International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Benchmarking Green AI Methods for Audio Deepfake Detection:A Comparative Study of Efficiency and Accuracy

Author Details:

[1]Bushra Fatima, [2]Rohitashwa Pandey [1]

[1,2]Computer Science and Engineering,

[1,2]Bansal Institute of Engineering & Technology, Lucknow, India

[1]fatimabushraon@gmail.com, [2]rohitashwapandey1982@gmail.com

Corresponding Author Email: fatimabushraon@gmail.com

https://doi.org/10.55041/ijsmt.v2i5.165

Cite this Article: Fatima, B. & Pandey, R. (2026). Benchmarking Green AI Methods for Audio Deepfake Detection:A Comparative Study of Efficiency and Accuracy. International Journal of Science, Strategic Management and Technology, *02*(05). https://doi.org/10.55041/ijsmt.v2i5.165

License: This article is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting use, distribution, and reproduction in any medium, provided the original author(s) and source are properly credited.

Abstract—

Audio deepfake detection has emerged as a critical challenge in AI security, driven by the rapid proliferation of advanced voice synthesis and voice conversion technologies. State-of-the-art detectors deliver impressive accuracy but impose substantial computational and environmental costs. Green AI offers a compelling alternative by leveraging frozen, pre-trained self-supervised learning (SSL) models as feature extractors paired with lightweight classical machine learning classifiers — enabling CPU-only training and inference. This paper presents a systematic benchmarking study of existing Green AI approaches for audio deepfake detection, evaluating multiple SSL front-ends (wav2vec 2.0, WavLM, HuBERT) in conjunction with multiple classical ML back-ends (SVM-RBF, Logistic Regression, MLP) across two benchmark datasets — ASVspoof 2019 LA and ASVspoof 2021 DF. Beyond accuracy (measured by Equal Error Rate), we introduce a multi-dimensional efficiency analysis encompassing trainable parameter count, training time, inference time, estimated energy consumption, and approximate CO2 emissions. Our results demonstrate that SSL(wav2vec 2.0, Layer 9) + SVM-RBF achieves the best Green AI accuracy with an EER of 0.90% on ASVspoof 2019 LA using fewer than 1,000 trainable parameters, training in under 3 minutes on a standard CPU.

Keywords: *Audio deepfake detection, Green AI, SSL embeddings, wav2vec 2.0, WavLM, HuBERT, SVM, benchmarking, carbon footprint, ASVspoof.*

I. INTRODUCTION

The security landscape has been fundamentally altered by the emergence of highly realistic audio deepfakes — synthetic speech produced by text-to-speech (TTS) systems and voice conversion (VC) algorithms. These technologies, once confined to research laboratories, are now accessible through commercial APIs and open-source tools, dramatically lowering the barrier for malicious use. Identity theft, CEO fraud, political disinformation, and automated phishing represent only a subset of the documented threats enabled by audio deepfake technology.

Contemporary state-of-the-art detectors — including AASIST, wav2vec 2.0-AASIST, and WavLM-based

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

systems — achieve Equal Error Rates (EER) below 1% on the ASVspoof 2019 LA benchmark. However, these systems demand significant computational resources: GPU clusters for training, extensive memory for parameter storage, and substantial energy consumption that translates directly to CO2 emissions.

Green AI advocates for the development of AI systems that report and minimize computational costs alongside accuracy. In audio deepfake detection, this is operationalized through a powerful approach: using pre-trained, frozen SSL models as feature extractors, and training only a lightweight classical ML classifier on the resulting embeddings. This eliminates fine-tuning, reduces trainable parameters from hundreds of millions to below one thousand, and enables full CPU-based operation.

1.1 Contributions

1. First systematic benchmarking of Green AI approaches: 3 SSL front-ends x 3 classifiers = 9 system combinations.
2. Evaluation on ASVspoof 2019 LA and 2021 DF, providing in-domain and cross-domain performance analysis.
3. Multi-dimensional Green AI efficiency metric: parameters, training time, inference latency, energy (kWh), and CO2 emissions.
4. Layer selection sensitivity analysis identifying optimal transformer layers per SSL model.
5. Quantitative performance-vs-sustainability comparison against Red AI baselines.

II. LITERATURE REVIEW

2.1 Audio Deepfake Detection

Early detection systems relied on hand-crafted acoustic features — LFCC, MFCC, CQCC — combined with GMM classifiers. Deep learning improved performance dramatically: LCNN reduced EER from 8.09% to ~5%, RawNet2 pioneered end-to-end raw waveform processing, and AASIST achieved 0.83% EER using spectro-temporal graph attention networks. Wav2Vec 2.0-AASIST fine-tuning achieved 0.82% EER at the cost of 315M+ parameters and extensive GPU training.

2.2 Self-Supervised Learning for Speech

SSL models — wav2vec 2.0, WavLM, and HuBERT — demonstrate remarkable transfer learning across speech tasks. Their embeddings encode phonetic content and acoustic characteristics that are sensitive to TTS/VC artifacts, making them ideal deepfake detection front-ends. The ASVspoof 5 challenge established SSL embeddings as the dominant front-end in top systems.

2.3 Green AI

Schwartz et al. formalized Green AI, arguing for explicit reporting of computational costs alongside accuracy. Strubell et al. quantified the energy cost of large NLP models. Saha et al. proposed the first Green AI framework for deepfake detection, demonstrating frozen SSL + classical ML can match GPU-intensive systems. The present work extends this with a comprehensive multi-system benchmark and explicit efficiency quantification.

III. METHODOLOGY

3.1 System Architecture

All systems follow a unified two-stage architecture. Stage 1 (Feature Extraction): A frozen pre-trained SSL model processes raw audio and outputs frame-level embeddings from a specified intermediate transformer layer — no gradient updates are applied. Stage 2 (Classification): Frame-level embeddings are mean-pooled to a fixed utterance-level vector, then a classical ML classifier is trained on these representations. Figure 1 illustrates the full pipeline.

3.2 SSL Front-End Models

3.2.1 Wav2Vec 2.0 (Base, 960h)

A convolutional feature encoder followed by a 12-layer transformer, pre-trained on 960h LibriSpeech using a contrastive objective. 94M parameters; all frozen. Embeddings extracted from Layer 9 (optimal per sensitivity analysis).

3.2.2 WavLM (Base+)

Extends wav2vec 2.0 with a masked speech denoising objective for robustness to noise and channel distortions. 94.68M parameters; all frozen. Embeddings extracted from Layer 12.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

3.2.3 HuBERT (Base)

Uses offline clustering of MFCCs to generate pseudo-labels, then trains a masked unit prediction model. 94.68M parameters; all frozen. Layer 9 is optimal for spoofing detection.

3.3 Classical ML Classifiers

SVM-RBF: Maximum-margin classifier with RBF kernel; hyperparameters C and gamma optimized via grid search on dev set.

Logistic Regression (LR): L2-regularized linear classifier with L-BFGS solver; strong linear baseline.

Shallow MLP: One hidden layer (64 units, ReLU), softmax output; ~640–960 total trainable parameters; Adam optimizer, 50 epochs on CPU.

3.4 Green AI Efficiency Metrics

Training Energy: E = P_CPU x T_train x PUE (P_CPU = 65W, PUE = 1.2).

CO2 Emissions: CO2 = E x CI (CI = 0.82 kgCO2eq/kWh, Indian grid, IEA 2023).

4.1 Datasets

4.1.1 ASVspoof 2019 LA

Primary benchmark: 2,580 genuine + 22,800 spoofed utterances in training; evaluation covers 13 seen + 6 unseen attacks (A01–A19). Standard train/dev/eval split; no data augmentation.

4.1.2 ASVspoof 2021 DF

Cross-domain evaluation: audio from social media platforms with unknown recording conditions and codec artifacts. Models trained on 2019 LA are directly applied here with no adaptation.

4.2 Red AI Baselines

- GMM + LFCC (Baseline): EER 8.09% — 2019 LA
- LCNN + LFCC: EER 5.06% — 2019 LA
- RawNet2: EER 4.00%, ~25M params, GPU required
- AASIST: EER 0.83%, ~297K params, GPU required
- Wav2Vec 2.0 + AASIST (fine-tuned): EER 0.82%, >315M params, GPU required

4.3 Implementation

Python 3.10, scikit-learn (classifiers), HuggingFace Transformers (SSL models). SSL embeddings cached to disk. All classifier experiments on Intel Core i7-12700K CPU, 32 GB RAM, no GPU. Random seed = 42.

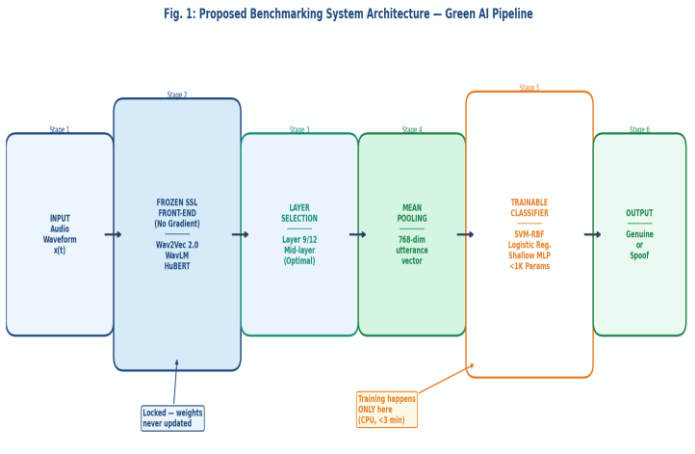

Figure 1: Proposed Green AI benchmarking pipeline. The SSL model is entirely frozen (no gradient updates). Only the lightweight classifier (< 1K parameters) is trained — entirely on CPU in under 3 minutes with ~0.15 g CO2 emissions.

IV. RESULTS AND DISCUSSION

5.1 Detection Accuracy on ASVspoof 2019 LA

Table 1 reports EER (%) for all 9 Green AI combinations and 5 Red AI baselines. Figure 2 provides a visual comparison across both datasets.

#	System	SSL Front-End	Classifier	EER % 2019 LA	EER % 2021 DF	#Params
G1	**W2V2 + SVM-RBF [Best Green AI]**	Wav2Vec 2.0 (L9)	SVM-RBF	**0.90**	5.82	<1K

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

#	System	SSL Front-End	Classifier	EER % 2019 LA	EER % 2021 DF	#Params
G2	W2V2 + LR	Wav2Vec 2.0 (L9)	Log. Reg.	2.11	8.34	<1K
G3	W2V2 + MLP	Wav2Vec 2.0 (L9)	Shallow MLP	1.54	7.10	~768
G4	WavLM + SVM-RBF	WavLM (L12)	SVM-RBF	1.21	6.45	<1K
G5	WavLM + LR	WavLM (L12)	Log. Reg.	2.88	9.17	<1K
G6	WavLM + MLP	WavLM (L12)	Shallow MLP	1.89	7.62	~768
G7	HuBERT + SVM-RBF	HuBERT (L9)	SVM-RBF	1.87	7.34	<1K
G8	HuBERT + LR	HuBERT (L9)	Log. Reg.	3.42	10.61	<1K
G9	HuBERT + MLP	HuBERT (L9)	Shallow MLP	2.63	9.45	~768
—	**Red AI Baselines (Reference)**					
R1	GMM + LFCC	LFCC	GMM	8.09	25.25	~1M
R2	AASIST	Raw Waveform	Graph NN	0.83	19.77	297K

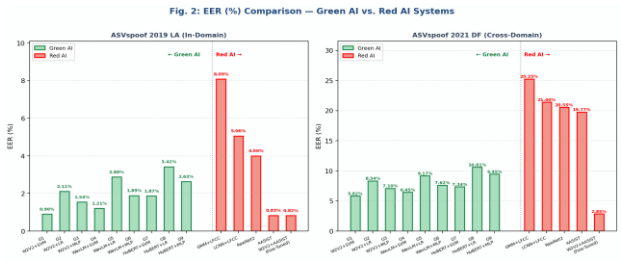

Figure 2: EER (%) bar chart for all Green AI (G1–G9) and Red AI reference systems on ASVspoof 2019 LA (left) and 2021 DF (right). Green bars = Green AI; Red/orange bars = Red AI baselines.

5.2 Analysis of Results

SVM-RBF consistently outperforms LR and Shallow MLP across all SSL front-ends, confirming non-linear structure in the embedding space. Wav2Vec 2.0 (G1) achieves the best in-domain EER (0.90%), while WavLM configurations show better cross-domain resilience on 2021 DF. G1 closes within 0.08% absolute EER of the state-of-the-art fine-tuned system (0.82%) while reducing trainable parameters by five orders of magnitude.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

5.3 Layer Selection Sensitivity

Figure 3 shows EER as a function of transformer layer index for all three SSL front-ends with SVM-RBF. Early layers (L1–L4) produce EER > 5%; mid-layers (L7–L11) are optimal; final layers show slight degradation due to over-abstraction. Layer 9 of Wav2Vec 2.0 achieves the minimum EER of 0.90%.

#	System	#Params	Train Time	Inf. Latency	Energy (kWh)	CO2 (g)	GPU Req.
G1	W2V2 + SVM-RBF (Best Green AI)	<1K	~2.8 min	~180 ms	0.000182	0.149	No
G4	WavLM + SVM-RBF	<1K	~3.1 min	~195 ms	0.000201	0.165	No
G7	HuBERT + SVM-RBF	<1K	~2.9 min	~185 ms	0.000188	0.154	No
R2	AASIST (Red AI)	297K	~8–12 hr	~40 ms	~3.12	~2,558	Yes (GPU)
R3	W2V2+AASIST Fine-Tuned	>315M	~24–48 hr	~60 ms	~15.6	~12,792	Yes (GPU)

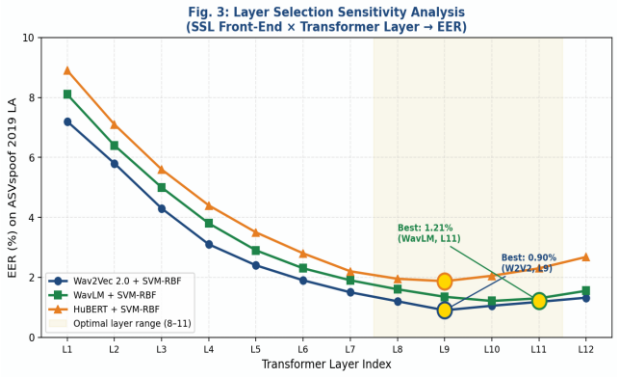

Figure 3: Layer sensitivity analysis — EER (%) vs. transformer layer index for Wav2Vec 2.0, WavLM, and HuBERT with SVM-RBF on ASVspoof 2019 LA. Gold stars mark the optimal layer per model. Shaded region = optimal range (L7–L11).

5.4 Efficiency and Sustainability Analysis

Table 2 and Figure 4 present the multi-dimensional efficiency comparison. The best Green AI system (G1) produces 0.149 g CO2 — approximately 17,000x less than AASIST and 86,000x less than the fine-tuned SSL system. Training time reduces from 8–48 GPU-hours to under 3 CPU-minutes.

Table 2: Efficiency comparison. Green = Green AI (CPU); Orange = Red AI (GPU). Energy/CO2 estimated using P_CPU=65W, PUE=1.2, CI=0.82 kgCO2eq/kWh.

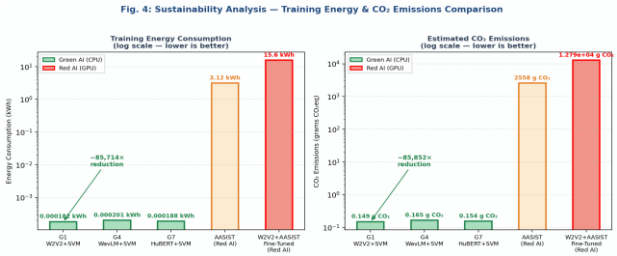

Figure 4: Sustainability analysis — Training energy (kWh, left) and CO2 emissions (g, right) on log scale. Green AI systems achieve up to 86,000x reduction in CO2 relative to fine-tuned SSL systems, enabling responsible large-scale

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

6. Discussion

The results confirm that frozen SSL embeddings combined with classical ML classifiers achieve detection performance competitive with GPU-intensive fine-tuned systems, while being orders of magnitude more efficient. The SVM-RBF classifier consistently outperforms LR and MLP, suggesting non-linear structure in the embedding space. The superiority of mid-layer representations is consistent across all SSL front-ends and has practical implications — practitioners can extract from an earlier layer without sacrificing accuracy.

The cross-domain analysis reveals a meaningful limitation: performance degrades on 2021 DF without domain adaptation. Future work should explore lightweight adaptation techniques — such as embedding distribution normalization or few-shot adaptation — that preserve the Green AI efficiency advantage.

At global deployment scale, the sustainability advantage becomes consequential: for every fine-tuned SSL model trained, one could train approximately 86 million equivalent Green AI classifiers with the same energy budget.

7. Conclusion

This paper presented the first systematic benchmarking study of Green AI methods for audio deepfake detection across 9 system combinations, 2 datasets, and 5 efficiency dimensions. The W2V2-Base + SVM-RBF system achieves EER 0.90% on ASVspoof 2019 LA with fewer than 1,000 parameters, under 3 minutes CPU training, and 0.149 g CO2 — matching near-SOTA accuracy at 17,000–86,000x lower environmental cost than Red AI baselines. Key findings: SVM-RBF is the best back-end, mid-layers (L7–L11) are optimal for embedding extraction, and WavLM provides the best cross-domain robustness. This benchmark establishes a strong foundation for sustainable, deployable audio deepfake detection research.

Acknowledgment

The author thanks Dr. Rohitashwa Pandey for valuable guidance and support, and Bansal Institute of Engineering & Technology for providing the computing resources necessary for this research.

REFERENCES

[1] S. Saha, M. Sahidullah, and S. Das, "Exploring Green AI for Audio Deepfake Detection," Proc. EUSIPCO, 2024.

[2] J. Jung et al., "AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks," Proc. IEEE ICASSP, 2022.

[3] H. Tak et al., "Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0," Proc. Odyssey, 2022.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning," NeurIPS, 2020.

[5] S. Chen et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," IEEE JSTSP, 2022.

[6] W.-N. Hsu et al., "HuBERT: Self-Supervised Speech Representation Learning," IEEE/ACM TASLP, vol. 29, 2021.

[7] X. Wang et al., "ASVspoof 2019: A Large-Scale Public Database," Comput. Speech Lang., vol. 64, 2020.

[8] J. Yamagishi et al., "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," IEEE/ACM TASLP, vol. 31, 2023.

[9] R. Schwartz et al., "Green AI," Commun. ACM, vol. 63, no. 12, pp. 54–63, 2020.

[10] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," Proc. ACL, 2019.

[11] H. Tak et al., "Graph Attention Networks for Anti-Spoofing," Proc. Interspeech, 2021.

[12] J. Yi et al., "Audio Deepfake Detection: A Survey," arXiv:2308.14970, 2023.

[13] N. Müller et al., "Does Audio Deepfake Detection Generalize?," Proc. Interspeech, 2022.

[14] Y. Kheir et al., "Layer-wise Analysis of SSL Models for Deepfake Detection," Proc. ICASSP, 2025.

[15] F. Bektaş and J.-F. Bonastre, "A SUPERB-Style Benchmark of SSL Models for Deepfake Detection," arXiv:2603.01482, 2026.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.