Robust deepfake speech algorithm recognition: classifying generative algorithms via speaker x-vectors and deep learning.

Robust Deepfake Speech Algorithm Recognition: Classifying Generative Algorithms via Speaker X-Vectors and Deep Learning

Harry Maltby[1], Julie Wall[2], Cornelius Glackin[3],
Mansour Moniri[1], Roman Shrestha[3], Nigel Cannings[3], Iwa Salami[4]
[1]*School of Architecture, Computing and Engineering, University of East London, London, UK*
[2]*School of Computing and Engineering, University of West London, London, UK*
[3]*Verint Systems UK Ltd., London, UK*
[4]*Royal Docks School of Business and Law, University of East London, London, UK*
u1606661@uel.ac.uk, julie.wall@uwl.ac.uk, neil.glackin@verint.com,
m.moniri@uel.ac.uk, nigel.cannings@verint.com, i.salami@uel.ac.uk, roman.shrestha@verint.com

Abstract*—The rapid advancement of deepfake voice technologies has resulted in alarming cases of impersonation and deception, highlighting the urgent need for robust tools that can not only distinguish real audio from fake but also recognise the generative algorithms responsible. The ability to not only detect deepfake audio but also recognise the generative methods used is essential for forensic investigations, legal proceedings, and regulatory enforcement. Without robust and explainable detection frameworks, legal professionals and investigators lack the tools needed to effectively monitor, investigate, and prosecute cases involving deepfake misuse. In this work, we take a voice biometrics approach, shifting the focus from identifying who is speaking to identifying which algorithm is speaking. Doing so allows our approach to inherently handle unseen classes while achieving competitive performance for deepfake speech algorithm recognition. Our system leverages a voice-focused ResNet101-based x-vector extraction model and combines diverse audio features, and our experimental novel feature LFCC-HF, enhanced with Linear Discriminant Analysis and cosine similarity clustering. This approach allows for a more transparent and interpretable decision-making process by usinga single voice similarity decision boundary compared to the ensemble-based methods commonly used in the literature. Unlike previous works that rely on an ensemble of models, which convolute the decision-making process, our method achieves comparable results while using a significantly lighter-weight architecture, with our model having 14.84 M parameters compared to 95 M and 317 M parameters for Wav2Vec2 base and large. Furthermore, we demonstrate the benefits of targeted data augmentation, which, combined with feature fusion and our novel feature, improves system robustness and adaptability, increasing our F1 Score from 0.624 to 0.763, a 22.275% increase over our best single feature, and a 40.775% increase over the best ADD 2023 Track 3 baseline. Importantly, the system achieves interpretability through its back-end classification process, where decisions are based on a transparent, learned threshold for voice similarity to known voiceprints. This work offers a foundation for advancing more robust and interpretable solutions in the field of deepfake speech detection.

***Index Terms*—Deepfake Detection, Deepfake Audio, Generative Algorithm Recognition, Synthetic Speech Detection**

I. Introduction

Advancements in deepfake technologies have led to increasingly realistic synthetic speech, posing significant challenges for security and trust in voice-based systems. Deepfake voice cloning and impersonations are already rampant, having been used in an elaborate $35 million dollar bank heist in the UAE [1], spreading malicious disinformation using the stolen voices of UK politicians [2], and framing a US school principal for racial abuse [3]. The latter case is particularly egregious, as it marks one of the first instances in the United States where a victim was compelled to appear in court, with detectives relying on audio analysis tools to expose the deepfake. This case highlights the need for robust tools capable of not only distinguishing real speech from fake but providing interpretable insights into the decision-making process.

Binary classification approaches, (real vs. fake), often fail to meet the requirements of high-stakes applications such as legal proceedings, where transparency and interpretability are paramount [4]. Furthermore, identifying the specific algorithm used to generate the deepfake, a process known as algorithm recognition, is essential for tracing the source of the forgery, improving forensic analysis, and providing transparent results. The ability to not only detect deepfake audio but also attribute it to specific generative methods is essential for forensic investigations, legal proceedings, and regulatory enforcement.

An interpretable framework for deepfake voice classification, capable of classifying deepfake speech to specific generative methods, offers several advantages. It provides clear, interpretable evidence of how a decision was reached, enabling stakeholders such as legal professionals, investigators, and regulators to make informed decisions. Moreover, algorithm classification supports efforts to track the evolution of deepfake techniques, aiding in the development of proactive defences against emerging threats.

In this work, we take a voice biometrics approach and propose an interpretable framework for the detection of deep-

fake speech and generative algorithm classification that can inherently handle unseen classes. By leveraging speaker x-vectors extracted using a voice focussed ResNet101 trained from an untrained state, combined with Linear Discriminant Analysis (LDA) for dimensionality reduction and cosine similarity clustering, our framework offers robust performance while maintaining an interpretable back-end. The system classification back-end is interpretable, utilising voice similarity to known voice prints, in this case each algorithm is a voiceprint. Additionally, the processing requirements for such a system are far lower than contemporaries who predominantly utilise the transformer architecture with models such as Wav2Vec2 [5] [6]. Finally, we demonstrate the benefits of targeted data augmentation, which, combined with feature fusion and our novel Linear Frequency Cepstral Coefficient-High-Frequency (LFCC-HF) feature which is designed to capture high-frequency information by focusing exclusively on higher frequency regions, doing so improves system robustness and adaptability.

II. Related Work

The majority of current research focuses solely on binary classification, aiming to distinguish real speech from fake. This approach is evident in challenges such as ASVSpoof [7], which emphasise the development of countermeasures for automatic speaker verification. While effective for real vs. fake classification, this binary perspective overlooks the importance of identifying the specific generative methods responsible for deepfake speech. Notably, the Audio Deepfake Detection Challenge (ADD) 2023 Track 3 introduced the task of algorithm recognition [8], an important first step towards understanding the origins and characteristics of different deepfake generative methods. Without classifying the generative algorithm used, it leads to a shallow comprehension of the threat landscape, and the opportunity to develop more adaptive and comprehensive detection systems is missed.

Deep learning models, particularly ResNet-based architectures, have demonstrated impressive performance in deepfake speech detection. Studies utilising ResNet variants have reported high accuracy on benchmark datasets such as ASVSpoof 2021 DeepFake (DF) dataset [9] and Logical Access (LA) dataset [10] as well as the ADD 2023 Track 3 dataset [5]. These deep learning models excel in extracting high-level features from spectrogram representations, making them highly effective for binary classification tasks like distinguishing real vs. fake speech Combining x-vectors extracted from such models for classifying deepfake voice algorithms remains under-explored.

Further, several works have explored the task of algorithm recognition utilising the ADD Track 3 dataset such as the previously mentioned Lu et al. [5] and Qin et al. [6], where both utilised model fusion, providing impressive performance but in doing so convolute the decision-making process. Zeng et al. [11] and Wang et al. [12] additionally provide strong results, but at the cost of using larger classification models such as Wav2Vec2 and WavLM.

While numerous deepfake detection methods have been proposed, including ensemble-based architectures and transformer models [6, 5], these approaches often lack interpretability, making them impractical for forensic and legal applications. Regulatory agencies and law enforcement require systems that not only provide accurate classifications but also justify their decisions in a transparent manner. Black-box AI models can create challenges in court proceedings, where evidence admissibility requires clear reasoning and traceability of decision-making.

Various cepstral-based features have been used to successfully classify real from deepfake speech such as LFCC [13], novel features from gamma tone cepstral coefficients (GTCC) [9], mel frequency cepstral coefficients (MFCC) [10] and constant Q cepstral coefficients (CQCC) [14]. Most existing approaches rely on the selection of single audio features, such as MFCCs, spectrograms, or x-vectors [15], based on varying justifications. This fragmented approach fails to leverage the complementary strengths of multiple feature types in a coherent and explainable manner. Recent work by Firc et al. has shown that no single feature set performs consistently well across all datasets or deepfake algorithms, as different features capture complementary aspects of audio manipulation and generative patterns [16]. This inconsistency underscores the importance of systematically combining multiple feature sets to leverage their strengths while mitigating their individual weaknesses. However, current practices often address this challenge by employing ensembles of classification models, which, while effective, significantly increase model complexity and reduce interpretability.

III. Methodology

A. Features and X-Vector Combinations

The proposed system leverages five distinct audio representations, LFCC, MFCC, CQCC, GTCC, and LFCC-HF, to capture complementary aspects of the audio signal. We introduce a novel variant of LFCC, named LFCC-HF, that is computed using only the portion of the audio that is greater than or equal to 3 kHz. The threshold of 3 kHz was chosen based work by Maltby et al. that showed that there exists more difference between human and deepfake speech at roughly 3 kHz and above [17], indicating the presence of vocoder artefacts due to being generated in the mel-spectrum.

ResNet101 has been shown to be effective for speaker diarisation (determining who is speaking and when) [18]. This motivated our approach to adapt it for classifying which algorithm is speaking rather than which person. The x-vector embedding representations are the neuron activations within convolutional neural network-based models [19], generally corresponding to the first fully connected layer of the deep learning model. The input audio is segmented into utterances using energy-based Voice Activity Detection (VAD). Each utterance is further split into 14 ms segments, which are processed by the ResNet101 model to extract speaker x-vectors. After all utterance segments for an audio file are

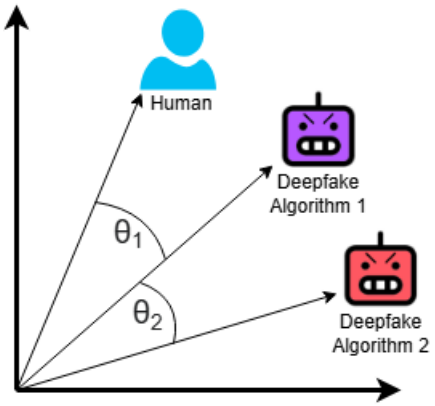

Fig. 1. Diagram showing cosine similarity-based classification between respective classes: Human, Deepfake Speech Algorithm 1, and Deepfake Speech Algorithm 2. The angles θ_1 and θ_2 represent the cosine distance and thus similarity between speaker vectors.

processed, the mean of the vectors is computed and used as the final speaker x-vector for that audio sample.

The audio features we included are designed to extract both speaker-specific and generative-method-specific characteristics, ensuring robust classification and algorithm recognition. LFCCs are a widely used feature set in speech processing, similar to MFCCs but based on a linear frequency scale. This linear scale allows LFCCs to retain high-frequency information, making them suitable for detecting subtle variations between synthetic and real voices.

B. Classification Methodology and System Architecture

For our classification system, we took a voice biometrics approach, treating each algorithm as a separate speaker identity. Similar to a speaker verification system, our system learns that x-vectors that are closer together are related to the same speaker, or generative algorithm in this case. Distinct clusters for different generative algorithm classes (e.g., Aliyun, Databaker, HiFiGAN), also known as voiceprints, are obtained from the cluster centroid of the known speaker vectors. New voice samples are then classified based on proximity to the nearest known voiceprint, with decision boundaries delineating regions of influence for each algorithm based on cosine similarity to the voiceprint as in Figure 1. Classes that are not similar to any known voiceprint are classified as ”unknown” via an optimal decision similarity threshold learned during training using the validation set. This method inherently allows the system to deal with unknown classes and out-of-distribution (OOD) data as voice prints that are not near a known class would be classified as unknown. This is different to separate dedicated OOD detection stages used to filter out OOD before giving it to the classification system like with Lu et al. [5]. These are unlike traditional deep learning models that use softmax-like functions to finalise a decision from a known set of seen classes.

The proposed classification system follows a structured pipeline to classify deepfake speech algorithms as seen in Figure 2. Initially, audio features are extracted from both train and test audio using four complementary representations: LFCC, MFCC, CQCC, and GTCC. These features are processed through a ResNet101-based x-vector extractor, producing 256-dimensional x-vectors for each feature type. The resulting x-vectors are then concatenated into a 1,024-dimensional composite x-vector, representing a combination of all feature types.

Once extracted, the x-vectors undergo LDA for dimensionality reduction, enhancing discriminative power while minimizing redundancy. LDA is particularly effective under the assumption that the classes, which are represented by voiceprint clusters in our case, are normally distributed with a shared covariance matrix. The reduced x-vectors are then classified using a cosine similarity-based thresholding mechanism, which determines the similarity of the test x-vector to known voiceprints or algorithm classes. This approach allows the system to classify test samples as belonging to one of several algorithm categories (e.g., Algorithm 1, Algorithm 2, or Human), or assign them to an Unknown category if no sufficient similarity is found based on a learned threshold during training.

C. Model and Training Methodology

For our ResNet101-based x-vector extractor we trained our model using a modified GitHub recipe by Landini et al. [20]. To do this we created a Kaldi-style directory for our ’wav.scp’ file to map utterance ID to the full path of the corresponding audio file and an ’utt2spk’ file that maps each utterance ID to the corresponding speaker ID, both files must be in the same directory. Input audio was then split into segments by energy-based VAD with each of these segments being labelled to the full path of the audio via the ’wav.scp’ file. These utterances were then further split into 14 ms segments and then fed to the model for training with 60 frequency bins being used for MFCC. During training of the x-vector model, the training data was augmented with the Room Impulse Response (RIR) [21] and MUSAN [22] noise and reverb datasets, doing so increases robustness by simulating diverse acoustic environments and enhances generalisation. With this, the x-vector model we used was pre-trained using the VoxCeleb1, VoxCeleb2 and CN-CELEB datasets for the purpose of automatic speaker recognition [18], the details of the data they used can be seen in Table I, with the model being trained until validation loss stagnated. After pre-training, our model was further fine-tuned on a very wide breadth of speech data including DIHARD-III, AISHELL-4 and VoxConverse. The Standup dataset used to pre-train our ResNet model was developed as an in-house proprietary dataset, the full breakdown of these datasets can be seen in Table II. We used an initial learning rate of 0.01 which progressively decayed to a final learning rate of 0.00005 using the stochastic gradient descent (SGD) optimiser.

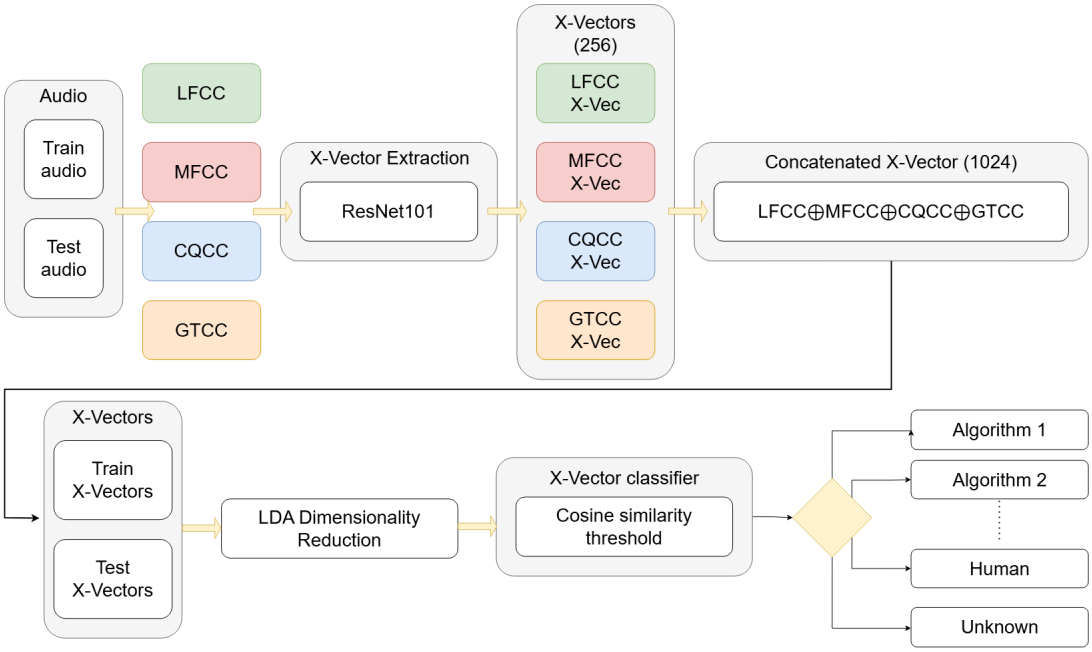

Fig. 2. Classification system pipeline.

TABLE I
DATASETS USED FOR INITIAL TRAINING OF RESNET101 MODEL.

Dataset	Duration	No. Speakers	Details
VoxCeleb1	323 hours	1,211	General speech dataset
VoxCeleb2	2,290 hours	5,994	Large speaker dataset
CN-Celeb	264 hours	973	Chinese speaker dataset

TABLE II
FINE-TUNING DATASETS USED FOR MODEL ADAPTATION.

Dataset	Language	Duration	Ref
DIHARD-III	English	61 hours	[23]
VoxConverse	English	50+ hours	[24]
Standup (internal)	English	45 mins	In-House
Zeroth	Korean	54 hours	[25]
Kokoro	Japanese	58 hours	[26]
CN-Celeb 1 & 2	Chinese	650k utterances	[27] [28]
Chime 6	English	40+ hours	[29]
Alimeeting	Chinese	109 hours	[30]
AISHELL-4	Chinese	120 hours	[31]

TABLE III
ADD 2023 TRACK 3 PARTITION COUNTS.

Class (Label)	#Training Set	#Development Set	#Testing
Human (6)	3,200	1,200	10,507
Aliyun (0)	3,200	1,200	9,512
Databaker (1)	3,200	1,200	10,474
Aispeech (2)	3,200	1,200	7,169
HiFiGAN (3)	3,200	1,200	10,461
WaveNet (4)	3,200	1,200	10,391
World (5)	3,200	1,200	10,507
Baidu (7)	3,200	1,200	10,469
Total Number	22,400	8,400	79,490

IV. EXPERIMENTAL SETUP

A. Datasets

For our experiments, we used the ADD 2023 Track 3 dataset [8]. Track 3 of the ADD 2023 dataset is intended for the subtask of algorithm classification, not just real vs fake but which generative algorithm was used to generate the deepfake speech while also including the human speech class. The training and development/validation sets of ADD 2023 Track 3 contains 7 classes while the test set contains the 7 classes seen in the training and development sets but also an unseen 8th class. This 8th class in the test set aims to test systems for robustness, and is more aligned to a real-world scenario. To contribute further to a real-world scenario and suboptimal conditions, the testing set contains clean, noisy, and compressed audio produced using undisclosed methods, the full counts and breakdown of the dataset can be seen in Table III. Further, we saw this dataset as a suitable way to test the inherent OOD detection capabilities of our approach.

B. Data augmentation

The test set of the ADD 2023 Track 3 dataset contains audio affected by noise and compression. To address this, we applied data augmentation to the training and development audio. For noise and reverberation, we selected samples from the MUSAN [22] and RIR [21] datasets, applying them systematically to introduce controlled variability into the audio. From these augmentation datasets we further split the RIR into real RIR and simulated RIR to further see the effects on classification performance. Further, we implemented a CutMix-style approach to generate new synthetic data [32]. Using audio from the training we generate new synthetic audio samples. We randomly sampled multiple audio files from the training set, from these audio files we separated them by

TABLE IV
ADD 2023 TRACK 3 BEST ABLATION RESULTS IN DESCENDING ORDER.

Features	Augmentation	F1 Score
All	Codec, Env, Cutmix, Baseline	0.763
All	Baseline + CutMix	0.761
All	Env Noise + Codec	0.750
All	Codec, Env, Cutmix	0.751
All	CutMix	0.722
All	All RIR + Environmental Noise	0.729
All	Environmental Noise	0.735
All	Baseline	0.690
All	Codec	0.703
All	Simulated RIR	0.698
All	Real RIR	0.697

utterance via energy-based VAD into smaller segments. We then took these segments and applied a random combination of data augmentation techniques to the audio such as noise, reverb, compression or any combination of the three. From these augmented segments we then stitch them together to form different permutations, producing new unseen synthetic data.

V. RESULTS DISCUSSION

A. Results

Table IV shows our ablation results for different augmentation strategies using the highest performing feature combination of LFCC + MFCC + CQCC + GTCC + LFCC-HF. Our best F1 score of 0.763, achieved using all features with the Codec, Environmental Noise, and CutMix augmentation, is competitive with single-model baselines in the literature, such as Lu et al. [5], which achieved similar performance without ensemble methods, shown in Table VI. Our lightweight audio focussed ResNet101-based architecture delivers high performance while maintaining interpretability through a transparent thresholding process. This is unlike ensemble-based methods, such as those employed by Qin et al. [6], which obscure decision-making through complex model fusion. Our approach emphasises clarity and computational efficiency. Comparatively, our voice-focussed ResNet101 model has far fewer parameters than contemporaries at 14.84 M parameters vs 95 M and 317 M parameters for Wav2Vec2 base and large.

Table V presents our ablation study results showing the F1 score improvements for all feature combinations under our best augmentation strategy (Codec, Env, Cutmix, Baseline) for the ADD 2023 Track 3 dataset. Our initial F1 score results using single features rival those of the challenge baselines. The ADD 2023 Track 3 baseline classifiers scored 0.5350 and 0.5416, the difference between them being that they used different types of thresholds for the detection of OOD data samples. Among the single feature combinations, our MFCC achieved the highest performance with an F1 score of 0.624, while our novel LFCC-HF feature scored the lowest at 0.517, this is probably due to the majority of the signal energy being below 3 kHz. This result aligns with the expectation that high-frequency components alone, which potentially obtain deepfake generation artefacts, as in LFCC-HF, may lack sufficient discriminatory information. However, when combined with other features, LFCC-HF provided complementary information, this is particularly evident in the final configuration, where adding LFCC-HF increased the F1 score from 0.747 to 0.763, displaying its complementary value. Without LFCC-HF, our results would fall short of Zeng et al. [11], where they reported an F1 score of 0.754, while employing Wav2Vec2, a significantly larger model. Overall, we improve upon our best single feature of MFCC by 22.275% when combining all feature together shown in Table V, and improve upon the best ADD 2023 Track 3 baseline of 0.542 by 40.775% shown in Table VI.

Upon reviewing our confusion matrix for our best feature and augmentation combination, we can see the performance and systems ability to handle unseen classes. Figure 3 presents the confusion matrix for our highest-performing feature combination (LFCC + MFCC + CQCC + GTCC + LFCC-HF). The matrix shows strong classification accuracy across most seen classes, such as Databaker, Wavenet, and World, each with dominant diagonal entries, indicating correct predictions. Of note, the OOD class (Baidu/Unk), which was not seen

TABLE V
F1 SCORE IMPROVEMENTS FOR FEATURE COMBINATIONS USING THE HIGHEST PERFORMING AUGMENTATION.

Combination	**F1 Score**	**Absolute**	**Relative (%)**
LFCC + MFCC + CQCC + GTCC + LFCC-HF	**0.763**	**0.246**	**47.58**
LFCC + MFCC + CQCC + GTCC	0.747	0.230	44.49
MFCC + CQCC + GTCC + LFCC-HF	0.737	0.220	42.55
LFCC + MFCC + CQCC + LFCC-HF	0.736	0.219	42.36
MFCC + CQCC + GTCC	0.734	0.217	41.97
LFCC + MFCC + GTCC + LFCC-HF	0.731	0.214	41.39
LFCC + MFCC + CQCC	0.730	0.213	41.20
LFCC + CQCC + GTCC + LFCC-HF	0.728	0.211	40.81
LFCC + CQCC + GTCC	0.728	0.211	40.81
LFCC + MFCC + GTCC	0.724	0.207	40.04
LFCC + CQCC + LFCC-HF	0.716	0.199	38.49
MFCC + CQCC + LFCC-HF	0.713	0.196	37.91
MFCC + GTCC + LFCC-HF	0.708	0.191	36.94
LFCC + MFCC + LFCC-HF	0.707	0.190	36.75
CQCC + GTCC + LFCC-HF	0.705	0.188	36.36
MFCC + CQCC	0.702	0.185	35.78
LFCC + GTCC + LFCC-HF	0.700	0.183	35.40
LFCC + CQCC	0.699	0.182	35.20
MFCC + GTCC	0.693	0.176	34.04
LFCC + MFCC	0.692	0.175	33.85
CQCC + GTCC	0.687	0.170	32.88
LFCC + GTCC	0.681	0.164	31.72
CQCC + LFCC-HF	0.664	0.147	28.43
MFCC + LFCC-HF	0.660	0.143	27.66
LFCC + LFCC-HF	0.657	0.140	27.08
GTCC + LFCC-HF	0.641	0.124	23.98
MFCC	0.624	0.107	20.70
CQCC	0.609	0.092	17.79
LFCC	0.594	0.077	14.89
GTCC	0.585	0.068	13.15
LFCC-HF	0.517	0.000	0.00

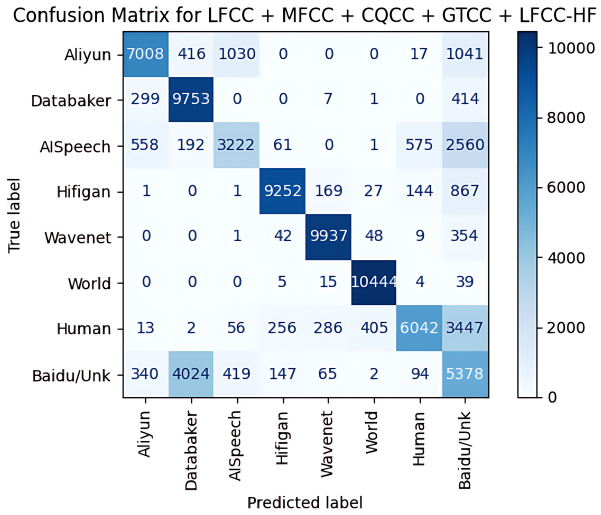

Fig. 3. Confusion matrix for the best feature combination.

during training, is largely clustered along the correct prediction axis, with 5,378 out of 10,469 samples correctly classified as "Unknown", displaying the inherent OOD detection capability of our system. This is unlike typical softmax-based classifiers that force all samples into one of the known categories, our similarity-based approach with a learned threshold allows the model to identify inputs that do not match any known generative voiceprint. While some misclassification into similar generative algorithms is observed, notably a mix-up between the OOD and Databaker classes likely due to overlapping voiceprints, the system still demonstrates a clear ability to distinguish novel algorithms from known ones without requiring a dedicated OOD pre-step.

B. Marginal Gain

To quantify the contribution of individual features in our ablation study we use marginal gain. We define marginal gain as $\Delta f(x, C)$ as the improvement in F1 score when a feature x is added to a feature combination C. Formally, the marginal gain is expressed as:

$$\Delta f(x, C) = f(C \cup \{x\}) - f(C)$$

where:

- C: set of features in the current combination (e.g., $C = \{\text{MFCC}, \text{CQCC}\}$),
- $f(C)$: F1 score of the model using the feature set C,
- $f(C \cup \{x\})$: F1 score of the model when feature x is added to the feature set C,
- x: the feature whose marginal gain is being calculated.

To compute the **average marginal gain** of a feature x across all combinations, we take the mean of $\Delta f(x, C)$ over all valid feature combinations C (excluding those that already include x):

$$\text{Average Marginal Gain of } x = \frac{1}{|S|} \sum_{C \in S} \Delta f(x, C)$$

where:

- S: set of all feature combinations C that do not include x,
- $|S|$: the number of such combinations.

Table VII presents the average marginal gain of individual features in improving the system's F1 score. This metric quantifies the contribution of each feature by measuring the improvement in classification performance when the feature is added to various combinations. First to note, is that all features have a positive gain, showing that each feature provides a positive contribution. Among the features, CQCC exhibited the highest average marginal gain (0.0583), followed closely by MFCC (0.0577) and LFCC (0.0524). These results indicate that CQCC and MFCC are the most complimentary for distinguishing between deepfake algorithms. In contrast, LFCC-HF demonstrated the lowest average marginal gain (0.0225), reflecting its limited standalone discriminative power. However, the complementary value of our novel LFCC-HF is shown when combined with other features, as our best F1 results were only attained using LFCC-HF.

VI. Conclusion

In this work, we have demonstrated an effective and more interpretable approach to deepfake speech recognition by leveraging a voice biometrics-inspired framework. Our system integrates a voice-focussed ResNet101-based x-vector extraction, LDA for dimensionality reduction and cosine similarity clustering for classification. This architecture offers both more robust performance as it can inherently deal with unseen classes and interpretability through voice similarity, making it a more suitable choice for forensic and regulatory applications. A voice biometrics similarity approach provides the ability to both detect deepfake audio and recognise specific generative algorithms, enhancing transparency in automated decision-making, a necessary factor for legal and regulatory applications.

Through the systematic combination of diverse feature representations, MFCC, LFCC, GTCC, CQCC, and our novel feature LFCC-HF, along with targeted data augmentation strategies, we achieved a significant improvement in classification performance in our ablation studies. Specifically, the F1 score increased by 22.275%, from 0.624 to 0.763 by combining all features versus our best single feature, and a 40.775% increase over the best ADD 2023 Track 3 baseline of 0.542. Notably, our system surpasses the performance of Zeng et al. [11], who reported an F1 score of 0.7541, while employing a significantly lighter-weight architecture compared to Wav2Vec2. Furthermore, our solution delivers results comparable to Lu et al.'s [5] models that did not incorporate dedicated OOD detection, but a lower score than their systems that did incorporate a dedicated OOD classification pre-stage. While our proposed system can inherently deal with unseen data, incorporating a

TABLE VI
ADD 2023 TRACK 3 F1 SCORE COMPARISON.

Ref	Model	Features	Augmentation	F1 Score	OOD
[5]	Model Fusion	Fusion	Codec+Env+CutMix	0.896	kNN
[6]	ResNet34SimAM-ASP, ResNet34-GSP, ResNet34SE-ASP, ECAPA-TDNN-ASP, LCNN, AASIST-SAP, Wav2Vec-ECAPA, wavlm-ECAPA	Wav2Vec2	-	0.831	Max Similarity
[5]	SENet18	STFT	-	0.779	-
Ours	**ResNet101**	**All**	**Codec, Env, CutMix**	**0.763**	**Voice Similarity**
[11]	ECAPA-TDNN	Wav2Vec2	Noise, reverb, mixup	0.754	Threshold
[12]	ResNet101-Temporal-Frequency-Transformer (TFT)	Log mel spec, WavLM	Noise, random sampling, time stretching, time masking, freq. masking	0.736	Threshold
[33]	RawNet2, SE-Res2Net50, HuBERT	Raw, LFCC, HuBERT	Noise, remove silence	0.735	Manifold-based multi-model fusion
Ours	**ResNet101**	**MFCC**	**Codec, Env, CutMix**	**0.624**	**Voice Similarity**
Ours	**ResNet101**	**CQCC**	**Codec, Env, CutMix**	**0.609**	**Voice Similarity**
Ours	**ResNet101**	**LFCC**	**Codec, Env, CutMix**	**0.594**	**Voice Similarity**
Ours	**ResNet101**	**GTCC**	**Codec, Env, CutMix**	**0.585**	**Voice Similarity**
[8]	ResNet	LFCC	-	0.542	Threshold
[8]	ResNet	LFCC	-	0.535	OpenMax
Ours	**ResNet101**	**LFCC-HF**	**Codec, Env, CutMix**	**0.517**	**Voice Similarity**

TABLE VII
AVERAGE MARGINAL GAIN OF INDIVIDUAL FEATURES IN IMPROVING F1 SCORE.

Feature	Average Marginal Gain
LFCC-HF	0.0225
GTCC	0.0458
LFCC	0.0524
MFCC	0.0577
CQCC	0.0583

dedicated OOD detection mechanism represents the next step in advancing performance, as current models in the literature, including ours, tend to plateau around an F1 score of 0.75-0.77 without a dedicated OOD pre-classification step to filter out OOD data samples. This hurdle highlights the potential of our lightweight and explainable framework as a foundation for further research into robust deepfake detection systems. Addressing these challenges remains a key direction for future work, as further advancements in this field will likely depend on developing more sophisticated methods for managing unseen data distributions. OOD data detection remains an open problem, and tackling it is essential for building systems that are both reliable and adaptable to new challenges. Finally, In the future we aim to perform further testing on different datasets to evaluate the capability of this approach for binary classification, multi-class classification and further classification tasks that involve OOD data. And while our primary focus in this work was building a lightweight and robust algorithm classification framework, the system's design also allows for enhanced interpretability. Specifically, the use of x-vectors extracted from a convolutional neural network opens the door to backtracking and showing influential regions in the various spectrograms via deconvolution techniques. This, coupled with cosine similarity-based classification, allows for potential reasoning about which parts of the audio most influenced a given classification. Although we do not fully explore this explainability component in this paper, it lays the groundwork for future work aimed at providing deeper insights into the model's decision-making process

REFERENCES

[1] Thomas Brewster. "Fraudsters Cloned Company Director's Voice In $35 Million Heist, Police Find". In: *Forbes* (Oct. 2021). Updated May 2, 2023. URL: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/.

[2] Ben Quinn. "Slew of deepfake video adverts of Sunak on Facebook raises alarm over AI risk to election". In: *The Guardian* (Jan. 2024). Published on 12 January 2024. URL: https://www.theguardian.com/technology/2024/jan/12/deepfake-video-adverts-sunak-facebook-alarm-ai-risk-election.

[3] Ben Finley. "Deepfake of principal's voice is the latest case of AI being used for harm". In: *The Independent* (Apr. 2024). Published on 29 April 2024. URL: https://www.independent.co.uk/news/ap-deepfake-maryland-people-experts-b2536677.html.

[4] Andrzej Porebski. "Looking for the Right Paths to Use XAI in the Judiciary: Which Branches of Law Need Inherently Interpretable Machine Learning Models and Why?" English. In: *Joint Proceedings of the xAI 2024 Late-breaking Work, Demos and Doctoral Consortium*. Vol. 3793. CEUR Workshop Proceedings. Valletta, Malta: CEUR-WS, 2024, pp. 129–136. URL: https://ceur-ws.org/Vol-3793/paper_17.pdf.

[5] Jingze Lu et al. "Detecting Unknown Speech Spoofing Algorithms with Nearest Neighbors." In: *DADA@ IJCAI*. 2023, pp. 89–94.

[6] Xiaoyi Qin et al. "From Speaker Verification to Deepfake Algorithm Recognition: Our Learned Lessons from ADD2023 Track 3." In: *DADA@ IJCAI*. 2023, pp. 107–112.

[7] 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. DOI: 10.1109/TASLP.2023.3285283.

[8] Jiangyan Yi et al. "Add 2023: the second audio deepfake detection challenge". In: *arXiv preprint arXiv:2305.13774* (2023).

[9] Mohit Dua, Swati Meena, Nidhi Chakravarty, et al. "Audio Deepfake Detection Using Data Augmented Graph Frequency Cepstral Coefficients". In: *2023 International Conference on System, Computation, Automation and Networking (ICSCAN)*. IEEE. 2023, pp. 1–6.

[10] A. Cohen et al. "A study on data augmentation in voice anti-spoofing". In: *Speech Communication* 141 (2022), pp. 56–67.

[11] Xiao-Min Zeng et al. "Deepfake Algorithm Recognition System with Augmented Data for ADD 2023 Challenge." In: *DADA@ IJCAI*. 2023, pp. 31–36.

[12] Ziqian Wang et al. "The NPU-ASLP System for Deepfake Algorithm Recognition in ADD 2023 Challenge." In: *DADA@ IJCAI*. 2023, pp. 64–69.

[13] Xinhui Chen et al. "UR Channel-Robust Synthetic Speech Detection System for ASVspoof 2021". In: *ArXiv* abs/2107.12018 (2021). URL: https://api.semanticscholar.org/CorpusID:236428417.

[14] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. "A comparative study on physical and perceptual features for deepfake audio detection". In: *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 2022, pp. 35–41.

[15] L. Wu and Y. Jiang. "Attentional Fusion TDNN for Spoof Speech Detection". In: *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*. IEEE. Aug. 2022, pp. 651–657.

[16] Anton Firc, Kamil Malinka, and Petr Hanáček. "Deepfake Speech Detection: A Spectrogram Analysis". In: *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. 2024, pp. 1312–1320.

[17] Harry Maltby et al. "A Frequency Bin Analysis of Distinctive Ranges Between Human and Deepfake Generated Voices". In: *2024 International Joint Conference on Neural Networks (IJCNN)*. 2024, pp. 1–7. DOI: 10.1109/IJCNN60899.2024.10650554.

[18] Roman Shrestha et al. "Speaker Recognition using Multiple X-Vector Speaker Representations with Two-Stage Clustering and Outlier Detection Refinement". In: *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. 2022, pp. 1–6. DOI: 10.1109/DASC/PiCom/CBDCom/Cy55231.2022.9927875.

[19] David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 5329–5333.

[20] Federico Landini et al. *Bayesian HMM Clustering of X-Vector Sequences (VBx) in Speaker Diarization: Theory, Implementation and Analysis on Standard Tasks*. https://github.com/phonexiaresearch/VBx-training-recipe. Accessed: 2025-01-09. 2022.

[21] Tom Ko et al. "A study on data augmentation of reverberant speech for robust speech recognition". In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pp. 5220–5224.

[22] David Snyder, Guoguo Chen, and Daniel Povey. "Musan: A music, speech, and noise corpus". In: *arXiv preprint arXiv:1510.08484* (2015).

[23] Neville Ryant et al. "The third DIHARD diarization challenge". In: *arXiv preprint arXiv:2012.01477* (2020).

[24] Joon Son Chung et al. "Spot the conversation: speaker diarisation in the wild". In: *INTERSPEECH*. 2020.

[25] Zeroth Project. *Zeroth-Korean Speech Corpus*. http://www.openslr.org/40/. Accessed: Jan. 15, 2025.

[26] Katsuya Iida. *Kokoro Speech Dataset*. https://github.com/kaiidams/Kokoro-Speech-Dataset. Accessed: Jan. 15, 2025. 2021.

[27] Yue Fan et al. "CN-CELEB: a challenging Chinese speaker recognition dataset". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 7604–7608.

[28] Lantian Li et al. *CN-Celeb: multi-genre speaker recognition*. 2020. arXiv: 2012.12468 [eess.AS].

[29] Shinji Watanabe et al. "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings". In: *CHiME 2020-6th International Workshop on Speech Processing in Everyday Environments*. 2020.

[30] Fan Yu et al. "Summary on the ICASSP 2022 multi-channel multi-party meeting transcription grand challenge". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 9156–9160.

[31] Yihui Fu et al. "AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario". In: *Interspeech*. 2021. URL: https://arxiv.org/abs/2104.03603.

[32] Sangdoo Yun et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6023–6032.

[33] Ye Tian et al. "Deepfake Algorithm Recognition through Multi-model Fusion Based On Manifold Measure." In: *DADA@ IJCAI*. 2023, pp. 76–81.