UDC: 004.93 https://doi.org/10.15407/jai2026.02.076

V. Tymoshchuk[1], **N. Shapoval**[2]
[1,2]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
37 Beresteysky Avenue, Kyiv, 03056
[1]timoschuk.vlad@gmail.com
[2]shovgun@gmail.com
[1]https://orcid.org/0009-0002-6565-135X
[2]https://orcid.org/0000-0002-8509-6886

DRIFT-AWARE DETECTION OF AI-GENERATED VOICES WITH A KAN-INSPIRED ADAPTER

Abstract. This paper presents a parameter-efficient approach to detecting AI-generated voices under domain shift. The task is formulated as binary classification of bona fide and spoofed speech. A frozen WavLM Base+ backbone is used as the feature encoder, while only a lightweight post-encoder adapter and a linear classifier are trained. The study follows a strict source-only protocol: models are trained on ASVspoof 2019 LA train, selected on ASVspoof 2019 LA development data, and evaluated on ASVspoof 2021 LA and the external In-the-Wild dataset. The main comparison includes no adaptation, an MLP adapter, and a KAN-inspired adapter under nearly identical trainable-parameter budgets. Experimental results show that both learned adapters substantially improve cross-domain performance compared with the no-adaptation baseline. The KAN-inspired adapter achieves the lowest mean EER on ASVspoof 2021 LA, while MLP and KAN-inspired adaptation remain close on In-the-Wild. The results indicate that lightweight post-encoder adaptation is effective for improving transfer under benchmark domain shift, whereas the advantage of KAN-inspired nonlinear adaptation is domain-dependent rather than universal.

Keywords: audio deepfake detection, speech anti-spoofing, domain shift, parameter-efficient adaptation, WavLM, KAN-inspired adapter, equal error rate.

Introduction

Recent progress in speech synthesis and voice conversion has made AI-generated speech increasingly realistic, accessible, and inexpensive to produce. As a result, synthetic and manipulated voices now pose a growing risk in biometric authentication, media verification, fraud prevention, and human-computer interaction. A central difficulty in this area is that detection performance often degrades under domain shift: systems that perform well on the distribution used for development may become substantially less reliable when evaluated on newer attacks, different recording conditions, or externally collected audio. For this reason, robust cross-domain detection is at least as important as strong in-domain benchmark performance.

This challenge is particularly relevant in logical-access spoofing scenarios, where attackers can generate or transform speech without replay artifacts and where benchmark conditions continue to evolve across challenge editions. In such settings, it is not sufficient to optimize a detector only for a fixed training distribution. Instead, a practically useful system should preserve discriminative performance when transferred to a shifted target benchmark and, ideally, to more weakly controlled external data. This motivates the use of evaluation protocols that separate source-domain training from target-domain testing and explicitly measure degradation under domain shift.

Self-supervised speech encoders provide a strong foundation for this problem because they can supply rich frame-level representations without requiring task-specific full-model fine-tuning. However, updating the entire backbone is computationally expensive and can complicate controlled comparison across adaptation strategies. A natural alternative is parameter-efficient adaptation, in which the pretrained backbone remains frozen and only a small number of trainable parameters are introduced downstream. This design is attractive both scientifically and practically: it reduces the optimization footprint, simplifies experiment management, and makes it easier to attribute performance differences to the adaptation

module rather than to broad changes in model capacity.

In this work, we study drift-aware detection of AI-generated voices using a frozen WavLM Base+ backbone and lightweight post-encoder adaptation. We focus on a strict source-only setup in which training is performed on ASVspoof 2019 LA train, model selection is based only on ASVspoof 2019 LA dev, and final evaluation is conducted on ASVspoof 2021 LA and In-the-Wild. Within this setup, we compare four configurations: no adaptation, an MLP adapter, a KAN-inspired adapter, and a preliminary LoRA-based comparator. The main controlled comparison is between the MLP and KAN-inspired adapters, which are matched to nearly identical trainable-parameter budgets to isolate the effect of adapter design.

The central question of this study is whether a KAN-inspired nonlinear adapter provides a meaningful advantage over a conventional MLP adapter for cross-domain synthetic speech detection under a matched parameter budget. Our results show that both adapters strongly improve over the no-adaptation baseline. The KAN-inspired adapter yields the strongest transfer to ASVspoof 2021 LA, where it achieves lower mean EER and lower run-to-run variability than the matched-budget MLP baseline. At the same time, this advantage does not clearly extend to the more challenging In-the-Wild evaluation, where the two adapters remain broadly comparable. Thus, the evidence supports a more specific claim: KAN-inspired post-encoder adaptation is beneficial for the target benchmark considered in this work, but it should not be interpreted as universally superior across all out-of-domain conditions.

The contributions of this paper are threefold. First, we present a protocol-safe source-only evaluation framework for cross-domain detection of AI-generated voices with ASVspoof 2019 LA as the source domain, ASVspoof 2021 LA as the target domain, and In-the-Wild as an external evaluation set. Second, we introduce and study a lightweight KAN-inspired post-encoder adapter on top of a frozen WavLM Base+ backbone. Third, we provide a controlled multi-seed comparison of no adaptation, MLP adaptation, and KAN-inspired adaptation under matched trainable-parameter budgets, complemented by efficiency analysis based on trainable parameter count, training time, and throughput.

Related Work

2.1. Audio Deepfake and Speech Anti-Spoofing Detection

Synthetic speech detection and speech anti-spoofing have been studied extensively through the ASVspoof challenge series, which has helped establish common datasets, protocols, and evaluation metrics for logical-access and related attack scenarios [1]-[3]. More recent benchmark-oriented systems increasingly rely on deep neural architectures rather than hand-crafted front ends, with strong representatives including architectures such as AASIST [8]. However, benchmark progress does not automatically imply robust deployment behavior: external generalization studies show that detectors trained under controlled benchmark conditions can degrade substantially on less controlled data [4]. Related analyses further indicate that ASVspoof-trained models may exploit unintended dataset-specific cues rather than robust synthesis artifacts [5].

2.2. Self-Supervised Speech Representations

The rise of self-supervised learning has changed the design space for synthetic speech detection. Large pretrained speech encoders such as WavLM provide general-purpose frame-level representations that can be reused across diverse downstream tasks. WavLM is designed for full-stack speech processing and is therefore a natural candidate backbone for synthetic speech detection [7]. Recent evidence also suggests that self-supervised representations can be particularly valuable for robust detection under shift: frozen self-supervised speech encoders paired with simple classifiers have been shown to provide strong cross-dataset behavior in synthetic speech detection [6]. These findings motivate the present study's decision to

keep the backbone frozen and concentrate adaptation capacity in lightweight downstream modules.

2.3. Parameter-Efficient Adaptation

Parameter-efficient adaptation offers a practical compromise between full-model fine-tuning and purely frozen-feature classification. Instead of updating all backbone parameters, a compact trainable module is introduced while the pretrained encoder remains fixed. This idea has been explored through several mechanisms, including bottleneck adapters, low-rank updates, and lightweight residual modules [9], [10]. In speech applications, parameter-efficient adaptation is appealing because pretrained encoders are relatively large, optimization can be expensive, and careful control of trainable capacity is important for matched-capacity comparison. The strong performance of frozen self-supervised representations in synthetic speech detection further supports this direction [6]. In the present work, this perspective is central: the backbone is frozen in all experiments, and only the adapter and classifier are trained. This allows the comparison to focus on the structure of the adaptation module rather than on large-scale end-to-end optimization.

2.4. KAN-Inspired Nonlinear Adaptation

Kolmogorov-Arnold Networks (KANs) have recently attracted attention as an alternative neural design in which learnable univariate function modeling plays a more explicit role than in standard multilayer perceptrons [11]. Although the original formulation is not specifically designed for speech anti-spoofing, it motivates the broader hypothesis that richer nonlinear basis functions may yield stronger adaptation than conventional pointwise activations under a constrained parameter budget. For this reason, KAN-inspired components are a natural candidate for lightweight post-encoder adaptation. However, the relevance of such nonlinear adaptation to cross-domain synthetic speech detection remains largely unexplored. This work addresses that gap by comparing a conventional MLP adapter with a KAN-inspired adapter under nearly identical trainable-parameter budgets and within a protocol-safe cross-domain evaluation setup.

Method

3.1. Problem Formulation

We consider binary classification of speech utterances into two classes: bona fide and spoofed. Let an input waveform be denoted by x, and let $y \ in \ \{0,1\}$ be the corresponding label, where $y = 1$ denotes bona fide speech and $y = 0$ denotes spoofed speech. The objective is to learn a detector that produces a scalar score $s(x)$ such that higher scores indicate stronger evidence for the bona fide class.

The study is conducted under a strict source-only protocol. Training uses only ASVspoof 2019 LA train, model selection uses only ASVspoof 2019 LA dev, and no target evaluation data are used for adaptation or checkpoint selection. Generalization is then assessed on two out-of-source evaluation domains: ASVspoof 2021 LA and In-the-Wild. The primary metric is equal error rate (EER), while additional operating-point, ranking, and efficiency metrics are reported as supporting evidence.

3.2. Detection Pipeline

The detector follows a fixed modular structure:

Figure 1 illustrates the overall detector pipeline and highlights the point at which the adapter choice is introduced.

First, the input waveform is passed through a pretrained WavLM Base+ encoder, which produces frame-level hidden representations.

The backbone remains frozen in all main experiments. During training, the backbone is also forced to remain in evaluation mode, preventing updates to its parameters and internal training-state behavior.

This isolates the effect of the adaptation module and reduces the number of trainable parameters

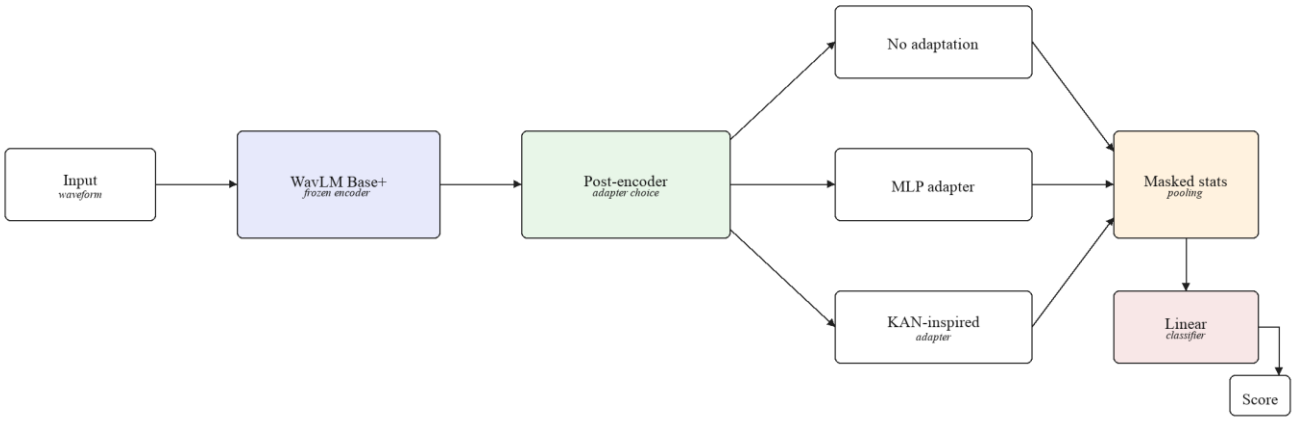

Figure 1. Overall detector architecture with post-encoder adaptation

The frame-level outputs are then processed by a post-encoder adapter. This design choice is deliberate: adaptation is introduced after the pretrained encoder rather than inside the backbone, which keeps the comparison between adapter families simple and auditable. After adaptation, masked statistics pooling computes the mean and standard deviation of valid frame representations and concatenates them into a single utterance-level embedding. A linear classifier finally maps this pooled representation to a scalar logit for binary decision making.

Training uses binary cross-entropy with logits. The training set is presented as random 4-second waveform crops, while development and evaluation use full utterances without random cropping. This preserves a lightweight and reproducible training procedure while allowing final metrics to reflect complete utterance-level inference.

3.3. MLP Adapter

The MLP adapter serves as the main conventional baseline for post-encoder adaptation. It is implemented as a residual bottleneck module operating on frame-level backbone features. Let $h_t \, in \, R^d$ denote the feature vector at frame t. The adapter applies a down-projection from d to a hidden bottleneck dimension m, followed by a GELU nonlinearity, dropout, and an up-projection back to d. The transformed output is then added to the original input through a residual connection.

This design provides a compact nonlinear transformation while preserving the dimensionality expected by the downstream pooling and classifier layers. Because the adapter is inserted after the frozen backbone, all learnable flexibility comes from this bottleneck transformation and the final classifier. In the experiments, the hidden dimension is not chosen arbitrarily; instead, it is derived from a target trainable-parameter budget to enable a matched-capacity comparison with the KAN-inspired adapter.

3.4. KAN-Inspired Adapter

The proposed KAN-inspired adapter keeps the same residual bottleneck structure as the MLP baseline but replaces the standard pointwise activation with a learnable fixed-grid basis activation. As in the MLP case, each frame representation is first projected from the backbone dimension d to a hidden bottleneck dimension m. Instead of applying GELU, the hidden representation is transformed by a learnable basis-response module defined over a fixed grid. The resulting transformed features are then passed through dropout, projected back to dimension d, and added to the input through a residual connection.

The basis-response module is intended to provide a richer nonlinear transformation than a

conventional MLP activation while remaining lightweight. In the implementation used here, each hidden channel has learnable coefficients over a fixed grid together with trainable scale and bias terms. The adapter therefore remains compact, but its nonlinearity is more structured than the standard GELU-based bottleneck. Because this module is inspired by, rather than identical to, the original KAN formulation, we refer to it as a KAN-inspired adapter throughout the paper.

Figure 2 visualizes the internal structure of the KAN-inspired adapter used in this work. This schematic clarifies the distinction between the proposed KAN-inspired adapter and a direct reimplementation of the original KAN architecture. The module preserves the residual bottleneck layout of the MLP adapter while replacing the standard activation stage with a learnable fixed-grid basis-response transformation.

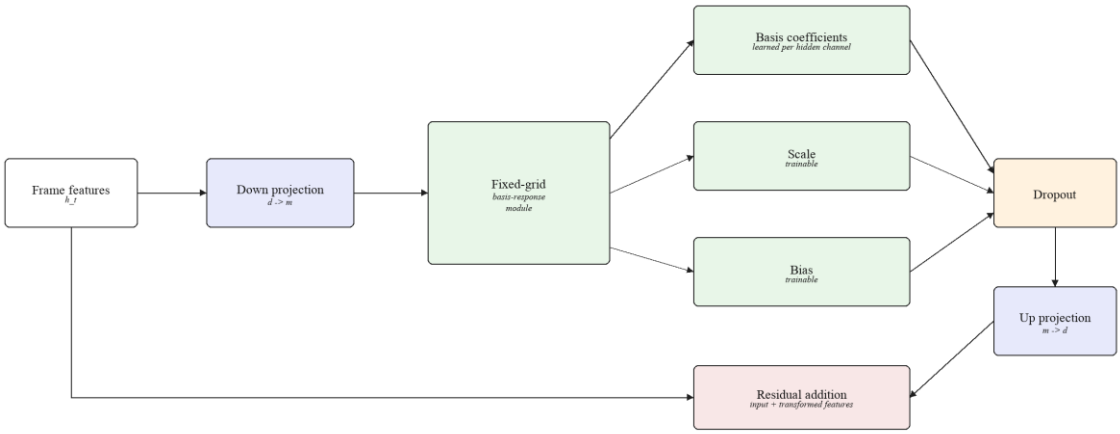

Figure 2. Internal structure of the KAN-inspired post-encoder adapter

In the main comparison, we use a single-block KAN-inspired adapter with basis size 8 and the same dropout rate as the MLP baseline. Additional variants, including stacked, gated, smaller, and larger KAN-inspired adapters, were explored separately as ablations, but the primary conclusions of this paper are based on the single-block matched-budget comparison against the MLP baseline.

3.5. Controlled Evaluation Protocol

Particular attention is paid in this study to the validity of comparing the developed model variants. To ensure a correct comparison, the experimental setup follows three constraints. First, the backbone is frozen by default in all main experiments, and only the adapter and classifier are trainable. Second, checkpoint selection is performed exclusively on ASVspoof 2019 LA dev EER; neither ASVspoof 2021 LA nor In-the-Wild is used for model selection. Third, the MLP and KAN-inspired adapters are matched to nearly identical trainable-parameter budgets through an explicit budget-construction procedure. When the hidden dimension is not specified manually, it is derived automatically from a shared target parameter budget, and the build is rejected if the resulting MLP and KAN-inspired parameter counts differ beyond a configured tolerance.

Under the final main setup, the learned adapters are matched closely enough that their exact parameter counts differ only marginally; the values are reported later in the efficiency analysis. This ensures that performance differences between the two adapter families cannot be attributed to a large mismatch in trainable capacity. In addition, all main comparisons are reported across three random seeds, which helps avoid conclusions based on a single favorable run. The result is a controlled experimental setting in which differences in

cross-domain performance can be interpreted primarily in terms of adaptation strategy rather than protocol leakage, backbone fine-tuning, or parameter-count imbalance.

Experimental Setup

4.1. Datasets and Evaluation Domains

The experiments follow a source-only cross-domain protocol with one source domain, one target benchmark domain, and one external evaluation domain. ASVspoof 2019 LA is used as the source dataset: the train split is used for optimization and the dev split is used for model selection. ASVspoof 2021 LA serves as the target-domain benchmark, and In-the-Wild serves as an additional external evaluation set intended to probe generalization beyond challenge-style conditions.

This partitioning is important for the scientific validity of the study. Neither ASVspoof 2021 LA nor In-the-Wild is used during training or checkpoint selection. The protocol logic is enforced in code at the manifest and dataloader levels, preventing target-domain leakage into the adaptation process. As a result, performance on ASVspoof 2021 LA and In-the-Wild reflects genuine transfer from the source-domain training setup rather than target-aware tuning.

The choice of evaluation domains allows us to analyze two different forms of shift. ASVspoof 2021 LA represents a benchmark target domain that is newer than the source domain but still belongs to the ASVspoof logical-access family. In contrast, In-the-Wild is treated as a more challenging external domain with less controlled conditions. This distinction is useful because it separates improved target-benchmark transfer from broader out-of-domain robustness.

4.2. Training Configuration

All main experiments use the same backbone, detector structure, optimizer family, and training schedule. The backbone is `microsoft/wavlm-base-plus`, kept frozen throughout training. The detector uses masked mean and standard deviation pooling followed by a linear binary classification head. Optimization is performed with AdamW, and a cosine annealing learning-rate schedule is applied over 10 epochs. The training batch size is 8, while the development and evaluation batch sizes are 4. Mixed-precision training is enabled on CUDA where available.

To reduce memory use and standardize the training procedure, each training example is presented as a random 4-second crop. Development and evaluation, however, are performed on full utterances without random cropping. The loss function is binary cross-entropy with logits, and gradient clipping is applied with a maximum norm of 5.0. The main experiments are repeated with three random seeds: 1337, 2024, and 3407.

The main controlled comparison includes three configurations: no adaptation, an MLP adapter, and a KAN-inspired adapter. The no-adaptation baseline trains only the final classifier. The MLP and KAN-inspired variants both train only the adapter and classifier on top of the frozen backbone. Their hidden dimensions are derived automatically from a shared target parameter budget of approximately 200k trainable parameters; exact trainable parameter counts are reported in the efficiency analysis. We also include a preliminary LoRA-based comparator [10], in which low-rank updates are attached to the frozen backbone while the post-encoder adapter is removed. In the current setup, this LoRA configuration targets the `q_proj` and `v_proj` modules.

4.3. Evaluation Metrics

The primary evaluation metric is equal error rate (EER), which is standard for spoofing detection and remains the main basis for model selection and comparison. Checkpoints are selected exclusively by source-domain development EER, and all main claims in the paper are anchored in EER on ASVspoof 2021 LA and In-the-Wild.

To complement EER, we also report several secondary metrics computed from the saved score files. These include ROC-AUC, accuracy, precision, recall, and F1 score, as well as FRR at 1% FAR and FRR at 0.1% FAR. For ASVspoof 2021 LA, we additionally compute

minimum tandem detection cost function (min t-DCF) using the official ASVspoof 2021 evaluation package. This metric is especially useful because it reflects the practical cost trade-off between missed bona fide trials and accepted spoof trials in an ASVspoof-style operating scenario.

Efficiency is assessed through trainable parameter count, total training time, and training throughput in examples per second. We treat these quantities as secondary but important measures because the study is motivated not only by cross-domain quality, but also by parameter efficiency and clean experimental comparison. For the main MLP-versus-KAN-inspired comparison, throughput is particularly informative because it helps determine whether any quality gain is achieved at a substantial computational cost.

Results and Discussion

5.1. Main Results

The no-adaptation baseline performs worst by a large margin, reaching 7.41% EER on ASVspoof 2019 LA dev, 21.29% on ASVspoof 2021 LA, and 42.41% on In-the-Wild. This confirms that a frozen backbone followed only by a linear classifier is insufficient for robust cross-domain spoofing detection in the present setting.

Table 1 reports the per-seed results for the two main learned adapters. This view is more informative than a purely aggregate table because it shows both central tendency and run-to-run behavior. The KAN-inspired adapter outperforms the MLP baseline on ASVspoof 2021 LA for all three tested seeds, whereas the In-the-Wild comparison remains mixed and much closer.

Table 1. Per-seed comparison of MLP and KAN-inspired adapters

Method	Seed	Dev EER	ASVspoof 2021 LA EER	In-the-Wild EER
MLP adapter	1337	1.76%	17.95%	26.21%
MLP adapter	2024	1.80%	16.90%	**24.72%**
MLP adapter	3407	1.73%	16.17%	27.30%
KAN-inspired adapter	1337	**1.65%**	15.56%	25.08%
KAN-inspired adapter	2024	1.69%	**15.30%**	26.79%
KAN-inspired adapter	3407	1.88%	15.72%	27.01%

The direct comparison between MLP and KAN-inspired adaptation should be interpreted carefully. On the source-domain dev set, the difference is negligible: 1.76% versus 1.74% EER. On ASVspoof 2021 LA, however, the KAN-inspired adapter improves mean EER by 1.48 percentage points relative to the MLP baseline, and it does so with considerably lower run-to-run variability. On In-the-Wild, the two methods remain effectively comparable within the observed variance, and the KAN-inspired adapter does not demonstrate a clear advantage.

5.2. Cross-Domain Generalization

The most important practical question in this study is not whether adaptation improves source-domain development performance, but whether it improves transfer under domain shift. From this perspective, the results are encouraging: both learned adapters reduce the large degradation observed for the no-adaptation baseline. The gain is especially pronounced on In-the-Wild, where a linear classifier on frozen features is clearly insufficient, while both learned post-encoder adapters remain substantially more robust.

Table 2 presents supporting metrics on ASVspoof 2021 LA. These values are useful because they show that the KAN-inspired adapter improves not only the primary EER metric, but also several additional target-domain indicators, including min t-DCF, ROC-AUC, and F1 score.

Table 2. Supporting metrics on ASVspoof 2021 LA

Method	EER	min t-DCF	ROC-AUC	F1	FRR @ 1% FAR
no adaptation	21.29%	0.7785	85.80%	42.91%	61.55%
MLP adapter	17.01% +/- 0.90%	0.6132 +/- 0.0317	92.35%	49.81%	46.27%
KAN-inspired adapter	**15.53% +/- 0.21%**	**0.5633 +/- 0.0112**	**93.41%**	**52.51%**	**45.70%**

These findings suggest that post-encoder adaptation is not merely improving source-domain fit, but is also learning transformations that help the frozen backbone transfer more effectively to shifted evaluation conditions. At the same time, the two evaluation domains tell slightly different stories. ASVspoof 2021 LA indicates that the KAN-inspired adapter is the stronger target-domain transfer mechanism in the present setup. In contrast, In-the-Wild suggests that this advantage is not universal: although both adapters generalize far better than no adaptation, neither clearly dominates the other on this more challenging external dataset.

This distinction matters for interpretation. The ASVspoof 2021 LA result is practically relevant because it shows improved transfer to a newer target benchmark under controlled logical-access shift. However, it should not be overstated as a complete proxy for real-world robustness. The In-the-Wild results act as a useful counterbalance by showing that strong target-benchmark transfer does not automatically translate into clearly superior performance in more weakly controlled external conditions.

5.3. Efficiency Analysis

Efficiency is an important part of the comparison because the study is motivated by parameter-efficient adaptation rather than by unrestricted model scaling. The no-adaptation baseline is the most compact configuration, but its quality is substantially weaker than that of the adapted models. The MLP and KAN-inspired adapters, in contrast, are intentionally matched by trainable budget, making the comparison effectively budget-matched.

Table 3 summarizes the main efficiency indicators. For the central MLP-versus-KAN-inspired comparison, the two adapted models are almost indistinguishable in parameter count, training time, and training throughput. This makes the target-domain quality difference easier to interpret as an effect of adaptation design rather than of computational scale.

Table 3. Efficiency summary

Method	Trainable Parameters	Mean Train Time	Train Throughput
no adaptation	1,537	NA	NA
MLP adapter	202,115	0.84h +/- 0.21h	88.42 +/- 23.72 ex/s
KAN-inspired adapter	201,868	**0.83h +/- 0.14h**	**88.61 +/- 17.11 ex/s**
LoRA (preliminary)	185,857	0.96h +/- 0.00h	73.54 +/- 0.00 ex/s

The efficiency results show that the MLP and KAN-inspired adapters are effectively matched not only in trainable parameter count, but also in training time and throughput. Therefore, the ASVspoof 2021 LA improvement of the KAN-inspired model cannot be explained by a larger trainable budget or by a favorable computational trade-off. In practical terms, the KAN-inspired adapter offers a stronger quality-efficiency balance on the target benchmark, while preserving the same operational profile as the MLP baseline.

We also conducted several exploratory KAN-inspired ablations. A smaller variant degraded dev performance, a larger variant did not clearly improve over the base configuration, and stacked or gated variants did not outperform the single-block base adapter. Although these ablations are not yet supported by the same full target-domain evaluation coverage as the main comparison, they suggest that the current base KAN-inspired configuration is already close to

the most favorable quality-efficiency operating point among the tested KAN-inspired variants.

5.4. Preliminary LoRA Comparison

The preliminary LoRA configuration performs weakly in the current setup. With 185,857 trainable parameters, it reaches 7.38% dev EER, which is close to the no-adaptation baseline and far worse than either post-encoder adapter. Because this result is currently available only for development evaluation and only for a single seed, it should be interpreted as a preliminary negative result rather than as a definitive statement about LoRA for speech anti-spoofing in general.

Nevertheless, the current outcome is informative. It suggests that, in the present frozen-backbone setup, lightweight adaptation applied after the encoder is substantially more effective than the tested low-rank updates to `q_proj` and `v_proj`. This makes the MLP and KAN-inspired post-encoder adapters the more relevant comparison pair for the central research question of the paper.

5.5. Limitations

Several limitations should be acknowledged. First, the study uses a single pretrained backbone, WavLM Base+, and therefore does not establish whether the same conclusions hold across other self-supervised speech encoders. Second, although the source-only protocol is strict and scientifically clean, it is still built around a particular benchmark family, and improved performance on ASVspoof 2021 LA should not be equated automatically with universal real-world robustness. Third, the LoRA comparison is still incomplete, since it currently lacks the same multi-seed and full cross-domain evaluation coverage as the main post-encoder adapter experiments.

Additional limitations concern the scope of architectural exploration. While several KAN-inspired variants were tested, they were evaluated primarily as exploratory checks rather than as full protocol-matched target-domain studies. Finally, although multiple secondary metrics are reported, the core evidence still comes from a relatively compact set of runs. Extending the study with additional seeds, backbones, and external domains would strengthen the generality of the conclusions.

Conclusion

This paper studied drift-aware detection of AI-generated voices using lightweight post-encoder adaptation on top of a frozen WavLM Base+ backbone. Under a strict source-only protocol, we compared no adaptation, an MLP adapter, and a KAN-inspired adapter, with the two learned adapters matched to nearly identical trainable-parameter budgets. The results show that post-encoder adaptation is highly effective in this setting: both learned adapters substantially outperform the no-adaptation baseline on ASVspoof 2021 LA and In-the-Wild.

The main comparative finding is that the KAN-inspired adapter provides the strongest transfer to ASVspoof 2021 LA. It achieves lower mean EER than the matched-budget MLP adapter and also improves supporting target-domain metrics such as min t-DCF, ROC-AUC, and F1, while maintaining essentially identical training throughput. This indicates that the gain is not explained by a larger trainable budget or an obvious efficiency trade-off. At the same time, the KAN-inspired advantage does not clearly extend to In-the-Wild, where the MLP and KAN-inspired adapters remain broadly comparable within the observed run-to-run variability.

Taken together, the findings support a focused rather than universal conclusion: KAN-inspired nonlinear post-encoder adaptation is a promising parameter-efficient strategy for target-domain transfer in synthetic speech detection, particularly on ASVspoof 2021 LA under the studied source-only setup. However, the current evidence does not justify a blanket claim of superiority across all out-of-domain conditions. Future work should extend the comparison to additional speech backbones, stronger parameter-efficient baselines, more extensive multi-seed evaluation, and broader external datasets to better

characterize when KAN-inspired adaptation offers the greatest practical benefit.

References

1. Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K. A., et al. (2020). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. Computer Speech & Language, 64, Article 101114.
https://doi.org/10.1016/j.csl.2020.101114

2. Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., Evans, N., & Delgado, H. (2021). ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. arXiv preprint, arXiv:2109.00537. https://doi.org/10.48550/arXiv.2109.00537

3. Delgado, H., Evans, N., Kinnunen, T., Lee, K. A., Liu, X., Nautsch, A., Patino, J., Sahidullah, M., Todisco, M., Wang, X., & Yamagishi, J. (2021). ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. arXiv preprint, arXiv:2109.00535.
https://doi.org/10.48550/arXiv.2109.00535

4. Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A., & Böttinger, K. (2022). Does audio deepfake detection generalize? Proceedings of Interspeech 2022, 2783-2787.
https://doi.org/10.21437/Interspeech.2022-108

5. Müller, N. M., Dieckmann, F., Czempin, P., Canals, R., Böttinger, K., & Williams, J. (2021). Speech is silver, silence is golden: What do ASVspoof-trained models really learn? Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 55-60.
https://doi.org/10.21437/ASVSPOOF.2021-9

6. Oneata, D., Stan, A., Pascu, O., Oneata, E., & Cucu, H. (2023). Towards generalisable and calibrated synthetic speech detection with self-supervised representations. arXiv preprint, arXiv:2309.05384.
https://doi.org/10.48550/arXiv.2309.05384

7. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., & Wei, F. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1505-1518.
https://doi.org/10.1109/JSTSP.2022.3188113

8. Jung, J., Heo, H., Tak, H., Shim, H., Chung, J. S., Lee, B., Yu, H. J., & Evans, N. (2022). AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 6367-6371.
https://doi.org/10.1109/ICASSP43922.2022.9747766

9. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. Proceedings of the 36th International Conference on Machine Learning, 97, 2790-2799.
https://proceedings.mlr.press/v97/houlsby19a.html

10. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv preprint, arXiv:2106.09685.
https://doi.org/10.48550/arXiv.2106.09685

11. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., & Tegmark, M. (2024). KAN: Kolmogorov-Arnold networks. arXiv preprint, arXiv:2404.19756.
https://doi.org/10.48550/arXiv.2404.19756

The article has been sent to the editors 28.04.26.
After processing 15.05.26.
Submitted for printing 30.06.26