



IndianDeepGuard: Enhancing Deepfake Detection through Cultural and Multi-Modal Context

Kiah Monpara¹, Jeeshia Savla¹, Mishree Sanura¹,
Manan Shah¹, Prachi Tawde¹, and Neha Agarwal¹

¹ Dept of Information Technology, D.J. Sanghvi College of Engineering,
Mumbai, India

kiahmonpara@gmail.com,
jsavla789@gmail.com

Abstract. Deepfake technology has advanced rapidly, threatening information integrity across diverse demographics. This paper proposes IndianDeepGuard, a novel multi-modal framework designed to address generalization gaps when Western-based deepfake detectors are applied in the Indian context. The framework compares GenConViT, EfficientNetB4+Bi-LSTM, ResNeXt-101+Bi-LSTM for spatial-temporal feature learning, and incorporates cultural modifications such as Contrast Limited Adaptive Histogram Equalization for illumination normalization, a Cultural Context Extractor for analyzing cultural attire, and Multilingual Acoustic Anchoring (Wav2Vec2-XLS-R) for phoneme-viseme learning. Domain-Adversarial Neural Networks enable device-invariant feature learning. ResNeXt-101 achieves peak accuracy of 96.10% with a Macro-F1 score of 94.22%, outperforming the baseline (58.00%) by a substantial margin. EfficientNet-B4 achieves 95.67% accuracy, while GenConViT attains 93.00%. Ablation analysis confirm that Indian-specific modules consistently improve performance across all models.

Keywords: Deepfake detection, multi-modal learning, domain adaptation, cultural context analysis, CLAHE, DANN, Indian multimedia.

1 Introduction

The rapid growth of generative artificial intelligence has made possible the generation of highly realistic deepfake content by utilizing deep learning algorithms such as Generative Adversarial Networks (GANs) and autoencoder models. With their continued development, automated detection systems and humans increasingly struggle to reliably distinguish authentic from manipulated content [1]. This technology has been implicated in an array of issues including political misinformation, financial fraud, and violations of privacy [2].

According to Murphy et al. [3], even a relatively low-cost political deepfake can result in an approximately 20% reduction in individuals' likelihood to vote. While the

performance of CNN-based systems has reached a near maximum level under exhibits severe degradation, a phenomenon known as the generalization gap [1].

The Generalization Gap in India. Architectures like GenConViT [5] and GenConViT+ [6], which integrate convolutional frame extraction and vision transformers, have shown advancements in improving robustness; however, there has been limited research undertaken regarding cultural, demographic, and linguistic diversity. The majority of the current video systems have primarily been trained on Western-origin video data. This results in a performance decrease of over 50% when applied to unconstrained real-life video data from India [7]. The situation is further worsened by India's 1.4 billion people, hundreds of regional languages, wide variety of skin tones and different socio-cultural indicators.

2 Related Work

Deep fake detection is a swiftly growing research area, perhaps in reaction to generative Media that now look like an actual video are generated by models. [8]. First efforts were primarily directed towards pipelines based on CNN. Located objects in the face region. Obviously there was a big breakthrough with Face X-Ray: [9] demonstrated blending in boundaries at the manipulation site was carried. There are universal forgery cues that are applicable regardless of the type of synthesis. The release of The community got a structured benchmark from FaceForensics++ [4]. to train and evaluate models, and the newly announced Deepfake Detection Challenge dataset. This was extended in further work with more expert and greater diversity by [10]. However, they have achieved suboptimal results on real-world videos without any constraints when evaluated against the videos in the wilddeepfake dataset [7]. The government's claim that there is a problem has been substantiated by government testing conducted on Celeb-DF, which showed that The performance is seriously degraded when a model is moved outside its training distribution [11].

It having been recognized that this brittleness has motivated designs that aren't restricted to local CNNs. features. Transitional network to generate video frames. GenConViT [5] used a convolutional network coupled with a Transitional network to create video frames. We propose latent space regularization for boosting robustness and Vision Transformer backbone for enhanced strength of representation. across datasets. Extending this work was done by Patil et al. [6] with their GenConViT+ model, Incorporating 3D-CNN temporal modelling to address sequential aspect of video evidence. On the efficiency front, EfficientNet worked along with a GRU encoder [12] offers a practical compromise to detect anomalies at the frame level while avoiding real-time loss of performance. throughput. For high-definition content, where the partitioning of a video's space is significant, ResNeXt with Sequential models that use LSTM models, such as ResNeXt [13] have given good performance. That said, these more advanced models don't fare well even with testing data that varies by ethnicity, this could be different than what they observed during training, in terms of language, lighting, or recording equipment [7].

While visual deepfakes receive the lion’s share of the limelight, audio deepfakes are not as widely discussed. [15]. Traditional methods were based on MFCC features [16] for speech recognition, However, with the development of neural TTS and voice cloning, these approaches are being increasingly replaced by more recent ones. Linking inconsistencies in the audio-visual modality across time is bigger, and known as multimodal detection. A Dense Swin Transformer was investigated by Ilyas et al. as a promising direction on the FakeAVCeleb corpus [18]. However, there has been relatively no effort to examine how these detection models perform in the context of culturally diverse populations, especially in the case of Indian languages and people. The present work aims

Table 1. Literature Review of Deepfake Detection Methods

Author	Method	Dataset	Acc.	Key Limitation
Deressa (2025) [5]	Hybrid CNN+ViT	FF++, Celeb-DF	~99%	Real-world drop
Harshit (2025) [12]	EffNet+GRU	FF++	96.8%	Compression issues
Kamil (2025) [13]	ResNeXt+LSTM	FF++, DFDC	97.4%	High compute cost
Bhandarkawthekar (2025) [14]	ResNet+LSTM	DFDC	95.2%	Limited generalization
Patil (2024) [6]	GenConViT+3D-CNN	FF++, Celeb-DF	AUC > 0.97	Memory-intensive
Sun (2024) [17]	Diffusion Aug.	Multi-source	Reported	Not standalone
Ilyas (2023) [18]	Dense Swin Trans.	FakeAVCeleb	94.8%	Cultural adaptation
Zi (2021) [7]	Cross-dataset Eval	WildDeepfake	<50% drop	Poor cross-dataset
Rössler (2019) [4]	CNN Artifacts	FF++	>99%	Unseen fragility

3 Methodology

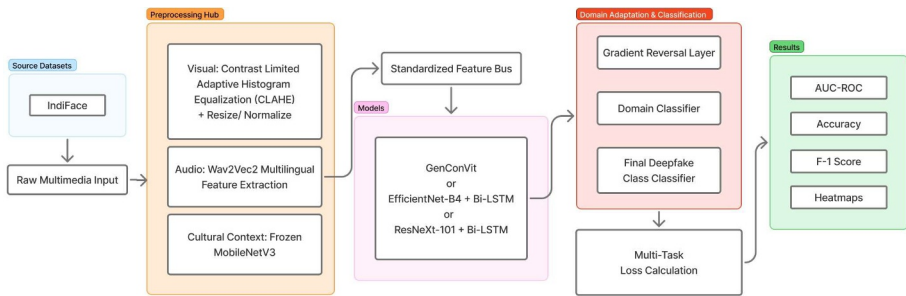


Fig. 1. Overall architecture of the IndianDeepGuard framework

3.1 Dataset

We used the IndiFace dataset [19], a pre-existing dataset (access requested from the original authors) for training purposes. It has 2,303 videos (1,836 fake, 467 real) representing diverse Indian demographics, lighting conditions, and recording environments. It features manipulation techniques like face-swapping and face reenactment. For EfficientNet-B4 and ResNeXt-101, a 80/10/10 stratified split was used: 1,853 training (1,470 fake, 383 real), 219 validation (182 fake, 37 real), and 231 test (184 fake, 47 real) videos. For GenConViT, 5-fold cross-validation was employed with 450 videos/fold (366 fake, 84 real). The rationale behind these splitting strategies is discussed further in Section 3.3.

3.2 Framework Architecture

We evaluated three distinctive high-capacity architectures to identify the best approach to detect deepfakes:

1. **GenConViT.** It combines ConvNeXt blocks for capturing the local texture of a still image with Swin Transformers to provide global attention across a temporal sequence of images. The internal autoencoders act as an additional signal for detecting forgery by penalizing the model if it cannot correctly reconstruct the features of a matching real face from the input. Our pipeline is optimized on this constrained feature space to jointly perform forgery detection and domain-adversarial tasks.
2. **EfficientNet-B4 + Bi-LSTM.** EfficientNet-B4 uses compound scaling to extract rich, frame-level representative features from multiple frames in a video. Adding a bidirectional long short-term memory (Bi-LSTM) layer to EfficientNet-B4 creates helps identify temporal abnormalities that arise, in facial motion (flicker), inter-frame distortions, and muscle movements.
3. **ResNeXt-101 + Bi-LSTM.** ResNeXt-101 uses high-cardinality convolutional filters (with 32 parallel paths per layer) to help pinpoint the location of forgery characteristics. We use this alongside the temporal frame data collected from the Bi-LSTM to identify small changes in both the spatial and temporal dimensions of the frames being analyzed.

To address Indian-specific challenges, four adaptation modules were implemented:

1. **CLAHE.** Applies grid-based local contrast-enhancement to highlight facial alterations hidden under extreme variation of luminance. CLAHE reduces noise by clipping instead of peaking, but retains manipulation cues by clipping each local tile, as opposed to the entire image, thus being more appropriate to Indian luminance conditions than histogram equalization.
2. **CCE.** Western-trained models often misinterpret these items as visual artifacts. Western deepfake models focus strictly on facial manipulation, ignoring non-facial context and frequently mistake Indian cultural markers - such as sarees, turbans, or bindis - for deepfake artifacts. Similarly, culturally-specific gestures like a lateral head shake could be interpreted as unnatural face movement. When a face has been manipulated, surrounding cultural markers have not been affected. Therefore, there

will be an inconsistency between the manipulated face and the culture surrounding it. CCE uses MobileNetV3 [20] to extract features from non-face regions.

3. **Multilingual Acoustic Anchoring (Wav2Vec2-XLS-R)**. Utilizes a multilingual self-supervised speech model [21]. The XLS-R variant [22] enables detection of unsynchronized phoneme-viseme in dubbed or audio-manipulated videos across Indian languages. The raw audio is processed through Wav2Vec2XLS-R to extract acoustical features, resulting in a sequence of contextualised embeddings $H_a = \{h_1, h_2, \dots, h_T\} \in R^{T \times 1024}$. A temporal mean-pooling operation aggregates these into a single audio representation:

$$\mathbf{a} = \frac{1}{T} \sum_{t=1}^T h_t, \quad \mathbf{a} \in \mathbb{R}^{1024} \quad (1)$$

This embedding is integrated with the Bi-LSTM visual-temporal feature $\mathbf{v} \in R^{1024}$ to form the joint representation passed to the classifier:

$$\mathbf{f} = \mathbf{v} \oplus \mathbf{a}, \mathbf{f} \in R^{2048} \quad (2)$$

where $\oplus$ denotes integration. In fake videos, phoneme-viseme desynchronization is characterized by an inconsistency between audio embedding $\mathbf{a}$ and the corresponding visual-temporal embedding $\mathbf{v}$. The level of inconsistency is represented using cosine similarity:

$$S(\mathbf{a}, \mathbf{v}) = \frac{\mathbf{a} \cdot \mathbf{v}}{\|\mathbf{a}\| \|\mathbf{v}\|} \quad (3)$$

with lower cosines representing weaker audio-visual consistency. In the proposed framework, these inconsistencies are learned implicitly through supervised feature fusion rather than explicit frame-by-frame synchronization or contrastive alignment [18].

4. **DANN**: Uses a Domain-Adversarial Neural Network with a Gradient Reversal Layer to ensure the model detects actual forgeries without relying on the specific camera or device used to record the video [23]. This prevents over-relying on device-specific noise patterns, promoting generalization across cameras and smartphones used throughout India. The five folds of the IndiFace dataset are treated as independent adversarial domains. Since each fold contains sessions recorded with varying sensors (CMOS noise profiles) and light conditions, this forces the model to suppress individual device fingerprints while identifying generalized forgery across domains. In the absence of explicit device-specific metadata, these folds serve as a practical proxy for device-based distribution shifts.

Indian DeepGuard is structured as a five-stage modular pipeline as follows:

1. **Source Input**. Normalization and standardization of raw multimedia inputs from the IndiFace dataset.
2. **Preprocessing Hub**. Parallel execution of visual enhancement via CLAHE, multilingual audio feature extraction via Wav2Vec2-XLS-R, and cultural context analysis via MobileNetV3 (CCE).

3. **Experimental Model Bench.** Three independent detection pipelines corresponding to the architectures described above, each trained under a progressive unfreezing schedule.
4. **Domain Adaptation.** A Gradient Reversal Layer is inserted to form a Domain-Adversarial Neural Network, enforcing camera- and device-agnostic feature representations. The domain discriminator is a three-layer MLP branching from the shared multi-modal embedding, with $\alpha = 0.1$ selected via validation performance. To stabilize adversarial training, the GRL adaptation parameter follows a sigmoid-based ramp-up schedule:

$$\lambda_p = \frac{2}{1 + \exp(-10 \cdot p)} - 1 \quad (4)$$

Table 2. Ablation Results

Configuration	GenConViT		EfficientNet-B4		ResNeXt-101	
	Acc.	M-F1	Acc.	M-F1	Acc.	M-F1
Baseline (No Features)	87.78%	72.17%	56.00%	53.00%	58.00%	56.00%
Wav2Vec2 (Audio) only	88.00%	73.20%	69.00%	59.00%	60.00%	58.00%
CLAHE only	89.11%	76.27%	82.00%	65.00%	61.00%	58.00%
CCE only	90.00%	85.00%	86.00%	68.00%	63.46%	61.00%
DANN only	91.00%	81.00%	91.32%	86.94%	69.00%	63.00%
All features	93.00%	88.00%	95.67%	93.43%	96.10%	94.22%

where p is the normalized training progress. We capped this at 0.5 to ensure the adversarial signal does not destabilize the visual backbones during the early stages of fine-tuning.

5. **Evaluation Framework.** Quantitative evaluation using AUC-ROC, accuracy, and Macro-F1 metrics. Given the 80/20 class imbalance, we used MacroF1 as the primary evaluation metric over raw accuracy so that the majority class (fake videos) does not skew the perceived overall model performance.

All models were optimized using the Adam optimizer with a starting learning rate of 1×10^{-4} . The complete objective is designed to enforce classification accuracy and domain invariance simultaneously:

$$L_{total} = L_{class} + \alpha L_{domain} \quad (5)$$

where $\alpha = 0.1$ is the domain loss weight.

3.3 Training

The backbone was frozen for the first three epochs to stabilize training, and then followed by end-to-end fine-tuning. EfficientNet-B4 and ResNeXt-101 used a stratified split of 80/10/10, whereas GenConViT was evaluated using fivefold cross-validation because to its sensitivity to facial distribution shifts. To address the 80/20 class

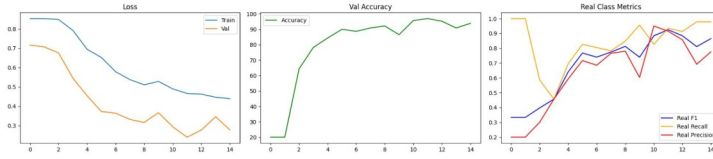
imbalance, minority-targeted augmentation and a 0.1 label smoothing were applied, with the macro-F1 score of real class used for checkpoint selection and early stopping.

4 Results and Discussion

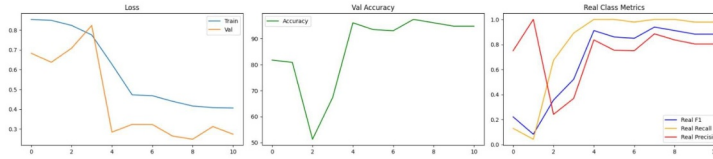
To assess the role of each component of the model, ablation tests were run. Based on those tests, for the Indian Adaptation Module, each base model was only compared against its own baseline, as opposed to all of the other base models. As shown in Table 2, the complete feature set yielded statistically significant results in all cases that were substantially better than the baseline models were.

For the Indian Adaptation Module, ResNeXt-101 had the most accurate classification with 96.10% and had the highest mean F1 score with 94.22%. This is also a considerable improvement over the baseline of ResNeXt-101 with an accuracy of only 58.00%. EfficientNet-B4 had an accuracy of 95.67% and GenConViT had an accuracy of 93.00%. The consistent overall improvements in every base model demonstrate that the proposed Indian Adaptation Modules are effective in bridging the domain gap created by the cultural differences in the content of the images and videos included in each base model.

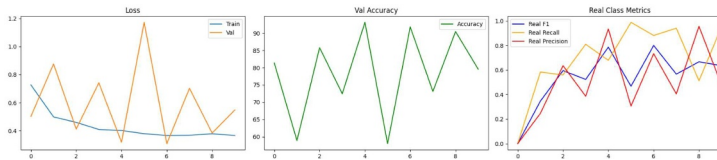
The per-module ablation in Table 2 shows that the DANN and CCE modules contributed the most to the improvement of ResNeXt-101 and may be more valuable than the other modules due to the importance of device-invariant feature learning and culturally grounded non-facial features.



(a) EfficientNet-B4 Training Dynamics



(b) ResNeXt-101 Training Dynamics



(c) GenConViT Training Dynamics

Fig. 2. Comparative training performance across 15 epochs.

Looking at the training behaviour(Figure 2) for EfficientNet-B4 and ResNeXt101 showed stable convergence within 15 epochs with EfficientNet-B4 converging in 6 epochs, however, GenConViT showed larger fluctuations in validation accuracy and loss than the previous two models and indicates that the generative regularization for this base model was not fully optimised and will require further, more stringent parameter tuning in order to achieve stable convergence.

Table 3 shows a comparison of computational complexity of different backbone networks. EfficientNet-B4 had the best balance of detection accuracy and computational efficiency for use in near real-time applications; therefore, it is the best candidate for near real-time applications. ResNeXt-101 had increased overhead from the use of grouped convolutions and aggregated residual transformations, but provided a more robust discriminative feature representation than EfficientNet-B4 and was able to provide more accurate results for fine-grained forgery detection. The architecture of GenConViT entails a hybrid convolutional

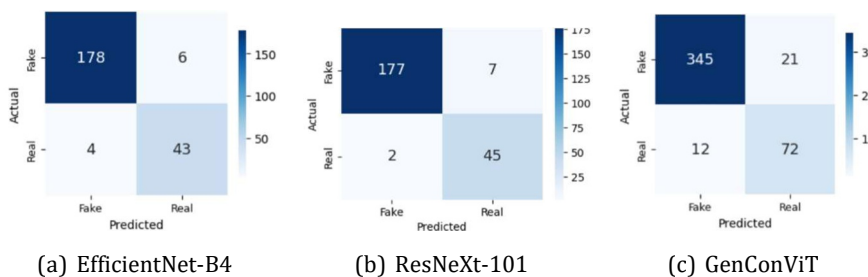


Fig. 3. Confusion matrices for the final IndianDeepGuard configuration.

and transformer-based model [5] that poses a much higher computational cost of execution due to its use of generative regularisation methods [6].

While no actual measurement of the latency per frame was conducted, the computational characteristics of the tested backbone architectures were consistent with previously reported results. Future work should look at profiling end-to-end throughput on deployment hardware to evaluate real-time performance.

Table 3. Computational Complexity Analysis

Architecture	Params	FLOPs	Complexity	Deployment
EfficientNet-B4	~19M	~4.2B	Moderate	High
ResNeXt-101	~88M	~32B	High	Moderate
GenConViT	~130M+	~45B+	Very High	Low

The confusion matrices (Figure 3) provide a fine-grained view of each model’s reliability. ResNeXt-101 achieved the highest sensitivity on the “Real” class, misclassifying only two genuine samples as fake. EfficientNet-B4 also performed well, with four genuine samples misclassified. GenConViT incurred more substantial errors, misclassifying 21 fake samples as real and 12 real samples as fake. Collectively, these results confirm that ResNeXt-101 with the proposed Indian adaptation modules is the

most reliable configuration for protecting against deepfake content in the Indian context.

5 Conclusion and Future Work

This paper presents IndianDeepGuard, which is a framework that detects deepfakes using multiple context clues (e.g., visual frames, audio, cultural context, etc.) and thereby provides broader generalization across different types of Indian multimedia. Within the framework, three neural network models were implemented separately as feature extractors (GenConViT, EfficientNet-B4, and ResNeXt-101) using bi-directional long short-term memory (Bi-LSTM) layers, along with four modules specifically adapted for the scenario of deepfake detection in India (i.e., CLAHE, CCE, Wav2Vec2-XLS-R, and DANN). The experimental results of the study demonstrated through both the use of ablation studies, multiple lighting conditions, multiple skin tones, various languages, and multiple recording devices that the four proposed modules consistently provided improved deepfake detection performance relative to control situations.

Whereas ResNeXt-101 demonstrated superior performance across all evaluation metrics (i.e., accuracy, precision, recall, F1-score), demonstrating that culturally adaptive feature representations can be effective for deepfake detections in India. Future goals include the exploration of the integration of Vision-Language Models, physiological signals (i.e., rPPG), and eye-tracking signals into this framework and the creation of a multilingual Indian deepfake database to increase robustness and generalization.

Ethical Statement.

The IndiFace dataset is comprised of synthetic video clips created from real faces. Usage of IndianDeepGuard is subject to the Digital Personal Data Protection Act (2023) of India and requires informed consent of the individuals whose likenesses are used. The Cultural Context Extractor (CCE) should not be used for surveillance, religious profiling, or any other purpose besides deepfake detection. As with all deepfake research, this knowledge is dual-use and should be tightly controlled. While IndianDeepGuard targets underrepresented Indian demographics, residual dataset biases may remain; wider demographic verification is recommended before high-stakes deployment in domains such as journalism or law.

Declarations.

Funding. The authors did not receive support from any organization for the submitted work.

Data Availability. The IndiFace dataset analyzed during the current study is available from the original authors (Kuckreja et al., 2024) on request.

Author Contributions. All authors contributed to the study conception and design.

Material preparation, data collection, and analysis were performed by Kiah Monpara,

Jeesha Savla, Mishree Sanura, and Manan Shah. All authors read and approved the final manuscript.

Acknowledgments. The authors would like to thank the management and administration of D.J. Sanghvi College of Engineering for providing the necessary facilities and support to carry out this research work.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Heidari, A., Jafari Navimipour, N., Dag, H., Unal, M.: Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs Data Mining Knowl. Discov.* 14(2) (2023). doi:10.1002/widm.1520
2. Labuz, M.: A teleological interpretation of the definition of deepfakes in the EU AI Act. *Policy & Internet* 17(1) (2025). doi:10.1002/poi3.435
3. Murphy, G., et al.: An average Joe, a laptop, and a dream: Assessing the potency of homemade political deepfakes. *Appl. Cogn. Psychol.* 39(2) (2025). doi:10.1002/acp.70061
4. Roßler, A., et al.: FaceForensics++: Learning to Detect Manipulated Facial Images. In: *ICCV*, pp. 1–11 (2019). doi:10.1109/ICCV.2019.00009
5. Deressa, D.W., et al.: GenConViT: Deepfake Video Detection Using Generative Convolutional Vision Transformer. *Appl. Sci.* 15, 6622 (2025). doi:10.3390/app15126622
6. Patil, M., et al.: GenConViT+: Advanced Hybrid Framework for Deepfake Detection for Safeguarding Digital Media Integrity. *J. Integr. Sci. Technol.* 12 (2024). doi:10.62110/sciencein.jist.2024.v12.820
7. Zi, B., et al.: WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In: *ACM MM*, pp. 2382–2390 (2020). doi:10.1145/3394171.3413769
8. Tolosana, R., et al.: DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. *Inf. Fusion* 64, 131–148 (2020). doi:10.1016/j.inffus.2020.06.014
9. Li, L., et al.: Face X-ray for More General Face Forgery Detection. In: *CVPR*, pp. 5001–5010 (2020). doi:10.1109/CVPR42600.2020.00505
10. Dolhansky, B., et al.: The Deepfake Detection Challenge (DFDC) Dataset. *arXiv:2006.07397* (2020)
11. Li, Y., et al.: Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In: *CVPR*, pp. 3207–3216 (2020). doi:10.1109/CVPR42600.2020.00327
12. Harshit, G., Chakradhar Reddy, M., Sarath, Y.V.S.: Deepfake Detection Using EfficientNet-B0 and GRU. *IJERT* 14(4) (2025)
13. Kamil, F., Maharani, E.M., Rahmania, R.: ResNeXt-50 and LSTM for Deepfake Detection. In: *ICoAILO, Bali*, pp. 113–118 (2025). doi:10.1109/ICoAILO66760.2025.11156046
14. Bhandarkawthekar, V., Navamani, T.M., Sharma, R., Shyamala, K.: Design and development of an efficient RNet prediction model for deepfake video detection. *Front. Big Data* 8, 1569147 (2025). doi:10.3389/fdata.2025.1569147
15. Almutairi, Z., Elgibreen, H.: A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. *Algorithms* 15(5), 155 (2022). doi:10.3390/a15050155
16. Hamza, A., Javed, A.R., Iqbal, F., Kryvinska, N., Almadhor, A.S., Jalil, Z., Borghol,

- R.: Deepfake Audio Detection via MFCC Features Using Machine Learning. *IEEE Access* 10, 134018–134028 (2022). doi:10.1109/ACCESS.2022.3231480
17. Sun, K., Chen, S., Yao, T., Liu, H., Sun, X., Ding, S., Ji, R.: DiffusionFake: Enhancing generalization in deepfake detection via guided stable diffusion. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 101474–101497 (2024). <https://arxiv.org/abs/2410.04372>
 18. Ilyas, H., Javed, A., Malik, K.M.: AVFakeNet: A Unified End-to-End Dense Swin Transformer for Audio–Visual Deepfake Detection. *Appl. Soft Comput.* 136, 110124 (2023). doi:10.1016/j.asoc.2023.110124
 19. Kuckreja, K., Hoque, X., Poddar, N., Reddy, S., Dhall, A., Das, A.: INDIFACE: Illuminating India's Deepfake Landscape with a Comprehensive Synthetic Dataset. In: *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–9 (2024). doi:10.1109/FG59268.2024.10582046
 20. Howard, A., et al.: Searching for MobileNetV3. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324 (2019). doi:10.1109/ICCV.2019.00140
 21. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*, pp. 12449–12460 (2020). doi:10.48550/arXiv.2006.11477
 22. Babu, A., et al.: XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale. *arXiv:2111.09296* (2021)
 23. Ganin, Y., et al.: Domain-Adversarial Training of Neural Networks. *JMLR* 17(59), 1–35 *arXiv:1505.07818* (2016)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

