Leveraging Self-Supervised Learning as Features in Audio Deepfake Detection

*Note: Sub-titles are not captured in Xplore and should not be used

***Abstract*—This paper investigates the relative effectiveness of traditional acoustic features and self-supervised learning (SSL) embeddings for audio deepfake detection on the Fake-or-Real (FoR) corpus. We evaluate nine feature sets — MFCC, CQCC, RMS, ZCR, Teager, Wav2Vec2, HuBERT, WavLM, and Whisper, using Accuracy, F1-score, and Equal Error Rate (EER) as metrics. At the family level, Handcrafted and SSL features exhibit very similar distributions; quartiles and Mann–Whitney U tests reveal no statistically significant differences in Accuracy, F1, or EER ($p > 0.90$), indicating that SSL embeddings do not globally outperform handcrafted representations.**

At the feature level, however, clear patterns emerge, Whisper and MFCC are the most reliable features, with median accuracies of 0.9602 and 0.9571 and median EER of 0.0375 and 0.0391, respectively, without significant differences between them. WavLM forms a second tier with competitive but weaker results. Overall, the results show that cepstral features such as MFCC and CQCC remain robust options and that SSL embeddings, particularly Whisper and WavLM, complement rather than replace well-designed handcrafted features in audio deepfake detection.

***Index Terms*—audio deepfake detection, self-supervised learning, deep embeddings, handcrafted acoustic features, Fake-or-Real dataset**

I. Introduction

The recent growth of techniques for generating and synthesizing counterfeit versions of legitimate audio, has become a reality in contemporary society. These developments have raised significant concerns regarding security, institutional trust, and public credibility. Reports such as the one published by Reuters [25] highlight that compelling synthetic media, including cloned speech, pose a threat to both electoral integrity and financial stability. Major news outlets have repeatedly highlighted the problem, underscoring the urgency of developing robust countermeasures. This concern is echoed in the academic community, where interest in efficient detection techniques has been steadily growing. [29]. These concerns illustrate that the challenge is not merely technical but also deeply social, requiring solutions that can counter malicious manipulation and exploitation of legitimate audio content.

Motivated by this scenario, we investigate how different feature representations perform in detecting audio deepfakes. Our study is conducted on the Fake-or-Real (FoR) dataset [24], which combines genuine speech with synthetic audio generated by state-of-the-art text-to-speech (TTS) systems. In addition to offering a large-scale, balanced corpus suitable for robust training and testing, FoR includes audio synthesized by multiple TTS platforms, which introduces variability and creates more realistic, challenging experimental conditions compared to single-system corpora.

Regarding the feature representations, we focus on two sets of features, which are acoustic features handcrafted, such as Mel-Frequency Cepstral Coefficients (MFCC), Root Mean Square (RMS), Zero Cross Rate (ZCR), Constant-Q Cepstral Coefficients (CQCC), and Teager Energy—valued for their interpretability and low computational cost [28]. Second group *deep embedding representations* extracted from Self-Supervised Learning (SSL) models, namely **HuBERT**, **Wav2Vec2**, **WavLM**, and **Whisper**, recognized for capturing high-level and long-term speech characteristics [18].

To ensure a consistent evaluation, we apply the same experimental protocol and employ classifiers with complementary properties. Logistic Regression (LR) and Linear Support Vector Machine (SVM) provide robust linear baselines. Decision Trees (DT) offer interpretable non-linear models. Random Forests (RF) enhance robustness via ensemble learning, and Naïve Bayes (NB) delivers a lightweight probabilistic reference, often competitive in speech-related tasks.

To evaluate the experimental performance, we used Accuracy, F1-score, and Equal Error Rate (EER). Accuracy provides an overall measure of correctness; however, it is limited in the presence of imbalanced datasets and serves mainly as a baseline. The F1-score balances precision and recall [21], whereas the EER is the standard metric in biometric anti-spoofing systems, enabling fair comparison across systems regardless of the chosen operating threshold [27].

The main objective of this work is to compare handcrafted acoustic features with SSL features for audio deepfake detection. By using the same dataset, classifiers, and evaluation metrics, we ensure a transparent, controlled analysis that highlights the strengths and weaknesses of each approach. To achieve this overarching goal, we formulate the following research questions: **(i)** Do deep embeddings outperform traditional acoustic features (e.g., MFCC, CQCC, ZCR) in terms of consistency and robustness? **(ii)** Which classifiers achieve the best trade-off across Accuracy, F1-score, and EER? **(iii)** Among the evaluated feature sets, which emerge as the most reliable and effective for detection tasks? **(iv)** Is there evidence of certain methods prevailing regardless of the scenario? **(v)** Is it possible to claim that SSL outperforms Handcraft?

The remainder of this paper is organized as follows: Section II presents the theoretical background on voice spoofing

Identify applicable funding agency here. If none, delete this.

detection and feature representations. Section III details the Fake-or-Real (FoR) dataset, preprocessing steps, and experimental setup. Section IV reports and analyzes the comparative results between traditional features and embeddings across classifiers. Finally, Section V summarizes the main findings, and outlines directions for future work.

II. Background

Audio deepfakes, generated by state-of-the-art text-to-speech (TTS) and voice-conversion systems, pose a rapidly growing threat to contemporary society, increasing the risks of fraud, disinformation, and identity misuse. Although these technologies enable beneficial practical applications, humans are generally unable to distinguish original speech from synthesized speech reliably. When employed maliciously, this capability necessitates the development of automated countermeasures [20]. Audio deepfake detection remains technically challenging due to limited generalization to previously unseen synthesis methods, variability in channel and codec conditions, and the persistent gap between controlled laboratory settings and real-world scenarios [16].

A. Traditional Acoustic Features for Spoofing Detection

Handcrafted features have long served as baselines in spoofing detection, valued for their interpretability and efficiency. MFCC and their temporal derivatives (Δ, $\Delta\Delta$) capture spectral envelopes and dynamic variations of speech [5, 7]. RMS energy ZCR provides simple yet effective descriptors of periodicity and energy fluctuations [23]. CQCC, proposed as an alternative based on the constant-Q transform, became a firm baseline in the ASVspoof challenges [26]. Finally, the Teager Energy Operator highlights micro-modulations often smoothed in synthetic signals, adding complementary cues [12]. These features remain relevant as lightweight and interpretable benchmarks.

B. Self-Supervised Learning for Deep Embedding Extraction

Self-supervised learning (SSL) has changed how speech representations are obtained, making it possible to learn general-purpose embeddings without manual annotation while still encoding fine-grained phonetic information and longer-term temporal structure [20]. In this context, Wav2Vec2.0 adopts a contrastive objective over quantized latent units and has proved effective in a wide range of speech tasks [1]. HuBERT, unlike contrastive learning, predicts masked hidden units and tends to yield more detailed phonetic and prosodic cues [10]. WavLM further incorporates denoising and multitask training, which increases robustness to background noise and speaker variability [3]. Whisper differs from these models by being trained on large-scale multilingual data, producing embeddings that transfer reasonably well across domains and have already been explored in spoofing detection scenarios [22]. Overall, compared with traditional handcrafted features, these embeddings usually provide higher-level abstractions and better generalization to previously unseen attacks. However, their performance still depends on the alignment between training and deployment conditions.

C. Datasets for Deepfake Detection

Benchmark datasets are essential for progress in spoofing detection. The ASVspoof series (2015–2021) established standard corpora mixing genuine and synthetic speech, becoming the reference for countermeasure evaluation [27, 16]. Other corpora include FakeAVCeleb, which targets multimodal audio-visual deepfakes [13], and WaveFake, focusing on generative models such as WaveGlow and MelGAN [2]. This work utilizes the FoR dataset, which comprises over 195,000 utterances from diverse speakers, recording conditions, and TTS systems, providing both scale and variety [24]. Unlike ASVspoof, FoR emphasizes diversity over competition design, making it well-suited for comparative studies of features and embeddings.

D. Machine Learning Approaches

Classical machine learning algorithms provide strong and interpretable baselines for spoofing detection. Logistic Regression offers linear probabilistic modeling, while Support Vector Machines (SVM) are effective in high-dimensional feature spaces such as cepstral coefficients and embeddings. Decision Trees model non-linear patterns with rule-based interpretability, and Random Forests extend them via ensembles for robustness. Naïve Bayes, despite its independence assumption, remains competitive in lightweight speech classification tasks [17]. These models enable systematic assessment of features and embeddings under complementary paradigms.

E. Evaluation Metrics

A comprehensive evaluation of audio deepfake detection requires multiple complementary metrics. The most common is **Accuracy**, which measures the proportion of correctly classified samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. While widely used, Accuracy can be misleading in imbalanced datasets.

Capturing the proportion of actual positives correctly identified. Their harmonic mean yields the **F1-score**, which balances both aspects:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2}$$

These metrics provide a more robust view of system performance in imbalanced conditions .

Additionally, the **EER** is widely adopted in spoofing and biometric verification. It is defined as the operating point where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR):

$$\text{EER} = \text{FAR}(\tau^*) = \text{FRR}(\tau^*), \tag{3}$$

Where τ^* denotes the threshold at which both error rates are equal. EER has become the standard evaluation metric

in the ASVspoof challenges and remains the most relevant benchmark for comparing countermeasures [27].

F. Mann–Whitney U Test

To compare the performance distributions of feature families (SSL vs. Handcraft) and of individual representations, the non-parametric Mann–Whitney U test (also known as the Wilcoxon rank-sum test) was used. This test is particularly appropriate for machine-learning evaluations because it requires no assumption of normality, is robust to outliers, and is the most powerful rank-based method for detecting stochastic dominance between two independent samples [4].

All tests were two-sided, with the null hypothesis that the two samples come from the same distribution. The significance level was set at $\alpha = 0.05$.

G. Related Work

Early works relied on handcrafted features such as MFCC and CQCC combined with Gaussian Mixture Models (GMM), which proved effective in ASVspoof baselines [26, 27]. More recent studies highlight the advantages of SSL deep embeddings (e.g., Wav2Vec2.0, HuBERT, WavLM) in improving robustness to unseen spoofing methods and channel variability [10, 3, 20]. Whisper embeddings have also been shown to transfer effectively to cross-lingual and noisy scenarios [22]. Hybrid approaches, which combine cepstral features with deep embeddings, consistently report gains in cross-corpus evaluation [19]. However, systematic comparisons between handcrafted features and embeddings under identical setups with classical classifiers remain limited. This gap motivates the present work, which provides a unified and controlled evaluation of both paradigms.

A summary of representative studies, covering handcrafted approaches, SSL deep embeddings, multimodal datasets, and hybrid strategies, is provided in Table I. This overview illustrates the evolution of the field from cepstral baselines to self-supervised learning and fusion approaches, reinforcing the relevance of our comparative analysis.

III. Methodology

We adopt an experimental setup that enables systematic, reproducible evaluation of audio deepfake detection using both handcrafted features and deep embeddings. The process is divided into four stages: audio preprocessing, feature extraction, experimental configuration, and repeated sampling with evaluation.

A. Audio Preprocessing

To ensure consistency across samples and reduce variability unrelated to spoofing artifacts, all audio signals were normalized to a fixed duration of 4 seconds (64,000 samples at 16 kHz). Segments longer than this duration were centrally cropped, since peripheral regions often contain silence or background noise instead of informative speech. This preprocessing step is consistent with prior work that emphasizes temporal normalization in spoofing detection pipelines [27].

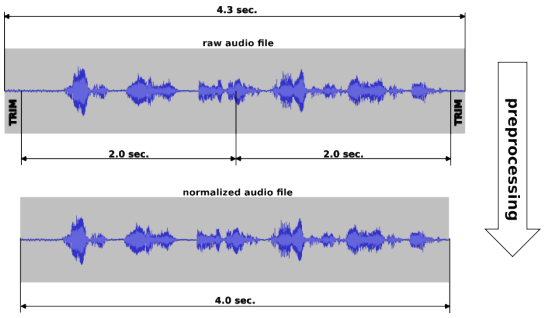

Fig. 1: Illustration of the preprocessing step: raw audio is normalized to 4 seconds by trimming central segments or applying zero-padding.

For utterances shorter than 4 seconds, symmetric zero-padding was applied to match the target length, avoiding information loss while maintaining fixed-size input representations. The resulting standardized signals make comparisons across different features and classifiers more consistent and simplify batch processing during feature extraction. Figure 1 illustrates the normalization procedure, in which the raw audio is trimmed and aligned to produce a uniform 4-second signal.

B. Feature Extraction

Handcrafted features such as MFCC, CQCC, RMS, and ZCR are traditionally extracted in a short-term manner, using windowing and overlap to capture local dynamics of the speech signal. As a consequence, these features naturally yield variable-length matrices whose dimensions depend on the duration of each utterance, which is inconvenient for classical classifiers that expect fixed-size input vectors.

A standard solution in the literature is to apply simple pooling strategies (e.g., mean, maximum, or summation across frames) to obtain a fixed-length representation. However, these pooling strategies can oversimplify the temporal dynamics of the signal and may reduce the energetic variations relevant to spoofing detection [11].

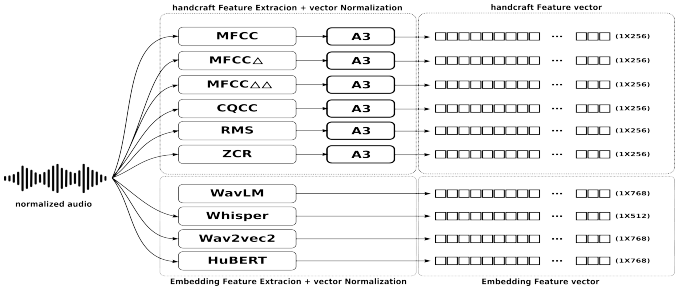

Fig. 2: Overview of the feature extraction pipeline. Handcrafted short-term features are normalized using the A3 energy-based approach, while embeddings are directly aggregated into fixed-length representations.

We therefore adopt the A3 energy-based normalization method [8], which aggregates frame-level descriptors into

TABLE I: Summary of representative studies on audio deepfake detection.

Reference	Dataset	Features / Embeddings	Main Contribution
[26]	ASVspoof 2017	CQCC + GMM	Introduced Constant-Q Cepstral Coefficients as a baseline, widely adopted for spoofing detection.
[27]	ASVspoof 2019	CQCC, LFCC, MFCC + GMM/SVM	Large-scale benchmark; confirmed handcrafted features effective but limited against unseen attacks.
[15]	ASVspoof 2021	CQCC + classical ML	Overview and analysis of ASVspoof 2021, highlighting generalization challenges across attack types and codecs.
[13]	FakeAVCeleb	Audio-visual embeddings	First multimodal dataset (audio + video) for deepfake detection, stressing cross-modal vulnerabilities.
[2]	WaveFake	Spectrograms + CNNs	Proposed evaluation with modern neural vocoders (WaveGlow, MelGAN), exposing weaknesses of feature-only methods.
[24]	FoR	MFCC, CQCC, embeddings	Large-scale corpus with 195k utterances; emphasized diversity and realism beyond challenge datasets.
[10]	LibriSpeech, LibriLight	HuBERT embeddings	Introduced hidden-unit prediction, achieving robust phonetic/prosodic SSL embeddings transferable to spoofing detection.
[1]	LibriSpeech	Wav2Vec2 embeddings	Pioneered contrastive SSL for speech, later applied in spoofing tasks with significant generalization gains.
[3]	Multiple corpora	WavLM embeddings	Enhanced robustness to noise and unseen conditions using denoising and multitask SSL objectives.
[19]	ASVspoof + In-The-Wild	MFCC + LFCC + Spectrogram	Showed feature fusion improves robustness and cross-corpus performance.
[6]	FoR	Spectro-Temporal Graph Attention Network	The best result in terms of EER was obtained by Spectro-Temporal Graph Attention Network (GAT) in an end-to-end approach, achieving an EER of 0.70%.

fixed-length vectors weighted by the signal's energy distribution. This preserves both spectral information and the overall energetic balance across frames, while producing comparable representations for all samples regardless of their original duration. In this way, handcrafted and embedding-based features can be evaluated under identical conditions within a controlled framework.

C. Experimental Setup

We use the FoR corpus and, specifically, its *original* subset as our evaluation benchmark [24]. The *original* release preserves the raw signal characteristics (amplitude scale, channel/coloration, and synthesis artifacts) without additional re-recording or amplitude normalization. This avoids extra confounding factors and makes it easier to attribute performance differences to the representations under study (handcrafted features vs. embeddings). It also follows the common practice of first evaluating countermeasures on source-domain signals before introducing channel shifts.

In the original subset of the FoR corpus, there is a substantial imbalance between classes: 33,600 genuine utterances versus 6,156 synthetic ones. This distribution tends to bias learning towards the majority class, inflating accuracy while degrading the detection of synthetic audio, which is the main target in spoofing detection. To mitigate this, we undersample the genuine class to match the number of fake samples. Thus, both classes contribute equally to training, following standard practice in imbalanced learning [9].

To obtain reliable estimates while controlling class prevalence (*real* vs. *fake*), we adopt a repeated, stratified holdout protocol [14]. For each of R=20 repetitions $r = 1, \ldots, R$:

1) We draw a fresh, **stratified** random subsample containing **90%** of the FoR-*original* utterances, preserving the real/fake proportion of the whole subset.
2) Within this 90% subsample, we perform a **stratified 70/30 holdout** split to obtain training and test sets, respectively. The remaining 10% of utterances from the full subset are not used in repetition r, ensuring independence across repeats.
3) Preprocessing and feature extraction are applied exactly as specified (4 s normalization; handcrafted short-term features aggregated to the utterance level via energy-based A3 normalization; utterance-level SSL embeddings).
4) A classifier is trained on the 70% training portion and evaluated on the 30% test portion for that repetition.

For clarity, class balancing and resampling are handled as follows. In each repetition r, the stratified 70% training split is *first* obtained and *then* balanced by random undersampling of genuine and spoofed utterance matches. The test set always preserves the original class distribution of the FoR corpus, ensuring an unbiased evaluation of the classifiers. We do not enforce disjointness of utterances across different repetitions: the independence we refer to is between training and test sets *within* each repetition, whereas repetitions themselves are treated as standard resampling of the corpus to obtain more stable performance estimates [14].

For each $(\text{dataset}, \text{feature type}, \text{classifier})$ tuple and each repetition r, we record the fundamental evaluation metrics: **Accuracy**, which reflects the overall proportion of correct classifications; the **F1-score**, the harmonic mean between precision and recall; Furthermore, the **EER**, a threshold-independent measure widely adopted in the anti-spoofing literature. From R=20 repetitions, we compute summary statistics for each metric. Finally, we apply pairwise Wilcoxon (Mann–Whitney U) tests both between feature families (handcrafted vs. SSL) and between individual feature representations to assess the statistical significance of performance differences.

D. Materials

For signal processing and feature extraction, we used `librosa` and `scipy`. Classical machine learning models

(Logistic Regression, SVM, Decision Tree, Random Forest, and Naïve Bayes) were implemented using `scikit-learn`. Embedding-based representations (WavLM, Whisper, Wav2Vec2, and HuBERT) were extracted using the `transformers` library from Hugging Face, which provides pretrained models through its model hub. Data manipulation and analysis were supported by `numpy` and `pandas`, and result visualization used `matplotlib` and `seaborn`. Because embedding extraction is computationally intensive, we used an NVIDIA RTX 3060 GPU to accelerate processing of the large audio set.

IV. Results

This section addresses the research questions **(i)–(v)** outlined in the introduction by analyzing the comparative performance of feature sets and classifiers across multiple evaluation metrics.

As illustrated by the boxplots in Fig. 3, among the best-performing methods, we highlight the handcrafted feature *MFCC* and the SSL embedding *Whisper*, whose distributions of *Accuracy*, F1, and EER are visibly close and have narrow boxes, indicating more stable results across repetitions. The same boxplots also show that *WavLM* and *CQCC* form a second group, with slightly lower performance but still concentrated in relatively narrow ranges.

Looking in more detail, the quartiles reported in Table II confirm this observation. The medians of *accuracy* are 0.9571 for *MFCC* and 0.9602 for *Whisper*, while the medians of F1 are 0.9580 (*MFCC*) and 0.9602 (*Whisper*). The median EER values are also close: 0.0391 for *MFCC* and 0.0375 for *Whisper*, both associated with narrow interquartile ranges. In the second group, *WavLM* shows medians of 0.9219 (Accuracy), 0.9234 (F1), and 0.0720 (EER), whereas *CQCC* reaches 0.9165 (Accuracy), 0.9194 (F1), and 0.0796 (EER), again with moderate variability.

On the less performant side, Fig. 3 and Table II reveal weaker methods in each family. Among handcrafted features, *RMS*, *ZCR*, and *Teager* exhibit median *Accuracy* values of 0.8721, 0.8691, and 0.8530, respectively, and median EER values of 0.1194, 0.1210, and 0.1516. Among embeddings, *Wav2Vec2* attains a median *Accuracy* of 0.8247 and a median EER of 0.1746, with higher dispersion. These results indicate that it is not appropriate to claim that "all deep embeddings outperform all traditional features."

In answer to research question RQ (i), we conclude that SSL embeddings do not exhibit uniform superiority over traditional acoustic features. Some SSL methods (in particular *Whisper* and, to a lesser extent, *WavLM*) tend to achieve slightly better metrics on average, but *MFCC* — and, to a lesser extent, *CQCC* — display comparable performance and stability. Both groups include descriptors with weak performance and higher variability. Therefore, in the evaluated scenario, SSL embeddings offer a slight average advantage, while well-designed cepstral features remain robust and consistent options.

Regarding the classifiers, the boxplots in Fig. 4 and the quartiles indicate that *Random Forest* achieves the best trade-off between *Accuracy*, F1, and EER: its medians are 0.9533 (Accuracy), 0.9525 (F1), and 0.0475 (EER), with a relatively narrow interquartile range. *Decision Tree* appears as the second-best option, with medians of 0.9135 (Accuracy), 0.9148 (F1), and 0.0865 (EER). *Logistic Regression* and *SVM* occupy an intermediate position, whereas *Naïve Bayes* yields the worst medians (0.8783 accuracy, 0.8859 F1, and 0.1187 EER), thus providing the weakest compromise among the metrics. These observations answer research question RQ (ii) by showing that Random Forest attains the most favorable balance across the three evaluation criteria in our setting.

In response to research question RQ (iii), the answer is constructed from three main pieces of evidence: (a) the quartiles of *Accuracy*, F1-score, and EER for each feature set, shown in Table II; (b) the boxplots per feature set in Fig. 3; and (c) the average *ranking* of *Accuracy* and EER obtained with five classifiers (Decision Tree — DT, Logistic Regression — LR, Naïve Bayes — NB, Random Forest — RF, and Support Vector Machine — SVM), reported in Tables III and IV. The joint analysis of these artefacts shows that three feature sets stand out simultaneously in terms of higher effectiveness, expressed by more favourable central values, and higher reliability, expressed by lower variability and better average position across classifiers:

1) **Whisper**, which attains the best medians in *Accuracy* equal to 0.9602, F1-score equal to 0.9602, and EER equal to 0.0375. It also achieves the lowest average *ranking* in EER (1.8) and the second-best average *ranking* in *Accuracy* (2.0), while maintaining high performance across all five classifiers.
2) **MFCC**, which yields the second-best medians, with *Accuracy* equal to 0.9571, F1-score equal to 0.9580, and EER equal to 0.0391. It presents the smallest interquartile range for EER and an average *ranking* of 1.8 in both *Accuracy* and EER, indicating high stability across classifiers.
3) **WavLM**, the best among the self-supervised embedding models, with *Accuracy* equal to 0.9219 and EER equal to 0.0720. It exhibits a relatively compact distribution and average *rankings* of 4.6 in *Accuracy* and 5.0 in EER, consistently outperforming the other SSL methods HuBERT and Wav2Vec2.

The remaining feature sets, namely CQCC, HuBERT, RMS, ZCR, Teager, and Wav2Vec2, occupy lower positions in the average *rankings* and/or exhibit greater dispersion across at least two metrics, which limits their reliability for general use under the evaluated conditions.

Thus, in light of these results, we conclude that the most reliable and effective feature sets for the detection tasks analysed are, in this order, **Whisper**, **MFCC**, and **WavLM**.

In response to RQ (iv), we compared the feature families in Fig. 3 , which do not indicate an apparent prevalence of either family: the medians of *Accuracy* are 0.9074 for *Handcraft* and 0.9081 for *SSL*, those of F1 are 0.9078 and 0.9106, respectively, and the median EER values are very similar (0.0926 for *Handcraft* and 0.0946 for *SSL*), with

TABLE II: Performance quartiles by feature set.

	Accuracy			F1-score			EER		
Feature	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
CQCC	0.9158	0.9165	0.9548	0.9184	0.9194	0.9549	0.0452	0.0796	0.0827
HuBERT	0.9035	0.9074	0.9089	0.9103	0.9103	0.9110	0.0911	0.0980	0.0995
MFCC	0.9464	0.9571	0.9587	0.9580	0.9580	0.9593	0.0337	0.0391	0.0536
MS	0.8706	0.8721	0.9265	0.8785	0.8785	0.9271	0.0735	0.1194	0.1248
TEAGER	0.8243	0.8530	0.9074	0.8625	0.8625	0.9078	0.0926	0.1516	0.1554
Wav2Vec2	0.8201	0.8247	0.8292	0.8297	0.8297	0.8352	0.1708	0.1746	0.1746
WavLM	0.9135	0.9219	0.9250	0.9234	0.9234	0.9265	0.0689	0.0720	0.0865
Whisper	0.9595	0.9602	0.9694	0.9602	0.9602	0.9696	0.0299	0.0375	0.0605
ZCR	0.8622	0.8691	0.8737	0.8756	0.8756	0.8796	0.1194	0.1210	0.1378

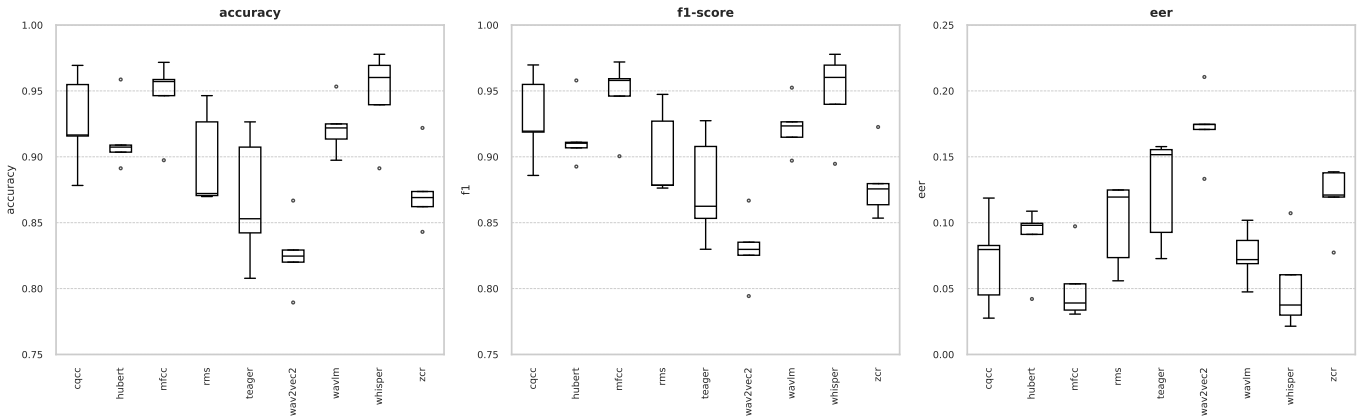

Fig. 3: Boxplots of feature performance across accuracy, F1-score, and EER.

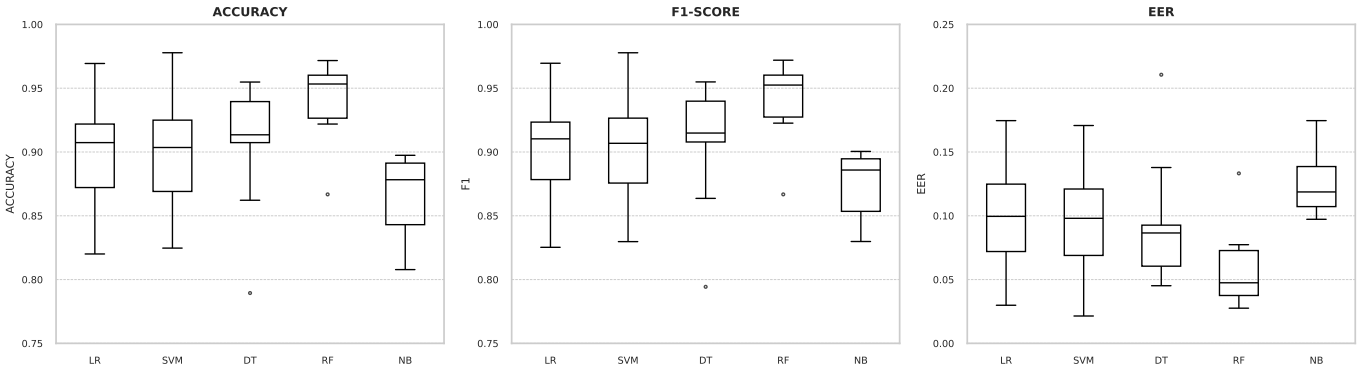

Fig. 4: Boxplots of classification performance across different classifiers considering Accuracy, F1-score, and EER.

comparable interquartile ranges. Thus, there is no support for claiming that one feature family dominates the other across all scenarios. When we examine specific methods, some patterns of prevalence become visible. From the classifier perspective, Fig. 4 shows that *Random Forest* attains the highest medians in *Accuracy* (0.9533) and F1 (0.9525) and the lowest median EER (0.0475), with moderate dispersion, whereas *Naïve Bayes* concentrates the worst medians in all three metrics. Among the feature sets, Fig. 3, Table II, and the average *rankings* in Tables III and IV indicate that *Whisper* and *MFCC* tend to occupy top positions for most classifiers, with median *Accuracy* of 0.9602 and 0.9571, F1 of 0.9602 and 0.9580, and EER of 0.0375 and 0.0391, respectively. Therefore, in addressing RQ (iv), we find evidence of prevalence only at the level of specific methods: *Random Forest* and the feature sets *Whisper* and *MFCC* recurrently appear among the best combinations of metrics and classifiers. By contrast, no single method or feature family performs best across all scenarios considered.

To examine whether SSL embeddings systematically outperform traditional Handcrafted features, the two families were compared on Accuracy, F1-score, and EER. The boxplots in Figure 3 and the corresponding quartiles reveal highly similar distributions: median Accuracy is 0.9074 for Handcraft and 0.9081 for SSL; median F1-score is 0.9078 versus 0.9106; and median EER is 0.0926 versus 0.0946, respectively. Across

TABLE III: Accuracy ranking by classifier performance in parentheses.

	DT	LR	NB	RF	SVM	Rank
cqcc	0.9663 (2)	0.9449 (3)	0.9051 (5)	0.9893 (1)	0.9433 (3)	2.8
hubert	0.9479 (5)	0.9342 (5)	0.9311 (1)	0.9755 (4)	0.9357 (5)	4.0
mfcc	0.9694 (1)	0.9786 (2)	0.9158 (3)	0.9893 (1)	0.9832 (2)	1.8
rms	0.9525 (4)	0.8928 (7)	0.8913 (6)	0.9663 (5)	0.8913 (6)	5.6
teager	0.9250 (7)	0.8652 (8)	0.8208 (9)	0.9541 (7)	0.8515 (8)	7.8
wav2vec2	0.8132 (9)	0.8515 (9)	0.8484 (8)	0.9051 (9)	0.8469 (9)	8.8
wavlm	0.9311 (6)	0.9418 (4)	0.9112 (4)	0.9663 (5)	0.9418 (4)	4.6
whisper	0.9556 (3)	0.9877 (1)	0.9265 (2)	0.9801 (3)	0.9893 (1)	2.0
zcr	0.8974 (8)	0.8959 (6)	0.8591 (7)	0.9387 (8)	0.8882 (7)	7.2

TABLE IV: EER ranking by classifier performance in parentheses.

	DT	LR	NB	RF	SVM	Rank
cqcc	0.0337 (2)	0.0551 (3)	0.0964 (5)	0.0184 (2)	0.0521 (3)	3.0
hubert	0.0521 (5)	0.0551 (3)	0.072 (1)	0.0245 (4)	0.0613 (5)	3.6
mfcc	0.0306 (1)	0.0214 (2)	0.0812 (3)	0.0123 (1)	0.0214 (2)	1.8
rms	0.0475 (4)	0.1149 (7)	0.1133 (6)	0.0337 (5)	0.1194 (7)	5.8
teager	0.0751 (7)	0.1378 (8)	0.1654 (9)	0.0490 (7)	0.1470 (8)	7.8
wav2vec2	0.1868 (9)	0.1516 (9)	0.1501 (8)	0.0873 (9)	0.1501 (9)	8.8
wavlm	0.0689 (6)	0.0628 (5)	0.0934 (4)	0.0368 (6)	0.0551 (4)	5.0
whisper	0.0444 (3)	0.0092 (1)	0.0720 (1)	0.0214 (3)	0.0092 (1)	1.8
zcr	0.1026 (8)	0.0949 (6)	0.1363 (7)	0.0613 (8)	0.1011 (6)	7.0

all three metrics, the interquartile ranges exhibit substantial overlap and, as noted in RQ (1), a slight advantage.

Given this close correspondence, a non-parametric Mann–Whitney U test was applied to compare the two families of each experimental condition. The results were $U = 244.5$, $p = 0.9091$ for Accuracy; $U = 248.0$, $p = 0.9727$ for F1-score; and $U = 252.0$, $p = 0.9727$ for EER. In every case, the p-values greatly exceed the conventional significance level of $\alpha = 0.05$, providing no basis for rejecting the null hypothesis that the performance distributions of the two families are equivalent.

The same analysis was repeated for the top-performing representative of each family — MFCC and Whisper. Again, the medians are remarkably close (e.g., Accuracy 0.9571 vs. 0.9602; EER 0.0391 vs. 0.0375), and Mann–Whitney tests yielded $U = 11.0$, $p = 0.8413$ for both Accuracy and F1-score, and $U = 14.0$, $p = 0.8413$ for EER — again revealing no statistically significant difference.

Thus, although SSL-based representations show marginally higher medians on specific metrics, both descriptive statistics and formal statistical testing indicate that, under the present experimental conditions, there is no evidence to conclude that the SSL family consistently outperforms the handcrafted feature family.

V. Conclusion and Future Work

In conclusion, the results obtained under the experimental conditions of this study do not provide statistically significant evidence that self-supervised (SSL) representations outperform traditional Handcrafted features (Handcrafted) as a group. Mann–Whitney U tests returned p-values greater than 0.84 for Accuracy, F1-score, and EER, and the interquartile ranges of the two families show substantial overlap.

Nevertheless, specific observations merit consideration. The SSL group exhibited slightly lower performance variability, which may indicate potential advantages in settings not examined here, such as highly noisy recordings, cross-domain generalization, or severely limited training data. These possibilities remain to be evaluated in future work.

Among individual methods, Whisper recorded the highest median values across all three metrics (Accuracy 0.9602, F1-score 0.9602, and EER 0.0375). The latter figure is consistent with the best results currently reported in the literature for comparable controlled-environment audio detection tasks (e.g., [6]).

At the same time, MFCC and CQCC continue to offer viable performance with considerably lower computational requirements. Consequently, the choice between self-supervised embeddings and Handcrafted features remains context-dependent: resource-rich deployments can reasonably favor models such as Whisper, whereas environments constrained by memory, latency, or energy consumption are still well served by established Handcrafted representations.

References

[1] Alexei Baevski et al. “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 12449–12460.

[2] RC Barik et al. “A Novel and Intelligent Approach for Indian Locale Based Text-to-Speech Model by Hybridizing Wave Net and Wave Glow with Mel-Spectrogram Analysis”. In: *International Conference on Artificial Intelligence and Speech Technology*. Springer. 2023, pp. 365–378.

[3] Sanyuan Chen et al. “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. Vol. 30. 2022, pp. 825–839. DOI: 10.1109/TASLP.2022.3148528.

[4] William Jay Conover. *Practical nonparametric statistics*. john wiley & sons, 1999.

[5] S. B. Davis and P. Mermelstein. “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing*. Vol. 28. 4. 1980, pp. 357–366. DOI: 10.1109/TASSP.1980.1163420.

[6] Anton Firc, Kamil Malinka, and Petr Hanáček. “Evaluation framework for deepfake speech detection: a comparative study of state-of-the-art deepfake speech detectors”. In: *Cybersecurity* 8.1 (2025), p. 50.

[7] Sadaoki Furui. “Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34.1 (1986), pp. 52–59. DOI: 10.1109/TASSP.1986.1164788.

[8] Rodrigo Capobianco Guido. “A tutorial on signal energy and its applications”. In: *Neurocomputing* 179 (2016), pp. 264–282.

[9] Haibo He and Edwardo A. Garcia. “Learning from Imbalanced Data”. In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284.

[10] Wei-Ning Hsu et al. “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”. In: *Proceedings of Interspeech*. 2021, pp. 1633–1637. DOI: 10.21437/Interspeech.2021-259.

[11] Lianyu Hu et al. “Temporal lift pooling for continuous sign language recognition”. In: *European conference on computer vision*. Springer. 2022, pp. 511–527.

[12] Vivek Kandpal, Md Sahidullah, and Tomi Kinnunen. “The Effectiveness of the Teager Energy Operator in Detecting Synthetic Speech Attacks”. In: *Proceedings of Interspeech*. 2020, pp. 1713–1717. DOI: 10.21437/Interspeech.2020-2411.

[13] Hasam Khalid and Simon S. Woo. “FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset”. In: *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. 2021, pp. 1–6. DOI: 10.1109/ICME51207.2021.9428409.

[14] Ron Kohavi. “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection”. In: *Proc. IJCAI*. 1995, pp. 1137–1145.

[15] Xuechen Liu et al. “ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023). Panorama do desafio; DF usa EER como métrica. DOI: 10.1109/TASLP.2023.3285283.

[16] Xuechen Liu et al. “Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 2507–2522.

[17] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[18] Marco Matassoni, Seraphina Fong, and Alessio Brutti. “Speaker anonymization: Disentangling speaker features from pre-trained speech embeddings for voice conversion”. In: *Applied Sciences* 14.9 (2024), p. 3876.

[19] Sharmin Akter Momu et al. “A Comprehensive Approach to Deepfake Audio Detection: Using Feature Fusion and Deep Learning”. In: *2024 27th International Conference on Computer and Information Technology (ICCIT)*. IEEE. 2024, pp. 351–356.

[20] Rami Mubarak et al. “A survey on the detection and impacts of deepfakes in visual, audio, and textual formats”. In: *Ieee Access* 11 (2023), pp. 144497–144529.

[21] Minh Nguyen-Duc et al. “A Comparative Study of Deep Audio Models for Spectrogram-and Waveform-based SingFake Detection”. In: *IEEE Access* (2025).

[22] Lam Pham et al. “Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models”. In: *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE. 2024, pp. 1–5.

[23] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993. ISBN: 9780130151575.

[24] Ricardo Reimao and Vassilios Tzerpos. “For: A dataset for synthetic speech detection”. In: *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE. 2019, pp. 1–10.

[25] Reuters. *UN report urges stronger measures to detect AI-driven deepfakes*. United Nations ITU report warns of deepfake audio threats and calls for robust verification tools. July 2025. URL: https://www.reuters.com/business/un-report-urges-stronger-measures-detect-ai-driven-deepfakes-2025-07-11 (visited on 09/20/2025).

[26] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. “Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification”. In: *Proceedings of Interspeech*. 2017, pp. 23–27. DOI: 10.21437/Interspeech.2017-111.

[27] Massimiliano Todisco et al. “ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection”. In: *Proc. Interspeech*. 2019, pp. 1008–1012.

[28] Zhizheng Wu et al. “Spoofing and countermeasures for speaker verification: A survey”. In: *Speech Communication* 66 (2015), pp. 130–153.

[29] J. Yi et al. “A Survey on Deepfake Audio Detection: State of the Art, Challenges and Future Trends”. In: *IEEE Signal Processing Magazine* 40.6 (2023), pp. 85–97. DOI: 10.1109/MSP.2023.3298123.