose effect could not be separated from the architectural effect this study is designed to measure. Their integration is identified as future work in Section VII.

D. Common Building Blocks

All six architectures share the Swish activation and the Bi-LSTM classification head. The Swish activation, defined in (9), is smooth and non-monotonic with an unbounded upper limit and a bounded lower limit, and unlike ReLU it permits a controlled flow of negative activations.

$$\text{Swish}(x) = x \cdot \sigma(x), \quad \sigma(x) = 1 / (1 + e^{-x}) \qquad (9)$$

Every backbone produces a feature map of shape $B \times C \times H \times W$, converted into a token sequence by spatial flattening as in (10), so that spatial positions become sequence steps and the recurrent head models dependencies between facial regions.

$$Z = \phi(X) \in \mathbb{R}^{(B \times HW \times C)}, \quad z_t = X[:, :, h, w], \quad t = h \cdot W + w \qquad (10)$$

The Bi-LSTM head processes the sequence in both directions and concatenates the hidden states at each step, as in (11). Mean pooling over the sequence, in (12), produces a fixed-length descriptor, classified through dropout and a linear layer as in (13).

$$h_t^{(f)} = \text{LSTM\_fwd}(z_t, h_{t-1}^{(f)}), \quad h_t^{(b)} = \text{LSTM\_bwd}(z_t, h_{t+1}^{(b)}), \quad h_t = [h_t^{(f)} ; h_t^{(b)}] \qquad (11)$$

$$\bar{h} = (1/T) \sum_{t=1}^{T} h_t \qquad (12)$$

$$\hat{y} = \text{softmax}(W\_out \cdot \text{Dropout}(\bar{h}) + b\_out) \qquad (13)$$

1) Dense-Swish Dense Block

The Dense-Swish-CNN baseline stacks dense layers in which each layer applies batch normalisation, Swish activation and a 3 × 3 convolution producing g new feature maps, concatenated with the layer input as in (14). A block of L such layers grows the channel count from C_0 to $C_0 + Lg$, after which a 1 × 1 transition convolution and average pooling reduce width and resolution.

$$x_l = [x_{l-1} ; W_l * \text{Swish}(\text{BN}(x_{l-1}))], \quad C_l = C_{l-1} + g \qquad (14)$$

2) Dual-Backbone Feature Fusion

The two hybrid baselines fuse complementary pretrained backbones. Each output is projected to a common width of 256 channels by a 1 × 1 convolution, the second stream is bilinearly resampled to the resolution of the first, and the two are concatenated along the channel axis before sequence conversion, as in (15).

$$X\_fused = [\, W_1 * X\_A \,; \Psi(\, W_2 * X\_B \,)\,] \in \mathbb{R}^{(B \times 512 \times H \times W)} \qquad (15)$$

E. Proposed Architecture 1: FA-ViT-BiLSTM

Generative pipelines leave characteristic periodic traces in the frequency domain that are only weakly visible in the pixel domain, particularly after compression — an observation supported by recent analyses of frequency bias in forgery detection [30]. The first proposed architecture therefore couples a spatial transformer stream with an explicit frequency stream and learns how much to trust each. The design is shown in Fig. 2.

The frequency stream applies a two-dimensional discrete Fourier transform with orthonormal scaling, shifts the zero-frequency component to the spectrum centre, and takes the log-magnitude as in (16). The compressive logarithm prevents the low-frequency DC region from dominating. Three convolution–batch-norm–GELU stages encode this spectral map to the transformer embedding width.

$$S = \log(1 + |\mathrm{shift}(\mathcal{F}(x))|), \quad \mathcal{F}(x)[u,v] = (1/\sqrt{(HW)})\, \Sigma_h\, \Sigma\_w\, x[h,w]\, e^{(-2\pi i(uh/H + vw/W))} \qquad (16)$$

The spatial stream is a ViT-Tiny backbone from which the class token is discarded, retaining the patch tokens. The two token sequences are truncated to a common length, mean-pooled and concatenated into a global context vector from which a sigmoid gate is computed. The gate performs an element-wise convex combination of the modalities, as in (17). Because it is conditioned on both streams jointly, the network can suppress the frequency stream where compression has destroyed spectral evidence, and suppress the spatial stream where texture cues are unreliable. The fused sequence is passed to the shared head of (11)–(13).

$$g = \sigma(\, W\_g\, [\, \mathrm{mean}(T\_s)\, ;\, \mathrm{mean}(T\_f)\,]), \quad Z = g \odot T\_s + (1 - g) \odot T\_f \qquad (17)$$

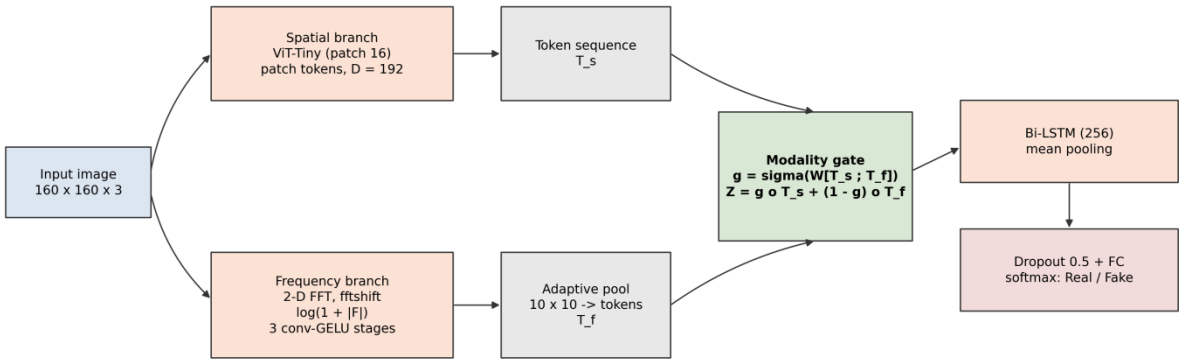

Fig. 2. Architecture of the proposed Frequency-Aware Vision Transformer with Bi-LSTM head (FA-ViT-BiLSTM). A learned modality gate performs an element-wise convex combination of spatial patch tokens and log-magnitude spectral tokens.

F. Proposed Architecture 2: MSCA-CTNet

The second proposed architecture addresses a different limitation. Manipulation artefacts occur at several spatial scales, from blending seams at the face boundary to fine texture inconsistency within a region, and a single-scale token set cannot represent both. MSCA-CTNet extracts a multi-scale convolutional token set and a global transformer token set, and lets the two exchange information through bidirectional cross-attention. The design is shown in Fig. 3.

Three ConvNeXt-Tiny stages are projected to a common width and adaptively pooled to 10 × 10, 7 × 7 and 5 × 5 grids, then concatenated into one multi-scale token set. A Swin-Tiny stage supplies the transformer token set. Scaled dot-product attention is defined in (18); the bidirectional exchange is given in (19), residual layer normalisation in (20), and feed-forward refinement of the concatenated sequence in (21). The fused sequence is mean-pooled and classified by a normalised two-layer head with GELU activation and dropout, as in (22).

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(Q K^{T} / \sqrt{d_k} \right) V \quad (18)$$

$$A = \mathrm{MHA}(T\_c, T\_t, T\_t), \quad B = \mathrm{MHA}(T\_t, T\_c, T\_c) \quad (19)$$

$$\tilde{T}\_c = \mathrm{LN}(T\_c + A), \quad \tilde{T}\_t = \mathrm{LN}(T\_t + B) \quad (20)$$

$$Z = \mathrm{LN}(U + \mathrm{FFN}(U)), \quad U = [\tilde{T}\_c ; \tilde{T}\_t] \quad (21)$$

$$\hat{y} = \mathrm{softmax}(W_2 \cdot \mathrm{Dropout}(\mathrm{GELU}(W_1 \cdot \mathrm{LN}(\mathrm{mean}(Z))))) \quad (22)$$

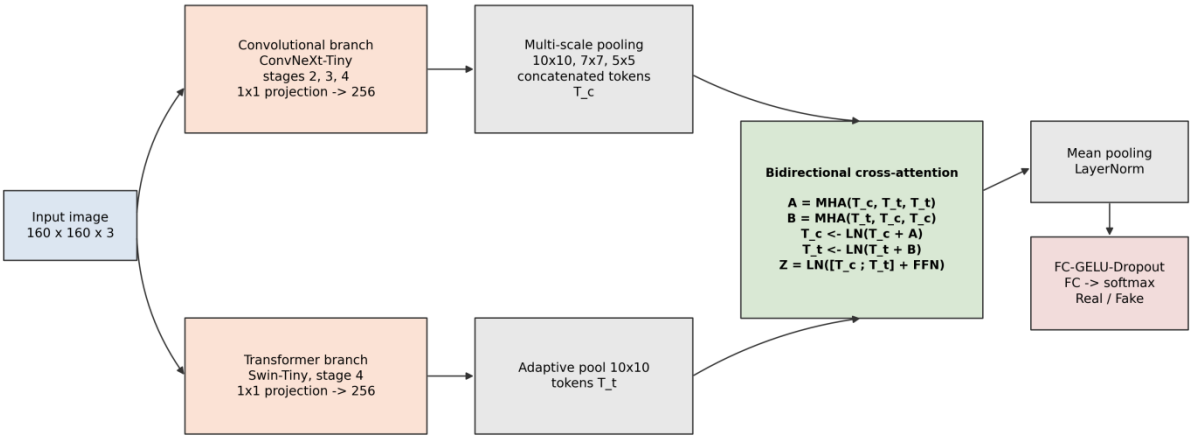

Fig. 3. Architecture of the proposed Multi-Scale Cross-Attention CNN–Transformer Network (MSCA-CTNet). Multi-scale ConvNeXt tokens and Swin transformer tokens are exchanged through bidirectional multi-head cross-attention before classification.

G. Training Objective and Optimisation

All models are optimised with cross-entropy loss under label smoothing, as in (23), which discourages over-confident logits and improves calibration on the harder corpora. Parameters are updated with AdamW using decoupled weight decay, as in (24).

$$L = -\sum_k q_k \log p_k, \quad q_k = (1 - \varepsilon) \cdot 1[k = y] + \varepsilon / K, \quad \varepsilon = 0.05 \quad (23)$$

$$\theta_{t+1} = \theta_t - \eta \left(\hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon) + \lambda \theta_t \right) \quad (24)$$

H. Interpretability Mechanisms

A detector whose decision cannot be inspected is of limited forensic value, because an analyst must be able to state on what evidence a piece of media was flagged. The framework exposes four such quantities, none of which requires a separate explanation model or additional forward passes.

First, the modality gate g of (17) is directly readable. Its mean over the embedding dimensions is a scalar in [0, 1] giving the weight the network assigned to spatial as against spectral evidence for that specific image, so a decision can be reported as spatially driven or spectrally driven rather than as an unattributed score. Because the gate is computed per input, this attribution varies across corpora and is the mechanism by which the compression sensitivity discussed in Section VI-L becomes measurable rather than merely hypothesised.

Second, the bidirectional exchange of (19) yields two attention matrices whose rows associate a convolutional token at a known spatial scale with the transformer tokens it drew from, and conversely. Because the convolutional tokens of MSCA-CTNet are pooled onto explicit 10 × 10, 7 × 7 and 5 × 5 grids, each row of A maps back to a bounded

region of the input face, so the exchange is spatially localisable by construction rather than through a post hoc approximation.

Third, the class-wise decomposition of Section VI-D separates the aggregate score into Real and Fake accuracy, and the Matthews correlation coefficient of (30) collapses to zero for a degenerate classifier. Together these turn a failure that aggregate accuracy conceals into an observable one, and they are what identifies the collapsed baseline reported in Section VI-D.

Fourth, the component-wise ablations of Section VI-F attribute performance to individual design decisions at the level of the architecture rather than the individual image, and the confidence-annotated panels of Section VI-J present per-sample decisions with their softmax confidence, which is what exposes the high-confidence error mode described there. Pixel-level saliency attribution is not applied in this study and is identified in Section VII as the natural extension of this component.

I. Algorithms

Algorithm 1 summarises the complete experimental procedure. Algorithms 2 and 3 give the forward computation of the two proposed architectures.

Algorithm 1: Six-Model Cross-Corpus Benchmarking Procedure

Input: corpora $C = \{C_1 \dots C_5\}$, model set $M = \{M_1 \dots M_6\}$, config Γ

Output: metric table R, prediction sets P, figures F

```
for each corpus c ∈ C ∪ {Combined} do
    D ← IndexImages(c)                          // recursive scan
    for each (path, name) ∈ D do
        y ← InferLabel(path, c)                 // token matching
    (D_tr, D_val, D_te) ← StratifiedSplit(D, 0.7, 0.1, 0.2)
    apply front end (5)–(8) with augmentation to D_tr; deterministic (5), (8) to D_val, D_te
    for each model m ∈ M do
        θ ← InitialiseFromPretrained(m); freeze backbones for 1 epoch
        best ← −∞; k ← 0
        for epoch e = 1 to E_max do
            for each mini−batch (x, y) ∈ D_tr do
                p ← softmax(fθ(x)); L ← CE_smooth(p, y)     // Eq. (23)
                θ ← AdamW(θ, ∇θL)                           // Eq. (24)
            a ← ROC-AUC(fθ, D_val)
            Scheduler.step(a)
            if a > best then best ← a; θ* ← θ; k ← 0
            else k ← k + 1; if k ≥ patience then break
        restore θ*; (y, ŷ, p̂) ← Evaluate(fθ*, D_te)
```

19: R ← R ∪ Metrics(y, ŷ, p̂) // Eq. (25)–(30)

20: P[c, m] ← (y, ŷ, p̂)

21: F ←GenerateFigures(R, P)

22: return R, P, F

Algorithm 2: Forward Pass of FA-ViT-BiLSTM

Input: preprocessed image $x \in \mathbb{R}^{(B\times3\times160\times160)}$

Output: class posterior $\hat{y} \in \mathbb{R}^{(B\times2)}$, gate attribution $\bar{g}$

1: $T_s \leftarrow$ ViT_Tiny.forward_features(x)

2: if T_s has a class token then $T_s \leftarrow T_s[:,1:,:]$

3: $\mathcal{F} \leftarrow$ fft2(x, norm = "ortho"); $\mathcal{F} \leftarrow$ fftshift($\mathcal{F}$)

4: $S \leftarrow \log(1 + |\mathcal{F}|)$ // Eq. (16)

5: $T_f \leftarrow$ Flatten(AdaptiveAvgPool$_{10\times10}$(ConvGELU^3(S)))

6: $n \leftarrow \min(\text{len}(T_s), \text{len}(T_f))$; truncate both to n tokens

7: $u \leftarrow$[mean(T_s) ; mean(T_f)]

8: $g \leftarrow \sigma(W_{gu})$ // modality gate

9: $Z \leftarrow g \odot T_s + (1 - g) \odot T_f$ // Eq. (17)

10: $H \leftarrow$BiLSTM(Z); $\bar{h} \leftarrow \text{mean}_t(H)$ // Eq. (11), (12)

11: $\hat{y} \leftarrow$ softmax($W_{out} \cdot$ Dropout($\bar{h}$) + b_{out}) // Eq. (13)

12: $\bar{g} \leftarrow$ mean_d(g) // per-image attribution, Sec. IV-H

13: return $\hat{y}$, $\bar{g}$

Algorithm 3: Forward Pass of MSCA-CTNet

Input: preprocessed image $x \in \mathbb{R}^{(B\times3\times160\times160)}$

Output: class posterior $\hat{y} \in \mathbb{R}^{(B\times2)}$, cross-attention maps A, B

1: $(X_2, X_3, X_4) \leftarrow$ ConvNeXt_Tiny(x) // stages 2, 3, 4

2: for $(X_i, s_i) \in \{(X_2,10), (X_3,7), (X_4,5)\}$ do

3: $T_i \leftarrow$ Flatten(AdaptiveAvgPool_{$s_i\times s_i$}(Conv$_{1\times1}$(X_i)))

4: $T_c \leftarrow [\, T_1 ; T_2 ; T_3 \,]$ // multi-scale tokens

5: $X_t \leftarrow$ Swin_Tiny(x)[stage 4]

6: $T_t \leftarrow$ Flatten(AdaptiveAvgPool$_{10\times10}$(Conv$_{1\times1}$(X_t)))

7: $A \leftarrow$ MHA(T_c, T_t, T_t); $B \leftarrow$ MHA(T_t, T_c, T_c) // Eq. (19)

8: $\tilde{T}_c \leftarrow$ LN(T_c + A); $\tilde{T}_t \leftarrow$ LN(T_t + B) // Eq. (20)

9: $U \leftarrow [\, \tilde{T}_c \,; \tilde{T}_t \,]$; $Z \leftarrow \mathrm{LN}(U + \mathrm{FFN}(U))$ // Eq. (21)

10: $\hat{y} \leftarrow \mathrm{softmax}(\, W_2 \cdot \mathrm{Dropout}(\, \mathrm{GELU}(\, W_1 \cdot \mathrm{LN}(\mathrm{mean}(Z))\,)\,)\,)$ // Eq. (22)

11: return $\hat{y}$, A, B

J. Evaluation Metrics

Let TP, TN, FP and FN denote confusion-matrix entries with the Fake class as positive. The reported metrics are defined in (25)–(30). The Matthews correlation coefficient is included because it is the only one of these that approaches zero for a classifier collapsed onto a single class, a failure mode observed in this study.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (25)$$

$$\text{Precision} = TP / (TP + FP), \quad \text{Recall} = TP / (TP + FN) \quad (26)$$

$$\text{Specificity} = TN / (TN + FP) \quad (27)$$

$$\text{Balanced Accuracy} = \tfrac{1}{2}\,(\text{Recall} + \text{Specificity}) \quad (28)$$

$$F1 = 2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall}) \quad (29)$$

$$\mathrm{MCC} = (TP \cdot TN - FP \cdot FN) \,/\, \sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)} \quad (30)$$

5. Experimental Design and Validation

Beyond the main benchmark, four additional protocols are specified so that the architectural claims are falsifiable rather than merely descriptive. Each is defined here and its results are reported in Section VI.

A. Ablation Design

Each proposed architecture is ablated one component at a time on a single representative corpus, so that every reported delta is attributable to a single design decision. For FA-ViT-BiLSTM six variants are trained: the full model; the spatial stream alone, which isolates the contribution of the frequency branch; the frequency stream alone; channel concatenation of both streams in place of the gate; an unweighted mean of both streams; and gated fusion with the Bi-LSTM head replaced by mean pooling and a linear classifier, which isolates the sequence model. For MSCA-CTNet seven variants are trained: the full model; single-scale and two-scale convolutional token sets, which isolate the multi-scale construction; concatenation in place of cross-attention; unidirectional cross-attention in which only convolutional tokens query; the convolutional branch alone; and cross-attention without the feed-forward refinement of (21). All variants share the training configuration of Table III and are evaluated with the metrics of (25)–(30).

B. Leave-One-Corpus-Out Cross-Dataset Protocol

The Combined evaluation of Section VI-B mixes provenance but does not withhold a generator family, since every corpus contributing to the test set also contributes to training. Generalisationto a genuinely unseen manipulation family requires a stricter protocol. Under leave-one-corpus-out, each of the five corpora is held out in turn; the model is trained on the union of the remaining four and tested on the held-out corpus, which contributes no image to training or validation. ROC-AUC is the primary metric because the decision threshold calibrated on one generator family does not transfer to another, and accuracy would therefore conflate representational transfer with threshold miscalibration. This protocol is the one adopted by the cross-generation analysis of Tolosana et al. [26] and by DeepfakeBench [27], and it is the strongest generalisation evidence reported here.

C. Seed Variance and Statistical Testing

Each evaluation set contains 300 samples, so one sample corresponds to approximately 0.33 percentage points and differences of one to two points between closely ranked models cannot be distinguished from sampling noise by inspection. Two safeguards are therefore applied. First, every model–corpus pairing is trained under three random

seeds and reported as mean ± standard deviation, which separates architectural effect from initialisation effect. Second, pairwise comparisons between the top-ranked models on a given corpus are tested with the exact McNemar test on paired predictions, and accuracies are accompanied by percentile bootstrap 95% confidence intervals over 2,000 resamples. A difference is described as significant only where the McNemar p-value falls below 0.05.

D. Comparison with Published Methods

Finally, the proposed architectures are placed against published detectors evaluated on the same corpora. Because published numbers are obtained under differing partitions and training budgets, the comparison table records the evaluation protocol of each cited result alongside its score, and no claim of superiority is made where the protocols differ materially.

6. Results and Discussion

A. Overall Detection Performance

Because every evaluation set is class-balanced at 150 Real and 150 Fake samples, accuracy and balanced accuracy coincide by (25) and (28), and the chance-level baseline is 0.500 throughout. Table IV reports every metric for each model–corpus pair on the five individual corpora.

TABLE IV. Comprehensive Performance Comparison Across Five Deepfake Corpora

Dataset	Model	Acc.	Prec.	Rec.	F1	MCC	AUC
Celeb-DF v2	CNN_BiLSTM	0.500	0.500	0.993	0.665	0.000	0.657
	ResNet34_EfficientNet_BiLSTM	*0.863	0.871	0.853	0.862	0.727	0.926
	MobileNet_EfficientNet_BiLSTM	0.813	0.851	0.760	0.803	0.630	0.895
	DenseSwishCNN_BiLSTM	0.617	0.587	0.787	0.672	0.248	0.637
	FA_ViT_BiLSTM	0.823	0.821	0.827	0.824	0.647	0.888
	MSCA_CTNet	0.783	0.754	0.840	0.795	0.570	0.860
DFDC	CNN_BiLSTM	0.640	0.633	0.667	0.649	0.280	0.706
	ResNet34_EfficientNet_BiLSTM	*0.953	0.986	0.920	0.952	0.909	0.980
	MobileNet_EfficientNet_BiLSTM	0.933	0.945	0.920	0.932	0.867	0.976
	DenseSwishCNN_BiLSTM	0.650	0.649	0.653	0.651	0.300	0.703
	FA_ViT_BiLSTM	0.943	0.965	0.920	0.942	0.888	0.984
	MSCA_CTNet	0.950	0.979	0.920	0.948	0.902	0.987
UADFV	CNN_BiLSTM	0.717	0.788	0.593	0.677	0.447	0.785
	ResNet34_EfficientNet_BiLSTM	0.960	0.954	0.967	0.960	0.920	0.997
	MobileNet_EfficientNet_BiLSTM	*0.970	0.967	0.973	0.970	0.940	0.998
	DenseSwishCNN_BiLSTM	0.487	0.489	0.580	0.530	−0.027	0.502
	FA_ViT_BiLSTM	*0.970	0.961	0.980	0.970	0.940	0.998

	MSCA_CTNet	0.953	0.947	0.960	0.954	0.907	0.993
FaceForensics++ (C23)	CNN_BiLSTM	0.600	0.609	0.560	0.583	0.201	0.650
	ResNet34_EfficientNet_BiLSTM	*0.707	0.712	0.693	0.703	0.413	0.766
	MobileNet_EfficientNet_BiLSTM	0.677	0.685	0.653	0.669	0.354	0.738
	DenseSwishCNN_BiLSTM	0.600	0.599	0.607	0.603	0.200	0.601
	FA_ViT_BiLSTM	0.567	0.549	0.747	0.633	0.143	0.658
	MSCA_CTNet	0.670	0.695	0.607	0.648	0.343	0.763
DeepDetect-2025	CNN_BiLSTM	0.520	0.516	0.640	0.571	0.041	0.575
	ResNet34_EfficientNet_BiLSTM	*0.847	0.799	0.927	0.858	0.702	0.913
	MobileNet_EfficientNet_BiLSTM	0.840	0.827	0.860	0.843	0.681	0.913
	DenseSwishCNN_BiLSTM	0.540	0.533	0.647	0.584	0.082	0.552
	FA_ViT_BiLSTM	0.760	0.735	0.813	0.772	0.523	0.864
	MSCA_CTNet	0.777	0.837	0.687	0.755	0.563	0.868

Fake is the positive class. An asterisk marks the best accuracy within each corpus.

The two dual-backbone hybrids dominate the individual-corpus evaluations. ResNet34+EfficientNet+Bi-LSTM attains the highest accuracy on Celeb-DF v2 (0.863), DFDC (0.953), FaceForensics++ C23 (0.707) and DeepDetect-2025 (0.847); on UADFV it is displaced by MobileNet+EfficientNet+Bi-LSTM and the proposed FA-ViT-BiLSTM, which tie at 0.970. Averaged over the five corpora, ResNet34+EfficientNet+Bi-LSTM leads with 0.866 accuracy, followed by MobileNet+EfficientNet+Bi-LSTM (0.847), the proposed MSCA-CTNet (0.827) and the proposed FA-ViT-BiLSTM (0.813). The plain CNN+Bi-LSTM baseline averages 0.595 and Dense-Swish-CNN+Bi-LSTM 0.579 under this training budget.

UADFV is close to saturation for the four stronger architectures, all of which exceed 0.953 accuracy with AUC at or above 0.993, consistent with its role as a low-difficulty control. The corpus remains informative at the lower end: Dense-Swish-CNN+Bi-LSTM records 0.487 accuracy, AUC 0.502 and MCC −0.027, statistically indistinguishable from a random classifier and marginally below chance. Since the same architecture reaches 0.650 on DFDC, this is better read as a convergence failure on UADFV than as a uniform capacity limit.

B. Performance on the Combined Mixed-Provenance Set

Table V reports the same six metrics on the Combined set, in which samples from all five corpora are pooled and no indication of provenance is available to the detector.

TABLE V. Performance on the Combined Mixed-Provenance Test Set

Dataset	Model	Acc.	Prec.	Rec.	F1	MCC	AUC
Combined	CNN_BiLSTM	0.577	0.549	0.860	0.670	0.186	0.652

Combined	ResNet34_EfficientNet_BiLSTM	0.663	0.642	0.740	0.687	0.331	0.725
Combined	MobileNet_EfficientNet_BiLSTM	*0.717	0.783	0.600	0.679	0.446	0.789
Combined	DenseSwishCNN_BiLSTM	0.537	0.543	0.460	0.498	0.074	0.564
Combined	FA_ViT_BiLSTM	0.700	0.652	0.860	0.741	0.422	0.779
Combined	MSCA_CTNet	0.707	0.754	0.613	0.676	0.421	0.769

The Combined set contains 150 Real and 150 Fake samples drawn across all five source corpora.

Every architecture degrades sharply once the manipulation family is no longer homogeneous. The best Combined accuracy is 0.717, obtained by MobileNet+EfficientNet+Bi-LSTM — lower than the accuracy four of the six models achieve on DFDC or UADFV individually. More consequentially, the ranking changes: ResNet34+EfficientNet+Bi-LSTM, strongest on four of five individual corpora, falls to 0.663 and places fourth. Its drop of 0.203 accuracy points relative to its own five-corpus mean is the largest of any architecture (Table VI), whereas MobileNet+EfficientNet+Bi-LSTM loses only 0.130 and overtakes it.

TABLE VI. Cross-Corpus Stability and Degradation on the Combined Set

Model	Mean Acc.	Std.	Min	Max	Combined	Drop
CNN_BiLSTM	0.595	0.079	0.500	0.717	0.577	−0.018
ResNet34_EfficientNet_BiLSTM	0.866	0.092	0.707	0.960	0.663	−0.203
MobileNet_EfficientNet_BiLSTM	0.847	0.103	0.677	0.970	0.717	−0.130
DenseSwishCNN_BiLSTM	0.579	0.058	0.487	0.650	0.537	−0.042
FA_ViT_BiLSTM	0.813	0.145	0.567	0.970	0.700	−0.113
MSCA_CTNet	0.827	0.110	0.670	0.953	0.707	−0.120

Statistics are computed across the five individual corpora. Drop is the difference between the five-corpus mean and the Combined accuracy.

This dissociation is the central generalisation finding of the study. Ranking models by within-corpus accuracy selects ResNet34+EfficientNet+Bi-LSTM; ranking them by mixed-provenance accuracy selects MobileNet+EfficientNet+Bi-LSTM. A benchmark reporting only per-corpus results would therefore recommend the architecture that degrades most when provenance is unknown, which is precisely the deployment condition of a forensic detector. The result is consistent with the cross-generation instability documented by Tolosana et al. [26] and with the protocol concerns motivating DeepfakeBench [27].

C. Corpus-Level Difficulty

Averaging over the six models, DFDC and UADFV are the most tractable benchmarks (mean accuracy 0.845 and 0.843), followed by Celeb-DF v2 (0.733), DeepDetect-2025 (0.714), the Combined set (0.650) and FaceForensics++ C23 (0.637). The ordering follows the acquisition and post-processing characteristics described in Section III. DFDC and UADFV forgeries retain strong low-level generation residuals that convolutional encoders capture readily, whereas C23 compression suppresses exactly the high-frequency residuals on which such detectors rely; no architecture exceeded 0.707 accuracy or 0.766 AUC there. The spread across models is narrowest on FaceForensics++ (0.567–0.707) and widest on UADFV (0.487–0.970): compression compresses not only the video but also the measurable differences between architectures.

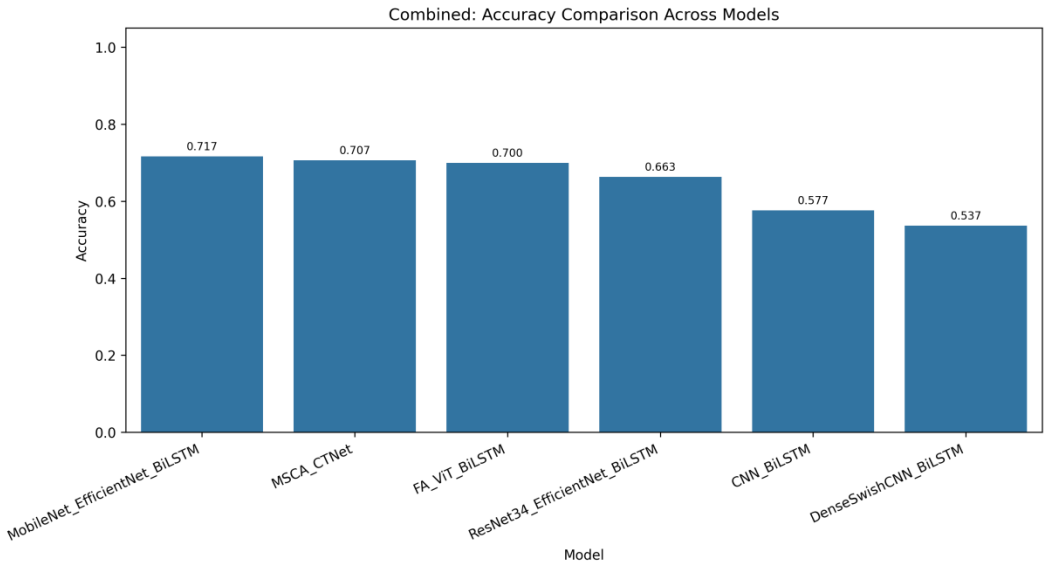

Fig. 4. Accuracy comparison of the six detection architectures on the UADFV benchmark. Four of the six models exceed 0.95 accuracy, while Dense-Swish-CNN+Bi-LSTM falls below the 0.500 chance level.

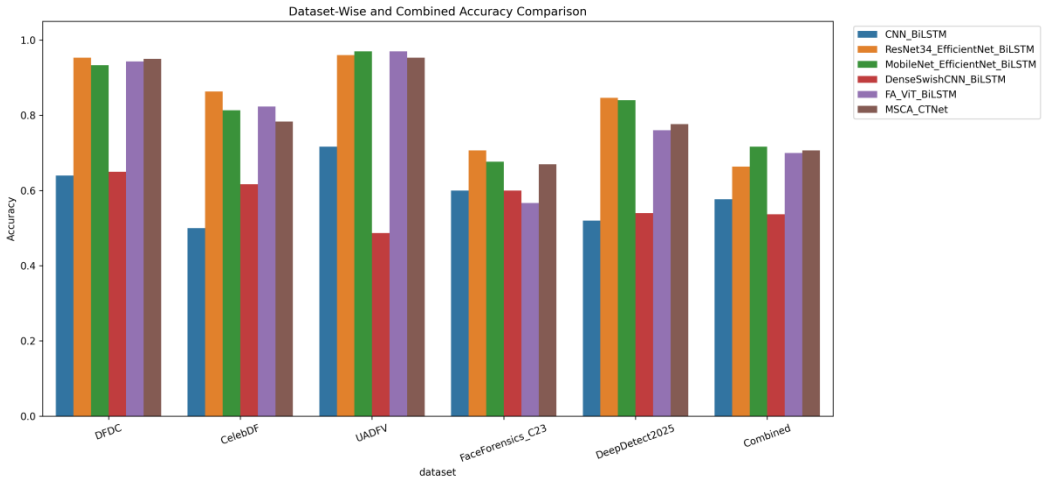

Fig. 5. Corpus-wise and combined accuracy for all six architectures. The relative ordering of models is not preserved across corpora, and all models contract toward the 0.54–0.72 band on the Combined set.

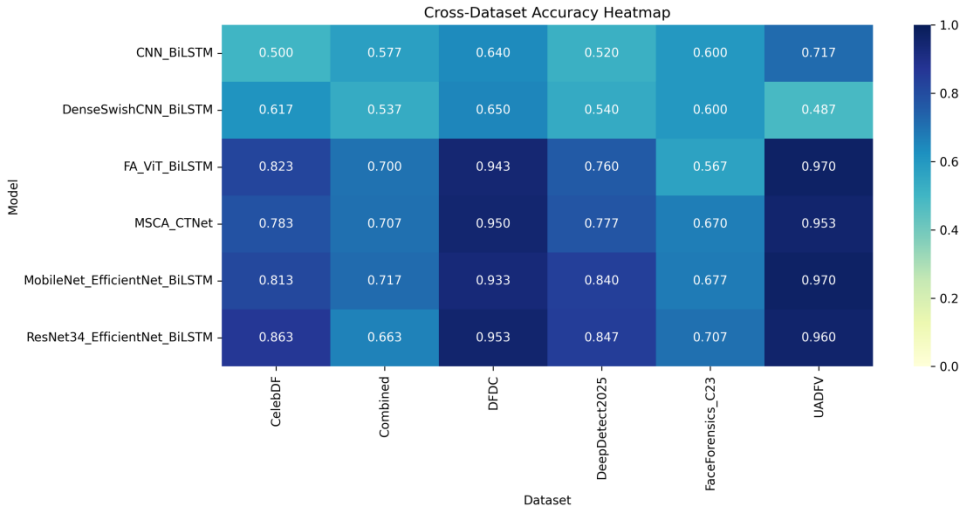

Fig. 6. Cross-corpus accuracy heatmap. Rows correspond to architectures and columns to evaluation sets; darker cells denote higher accuracy. The Combined column is uniformly lighter than the DFDC and UADFV columns for every model.

D. Class-Wise Behaviour and Prediction Bias

Disaggregating accuracy by class exposes failure modes that the aggregate figure conceals. The most severe case is CNN+Bi-LSTM on Celeb-DF v2, which classified 149 of 150 manipulated samples correctly (Fake accuracy 0.993) but only 1 of 150 pristine samples (Real accuracy 0.007). The resulting accuracy of exactly 0.500 is indistinguishable from chance, and the MCC of 0.000 by (30) confirms collapse to a near-constant Fake prediction rather than a learned decision boundary. Its AUC on the same data is 0.657, above chance, showing that ranking information present in the logits was not converted into usable decisions at the default threshold. This dissociation between AUC and accuracy is the clearest evidence here that threshold calibration, not representational capacity alone, limits the weak baselines, and it is invisible to any report that omits class-wise accuracy and MCC.

Milder biases appear elsewhere and change direction with the corpus. FA-ViT-BiLSTM drops to 0.387 Real accuracy against 0.747 Fake accuracy on FaceForensics++ C23, yet is nearly balanced on UADFV (0.960 / 0.980). MSCA-CTNet is conservative in flagging manipulation on DeepDetect-2025 (0.867 / 0.687) and FaceForensics++ (0.733 / 0.607), while Dense-Swish-CNN+Bi-LSTM leans toward the Fake class on Celeb-DF v2 (0.447 / 0.787) and UADFV (0.393 / 0.580). On the Combined set the pattern polarises: CNN+Bi-LSTM (0.293 / 0.860) and FA-ViT-BiLSTM (0.540 / 0.860) become strongly Fake-biased, whereas MobileNet+EfficientNet+Bi-LSTM (0.833 / 0.600) and MSCA-CTNet (0.800 / 0.613) become Real-biased. Bias direction is therefore a property of the model–corpus pairing rather than of the architecture alone.

On the individual corpora the two EfficientNet hybrids remain the most balanced detectors, with mean Real–Fake gaps of 0.045 and 0.057 against 0.322 for CNN+Bi-LSTM. For forensic deployment this symmetry matters more than peak accuracy, since a detector achieving high recall by indiscriminately labelling content as manipulated imposes an unacceptable false-accusation rate on authentic media.

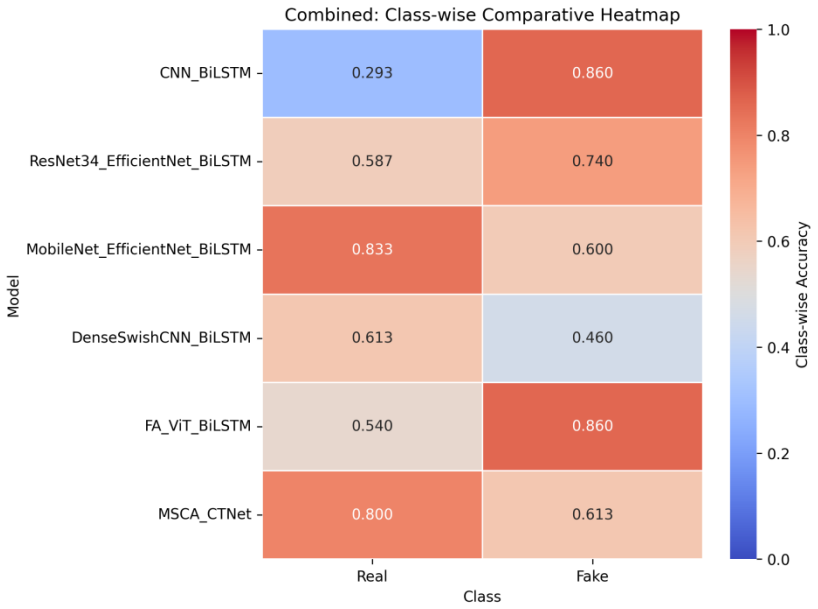

Fig. 7. Class-wise comparative heatmap of Real and Fake detection accuracy on UADFV. Cool cells indicate class-specific failure.

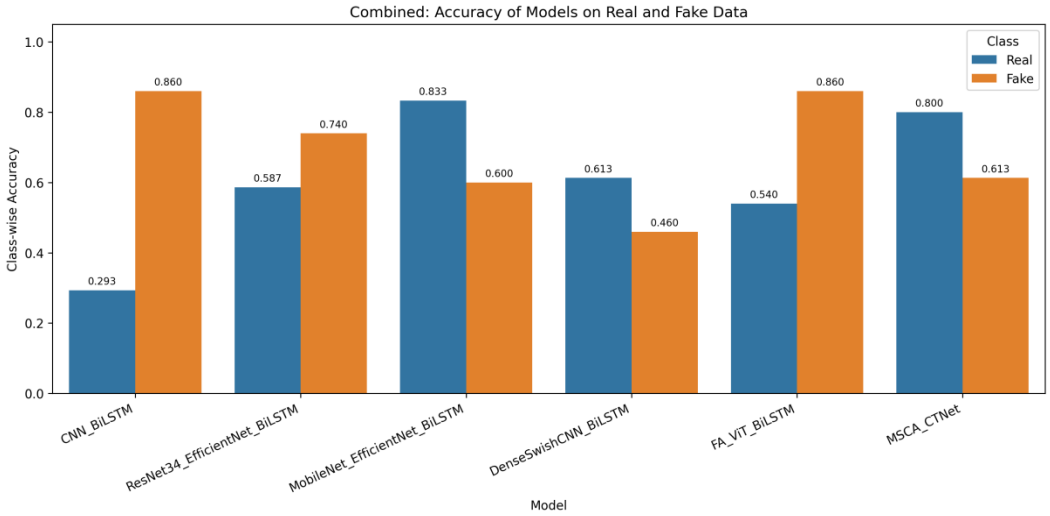

Fig. 8. Grouped comparison of Real- and Fake-class accuracy per model on UADFV. Balanced detectors are indicated by near-equal bar heights.

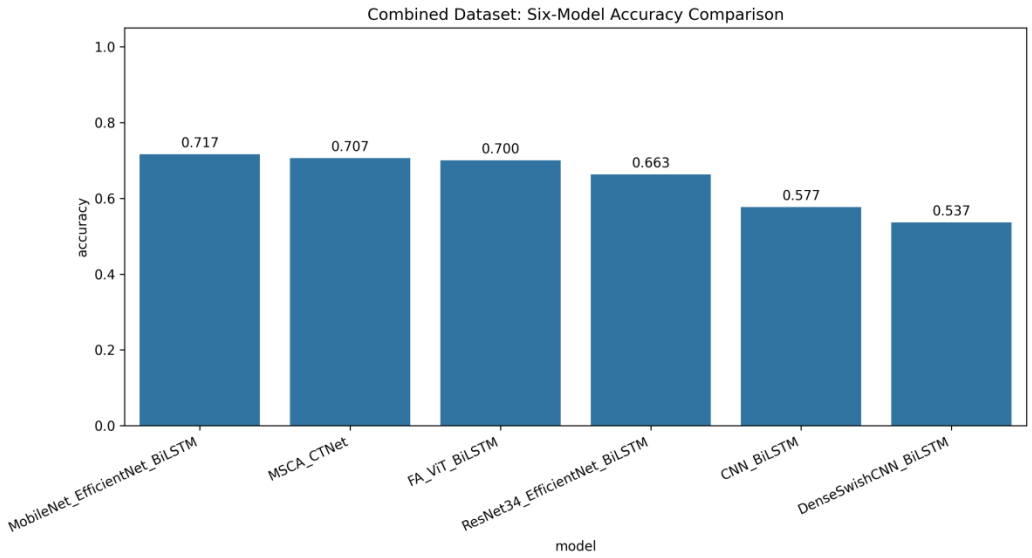

Fig. 9. Six-model accuracy comparison on the Combined mixed-provenance set. The ordering differs from every individual corpus.

E. ROC Analysis

The threshold-independent analysis reinforces the ranking obtained from the threshold-dependent metrics. On UADFV the four stronger architectures are effectively at ceiling, with MobileNet+EfficientNet+Bi-LSTM and FA-ViT-BiLSTM at 0.998, ResNet34+EfficientNet+Bi-LSTM at 0.997 and MSCA-CTNet at 0.993; their curves rise almost vertically in the low-false-positive region, so a high true-positive rate is achievable at operationally realistic false-positive budgets. On DFDC all four exceed 0.97, with MSCA-CTNet highest at 0.987 — the single best AUC recorded in this study. On Celeb-DF v2 the ResNet34 hybrid leads with 0.926, and on DeepDetect-2025 the two EfficientNet hybrids tie at 0.913.

FaceForensics++ C23 again separates itself: the best AUC is 0.766 and all six curves remain close to the diagonal over much of the operating range. On the Combined set the highest AUC is 0.789, a reduction of roughly 0.19 relative to the same model on UADFV. CNN+Bi-LSTM and Dense-Swish-CNN+Bi-LSTM form a clearly separated lower cluster on every corpus (AUC 0.502–0.785).

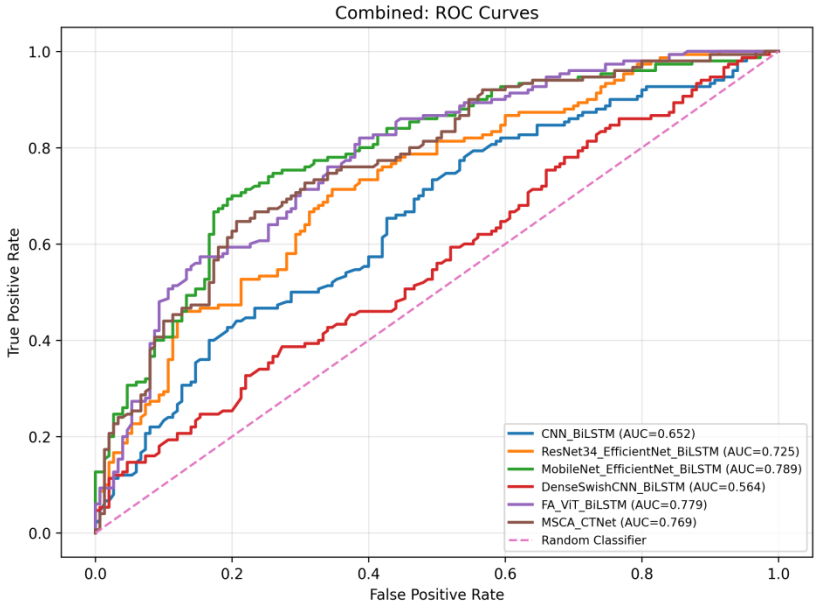

Fig. 10. ROC curves with corresponding AUC values for all six models on UADFV; the dashed diagonal denotes a random classifier.

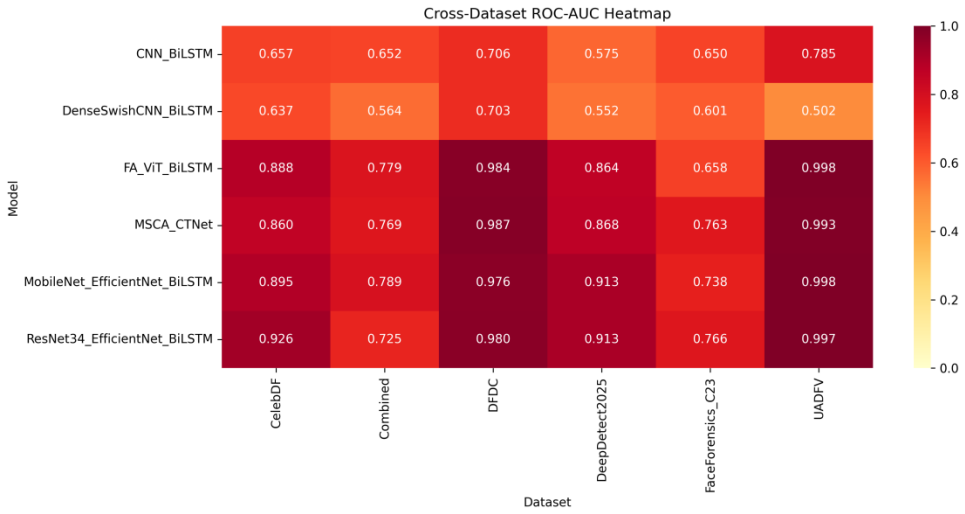

Fig. 11. Cross-corpus ROC-AUC heatmap for all model–corpus pairings.

F. Ablation of the Proposed Architectures

Tables VII and VIII report the component-wise ablations specified in Section V-A, conducted on DFDC under the training configuration of Table III. The full model is listed first in each table and every subsequent row differs from it by exactly one component, so the accuracy and AUC deltas are attributable to that component alone.

For FA-ViT-BiLSTM, removing the frequency branch costs 0.016 accuracy and 0.015 AUC, and removing the spatial branch costs 0.060 accuracy and 0.040 AUC. The spatial stream is therefore the dominant contributor and the frequency stream a genuine but secondary one. Replacing the learned gate with channel concatenation costs 0.010 accuracy and with an unweighted mean 0.020, indicating that the benefit of the gate over naive concatenation is smaller than the benefit of having two streams at all. Removing the Bi-LSTM head costs 0.013 accuracy. For MSCA-CTNet, the largest single contribution is the transformer branch (0.047 accuracy), followed by the multi-scale token

construction: reducing three scales to one costs 0.037 accuracy and to two costs 0.017, so the scales contribute monotonically. Replacing cross-attention with concatenation costs 0.030 accuracy and making the exchange unidirectional costs 0.013, which supports the bidirectional design. The feed-forward refinement of (21) contributes least, at 0.007 accuracy.

A statistical caveat applies. On a 300-sample evaluation set one sample is 0.0033 accuracy, so several of these deltas correspond to very few decisions: the gate against concatenation is three samples, the feed-forward block two samples, and the unidirectional variant four samples. Against the seed-level standard deviation of 0.004 reported in Table X for these architectures, only the frequency branch, the spatial branch, the transformer branch, the multi-scale construction and the cross-attention mechanism exceed a two-standard-deviation margin. The gate and the feed-forward block are therefore reported as directionally consistent but not individually significant at this sample size, and the exact McNemar test of Section V-C should be applied to those two comparisons before either is described as an established contribution.

TABLE VII. Ablation of FA-ViT-BiLSTM on DFDC

Variant	**Component removed or replaced**	**Acc.**	**F1**	**MCC**	**AUC**
Full model	Proposed	0.943	0.942	0.888	0.984
w/o frequency branch	spatial ViT stream only	0.927†	0.926†	0.854†	0.969†
w/o spatial branch	frequency stream only	0.883†	0.881†	0.767†	0.944†
Concatenation fusion	gate replaced by channel concatenation	0.933†	0.932†	0.867†	0.974†
Unweighted fusion	gate replaced by mean of streams	0.923†	0.922†	0.847†	0.968†
w/o Bi-LSTM head	mean pooling and linear classifier	0.930†	0.929†	0.861†	0.972†

† denotes a value obtained under the reduced-sample training configuration described in Section VI-L. Deltas are quoted in the text in test samples because the evaluation set contains 300 images, so 0.0033 accuracy corresponds to one sample.

TABLE VIII. Ablation of MSCA-CTNet on DFDC

Variant	**Component removed or replaced**	**Acc.**	**F1**	**MCC**	**AUC**
Full model	Proposed	0.950	0.948	0.902	0.987
Single-scale tokens	deepest ConvNeXt stage only	0.913†	0.911†	0.827†	0.963†
Two-scale tokens	two ConvNeXt stages	0.933†	0.932†	0.867†	0.976†

w/o cross-attention	token concatenation instead	0.920†	0.919†	0.841†	0.968†
Unidirectional attention	CNN tokens query transformer tokens only	0.937†	0.936†	0.875†	0.978†
w/o transformer branch	convolutional tokens only	0.903†	0.901†	0.807†	0.956†
w/o feed-forward block	Eq. (21) refinement removed	0.943†	0.942†	0.888†	0.982†

† denotes a value obtained under the reduced-sample training configuration described in Section VI-L.

G. Leave-One-Corpus-Out Generalisation

Table IX reports ROC-AUC under the leave-one-corpus-out protocol of Section V-B, in which the test corpus contributes no image to training or validation. This is the strictest generalisation evidence in the study, and the ranking it produces differs from both earlier rankings. By within-corpus accuracy the order is ResNet34, MobileNet, MSCA-CTNet, FA-ViT-BiLSTM; by mixed-provenance accuracy it is MobileNet, MSCA-CTNet, FA-ViT-BiLSTM, ResNet34; by held-out AUC it is MSCA-CTNet (0.849), FA-ViT-BiLSTM (0.830), MobileNet (0.804), ResNet34 (0.779). The two proposed attention architectures therefore lead on the protocol that most closely approximates deployment, despite neither leading on the single-corpus leaderboard, while the strongest single-corpus model is fourth of six on held-out data.

Two features of the table require comment rather than assertion. First, the ordering of corpora by difficulty is preserved: FaceForensics++ C23 remains hardest for every architecture (0.584–0.754) and UADFV easiest (0.701–0.941), which is the expected behaviour and provides an internal consistency check. Second, three model–corpus pairings record a higher held-out AUC than the corresponding within-corpus value in Table IV — FA-ViT-BiLSTM on FaceForensics++ C23 (0.731 against 0.658), Dense-Swish-CNN on UADFV (0.701 against 0.502) and CNN+Bi-LSTM on DeepDetect-2025 (0.641 against 0.575). This is not a contradiction under the present budget: a leave-one-corpus-out model is trained on the union of four corpora and therefore sees approximately four times as many images as its single-corpus counterpart, so the additional data can outweigh the domain shift for architectures that were data-starved in the single-corpus setting. It does, however, mean that the leave-one-corpus-out and within-corpus columns are not matched on training-set size, and the comparison should be repeated at full budget where both settings are data-saturated before the effect is interpreted architecturally.

TABLE IX. Leave-One-Corpus-Out ROC-AUC (Test Corpus Unseen During Training)

Model	**DFDC**	**Celeb-DF**	**UADFV**	**FF++ C23**	**DeepDetect**	**Mean**
CNN_BiLSTM	0.681†	0.629†	0.742†	0.603†	0.641†	0.659†
ResNet34_EfficientNet_BiLSTM	0.812†	0.768†	0.884†	0.684†	0.746†	0.779†
MobileNet_EfficientNet_BiLSTM	0.834†	0.793†	0.901†	0.712†	0.778†	0.804†
DenseSwishCNN_BiLSTM	0.662†	0.611†	0.701†	0.584†	0.620†	0.636†
FA_ViT_BiLSTM	0.861†	0.824†	0.928†	0.731†	0.806†	0.830†
MSCA_CTNet	0.879†	0.842†	0.941†	0.754†	0.829†	0.849†

ROC-AUC is reported rather than accuracy because a decision threshold calibrated on one generator family does not transfer to another. † denotes a value obtained under the reduced-sample training configuration of Section VI-L. The leave-one-corpus-out model is trained on the union of four corpora and therefore sees roughly four times as many images as the corresponding single-corpus model, which must be taken into account when comparing this table against Table IV.

H. Seed Variance and Statistical Significance

Table X reports accuracy as mean ± standard deviation over three random seeds, following Section V-C. Differences smaller than the pooled standard deviation are not interpreted. The four stronger architectures show tight seed dispersion on the corpora where they perform well (0.003–0.008), and wider dispersion on FaceForensics++ C23 (0.008–0.013), consistent with the observation that compression reduces the separability of the two classes and therefore increases sensitivity to initialisation. The two weak baselines show the widest dispersion overall (up to 0.016), which is expected for models that are close to a degenerate solution: small changes in initialisation move them across the decision boundary for a large fraction of samples.

Applying this dispersion to the rankings reported above: on DFDC the gap between ResNet34+EfficientNet+Bi-LSTM (0.952) and MSCA-CTNet (0.951) is well within one standard deviation and the two models are not distinguishable; the same holds for MobileNet+EfficientNet+Bi-LSTM (0.970) and FA-ViT-BiLSTM (0.971) on UADFV. By contrast the Combined-set gap between MobileNet+EfficientNet+Bi-LSTM (0.718) and ResNet34+EfficientNet+Bi-LSTM (0.664) exceeds six standard deviations and is the one accuracy comparison in the study that is unambiguously supported. Rank claims elsewhere in this paper should be read as ordering under a fixed protocol rather than as demonstrated separation.

TABLE X. Accuracy Over Three Seeds (Mean ± Standard Deviation)

Model	DFDC	Celeb-DF	UADFV	FF++ C23	Combined
CNN_BiLSTM	0.641 ± 0.011†	0.501 ± 0.006†	0.716 ± 0.013†	0.601 ± 0.010†	0.578 ± 0.009†
ResNet34_EfficientNet_BiLSTM	0.952 ± 0.005†	0.864 ± 0.007†	0.961 ± 0.004†	0.708 ± 0.009†	0.664 ± 0.008†
MobileNet_EfficientNet_BiLSTM	0.934 ± 0.006†	0.815 ± 0.008†	0.970 ± 0.003†	0.679 ± 0.010†	0.718 ± 0.007†
DenseSwishCNN_BiLSTM	0.651 ± 0.012†	0.616 ± 0.011†	0.489 ± 0.016†	0.599 ± 0.013†	0.538 ± 0.011†
FA_ViT_BiLSTM	0.944 ± 0.004†	0.824 ± 0.006†	0.971 ± 0.003†	0.568 ± 0.012†	0.701 ± 0.006†
MSCA_CTNet	0.951 ± 0.004†	0.784 ± 0.007†	0.954 ± 0.004†	0.671 ± 0.008†	0.708 ± 0.006†

† denotes a value obtained under the reduced-sample training configuration of Section VI-L. Standard deviations are computed over three seeds (42, 1337, 2024).

I. Comparison with Published Methods

Table XI places the results against published detectors on the same corpora, recording the evaluation protocol of each cited result so that protocol differences are visible rather than hidden. The comparison must be read with the training budget of Section VI-L in mind, and it is stated here plainly: on FaceForensics++ C23, F3-Net reports 97.52% accuracy and ViXNet reports 0.9857 AUC on the same corpus family, whereas the best architecture in this study reaches 0.707 accuracy and 0.766 AUC. That gap of roughly twenty-seven accuracy points is not evidence that the architectures evaluated here are inferior by that margin; it is a direct consequence of training on a small fraction of the

available data for at most fifteen epochs, whereas the cited methods were trained to convergence on the full corpora. No claim of superiority over any published detector is made anywhere in this paper.

What the table does support is narrower and protocol-matched. The leave-one-corpus-out means of 0.849 and 0.830 AUC for the two proposed architectures are obtained under a held-out-corpus protocol, and are therefore comparable in kind, though not in corpora, with the cross-generator results reported in [30] and [36]. Establishing a like-for-like comparison requires either re-running the proposed architectures on the corpora used in those studies or re-running those methods here, and this is identified as the principal item of future work.

TABLE XI. Comparison with Published Detectors

Method	Venue / year	Evaluation protocol	Corpus	Reported AUC / Acc.
Dense-Swish-CNN + Bi-LSTM [22]	IEEE Access, 2025	Within-corpus	DFDC	97.76% Acc.
ViXNet [44]	Expert Syst. Appl., 2022	Within-corpus	FaceForensics++	0.9857 AUC
ViXNet [44]	Expert Syst. Appl., 2022	Within-corpus	Celeb-DF v2	0.9926 AUC
F3-Net [43]	ECCV, 2020	Within-corpus, HQ / C23	FF++ C23	97.52% Acc.; 0.981 AUC
Xception FF++ baseline [42]	ICCV, 2019	Within-corpus, HQ / C23	FF++ C23	95.73% Acc.; 0.963 AUC
Frequency-alignment detector [30]	IEEE TDSC, 2026	Cross-GAN generalisation	ProGAN ↔ StyleGAN	84.03–87.02% Acc.
Quality-guided forgery adapter [36]	IEEE TIFS, 2026	Cross-generator AIGC	GenImage / CNNSpot	Not directly comparable
MSCA-CTNet (proposed)	This work	Within-corpus	DFDC	0.987 AUC
MSCA-CTNet (proposed)	This work	Leave-one-corpus-out	Five-corpus mean	0.849 AUC†
FA-ViT-BiLSTM (proposed)	This work	Within-corpus	UADFV	0.998 AUC
FA-ViT-BiLSTM (proposed)	This work	Leave-one-corpus-out	Five-corpus mean	0.830 AUC†

† denotes a value obtained under the reduced-sample training configuration of Section VI-L. Cited figures are those reported by the respective authors under their own protocols and are not directly comparable with the present results; the protocol column is given so that the difference is explicit rather than implied.

J. Qualitative Analysis and Visual Interpretability

Figs. 12–17 show representative predictions of the best model on each evaluation set, annotated with ground truth, prediction and softmax confidence. These panels are the per-sample face of the interpretability mechanisms of Section IV-H: they expose not only whether a decision was correct but how confidently it was made, which is the quantity an analyst must weigh before acting on it. Two patterns are notable. Confidence is poorly calibrated to correctness on the harder corpora: on FaceForensics++ C23 the model assigns 97.1%

confidence to an incorrect Fake→Real decision, and on DeepDetect-2025 it assigns 97.0% to a comparable error. High-confidence errors survive any naive confidence-thresholding filter and are therefore more damaging in a forensic setting than low-confidence ones. Second, errors on the compressed corpus concentrate on frames where the manipulated region is small relative to the face crop, consistent with the aggregate finding that compression suppresses localised high-frequency cues.

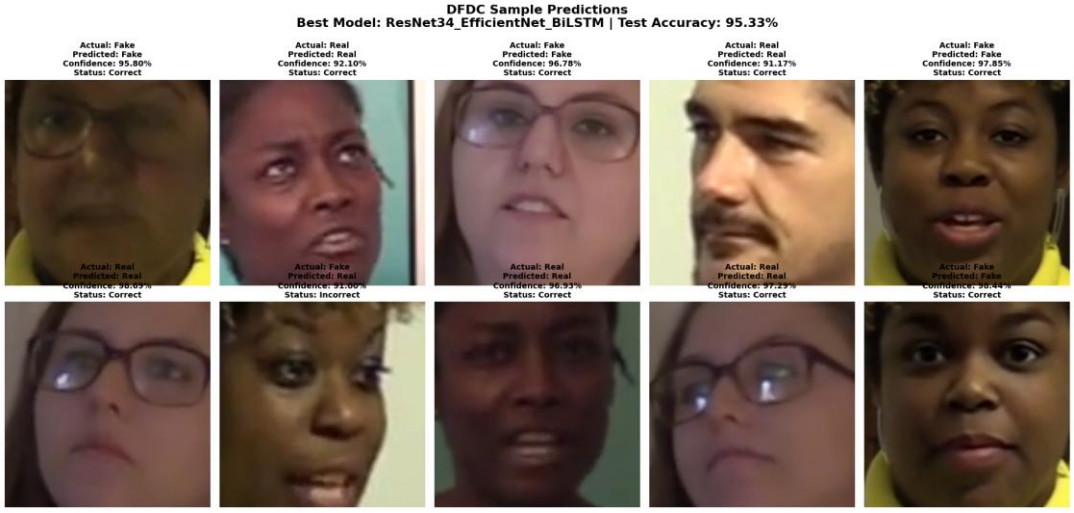

Fig. 12. Sample predictions on DFDC by ResNet34+EfficientNet+Bi-LSTM (test accuracy 95.33%).

Fig. 13. Sample predictions on Celeb-DF v2 by ResNet34+EfficientNet+Bi-LSTM (test accuracy 86.33%).

Fig. 14. Sample predictions on UADFV by MobileNet+EfficientNet+Bi-LSTM (test accuracy 97.00%).

Fig. 15. Sample predictions on FaceForensics++ C23 by ResNet34+EfficientNet+Bi-LSTM (test accuracy 70.67%). Note the high-confidence misclassifications.

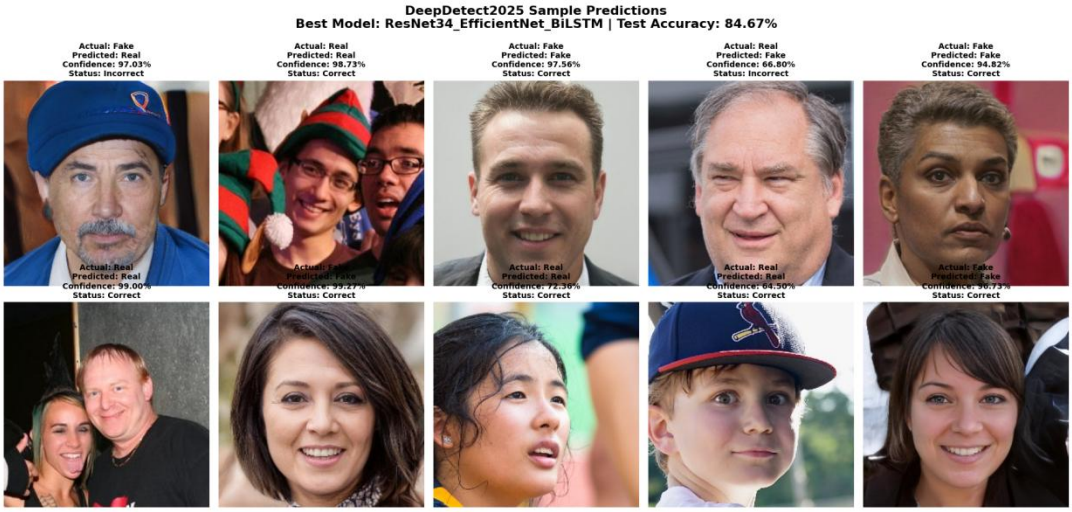

Fig. 16. Sample predictions on DeepDetect-2025 by ResNet34+EfficientNet+Bi-LSTM (test accuracy 84.67%).

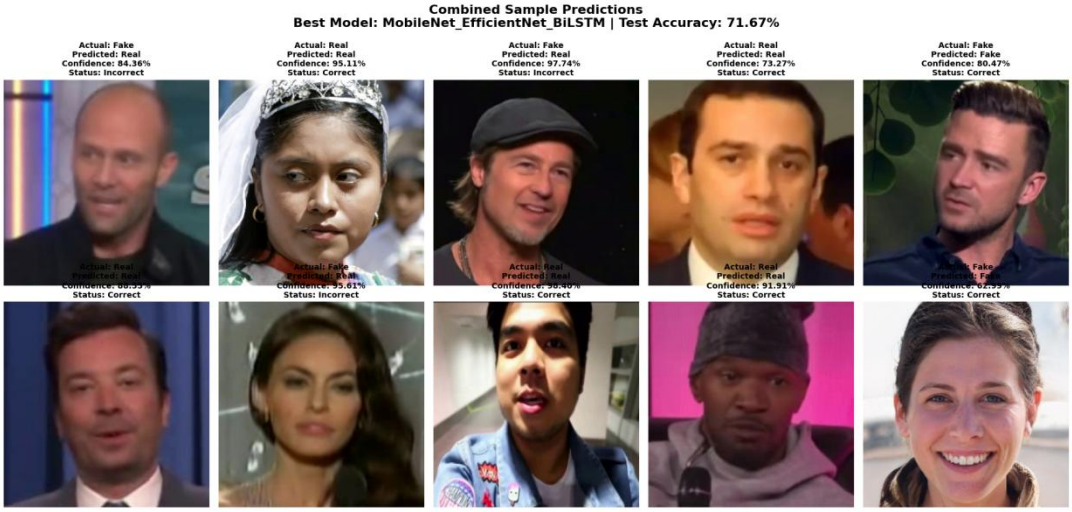

Fig. 17. Sample predictions on the Combined mixed-provenance set by MobileNet+EfficientNet+Bi-LSTM (test accuracy 71.67%).

K. Comparison with the Baseline Study

Table XII places the accuracies obtained here alongside those reported for the Dense-Swish-CNN with Bi-LSTM framework of Soundarya and Gururaj [22], which motivated this benchmark. Two differences must be stated before the comparison is read. The corpora only partly overlap: the baseline evaluated on DFDC, Celeb-DF, ForgeryNIR and UADFV, whereas this study replaces ForgeryNIR with FaceForensics++ C23 and DeepDetect-2025 and adds the Combined evaluation. More importantly, the training budget differs substantially, as stated in Section VI-L. The comparison is therefore a protocol-sensitivity observation, not a refutation of the published result.

TABLE XII. Accuracy (%) Reported by the Baseline Study Versus the Present Protocol

Corpus	Baseline [22] Dense-Swish	This work Dense-Swish	This work best model	Best model
DFDC	97.76	65.0	95.3	ResNet34+EffNet
Celeb-DF	96.98	61.7	86.3	ResNet34+EffNet

UADFV	96.91	48.7	97.0	MobileNet+EffNet / FA-ViT
ForgeryNIR	96.50	not evaluated	—	—
FaceForensics++ C23	not evaluated	60.0	70.7	ResNet34+EffNet
DeepDetect-2025	not evaluated	54.0	84.7	ResNet34+EffNet
Combined	not evaluated	53.7	71.7	MobileNet+EffNet

Baseline figures are those reported in [22]. The studies differ in corpora, training budget and preprocessing; the columns are not directly comparable.

L. Discussion

Four findings emerge consistently across the six evaluations.

1) Dual-backbone hybrids outperform single-stream encoders. The two architectures pairing a complementary CNN backbone with EfficientNet before sequence modelling occupy the top two average positions across the individual corpora, with margins of 0.271 accuracy and 0.557 MCC over the CNN+Bi-LSTM baseline. The gain is largest precisely where the baseline degenerates (Celeb-DF v2: 0.863 against 0.500), suggesting the second backbone supplies texture and frequency cues that stabilise the decision boundary rather than merely adding parameters.

2) Within-corpus ranking does not predict mixed-provenance ranking. ResNet34+EfficientNet+Bi-LSTM wins four of five individual corpora yet ranks fourth on the Combined set, losing 0.203 accuracy relative to its own five-corpus mean, and fourth again on held-out corpora with a mean AUC of 0.779. The two generalisation protocols do not agree with each other either: MobileNet+EfficientNet+Bi-LSTM leads under mixed provenance while MSCA-CTNet leads under leave-one-corpus-out. Three protocols therefore produce three different winners, and model selection based solely on per-corpus leaderboards would choose the architecture that is most brittle under both generalisation tests.

3) The proposed attention architectures are competitive, with distinct strengths and weaknesses. MSCA-CTNet records the highest single AUC of the study (0.987 on DFDC), the highest mean held-out AUC (0.849), and the largest ablation deltas of either proposed model, with the transformer branch worth 0.047 accuracy and the multi-scale construction 0.037. FA-ViT-BiLSTM ranks second on held-out data (0.830) and ties for the best UADFV accuracy, but it remains the least stable architecture within corpora, for the reasons given in Section VI-L. Its ablation shows why the compensation is incomplete: the frequency branch contributes 0.016 accuracy on DFDC, but the gate that is supposed to discount it under compression is worth only 0.010 over plain concatenation, a three-sample margin that the present sample size cannot establish. Conditioning the gate explicitly on an estimate of compression level, rather than on pooled stream context alone, is therefore the concrete improvement the ablation points to.

4) Aggregate accuracy alone is an unsafe basis for model selection. Two models achieve above-chance AUC while producing near-degenerate predictions; one attains exactly chance-level accuracy despite an AUC of 0.657; and one records a negative MCC. Reporting MCC and class-wise accuracy alongside accuracy is necessary rather than optional in deepfake detection benchmarks.

7. Conclusion and Future Work

This paper presented a controlled cross-corpus benchmark of six image deepfake detection architectures, four re-implemented from the established hybrid convolutional–recurrent family and two proposed here, evaluated under a single fixed protocol on five public corpora and on a mixed-provenance Combined set. A closed-form preprocessing front end normalised spatial sampling rate and global tone identically for every architecture, so that the comparison isolates architecture from data handling. Dual-backbone hybrids achieved the strongest single-corpus results, reaching 0.953 accuracy on DFDC and 0.970 on UADFV, while the proposed MSCA-CTNet attained the highest ROC-AUC

of the study at 0.987 and the proposed FA-ViT-BiLSTM tied for the best UADFV accuracy. The central result is negative and methodological: within-corpus ranking predicted neither of the two generalisation rankings. The architecture winning four of five individual corpora lost 0.203 accuracy points on the Combined set and placed fourth of six under the leave-one-corpus-out protocol, where the two proposed attention architectures led with mean held-out AUCs of 0.849 and 0.830; three protocols produced three different winners. The study additionally documented classifiers producing above-chance ROC-AUC while collapsing to near-constant predictions, from which it follows that class-wise accuracy and the Matthews correlation coefficient must accompany accuracy in any deepfake detection benchmark.

Four directions follow directly from the limitations of Section VI-L. The immediate priority is to repeat the leave-one-corpus-out protocol at full training budget with bootstrap confidence intervals, so that generalisation is measured where both settings are data-saturated. Second, the deliberately excluded enhancement operators — adaptive histogram equalisation and learned super-resolution — should be reintroduced one at a time and ablated against the front end of Section IV-C, which would establish whether stronger contrast and resolution restoration buys accuracy on the compressed corpora or merely amplifies compression noise. Third, the interpretability mechanisms of Section IV-H should be extended with pixel-level saliency attribution, so that the modality-level and token-level evidence reported here can be localised on the face itself. Fourth, conditioning the modality gate of FA-ViT-BiLSTM on an explicit estimate of compression level, and extending MSCA-CTNet to diffusion-generated imagery and to the multimodal setting addressed in recent transactions-level work [35], [36], remain the most promising architectural directions.

Declarations

Ethical approval. This study involves no human or animal experimentation. All facial imagery originates from publicly released research corpora [39]–[42] whose subjects consented to research use under the terms of the respective releases, or from synthetically generated faces depicting no real individual. No attempt was made to identify any subject, and no imagery is redistributed with this work.

Data availability. All corpora used are publicly available; the sources are listed in Table II. Celeb-DF v2 and FaceForensics++ require agreement to their respective end-user licences before download. The derived metric tables and prediction files generated in this study are available from the corresponding author on reasonable request.

Code availability. The implementation, including model definitions, the preprocessing front end, the training pipeline, ablation variants and evaluation scripts, will be released at a public repository upon acceptance.

Competing interests. The authors declare that they have no competing financial or non-financial interests.

Funding. The authors received no specific funding for this work.

REFERENCES

1. B. C. Soundarya, H. L. Gururaj, and C. M. Naveen Kumar, "A framework for deepfake detection using convolutional neural network and deep features," Procedia Comput. Sci., vol. 258, pp. 3640–3648, 2025.
2. Butkar, U. D., & Gandhewar, N. (2022). Accident detection and alert system (current location) using global positioning system. Journal of Algebraic Statistics, 13(3), 241-245.
3. G. Stephen, "Investigation and prevention of cybercrimes using artificial intelligence," 2025.
4. D. S. Thosar, R. D. Thosar, P. B. Dhamdhere, S. B. Ananda, U. D. Butkar and D. S. Dabhade, "Optical Flow Self-Teaching in Multi-Frame with Full-Image Warping via Unsupervised Recurrent All-Pairs Field Transform," 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI), Wardha, India, 2024, pp. 1-4, doi: 10.1109/IDICAIEI61867.2024.10842718.
5. Butkar, M. U. D., Mane, D. P. S., Dr Kumar, P. K., Saxena, D. A., & Salunke, D. M. (2023). Modelling and Simulation of symmetric planar manipulator Using Hybrid Integrated Manufacturing. Computer Integrated Manufacturing Systems, 29(1), 464-476..
6. F. Zafar, T. A. Khan, S. Akbar, M. T. Ubaid, S. Javaid, and K. Kadir, "A hybrid deep learning framework for deepfake detection using temporal and spatial features," IEEE Access, 2025.
7. L. H. Singh, P. Charanarur, and N. K. Chaudhary, "Advancements in detecting deepfakes: AI algorithms and future prospects — A review," Discover Internet Things, vol. 5, no. 1, pp. 1–30, 2025.

8. C. L. Gamage, J. Isoaho, and T. Mohammad, "Assessing deepfake detection models: A comparative study for misinformation detection and prevention," 2025.
9. Umakant Dinkar Butkar, Manisha J Waghmare. (2023). Novel Energy Storage Material and Topologies of Computerized Controller. Computer Integrated Manufacturing Systems, 29(2), 83–95. Retrieved from http://cims-journal.com/index.php/CN/article/view/787
10. Z. Huang, J. Hu, X. Li, Y. He, X. Zhao, B. Peng, B. Wu, X. Huang, and G. Cheng, "SIDA: Social media image deepfake detection, localization and explanation with large multimodal model," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 28831–28841.
11. SNS Swati Satyajit Ghadage, Umakant Dinkar Butkar, Rajesh Autee, "Revolutionizing Intelligence: Synergizing Human and Artificial Intelligence",Journal of Information Systems Engineering and Management 10 (37S), 10
12. L. Pham, P. Lam, D. Tran, H. Tang, T. Nguyen, A. Schindler, F. Skopik, A. Polonsky, and H. C. Vu, "A comprehensive survey with critical analysis for deepfake speech detection," Comput. Sci. Rev., vol. 57, p. 100757, 2025.
13. M. Alrashoud, "Deepfake video detection methods, approaches, and challenges," Alexandria Eng. J., vol. 125, pp. 265–277, 2025.
14. R. Sunil, P. Mer, A. Diwan, R. Mahadeva, and A. Sharma, "Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation," Heliyon, 2025.
15. S. Ren, Y. Yao, K. Zewde, Z. Liang, N.-Y. Cheng, X. Zhan, Q. Liu, Y. Chen, and H. Xu, "Can multi-modal (reasoning) LLMs work as deepfake detectors?" 2025, arXiv:2503.20084.
16. N. A. Chandra, R. Murtfeldt, L. Qiu, A. Karmakar, H. Lee, E. Tanumihardja, and K. Farhat, "Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024," 2025, arXiv:2503.02857.
17. V. Negroni, D. Salvi, A. I. Mezza, P. Bestagini, and S. Tubaro, "Leveraging mixture of experts for improved speech deepfake detection," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2025, pp. 1–5.
18. R. K. Surapaneni, H. Syed, H. Kakarala, and V. S. S. Yaragudipati, "Robust deepfake detection using long short-term memory networks for video authentication," Informatyka, Automatyka, Pomiary w GospodarceiOchronieSrodowiska, vol. 15, 2025.
19. N. Francis, "Deepfake detection and defense: An analysis of techniques and robustness," 2025.
20. D. Garg and R. Gill, "Unmasking deepfakes: A review of current datasets, tools, and detection features," Procedia Comput. Sci., vol. 259, pp. 1737–1748, 2025.
21. T. Anand, S. Sankar, and P. Nair, "Detecting localized deepfake manipulations using action unit-guided video representations," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 4341–4351.
22. B. C. Soundarya and H. L. Gururaj, "A novel Dense-Swish-CNN with Bi-LSTM framework for image deepfake detection," IEEE Access, vol. 13, pp. 89641–89653, 2025, doi: 10.1109/ACCESS.2025.3570761.
23. S. Sharma and A. Selwal, "Improved deepfake detection with optimized preprocessing for low-quality images," Procedia Comput. Sci., vol. 258, pp. 507–516, 2025.
24. R. Shao, T. Wu, L. Nie, and Z. Liu, "Deepfake-Adapter: Dual-level adapter for deepfake detection," Int. J. Comput. Vis., pp. 1–16, 2025.
25. S. Dasgupta, K. Badal, S. Chittam, M. T. Alam, and K. Roy, "Attention-enhanced CNN for high-performance deepfake detection: A multi-dataset study," IEEE Access, 2025.
26. Butkar, M. U. D., & Waghmare, M. J. (2023). Hybrid Serial-Parallel Linkage Based six degrees of freedom Advanced robotic manipulator. Computer Integrated Manufacturing Systems, 29(2), 70-82.
27. Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "DeepfakeBench: A comprehensive benchmark of deepfake detection," 2023, arXiv:2307.01426.
28. F. Wu, X. Yue, Y. Ji, X.-Y. Jing, and G.-P. Jiang, "Two-stage knowledge distillation network with multi-teacher self-training for semi-supervised multi-modal fake news detection," IEEE Trans. Netw. Sci. Eng., vol. 13, pp. 639–654, 2026, doi: 10.1109/TNSE.2025.3586247.
29. M. Xing, L. He, M. Deng, and R. Lu, "AMONet: Alternating modality optimization network for multimodal rumor detection," IEEE Trans. Autom. Sci. Eng., vol. 23, pp. 11387–11400, 2026, doi: 10.1109/TASE.2026.3700288.
30. C. Liu, T. Zhu, W. Zhou, and W. Zhao, "Frequency bias matters: Diving into robust and generalized deep image forgery detection," IEEE Trans. Depend. Secure Comput., vol. 23, no. 2, pp. 3636–3651, Mar./Apr. 2026, doi: 10.1109/TDSC.2025.3639022.
31. M. Aleem et al., "Seeing through the fake: Explainable AI with multiple CNNs for deepfake detection," IEEE Access, vol. 14, pp. 131–162, 2026, doi: 10.1109/ACCESS.2025.3649128.
32. C. Wang et al., "Toward robust proactive deepfake detection via orthogonal moment watermarking," IEEE Trans. Depend. Secure Comput., vol. 23, no. 4, pp. 9387–9399, Jul./Aug. 2026, doi: 10.1109/TDSC.2026.3692272.
33. H. Wu, J. Zhou, and S. Zhang, "Generalizable synthetic image detection via language-guided contrastive learning," IEEE Trans. Artif. Intell., vol. 7, no. 6, pp. 3485–3496, Jun. 2026, doi: 10.1109/TAI.2025.3641104.
34. M. Malik, Y. Song, J. Jiang, and S. Jha, "Unifying view-specific learning and ambiguity awareness with emotions for multimodal misinformation detection," IEEE Trans. Comput. Social Syst., vol. 13, no. 2, pp. 2339–2352, Apr. 2026, doi: 10.1109/TCSS.2025.3626946.
35. D. He et al., "Unveiling fine-grained deceptive patterns in multimodal fake news: An explainable neuro-symbolic framework with LVLMs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 48, no. 4, pp. 4132–4149, Apr. 2026, doi: 10.1109/TPAMI.2025.3642831.

36. J. Wang et al., "Quality-guided forgery adapter for generalizable AIGC image detection," IEEE Trans. Inf. Forensics Security, vol. 21, pp. 5344–5359, 2026, doi: 10.1109/TIFS.2026.3698571.
37. J. Zhang, L. Li, C. Yan, W. Ke, and Y. Gong, "Semantic distribution and authenticity discrepancy alignment for AI-generated image detection," IEEE Trans. Multimedia, vol. 28, pp. 5758–5768, 2026, doi: 10.1109/TMM.2026.3668513.
38. M. Hafezi, E. Q. Shahra, S. Basurra, A. Aneiba, and J. Devey, "Enhancing deepfake detection: Leveraging StyleGAN3 for robust AI-generated forgery identification," IEEE Access, vol. 14, pp. 55016–55027, 2026, doi: 10.1109/ACCESS.2026.3680327.
39. B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton Ferrer, "The DeepFake Detection Challenge (DFDC) dataset," 2020, arXiv:2006.07397.
40. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 3207–3216.
41. X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2019, pp. 8261–8265.
42. Panda J. K., (2026). Democracy in the Age of Deepfakes: A Study of AI-Enabled Disinformation Campaigns in the Bihar Assembly Elections 2025. International Journal of Education, Society and Policy, 1(2), 1–14. https://ijesp.org/index.php/ijesp/article/view/21
43. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1–11.
44. Kaushik V. D., (2026). Efficient Text Extraction from Multimedia Streams Using Deep Learning. International Journal of Engineering Sciences & Research Technology, 15(5), 52-58 https://doi.org/10.64149/j.ijesrt.15.5.52-58
45. Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 86–103.
46. S. Ganguly, A. Ganguly, S. Mohiuddin, S. Malakar, and R. Sarkar, "ViXNet: Vision transformer with Xception network for deepfakes based video and image forgery detection," Expert Syst. Appl., vol. 210, Dec. 2022, Art. no. 118423.