Deepfake Video Detection Using Deep Learning: A Comparative Study with a Novel Attention-CNN-LSTM Architecture

Abhishek Sharma
Department of Computer Science and Engineering
Rajasthan Institute of Engineering & Technology (RIET)
Jaipur, Rajasthan, India
abhishek.abhi567@gmail.com

Mr. Amit Kumar Sharma
Assistant Professor
Department of Computer Science and Engineering
Rajasthan Institute of Engineering & Technology (RIET)
Jaipur, Rajasthan, India
amitsharmacse09@gmail.com

***Abstract*—Deepfake videos present a growing and serious threat to the integrity of digital media. This paper makes three contributions toward reliable automated detection. First, it evaluates four established deep learning models—MesoNet, XceptionNet, EfficientNet-B4, and CNN-LSTM—on the FaceForensics++ benchmark under a single unified experimental setup, enabling a direct and fair comparison. Second, it proposes Attention-CNN-LSTM, a novel architecture that inserts a channel-wise spatial attention (CWSA) module between the XceptionNet feature extractor and the bidirectional LSTM, directing temporal reasoning toward the most discriminative facial feature channels. Third, it reports an ablation study, a cross-dataset generalization test, and statistical significance tests that together validate every design choice and reported ranking. The proposed model achieves 96.8% accuracy on FaceForensics++ (c23), outperforming EfficientNet-B4 (96.1%), XceptionNet (95.4%), CNN-LSTM (94.2%), and MesoNet (83.7%). All pairwise differences are statistically significant ($p < 0.05$). Cross-dataset evaluation on Celeb-DF shows that EfficientNet-B4 generalizes best (72.4%) to unseen data, highlighting a fundamental tension between in-domain accuracy and cross-domain robustness.**

***Index Terms*—Deepfake detection, attention mechanism, CNN-LSTM, FaceForensics++, EfficientNet-B4, XceptionNet, MesoNet, temporal modeling, ablation study, cross-dataset generalization**

I. Introduction

Deepfake technology uses generative adversarial networks (GANs) [2] and autoencoder-based pipelines to produce manipulated videos that are convincing enough to fool trained forensic experts in high-quality cases [1]. The harmful uses are wide: political disinformation, non-consensual intimate content, financial fraud through identity impersonation, and fabricated video evidence in legal proceedings. Social media platforms process millions of uploaded videos daily—a volume no human team can inspect—making reliable automated detection essential.

CNNs trained on real and fake video frames learn to detect visual manipulation artifacts such as unnatural face-boundary blending, skin-texture inconsistencies, and frequency-domain anomalies introduced by GAN upsampling. Temporal models extend this by finding inconsistencies in how expressions and textures evolve across frames—a signal invisible when each frame is examined alone. Most published methods, however, evaluate a single model against prior work using different preprocessing pipelines and training conditions, making fair comparison impossible.

The standard CNN-LSTM hybrid also has a structural weakness: it passes all spatial feature channels equally to the LSTM, even though not all channels carry equal manipulation evidence. A face-swap artifact concentrates near the face boundary; a texture-based manipulation spreads across the skin surface. An attention module that weights feature channels by relevance before the LSTM can therefore be expected to improve detection sensitivity.

This paper addresses both gaps. The contributions are: (1) a fair multi-model comparison of four baselines on FaceForensics++ [5] under a single pipeline; (2) a novel **Attention-CNN-LSTM** that adds a CWSA module between XceptionNet and the bidirectional LSTM; (3) an ablation study justifying the chosen frame sequence length and LSTM hidden units; (4) cross-dataset generalization testing on Celeb-DF [10]; and (5) pairwise statistical significance tests confirming all reported rankings.

II. Related Work

Afchar et al. [4] proposed MesoNet, a compact CNN that targets mesoscopic spatial features for fast deployment on limited hardware. Rossler et al. [5] released FaceForensics++ and showed that XceptionNet [6], which uses depthwise separable convolutions throughout, gives the best single-frame detection performance. Tan and Le [20] introduced EfficientNet, which scales CNN depth, width, and resolution together; EfficientNet-B4 achieves high deepfake detection accuracy while remaining computationally tractable.

On the spatial detection side, Li and Lyu [8] exploited face-boundary warping artifacts left by face-swap pipelines, and Yang et al. [9] detected reenactment deepfakes via 3D head-pose inconsistencies. Frank et al. [11] showed that GAN upsampling leaves periodic patterns in the Fourier spectrum that are invisible to humans but detectable by classifiers. Dang et al. [16] showed that gradient-based attention maps can

localize manipulated facial regions, improving both accuracy and interpretability.

For temporal modeling, Guera and Delp [3] were the first to pass CNN frame features to an LSTM to detect temporal inconsistencies. Masi et al. [14] extended this with a two-branch network that combines appearance features and optical flow. Coccomini et al. [17] combined EfficientNet with Vision Transformers, capturing local spatial detail alongside global temporal context.

For attention-guided detection, Zhao et al. [12] applied multiple attention modules to different facial regions simultaneously, showing that localized attention outperforms treating the entire face uniformly. Cozzolino et al. [15] used identity-aware attention to flag inconsistencies with a subject's known appearance. These two works directly motivate the attention module proposed here.

Li et al. [10] introduced Celeb-DF, revealing that models trained on FaceForensics++ suffer large accuracy drops on higher-quality celebrity deepfakes. Dolhansky et al. [13] organized the DeepFake Detection Challenge with over 100,000 clips. Bansal et al. [18] demonstrated real-time detection with efficient architectures, and Akhtar [19] provides a recent comprehensive survey.

No prior work combines, under a single experimental protocol: (a) a fair multi-model baseline comparison, (b) an attention-augmented CNN-LSTM, (c) a principled ablation study, (d) cross-dataset generalization testing, and (e) statistical significance validation. This paper fills all five gaps simultaneously.

III. Methodology

A. Dataset and Splits

FaceForensics++ [5] provides 1,000 original YouTube videos and manipulated counterparts using four methods: *Deepfakes* (autoencoder face swap), *Face2Face* (expression transfer), *FaceSwap* (3D graphics-based face replacement), and *NeuralTextures* (GAN-based texture modification preserving face geometry). Videos are available at three compression levels—raw, c23, and c40—and c23 is used as the primary condition. Cross-dataset testing uses Celeb-DF (590 real, 5,639 fake celebrity videos). The dataset is split **70% train / 15% validation / 15% test at the video level** to prevent data leakage.

B. Preprocessing

FFmpeg extracts one frame per second. MTCNN [22] detects and crops each face, which is resized to 256×256 and normalized with ImageNet statistics. Failed detections (extreme pose, occlusion) are discarded. The 4:1 fake-to-real class imbalance is addressed with class weights during training. Augmentation uses random horizontal flip, ± 10 rotation, and random brightness/contrast adjustment.

C. Baseline Architectures

MesoNet [4] (MesoInception-4 variant) is trained from scratch with Adam [23] at $\eta = 10^{-3}$, 30 epochs, early stopping.

XceptionNet [6] uses depthwise separable convolutions throughout. We fine-tune a pretrained model at $\eta = 10^{-4}$, 20 epochs, sigmoid output.

EfficientNet-B4 [20] applies compound scaling across depth, width, and resolution: $d = \alpha^{\phi}$, $w = \beta^{\phi}$, $r = \gamma^{\phi}$ ($\phi = 4$, input 380×380). Fine-tuned at $\eta = 10^{-4}$, 20 epochs.

CNN-LSTM [3] applies time-distributed XceptionNet to 10 consecutive frames (shared weights), feeding the feature sequence to a bidirectional LSTM (256 units). Trained end-to-end at $\eta = 10^{-4}$, 20 epochs.

D. Proposed Attention-CNN-LSTM

The plain CNN-LSTM gives equal weight to all C feature channels before the LSTM. We insert a **channel-wise spatial attention (CWSA)** module after XceptionNet and before the LSTM. For feature map $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$:

$$\mathbf{z} = \frac{1}{HW} \sum_{i,j} \mathbf{F}_{ij} \quad \text{(Global Average Pool)} \tag{1}$$

$$\mathbf{a} = \sigma\left(W_2\, \mathrm{ReLU}(W_1 \mathbf{z} + b_1) + b_2\right) \tag{2}$$

$$\tilde{\mathbf{F}} = \mathbf{F} \otimes \mathbf{a} \tag{3}$$

where $W_1 \in \mathbb{R}^{C/r \times C}$, $W_2 \in \mathbb{R}^{C \times C/r}$, reduction ratio $r = 16$, and $\otimes$ denotes channel-wise multiplication. This formulation is inspired by Squeeze-and-Excitation networks [24] but is embedded inside a temporal sequence model. The attended features $\tilde{\mathbf{F}}$ for all frames in the sequence pass to the bidirectional LSTM. The full pipeline is shown below (Eq. 4):

$$\underset{\text{frame}}{\mathbf{x}_t} \xrightarrow{\text{Xcep}} \mathbf{F}_t \xrightarrow{\text{CWSA}} \tilde{\mathbf{F}}_t \xrightarrow{\text{BiLSTM}} \mathbf{h} \xrightarrow{\text{FC}} \hat{y} \tag{4}$$

The module adds only ≈ 0.3M parameters ($< 2\%$ of total). Training uses Adam at 10^{-4}, 20 epochs, binary cross-entropy, early stopping (patience 5).

E. Evaluation

Accuracy, precision, recall, and F1-score are computed from the confusion matrix. Statistical significance between model pairs uses two-sided paired t-tests on per-video accuracy scores ($\alpha = 0.05$).

IV. Experimental Results

A. Implementation

Ubuntu 20.04, Python 3.8, Intel Core i7 3.2 GHz, 16 GB RAM, NVIDIA RTX 3060 (12 GB VRAM), TensorFlow/Keras, MTCNN, FFmpeg.

B. Overall Performance on FaceForensics++ (c23)

Table I shows results on the c23 test set (4,275 samples). Attention-CNN-LSTM achieves 96.8%—the highest accuracy and F1—outperforming all four baselines. Cross-validation scores (MesoNet 79.5%, XceptionNet 93.1%, EfficientNet-B4 94.2%, CNN-LSTM 91.8%, Attention-CNN-LSTM 95.4%) are consistently close to test accuracy, confirming no overfitting.

TABLE I
OVERALL PERFORMANCE ON FACEFORENSICS++ (C23)

Model	Acc. (%)	Prec.	Recall	F1
MesoNet	83.7	0.84	0.82	0.83
XceptionNet	95.4	0.95	0.96	0.95
EfficientNet-B4	96.1	0.96	0.96	0.96
CNN-LSTM	94.2	0.94	0.95	0.94
Attn-CNN-LSTM (Ours)	**96.8**	**0.97**	**0.97**	**0.97**

TABLE II
TRAINING ERROR, TEST ERROR, AND PREDICTION COUNTS (C23)

Model	Tr.Err (%)	Te.Err (%)	Correct	Wrong
MesoNet	14.2	16.3	3,579	696
XceptionNet	3.1	4.6	4,077	198
EfficientNet-B4	2.8	3.9	4,107	168
CNN-LSTM	4.5	5.8	4,024	251
Attn-CNN-LSTM	**2.5**	**3.2**	**4,137**	**138**

TABLE III
DETECTION ACCURACY (%) BY MANIPULATION TYPE

Type	Meso	Xcep	Eff-B4	C-LSTM	Attn (Ours)
Deepfakes	89.2	93.0	96.1	95.3	**97.1**
Face2Face	88.5	92.6	96.1	95.1	**96.9**
FaceSwap	87.1	91.3	96.2	95.4	**97.0**
NeuralTextures	86.0	90.7	96.1	95.1	**96.3**
Average	87.7	91.9	96.1	95.2	**96.8**

Table II further breaks down training error, testing error, and raw prediction counts. The train–test error gap for all five models stays within 2.5 pp, confirming that none of the models is overfitting to the training data. The proposed model produces the fewest incorrect predictions (138 out of 4,275 test samples), which is a 42% reduction in errors compared to the plain CNN-LSTM (251 incorrect) and a 16% reduction compared to EfficientNet-B4 (168 incorrect).

C. Performance Across Manipulation Types

Table III compares all five models across the four manipulation types. Attention-CNN-LSTM leads on every type. NeuralTextures is consistently the hardest type for all models because it modifies only facial texture and preserves face geometry, leaving no boundary artifacts. The attention module gives its largest gains on Deepfakes and FaceSwap (+1.8 pp over CNN-LSTM), where manipulation artifacts have clear spatial locality near the face boundary, and its smallest gain on NeuralTextures (+1.2 pp), where artifacts are spread uniformly across the skin surface—exactly as expected from the CWSA design.

TABLE IV
ABLATION STUDY: FRAME LENGTH T AND LSTM UNITS H

Var.	Val.	Acc. (%)	F1	Note
T	5	94.3	0.94	Short context
	10	**96.8**	**0.97**	**Optimal**
	15	96.2	0.96	Motion noise
H	128	94.9	0.95	Underpowered
	256	**96.8**	**0.97**	**Optimal**
	512	96.5	0.96	Overfits
Attn	Off (CNN-LSTM)	94.2	0.94	Baseline
	On (Ours)	**96.8**	**0.97**	+2.6 pp / +0.3M

TABLE V
DETECTION ACCURACY (%) AT DIFFERENT COMPRESSION LEVELS

Level	Meso	Xcep	Eff-B4	C-LSTM	Attn (Ours)
Raw	89.0	97.0	97.2	96.0	**97.6**
c23	83.7	95.4	96.1	94.2	**96.8**
c40	61.2	71.1	85.3	82.4	**83.6**

D. Ablation Study

Table IV reports the effect of frame sequence length T and LSTM hidden units H on Attention-CNN-LSTM (c23). All other settings are held fixed.

T=5 provides insufficient temporal context; deepfake inconsistencies often emerge only across longer windows. T=15 slightly underperforms T=10 because additional frames introduce irrelevant motion blur and natural head turns that dilute the temporal signal. H=128 is too small to hold the state needed across 10 attention-weighted frames; H=512 overfits despite dropout. The attention module adds only ≈0.3M parameters but delivers a consistent +2.6 pp improvement, confirming the gain is architectural, not parametric.

E. Compression Robustness

Table V shows accuracy across all three compression levels. MesoNet drops most severely (89.0 → 61.2%, −27.8 pp) because its mesoscopic feature detectors depend on fine pixel detail that compression destroys. EfficientNet-B4 (85.3%) and Attention-CNN-LSTM (83.6%) are the most compression-robust options and best suited for real-world use on social-media-compressed video.

F. Cross-Dataset Generalization on Celeb-DF

Table VI tests all five models on Celeb-DF without any fine-tuning. All models drop significantly, confirming that current detectors learn dataset-specific artifacts rather than general deepfake cues. EfficientNet-B4 generalizes best (72.4%), narrowly ahead of Attention-CNN-LSTM (70.1%). Despite losing on FaceForensics++, EfficientNet-B4's compound-scaled spatial features capture more transferable deepfake cues than the temporal and attention features learned by the proposed model

TABLE VI
CROSS-DATASET GENERALIZATION ON CELEB-DF (NO FINE-TUNING)

Model	FF++ Acc. (%)	Celeb-DF Acc. (%)
MesoNet	83.7	53.2
XceptionNet	95.4	65.3
EfficientNet-B4	96.1	**72.4**
CNN-LSTM	94.2	67.8
Attn-CNN-LSTM	**96.8**	70.1

TABLE VII
PAIRWISE PAIRED T-TEST RESULTS

Model A	Model B	t-stat	p-value
Attn-CNN-LSTM	CNN-LSTM	4.82	<0.001
Attn-CNN-LSTM	EfficientNet-B4	2.31	0.022
Attn-CNN-LSTM	XceptionNet	5.14	<0.001
EfficientNet-B4	XceptionNet	3.07	0.003
EfficientNet-B4	CNN-LSTM	3.51	<0.001
XceptionNet	MesoNet	9.23	<0.001

on FF++ alone. This finding has a direct practical implication: when the target domain is unknown, EfficientNet-B4 is the safer deployment choice.

G. Statistical Significance

Table VII reports pairwise two-sided paired t-tests on per-video accuracy scores. All differences are significant at α=0.05. The tightest gap—Attention-CNN-LSTM vs. EfficientNet-B4—yields t=2.31, p=0.022, confirming that even this 0.7 pp difference is genuine and not random variation.

H. Comparison with Prior Work

Table VIII places the proposed model in the context of the most closely related prior work on FaceForensics++. Where papers reported results under c23 compression, those figures are used; otherwise the nearest comparable setting is noted. The proposed Attention-CNN-LSTM achieves accuracy competitive with the best published single-dataset results while also introducing the first systematic ablation and statistical validation within this line of work.

The proposed model matches the accuracy of Zhao et al. [12]'s multi-attentional approach (97.6%) within 0.8 pp while using a temporal architecture, whereas Zhao et al. use a purely spatial multi-attention design. The two methods are therefore complementary and could potentially be combined in future work.

V. DISCUSSION

Why attention helps. The CWSA module directs the LSTM toward channels encoding face-boundary artifacts for swap-based deepfakes and skin-texture channels for texture-based ones. This explains the pattern in Table III: the largest gains

TABLE VIII
COMPARISON WITH RELATED WORK ON FACEFORENSICS++ (C23)

Reference	Method	Acc. (%)	Dataset
Afchar et al. [4]	MesoNet	83.0	FF++
Rossler et al. [5]	XceptionNet	95.7	FF++
Guera & Delp [3]	CNN-LSTM	97.1*	UADFV
Li & Lyu [8]	Warp artifact	92.3	FF++
Zhao et al. [12]	Multi-attention	97.6	FF++
Coccomini et al. [17]	EffNet+ViT	96.5	DFDC
Ours	Attn-CNN-LSTM	**96.8**	FF++

*Evaluated on a different dataset (UADFV).

appear for spatially localized artifacts (face swap) and the smallest for distributed artifacts (NeuralTextures).

Compression is the key deployment barrier. Videos on WhatsApp, Instagram, and YouTube undergo compression that erases the fine-grained pixel artifacts detectors rely on. Training on mixed-quality data or using compression simulation as augmentation are the most promising near-term mitigations.

In-domain vs. cross-domain accuracy diverge. The best in-domain model (Attention-CNN-LSTM, 96.8%) is not the best cross-domain model (EfficientNet-B4, 72.4%). This tells us that temporal and attention features learned on FaceForensics++ are partly dataset-specific, while compound-scaled spatial features transfer more broadly. For practical deployment: use Attention-CNN-LSTM when the source domain is known and matches training; use EfficientNet-B4 when the source domain is unknown or different. MesoNet is best suited as a fast first-stage filter before a heavier second-stage check.

Limitations. The CWSA module uses channel-wise attention; spatial attention, which localizes artifacts within a feature map, could provide finer discrimination. The cross-dataset evaluation used no fine-tuning; even a small amount of target-domain adaptation would likely recover much of the accuracy drop. The study is also limited to four manipulation types in FaceForensics++ and may not reflect newer GAN-free synthesis methods.

VI. CONCLUSION

This paper presented a fair multi-model comparison of five deep learning architectures for deepfake video detection on FaceForensics++, including the proposed Attention-CNN-LSTM. The proposed model achieved the highest in-domain accuracy (96.8%) by directing the LSTM toward the most discriminative facial feature channels before temporal reasoning. An ablation study confirmed the optimal frame sequence length (T =10) and LSTM hidden units (H=256). Statistical tests validated all reported rankings. Cross-dataset evaluation on Celeb-DF revealed that in-domain accuracy and cross-domain generalization do not always align, with EfficientNet-B4 generalizing best despite ranking second in-domain.

Key directions for future work include spatial (not just channel-wise) attention for finer artifact localiza-

tion; compression-robust training through mixed-quality and simulated-compression augmentation; domain adaptation to improve cross-dataset generalization; and interpretable detection that identifies which facial regions triggered the decision, supporting forensic use cases.

REFERENCES

[1] P. Korshunov and S. Marcel, "Deepfakes: A new threat to face recognition? Assessment and detection," *arXiv:1812.08685*, 2018.
[2] I. Goodfellow *et al.*, "Generative adversarial networks," *Proc. NeurIPS*, vol. 27, 2014.
[3] D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," *Proc. IEEE AVSS*, 2018.
[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," *Proc. IEEE WIFS*, pp. 1–7, 2018.
[5] A. Rossler *et al.*, "FaceForensics++: Learning to detect manipulated facial images," *Proc. IEEE/CVF ICCV*, pp. 1–11, 2019.
[6] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proc. IEEE CVPR*, pp. 1251–1258, 2017.
[7] Y. Li, M. Chang, and S. Lyu, "In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking," *Proc. IEEE WIFS*, 2018.
[8] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *Proc. IEEE CVPR Workshops*, 2019.
[9] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," *Proc. IEEE ICASSP*, pp. 8261–8265, 2019.
[10] Y. Li *et al.*, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," *Proc. IEEE/CVF CVPR*, pp. 3207–3216, 2020.
[11] J. Frank *et al.*, "Leveraging frequency analysis for deep fake image recognition," *Proc. ICML*, pp. 3247–3258, 2020.
[12] H. Zhao *et al.*, "Multi-attentional deepfake detection," *Proc. IEEE/CVF CVPR*, pp. 2185–2194, 2021.
[13] B. Dolhansky *et al.*, "The deepfake detection challenge (DFDC) dataset," *arXiv:2006.07397*, 2020.
[14] I. Masi *et al.*, "Two-branch recurrent network for isolating deepfakes in the wild," *Proc. ECCV*, pp. 667–684, 2020.
[15] D. Cozzolino *et al.*, "ID-Reveal: Identity-aware deepfake video detection," *Proc. IEEE/CVF ICCV*, pp. 15108–15117, 2021.
[16] H. Dang *et al.*, "On the detection of digital face manipulation," *Proc. IEEE/CVF CVPR*, pp. 5781–5790, 2020.
[17] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and vision transformers for video deepfake detection," *Proc. ICIAP*, pp. 219–229, 2022.
[18] N. Bansal *et al.*, "Real-time advanced computational intelligence for deep fake video detection," *Applied Sciences*, vol. 13, no. 5, p. 3095, 2023.
[19] Z. Akhtar, "Deepfakes generation and detection: A short survey," *Journal of Imaging*, vol. 9, no. 1, p. 18, 2023.
[20] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *Proc. ICML*, pp. 6105–6114, 2019.
[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[22] K. Zhang *et al.*, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, 2016.
[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. ICLR*, 2015.
[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proc. IEEE/CVF CVPR*, pp. 7132–7141, 2018.