Classical machine learning baselines for synthetic speech detection across modern voice cloning systems*

Khrystyna Ruda[1,*,†], Dmytro Sabodashko[1,†], Volodymyr Povoroznyk[1,†], and Yuriy Khoma[1,†]

[1] *Lviv Polytechnic National University, 12 Stepan Bandera str., 79000 Lviv, Ukraine*

Abstract

The rapid progress of neural text-to-speech and voice conversion technologies has enabled the generation of highly realistic synthetic speech, significantly increasing the risk of voice-based spoofing and social engineering attacks. This work investigates synthetic speech detection using classical machine learning approaches based on handcrafted acoustic features. A controlled dataset consisting of bona fide speech and synthetic audio generated by four modern voice cloning and text-to-speech systems—RVC, XTTS, ElevenLabs, and Tortoise—is constructed for experimental evaluation. A fixed-dimensional feature representation capturing cepstral, spectral, and temporal characteristics is used in combination with several classical classifiers, including support vector machines, ensemble methods, and distance-based models. Experimental results demonstrate that non-linear classifiers, particularly SVM with an RBF kernel, achieve strong detection performance under controlled conditions, with an Equal Error Rate as low as 5%. Vendor-wise analysis reveals differences in detectability across synthesis systems, with commercial neural TTS exhibiting slightly reduced artifact visibility. The presented results establish robust classical baselines and provide a reference point for comparison with more advanced deep learning–based detection method

Keywords

Audio deepfake; synthetic speech detection; spoofing countermeasures; machine learning; speaker verification.

1. Introduction

Recent advances in text-to-speech (TTS) and voice conversion technologies have enabled the generation of highly realistic synthetic speech that is increasingly difficult to distinguish from genuine human audio. Modern neural architectures, including transformer-based, diffusion-based, and autoregressive models, can accurately replicate speaker identity, prosody, and acoustic characteristics [1]. While these technologies support legitimate applications such as voice assistants and accessibility tools, their misuse introduces significant security and societal risks.

Synthetic speech can be exploited to bypass voice-based authentication systems, conduct voice phishing attacks, and disseminate misinformation through fabricated audio recordings. In forensic and legal contexts, the availability of convincing synthetic speech further complicates the assessment of audio authenticity and evidentiary reliability [2]. As a result, reliable detection of synthetic and manipulated speech has become a critical research problem in biometric security and audio forensics.

Early work on synthetic speech detection and spoofing countermeasures primarily focused on protecting automatic speaker verification (ASV) systems using handcrafted acoustic features and classical classifiers [3]. More recent research has introduced deep learning–based detectors, including convolutional, transformer-based, and end-to-end architectures, which achieve strong performance under controlled evaluation conditions [4]. However, several open challenges remain,

* *CPITS 2026: Cybersecurity Providing in Information and Telecommunication Systems, February 28, 2026, Kyiv, Ukraine*
* Corresponding author.
† These authors contributed equally.
khrystyna.s.ruda@lpnu.ua (K. Ruda); dmytro.v.sabodashko@lpnu.ua (D. Sabodashko); volodymyr.b.povoroznyk@lpnu.ua (V. Povoroznyk), yurii.v.khoma@lpnu.ua (Y. Khoma)
0000-0001-8644-411X (K. Ruda); 0000-0003-1675-0976 (D. Sabodashko); 0009-0001-7242-6989 (V. Povoroznyk); 0000-0002-4677-5392 (Y. Khoma)

particularly with respect to generalization across different synthesis methods, robustness to audio quality variations, and interpretability of detection decisions.

In this context, classical machine learning approaches remain relevant as computationally efficient and interpretable baselines. Despite being overshadowed by deep models, feature-based detectors provide valuable insights into synthesis artifacts and enable controlled analysis of classifier behavior under different conditions. Establishing strong classical baselines is therefore important for understanding the relative benefits of more complex architectures.

This paper investigates the detection of synthetic speech generated by multiple modern TTS and voice cloning systems. A systematic evaluation is conducted using a set of classical machine learning classifiers operating on handcrafted acoustic features. The study focuses on assessing detection performance across different synthesis methods and analyzing classifier behavior using standard metrics commonly employed in biometric and spoofing detection research.

In contrast to the prevailing trend toward increasingly complex neural architectures, this work deliberately focuses on classical machine learning methods. Such approaches remain highly relevant in security-sensitive and resource-constrained environments due to their computational efficiency, interpretability, and ease of deployment. Establishing strong classical baselines is essential for meaningful comparison with future deep learning–based detectors and for understanding which synthesis artifacts remain detectable without end-to-end representation learning.

The main contributions of this work are as follows:

1) construction of a controlled multi-vendor dataset comprising bona fide speech and synthetic audio generated by four modern TTS and voice conversion systems;

2) systematic evaluation of classical machine learning classifiers for synthetic speech detection using a unified handcrafted acoustic feature set;

3) vendor-wise analysis of detection performance highlighting generator-dependent artifacts and false rejection tendencies;

4) establishment of strong classical baselines suitable for benchmarking and ablation against deep learning–based spoofing detectors.

The presented results establish strong classical baselines and provide a reference point for subsequent comparison with more advanced deep learning–based detection approaches.

2. Related Work and Literature Review

The rapid advancement of neural text-to-speech (TTS) and voice conversion (VC) technologies has enabled the generation of highly realistic synthetic speech, substantially expanding the attack surface for speaker verification systems, voice-controlled services, and social engineering scenarios. Contemporary synthesis frameworks based on transformers, diffusion models, and neural vocoders are capable of producing speech with high perceptual quality and strong speaker similarity, rendering human judgment and many traditional signal-processing–based defenses increasingly unreliable. As highlighted in recent surveys, the primary challenge in audio deepfake detection has therefore shifted from identifying obvious synthesis artifacts toward achieving robust generalization across unseen generation methods, acoustic environments, and transmission channels [1].

To support systematic evaluation and comparison of detection approaches, the research community has introduced several standardized benchmarks and evaluation campaigns. The ASVspoof challenge series has emerged as the primary framework for assessing spoofing countermeasures in automatic speaker verification (ASV) systems. While early editions focused on logical access attacks based on TTS and VC, as well as physical access replay scenarios, more recent iterations emphasize realistic recording conditions, larger and more diverse datasets, and increased environmental variability [5]. In parallel, initiatives such as the Audio Deepfake Detection (ADD) challenge place additional emphasis on cross-domain generalization and the detection of modern neural synthesis methods in unconstrained settings [6]. Despite steady methodological progress,

evaluation reports consistently reveal a pronounced performance gap between in-domain and out-of-domain test conditions, underscoring the limited robustness of many existing detectors.

Early approaches to audio deepfake detection primarily relied on handcrafted acoustic features, including MFCC, CQCC, LFCC, and phase-based descriptors, combined with classical classifiers such as Gaussian mixture models and support vector machines. Although these methods are computationally efficient and offer a degree of interpretability, they often exploit generator-specific artifacts and dataset biases. As a result, their performance tends to degrade significantly when confronted with unseen synthesis techniques or real-world distortions [7]. At the same time, recent studies have demonstrated that individualized modeling can substantially affect speech system behavior, highlighting the importance of speaker-specific characteristics in both recognition and security-sensitive applications [8].

In contrast, current state-of-the-art detection systems are largely based on deep learning, with convolutional and transformer-based architectures dominating recent literature. A common strategy applies deep models to time–frequency representations such as Mel-spectrograms, CQT, or MFCC maps using CNNs, residual networks, LCNNs, or vision transformers adapted for audio analysis. Alternatively, end-to-end approaches, exemplified by RawNet-based architectures, operate directly on raw waveforms or minimally processed signals. While extensive experimental evidence shows that deep models substantially outperform classical baselines under controlled conditions, these gains are frequently accompanied by reduced interpretability and increased sensitivity to dataset-specific cues. Beyond standalone detection, integrated biometric systems based on speaker embeddings have been successfully applied for intruder detection and access control, demonstrating the practical relevance of voice-based security pipelines in real-world deployments [9, 10].

More recently, self-supervised learning (SSL) has emerged as a promising direction for improving the robustness of audio deepfake detection systems. Speech representation models pretrained on large-scale unlabeled corpora are increasingly employed as fixed feature extractors or fine-tuned backbones for spoofing detection. Several studies report improved transferability across datasets and attack types, particularly when SSL representations are combined with extensive data augmentation and regularization strategies. Nevertheless, recent surveys caution that SSL-based detectors are not immune to overfitting and may still rely on spurious correlations if evaluation protocols are insufficiently rigorous. Moreover, prior work has shown that the scalability and architectural design of neural voice biometric systems significantly influence authentication reliability, motivating further investigation into complementary spoofing detection mechanisms [11].

Given that many practical attacks explicitly target speaker verification systems, a substantial body of research has focused on integrating spoofing countermeasures directly into ASV pipelines rather than treating detection as an isolated classification task. Results from ASVspoof evaluations demonstrate that practical effectiveness depends not only on detection accuracy but also on the interaction between the countermeasure and ASV decision thresholds, as overly aggressive rejection of spoofed speech can lead to unacceptable false rejection rates for genuine users. This has motivated the development of deployment-aware evaluation metrics and joint optimization strategies. Recent surveys further emphasize that robustness and generalization remain open challenges despite rapid methodological progress [11–15].

In summary, existing research demonstrates considerable progress in audio deepfake detection; however, generalization, robustness, and real-world applicability remain persistent open challenges. The rapid evolution of speech synthesis technologies, together with channel effects such as noise, compression, and transmission artifacts, continues to undermine detector reliability. These limitations motivate further research into robust representations, evaluation beyond benchmark-specific conditions, and detection strategies that remain effective under realistic and evolving threat models.

3. Methodology and System Design

Recent advances in text-to-speech and voice conversion technologies have enabled the generation of highly realistic synthetic speech, introducing new security and societal risks. Synthetic audio can be exploited to bypass voice-based authentication systems, facilitate voice phishing and identity theft, and disseminate misinformation through fabricated audio evidence. In legal and forensic contexts, the availability of convincing synthetic recordings further complicates the assessment of audio authenticity. These developments highlight the growing need for reliable methods capable of distinguishing bona fide human speech from synthetically generated or manipulated audio.

From a technical perspective, audio deepfake detection can be formulated as a binary classification problem. Given an input audio signal $x \in R^N$, where N denotes the number of samples, the objective is to learn a classifier that estimates the probability of the signal being synthetic. The classifier outputs a scalar value in the range [0,1], which is interpreted as the likelihood that the input audio is artificially generated.

The primary objective of this work is to develop a robust synthetic speech detection system capable of identifying audio produced by multiple modern text-to-speech and voice cloning methods. In addition, this study aims to compare different architectural paradigms, including spectrogram-based models and raw waveform–based models, in order to assess their relative effectiveness. Particular emphasis is placed on evaluating the generalization capability of detection models across different synthesis systems and acoustic conditions. Finally, the work seeks to provide detection results that are interpretable and suitable for practical deployment in security-sensitive applications.

4. Dataset Description

The experiments are conducted on a custom dataset consisting of 3,600 audio recordings, including both authentic and synthetic speech samples. The dataset is designed to evaluate audio deepfake detection under controlled conditions with multiple state-of-the-art speech synthesis systems. Figure 1 demonstrates the dataset structure.

The dataset contains 400 real recordings from 20 speakers and 3200 synthetic recordings generated using four modern voice cloning and text-to-speech systems: RVC, XTTS, ElevenLabs, and Tortoise. Authentic and synthetic samples are balanced at the dataset level.

The data are divided into training, validation, and test subsets using a fixed split: training set: 50% (1,800 recordings), validation set: 25% (900 recordings), test set: 25% (900 recordings). The dataset is split into training (50%), validation (25%), and test (25%) subsets using a stratified procedure. The split preserves the distribution of (i) class labels (real vs. synthetic), (ii) generation method (RVC, XTTS, ElevenLabs, Tortoise), and (iii) speaker identity, i.e., samples from each speaker are present in all subsets while maintaining comparable per-speaker and per-method proportions. This controls for class and method imbalance and ensures consistent speaker coverage across all partitions.

Each recording is annotated with a binary label (real or synthetic). For synthetic samples, the corresponding generation method is also stored, enabling per-generator performance analysis.

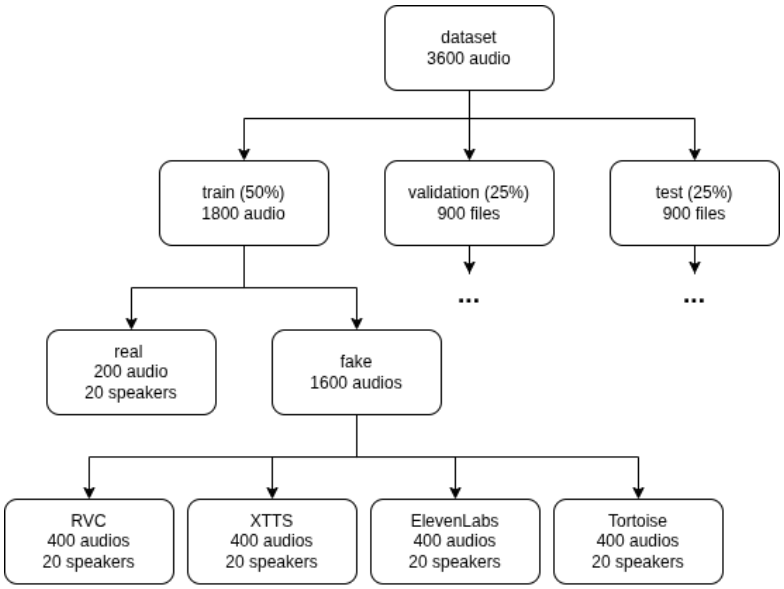

Figure 1: Structure of the proposed multi-vendor synthetic speech dataset

The dataset structure allows evaluation of detection performance across multiple synthesis techniques while maintaining consistent speaker coverage and controlled experimental conditions.

5. Access control system architecture

The proposed system architecture demonstrated on Figure 2 integrates a synthetic speech detection module with a conventional speaker verification (SV) system to improve robustness against audio deepfake and voice cloning attacks. The overall design follows a defense-in-depth paradigm, where spoofed or manipulated speech signals are filtered before reaching the biometric verification stage.

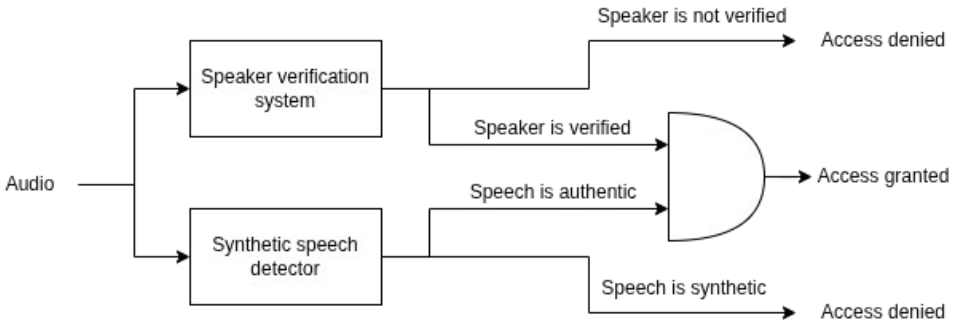

Figure 2: Access control system architecture

At the input, the system receives an audio utterance that may originate either from a genuine human speaker or from a synthetic source produced by text-to-speech or voice conversion models. This signal is first processed by the synthetic speech detector, which operates as an independent front-end module. The detector analyzes acoustic characteristics of the input signal and produces a binary or probabilistic decision indicating whether the speech is authentic or synthetically generated.

Based on the detector's output, a decision gating mechanism controls access to the downstream speaker verification system. If the input is classified as synthetic, the authentication process is terminated and the access request is rejected. Only signals that are classified as genuine are forwarded to the speaker verification system, which performs identity matching by comparing extracted speaker embeddings against enrolled speaker models.

The speaker verification subsystem follows a standard embedding-based paradigm, where speaker-specific representations are extracted from the input signal and compared using a similarity metric or decision threshold. By isolating the verification module from spoofed inputs, the architecture reduces the risk of false acceptances caused by high-quality voice cloning attacks.

This modular design provides several advantages. First, it allows the synthetic speech detector and the speaker verification system to be developed, trained, and evaluated independently. Second, it enables flexible deployment, as the detector can be updated or replaced without modifying the core biometric system. Finally, the architecture aligns with real-world authentication pipelines, where spoofing countermeasures are typically deployed as a front-end protection layer rather than as an integral part of the verification model.

6. Synthetic audio detection pipeline

The proposed detection pipeline shown on Figure 3 consists of two main stages: audio preprocessing and synthetic speech detection.

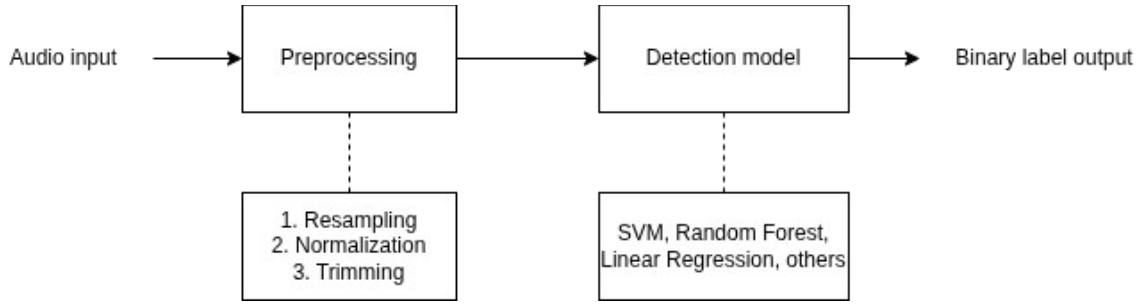

Figure 3: Classical synthetic speech detection pipeline

In the preprocessing stage, each input audio signal is first resampled to a unified sampling rate to ensure consistency across recordings. The signal is then normalized to reduce amplitude variability, followed by trimming to remove leading and trailing silence. These steps aim to minimize channel-related variability and ensure stable input conditions for all detection models.

The synthetic speech detection pipeline employed in this study consists of two main stages: audio preprocessing and feature-based classification. The pipeline is designed to ensure consistent input conditions across all samples and to enable a fair comparison of classical machine learning classifiers.

In the preprocessing stage, each input audio signal is resampled to a unified sampling rate to eliminate variability caused by differing recording formats. Amplitude normalization is applied to reduce dynamic range differences across recordings, followed by trimming of leading and trailing silence segments. These steps aim to minimize channel-related and recording-specific variability while preserving synthesis-related artifacts relevant for detection.

Following preprocessing, each audio signal is transformed into a fixed-dimensional handcrafted feature representation. This representation is intended to emphasize artifacts commonly introduced by text-to-speech and voice conversion systems, such as unnatural spectral envelopes, abnormal temporal dynamics, and inconsistencies in energy distribution. Several handcrafted features are motivated by known vocoder-related artifacts observed in synthetic speech [16].

The resulting feature vectors are provided as input to a set of classical machine learning classifiers, each trained to perform binary classification between bona fide and synthetic speech. All classifiers operate on the same feature representation and are trained and evaluated under identical data partitioning and preprocessing conditions to ensure comparability of results. Model hyperparameters are selected using the validation subset.

Each model independently produces a binary decision indicating whether the input signal is bona fide or synthetic. Model outputs are evaluated separately to enable a fair comparison of detection performance across different architectural paradigms.

This modular pipeline allows the preprocessing and detection components to be modified independently and supports systematic evaluation of multiple detection approaches under identical input conditions.

7. Approach

This study investigates a set of classical machine learning (ML) classifiers as baseline detectors for synthetic speech [17]. These methods follow a conventional two-stage pipeline consisting of handcrafted feature extraction followed by supervised classification, an approach widely adopted in earlier spoofing detection and audio forensics research.

The dataset includes synthetic speech generated using multiple modern text-to-speech and voice conversion systems, representing diverse synthesis paradigms.

Table 1
Overview of synthetic speech generation systems used in the dataset

Method	Type	Description
ElevenLabs	Neural TTS	Commercial high-quality TTS with voice cloning
RVC	Voice Conversion	Retrieval-based voice conversion using VITS [18]
Tortoise-TTS	Autoregressive TTS	Multi-speaker TTS with voice cloning [19]
XTTS	Neural TTS	Cross-lingual neural TTS system

Each audio signal is transformed into a fixed-dimensional feature vector designed to capture spectral, cepstral, and temporal characteristics of speech. These handcrafted features aim to emphasize artifacts commonly introduced by speech synthesis and voice conversion, such as irregular spectral envelopes, phase inconsistencies, and abnormal temporal dynamics. In total, 135 handcrafted features are extracted from each audio sample. The feature set is summarized in Table 2.

Table 2
Handcrafted acoustic feature categories and dimensionality

Feature Category	Description	Count
MFCCs	13 coefficients with mean, standard deviation, minimum, and maximum	52
Delta MFCCs	First-order derivatives with statistical descriptors	26
Mel-spectrogram	Global and per-band statistical features	44
Spectral	Centroid, bandwidth, roll-off, flatness (mean and std for each)	8
Temporal	Zero-crossing rate, RMS energy, frame energy (3)	5
Total		**135**

The extracted features are used as input to a range of classical ML classifiers to evaluate their effectiveness for synthetic speech detection. Both linear and non-linear models are considered. The evaluated classifiers and their main parameters are listed in Table 3.

Table 3
Classical machine learning classifiers and hyperparameters

Classifier	Parameters	Description
Random Forest	n_estimators = 100, max_depth = 20	Ensemble of decision trees
SVM (RBF)	kernel = rbf, class_weight = balanced	Non-linear support vector machine
SVM (Linear)	kernel = linear, class_weight = balanced	Linear support vector machine
Gradient Boosting	n_estimators = 100, max_depth = 5	Sequential boosting model
Logistic Regression	max_iter = 1000, class_weight = balanced	Linear probabilistic classifier
K-Nearest Neighbors	n_neighbors = 5	Distance-based classifier
Multilayer Perceptron	hidden layers = (256, 128)	Feedforward neural network

All classifiers are trained to perform binary classification, distinguishing between bona fide and synthetic audio samples. Hyperparameters are selected using the validation subset to ensure a fair and consistent comparison across methods.

Classical ML-based approaches offer several practical advantages, including low computational cost, fast training, and greater interpretability compared to deep neural networks. However, their performance is strongly dependent on feature selection and they are often sensitive to domain shift, particularly when evaluated on unseen synthesis methods.

8. Experimental Results and Analysis

This section presents the empirical evaluation of the implemented synthetic speech detection systems. The evaluation employs both primary and secondary performance metrics in order to provide a comprehensive assessment of detection accuracy and robustness.

The primary evaluation metric used in this study is the Equal Error Rate (EER), which is widely adopted in biometric authentication and spoofing detection research. EER represents the operating point at which the false acceptance rate equals the false rejection rate. Lower EER values indicate better detection performance and a more favorable trade-off between accepting synthetic speech and rejecting bona fide speech. Importantly, EER enables consistent comparison across different detection models without requiring the selection of a fixed decision threshold.

In addition to EER, several secondary metrics are reported to capture complementary aspects of classification performance. Accuracy reflects overall classification correctness, while precision and recall characterize the detector's behavior with respect to synthetic speech. The F1-score summarizes the balance between precision and recall, and the area under the ROC curve (AUC–ROC) provides a threshold-independent measure of class separability.

Together, these secondary metrics support a more detailed interpretation of model behavior and help identify potential sensitivity to class imbalance or threshold-dependent performance variations.

Table 4
Performance of classical machine learning classifiers on the evaluation dataset

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)	EER (%)
Logistic Regression	91.11	97.74	92.13	94.85	95.32	11.13
SVM RBF	95.22	97.73	96.86	97.30	97.54	7.38
SVM Linear	91.67	98.14	92.38	95.17	95.02	12.00
Random Forest	91.89	91.73	99.86	95.63	95.70	12.00
Gradient Boosting	94.00	93.78	99.86	96.73	95.36	12.00
KNN	94.67	94.76	99.50	97.97	93.69	10.81
MLP	93.89	94.50	98.86	96.64	94.49	12.31

Table 4 summarizes the performance of the evaluated classical machine learning classifiers for synthetic speech detection. Overall, the results indicate that classical approaches can achieve strong performance under controlled conditions, with several models demonstrating high discrimination capability between bona fide and synthetic speech.

Among all evaluated classifiers, the SVM with RBF kernel achieves the best overall performance. It attains the highest accuracy and AUC values while also yielding the lowest Equal Error Rate (EER), indicating a favorable balance between false acceptances and false rejections. This suggests that non-linear decision boundaries are effective in separating synthetic and real speech in the handcrafted feature space.

The KNN classifier also performs competitively, showing high accuracy and F1-score together with a relatively low EER. Its strong recall indicates effective detection of synthetic samples, although its distance-based nature may limit scalability to larger datasets.

Linear models, including logistic regression and linear SVM, exhibit stable but slightly inferior performance compared to their non-linear counterparts. While these models maintain high precision and recall, their higher EER values suggest reduced robustness at the decision boundary, likely due to the limited expressiveness of linear separators.

Tree-based ensemble methods, such as random forest and gradient boosting, achieve perfect or near-perfect recall, indicating strong sensitivity to synthetic speech. However, this behavior is accompanied by lower precision and higher EER, suggesting a tendency to over-predict the synthetic class and produce more false positives.

The MLP classifier demonstrates competitive recall but yields the highest EER among the evaluated methods, indicating less stable threshold behavior compared to other classifiers. This suggests that shallow neural architectures may not fully exploit the handcrafted feature representation in this setting.

Overall, these results show that while classical machine learning classifiers can achieve high accuracy and AUC values, their performance varies notably in terms of threshold stability and error trade-offs. The SVM with RBF kernel emerges as the most reliable classical baseline, providing a strong reference point for comparison with more advanced deep learning–based detection approaches.

9. Classifiers Performance by Vendor

This section analyzes the detection performance of classical machine learning classifiers across individual speech synthesis vendors, providing insight into vendor-dependent behavior and generalization characteristics. Figure 4 presents classification accuracy for all evaluated models, broken down by bona fide speech and by synthetic speech generated using ElevenLabs, RVC, Tortoise, and XTTS.

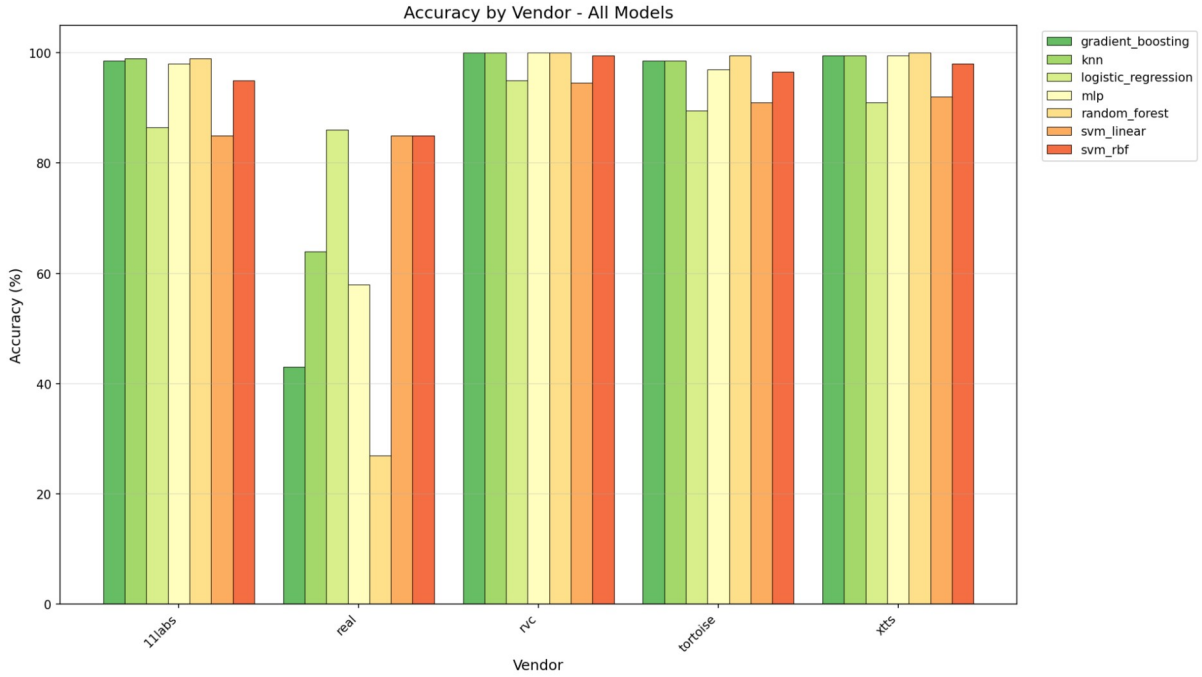

Figure 4: Classification accuracy of classical models by speech synthesis vendor

Overall, the results indicate that synthetic speech generated by neural TTS and voice conversion systems is detected with consistently high accuracy across most classifiers. For RVC, Tortoise, and XTTS, the majority of models achieve accuracy values close to or equal to 100%, suggesting that these synthesis methods introduce stable and detectable artifacts within the handcrafted feature space. In particular, non-linear classifiers such as SVM with RBF kernel, Random Forest, and Gradient Boosting demonstrate uniformly strong performance across all synthetic vendors.

In contrast, performance on bona fide (real) speech exhibits substantially higher variability across classifiers. While SVM-based models maintain relatively high accuracy on real samples, several classifiers—most notably Random Forest, Gradient Boosting, and MLP—show reduced accuracy in distinguishing real speech from synthetic samples. This behavior indicates a tendency toward over-detection of synthetic speech, consistent with the high recall values observed for these models. Such bias may lead to increased false rejection of genuine speech in practical deployment scenarios.

Vendor-specific differences are most pronounced for ElevenLabs-generated speech, where certain classifiers exhibit slightly lower accuracy compared to other synthetic sources. This observation suggests that commercial, high-quality neural TTS systems may produce fewer detectable artifacts than open or research-oriented synthesis frameworks. Nevertheless, even in this case, most models achieve accuracy above 95%, indicating effective detection under the evaluated conditions.

The KNN and SVM (RBF) classifiers demonstrate the most stable behavior across vendors, maintaining high accuracy for both synthetic and bona fide speech. In contrast, linear models and tree-based ensembles show stronger vendor sensitivity, particularly in the real speech category.

In summary, the vendor-wise analysis highlights that classical machine learning detectors can effectively identify synthetic speech produced by multiple modern TTS and voice cloning systems. However, the observed variability on real speech emphasizes the importance of evaluating detection systems not only on synthetic accuracy but also on their ability to preserve bona fide speech acceptance. These findings further motivate the use of balanced evaluation metrics and vendor-aware analysis when assessing synthetic speech detection systems.

Conclusions

This paper presented a systematic evaluation of classical machine learning approaches for synthetic speech detection using handcrafted acoustic features. A controlled dataset comprising bona fide speech and synthetic audio generated by four modern text-to-speech and voice conversion systems—RVC, XTTS, ElevenLabs, and Tortoise—was constructed to assess detection performance under consistent experimental conditions.

Experimental results demonstrate that classical classifiers can achieve strong performance when discriminating between genuine and synthetic speech in an in-domain setting. In particular, non-linear models such as SVM with an RBF kernel and K-nearest neighbors consistently outperformed linear classifiers, achieving low Equal Error Rates and high AUC values. Vendor-wise analysis revealed that while open and research-oriented synthesis systems introduce clearly detectable artifacts, high-quality commercial neural TTS systems produce speech that is more challenging to detect, though still distinguishable under the evaluated conditions.

At the same time, the results highlight important limitations of classical detection approaches. Several classifiers exhibit a tendency toward over-detection of synthetic speech, resulting in increased false rejection of bona fide samples. This behavior underscores the importance of threshold-aware evaluation and careful operating point selection when deploying spoofing countermeasures in real-world biometric systems.

Overall, the findings confirm that classical machine learning methods remain valuable as computationally efficient and interpretable baselines for synthetic speech detection. However, their sensitivity to feature design and potential lack of robustness under domain shift motivate further investigation into hybrid and deep learning–based approaches [19, 20].

Future work will focus on cross-dataset and cross-language generalization, integration of self-supervised speech representations, and direct comparison with end-to-end neural detectors under realistic acoustic and transmission conditions.

Declaration on Generative AI

While preparing this work, the authors used the AI programs Grammarly Pro to correct text grammar and Strike Plagiarism to search for possible plagiarism. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] Y. Wang, et al.: A survey on audio deepfake detection. ACM Computing Surveys, 56, 2024.

[2] Y. Mirsky, W. Lee, The creation and detection of deepfakes: A survey. ACM Computing Surveys, 54(1), 2021, 1–41, doi:10.1145/3425780

[3] T. Kinnunen, et al.: ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. arXiv preprint, 2021, arXiv:2109.00537

[4] M. Sahidullah, T. Kinnunen, C. Hanilçi, A comparison of features for synthetic speech detection. In: INTERSPEECH 2015, 2087–2091, doi:10.21437/Interspeech.2015-478

[5] H. Tak, et al.: End-to-end neural spoofing countermeasures for speaker verification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 2022, 1044–1057.

[6] ADD Challenge Organizers: The audio deepfake detection challenge 2023. In: CEUR Workshop Proceedings, 2023.
[7] V. Brydinskyi, et al. Enhancing automatic speech recognition with personalized models: Improving accuracy through individualized fine-tuning. IEEE Access, 12, 2024, 116649–116656, doi:10.1109/ACCESS.2024.3443811
[8] I. Zaiets, et al. Integrated system for speaker diarization and intruder detection using speaker embeddings. In: Proceedings of the CEUR Workshop Proceedings, 3654, 2024, 228–238.
[9] G. Lavrentyeva, et al.: Audio replay attack detection with deep learning frameworks. In: Proceedings of INTERSPEECH 2017, 82–86.
[10] H. Tak, et al.: End-to-end anti-spoofing with RawNet2. In: Proceedings of ICASSP 2021, 6369–6373.
[11] K. Ruda, et al. Scalability analysis of neural network architectures for voice biometric authentication. In: Cybersecurity Providing in Information and Telecommunication Systems, 2025, 349–357.
[12] O. Iosifova, et al., Techniques Comparison for Natural Language Processing, in: 2nd International Workshop on Modern Machine Learning Technologies and Data Science, vol. 2631, no. I, 2020, 57–67.
[13] I. Iosifov, et al., Transferability Evaluation of Speech Emotion Recognition Between Different Languages, Advances in Computer Science for Engineering and Education, 134, 2022, 413–426. doi:10.1007/978-3-031-04812-8_35
[14] O. Romanovskyi, et al., Accuracy improvement of spoken language identification system for close-related languages, Advances in Computer Science for Engineering and Education VII, vol. 242 (2025) 35–52. doi:10.1007/978-3-031-84228-3_4
[15] I. Iosifov, O. Iosifova, V. Sokolov, Sentence Segmentation from Unformatted Text using Language Modeling and Sequence Labeling Approaches, in: VII International Scientific and Practical Conference Problems of Infocommunications. Science and Technology (2020) 335–337. doi:10.1109/PICST51311.2020.9468084
[16] C. Sun, Synthetic voice detection – vocoder artifacts. GitHub repository (2025). https://github.com/csun22/Synthetic-Voice-Detection-Vocoder-Artifacts
[17] V. Poberezhnyk, V. Balatska, I. Opirskyy. Development of the Learning Management System Concept based on Blockchain Technology. CEUR Workshop Proceedings, 3550, 2023, 143–156.
[18] J. Kim, et al. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: Proceedings of ICML 2021 (2021).
[19] Betker, J.: TorToiSe: Text-to-speech. GitHub repository, 2022. https://github.com/neonbjb/tortoise-tts
[20] Tak, H., et al.: End-to-end anti-spoofing with RawNet2. In: Proceedings of ICASSP 2021, 6369–6373.