

Bridging Attribution and Open-Set Detection using Graph-Augmented Instance Learning in Synthetic Speech

Mohd Mujtaba Akhtar^{1*}, Girish^{2*}, Farhan Sheth^{3*} and Muskaan Singh^{4†}

¹Veer Bahadur Singh Purvanchal University, India

²UPES, India

³Manipal University Jaipur, India

⁴Ulster University, UK

Abstract

We propose a unified framework for not only attributing synthetic speech to its source but also for detecting speech generated by synthesizers that were not encountered during training. This requires methods that move beyond simple detection to support both detailed forensic analysis and open-set generalization. To address this, we introduce **SIGNAL**, a hybrid framework that combines speech foundation models (SFM) with graph-based modeling and open-set-aware inference. Our framework integrates Graph Neural Networks (GNN) and a k-Nearest Neighbor (KNN) classifier, allowing it to capture meaningful relationships between utterances and recognize speech that doesn't belong to any known generator. It constructs a query-conditioned graph over generator class prototypes, enabling the GNN to reason over relationships among candidate generators, while the KNN branch supports open-set detection via confidence-based thresholding. We evaluate **SIGNAL** using the DiffSSD dataset, which offers a diverse mix of real speech and synthetic audio from both open-source and commercial diffusion-based TTS systems. To further assess generalization, we also test on the SingFake benchmark. Our results show that **SIGNAL** consistently improves performance across both tasks, with Mamba-based embeddings delivering especially strong results. To the best of our knowledge, this is the first study to unify graph-based learning and open-set detection for tracing synthetic speech back to its origin.

1 Introduction & Background

Synthetic Speech Detection (SSD) plays a critical role in safeguarding digital communication, enabling systems to identify and mitigate the risks posed by highly realistic, machine-generated voices (Todisco et al., 2019; Wu et al., 2015). With the advent of advanced text-to-speech (TTS) and

voice conversion (VC) models, synthetic speech has reached a level of fidelity that closely mimics natural human prosody and timbre (Ren et al., 2021; Kong et al., 2021). While such advancements drive progress in accessibility and personalization (Cooper et al., 2020), they also introduce new vulnerabilities in the form of audio-based impersonation, fraud, and misinformation (Yi et al., 2023). Consequently, the ability to detect and analyze synthetic speech is vital not just for security but also for preserving trust in human-AI interaction. The past few years have witnessed remarkable advancements in neural speech synthesis, driven by diffusion-based models, expressive TTS systems, and multilingual voice conversion techniques. State-of-the-art models such as VALL-E (Ju et al., 2024), NaturalSpeech 3 (Ju et al., 2024), and Voicebox (Le et al., 2023) have demonstrated the ability to generate speech that not only mimics speaker identity but also captures fine-grained acoustic attributes such as emotion, prosody, and expressiveness. These models, often built upon large-scale speech-language pretraining, leverage powerful architectural backbones including transformers (Li et al., 2019), state-space models (Gu and Dao, 2024), and denoising diffusion processes (Ren et al., 2021). Their increasing accessibility through open-source implementations and commercial APIs has made synthetic speech generation more ubiquitous than ever. However, this rapid progress also underscores the growing complexity of the detection task, particularly in settings where the generation model is unknown or unseen at inference time.

Despite the increasing attention on synthetic speech detection, most existing methods frame SSD as a binary classification problem, focusing solely on distinguishing real from synthetic audio rather than identifying the generative source (Shin et al., 2024; Huang and Pun, 2024). However, most existing methods fail to generalize when faced

*Equal contribution as a first author.

†Corresponding: m.singh@ulster.uk.in

with previously unseen TTS or voice-conversion models, exhibiting significant performance degradation under open-set conditions (Guo et al., 2024; Stan et al., 2025). This limitation constrains their practical utility in real-world forensic scenarios, where source attribution and robust detection of out-of-distribution speech are critical. While recent work has advanced synthetic speech detection, a key gap remains: the ability to both identify the source TTS model and detect speech from unseen generators. Source attribution is increasingly important in forensic and regulatory settings, where knowing the origin of synthetic audio matters. At the same time, real-world systems must handle open-set scenarios, where new or unknown models may appear at test time. Bridging these two challenges is essential for building more reliable and generalizable detection frameworks. As the core focus of our study in DiffSSD, we explore a range of speech foundation models (SFMs), and *hypothesize that combining relational modeling via Graph Neural Networks (GNNs)—where a query-conditioned graph is formed over generator class prototypes (nodes) and refined through attention-based message passing—with open-set inference through k-Nearest Neighbors (KNN) offers an effective strategy for jointly tackling source attribution and unseen generator detection.* This approach is motivated by the complementary strengths of the two components: GNNs model class-level interactions among embedding nodes, capturing subtle relational cues for generator discrimination, while the KNN branch provides a lightweight yet robust mechanism for handling open-set scenarios via confidence-based thresholding. To validate this hypothesis, we perform extensive evaluations using diverse SFM embeddings—including Whisper, UniSpeech, ECAPA, and Mamba—under both seen and unseen generator conditions. Drawing inspiration from prior work on explainable source attribution (Mishra et al., 2025), the use of graph-based structures for deepfake detection (Febrinanto et al., 2025), and Graph Attention Networks for spoofing detection (Tan et al., 2025), To adress this, we introduce **SIGNAL: Speech Inference via Graph Networks and Augmented Learning** — a unified framework that performs GNN-based node classification alongside post-hoc KNN filtering to address both tasks simultaneously. To the best of our knowledge, this is the first work to explore a hybrid GNN–KNN approach for source tracing and open-set detection in synthetic speech.

Our key contributions are as follows:

- We propose **SIGNAL**, a novel hybrid framework combining Graph Neural Networks (GNNs) for relational modeling with k-Nearest Neighbors (KNN) for open-set inference, tailored for joint source attribution and unseen generator detection in synthetic speech.
- We perform a large-scale benchmarking of diverse Speech Foundation Models (SFMs)—including Whisper, UniSpeech, ECAPA, and Mamba—under both seen and unseen generator settings, revealing their complementary behavior across tasks.
- To the best of our knowledge, this is the first study to investigate graph-based posthoc reasoning on SFM embeddings for open-set synthetic speech forensics, establishing strong performance on both DiffSSD and SingFake benchmarks.

Resources for this study are available at: <https://github.com/Helixometry/SIGNAL.git>

2 Representations

In this section, we detail the speech foundation models (SFMs) used to extract utterance-level embeddings for downstream processing.

We utilize x-vector¹ (Snyder et al., 2018) and ECAPA-TDNN² (Desplanques et al., 2020) as speaker recognition models. Both are based on time-delay neural networks and are trained on the VoxCeleb1+2 datasets. ECAPA extends the standard x-vector architecture with Res2Net modules and SE blocks, yielding improved performance over x-vector, which itself significantly surpasses traditional i-vector baselines. x-vector contains approximately 4.2M parameters. For monolingual PTMs, we employ WavLM³ (Chen et al., 2022), UniSpeech-SAT⁴ (Chen et al., 2021), and wav2vec 2.0⁵ (Baevski et al., 2020), all in their base versions trained on 960 hours of English audio from LibriSpeech. WavLM and UniSpeech-SAT are top-performing PTMs on the SUPERB benchmark,

¹<https://huggingface.co/speechbrain/spkrec-xvect-voxceleb>

²<https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb>

³<https://huggingface.co/microsoft/wavlm-base>

⁴<https://huggingface.co/microsoft/unispeech-sat-base>

⁵<https://huggingface.co/facebook/wav2vec2-base>

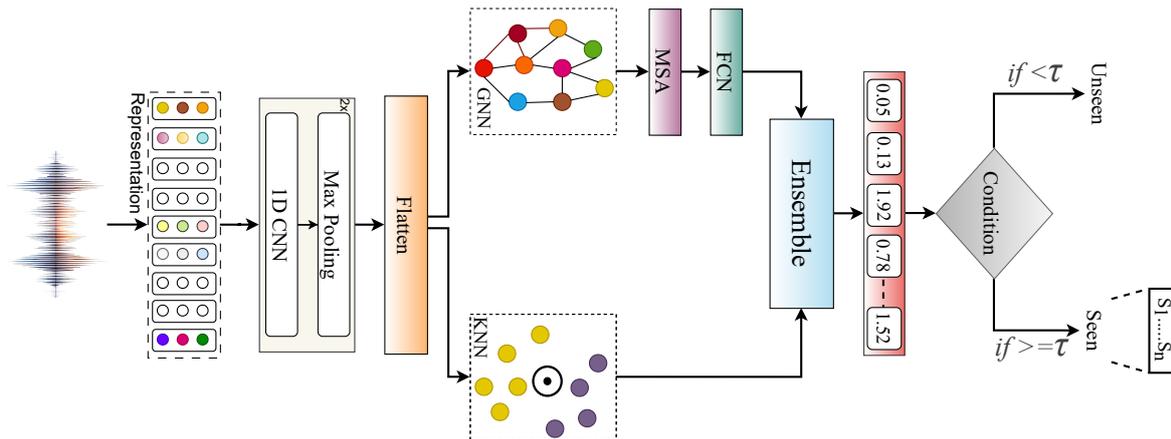


Figure 1: Proposed framework: **SIGNAL**. The model extracts representations, followed by parallel reasoning via a Graph Attention Network (GAT) and a K-Nearest Neighbors (KNN) module. The outputs are fused via an ensemble head. The final routing decision is based on a confidence threshold ($\tau = 0.5$), directing samples to either seen ($S_1 \dots S_n$) or unseen class predictions.

with WavLM trained using masked speech modeling and denoising objectives, while UniSpeechSAT adopts a multi-task framework incorporating speaker-aware contrastive learning. wav2vec 2.0 is optimized via a contrastive loss to distinguish true from distractor codebook entries. All three models have similar parameter sizes: WavLM (94.7M), UniSpeech-SAT (94.68M), and wav2vec 2.0 (95.04M). We further include Whisper⁶ (Radford et al., 2023), a multilingual ASR model trained on 680K hours of diverse audio. Whisper uses a transformer-decoder architecture and performs robustly across languages and tasks in a zero-shot setting. We use the base version with 74M parameters and extract representations from the encoder layer. In addition, we consider Audio-MAMBA⁷ (Yadav and Tan, 2024), a state space model trained on AudioSet to reconstruct masked spectrogram patches. Its SSM backbone captures both temporal dependencies and spectral detail effectively. We use its tiny (4.8M), small (17.9M), and base (69.3M) variants to study scalability effects.

Resampling to 16 KHz is done for the audio samples before passing it to the FMs. We obtain fixed-length embeddings by applying average pooling to the final hidden state of each frozen model. The resulting embedding dimensions are: 512 for Whisper and x-vector; 768 for WavLM, UniSpeech-SAT, and wav2vec 2.0; and 3840 for Audio-MAMBA (base version). ECAPA-TDNN produces 192-dimensional embeddings.

⁶<https://huggingface.co/openai/whisper-base>

⁷<https://github.com/SarthakYadav/audio-mamba-official>

3 Modeling Pipeline

In this section, we detail the modeling pipeline for individual representations and the proposed hybrid framework for source tracing and unseen generator detection. We use Fully Connected Networks (FCN) and Convolutional Neural Networks (CNN) as downstream models for individual representation-based modeling. Further, we propose **SIGNAL**, which integrates Graph Neural Networks (GNNs) with a K-Nearest Neighbors (KNN) module, enabling it to leverage both global relational structure and local instance-level similarity. Figure-1 illustrates the overall architecture.

3.1 Individual Representation Modeling

To establish strong baselines, we first evaluate standard classifiers with individual pre-trained speech representations, we use FCN and CNN backbones. The CNN consists of two 1D convolutional layers with 64 and 128 filters (kernel size = 3), followed by ReLU activation and max-pooling (pool size = 2). The output is flattened and passed through a dense layer of 128 neurons, followed by a softmax output layer. The FCN model uses the same dense block as the CNN, excluding convolutional layers.

Proposed framework : SIGNAL

We propose, **SIGNAL** for the joint task of source attribution and unseen generator detection in synthetic speech. The architecture of **SIGNAL** is shown in Figure 1.

Representation Encoding: Let $\mathbf{x} \in \mathbb{R}^{T \times F}$ denote the input audio signal, where T is the number of

PTMs	FCN									CNN								
	DEV			TEST			ID/OOD			DEV			TEST			ID/OOD		
	ACC ↑	F1 ↑	EER ↓	ACC ↑	F1 ↑	EER ↓	ACC ↑	F1 ↑	EER ↓	ACC ↑	F1 ↑	EER ↓	ACC ↑	F1 ↑	EER ↓	ACC ↑	F1 ↑	EER ↓
Baseline																		
Whisper	80.75	<u>79.3</u>	<u>7.78</u>	74.52	72.66	9.77	52.35	51.17	41.09	83.10	81.40	<u>5.71</u>	<u>79.45</u>	<u>78.13</u>	<u>7.81</u>	54.12	52.85	30.11
Unispeech	70.58	69.12	27.20	72.00	70.26	29.70	37.54	36.77	53.99	78.35	76.86	20.34	68.96	67.35	23.45	38.81	38.05	40.32
x-vector	66.91	65.42	19.86	69.47	67.52	22.67	43.83	42.80	52.20	74.15	72.90	16.42	67.82	65.58	21.39	45.50	44.39	37.78
wav2vec 2.0	76.82	75.45	8.73	69.46	67.31	12.88	41.90	41.08	44.33	78.13	76.42	6.49	79.19	77.44	9.40	43.43	42.62	32.66
wavLM	78.24	76.81	12.48	73.16	71.07	15.53	44.74	43.65	52.49	80.60	78.25	9.06	76.44	74.21	11.37	46.38	45.34	39.11
ECAPA	63.77	62.55	17.28	62.31	60.09	25.29	32.33	31.63	57.80	71.92	69.87	13.13	64.15	63.29	20.13	33.56	32.84	41.45
MAMBA-T	69.35	68.12	15.87	<u>76.01</u>	75.40	19.08	<u>59.23</u>	<u>58.13</u>	46.59	80.15	78.17	11.37	72.23	71.04	11.79	<u>60.94</u>	<u>59.92</u>	35.00
MAMBA-S	72.01	70.66	10.86	71.04	69.16	13.19	<u>57.33</u>	<u>56.21</u>	<u>40.56</u>	81.22	79.56	7.82	74.51	73.18	9.80	<u>59.32</u>	<u>58.11</u>	<u>29.64</u>
MAMBA-B	81.68	80.29	7.63	76.10	<u>74.35</u>	<u>10.74</u>	62.91	61.50	36.14	<u>82.90</u>	<u>80.11</u>	5.59	80.08	78.47	7.78	64.62	63.35	26.27

Table 1: Performance metrics (ACC, F1, EER) across different PTMs under FCN and CNN backbones. Top-performing scores are in **bold**, runner-up scores are *underlined*. All tables in the study follow the same formatting.

frames and F is the feature dimension. This signal is passed through a frozen Speech Foundation Model (SFM), yielding a fixed-length utterance embedding:

$$\mathbf{z}_0 = \text{SFM}(\mathbf{x}) \in \mathbb{R}^{d_0}$$

We then project \mathbf{z}_0 into a lower-dimensional latent space using a CNN encoder f_{cnn} :

$$\mathbf{z} = f_{\text{cnn}}(\mathbf{z}_0) \in \mathbb{R}^d, \quad \text{where } d = 64$$

The encoder f_{cnn} comprises two 1D convolutional layers with ReLU activations and max-pooling, followed by a dense projection layer.

GNN Head: To capture class-level relationships and enhance source attribution, we design a graph-based attention module centered around the input query and known generator classes. The graph consists of:

- A **query node** representing the encoded utterance embedding \mathbf{z} .
- **Class nodes** represented by N learnable prototype vectors $\{\mathbf{e}_1, \dots, \mathbf{e}_N\}$, each corresponding to a seen TTS generator.
- **Edges** connecting the query to each class node, allowing the model to assess similarity and perform reasoning over them.

Why a GNN for attribution: We adopt a GNN for attribution because source prediction benefits from reasoning over relationships among generator classes, not just per-class similarity scores. By propagating information across class nodes, the GNN captures relative structure that is difficult to represent with independent classifiers.

The query embedding \mathbf{z} is first projected into a latent space using a learnable transformation:

$$\mathbf{s} = \mathbf{W}_s \mathbf{z}, \quad \mathbf{s} \in \mathbb{R}^d$$

where $\mathbf{W}_s \in \mathbb{R}^{d \times d}$ is a trainable weight matrix. This projected vector \mathbf{s} is then *added to each class prototype* to produce query-aware node features:

$$\tilde{\mathbf{e}}_i = \mathbf{e}_i + \mathbf{s}, \quad \forall i \in \{1, \dots, N\}$$

These N modified node embeddings are passed through a **multi-head self-attention** mechanism, where each node attends to the others and refines its own representation:

$$\tilde{\mathbf{e}}'_i = \text{MultiHeadAttn}(\tilde{\mathbf{e}}_i, \{\tilde{\mathbf{e}}_j\}_{j=1}^N)$$

To compute class scores, each updated node is projected to a scalar logit:

$$\ell_i = \mathbf{w}^\top \tilde{\mathbf{e}}'_i, \quad \forall i \in \{1, \dots, N\}$$

The final class probabilities are obtained via a softmax over the logits:

$$\mathbf{p}_{\text{GNN}} = \text{softmax}([\ell_1, \dots, \ell_N])$$

To estimate the model’s uncertainty in attribution, we compute the attention entropy over the predicted distribution:

$$\mathcal{H}_{\text{attn}} = - \sum_{i=1}^N p_{\text{GNN},i} \log p_{\text{GNN},i}$$

Intuitively, low entropy implies confident attribution to a specific generator, while high entropy suggests the model is uncertain and attention was spread more uniformly across classes.

KNN Branch: To handle open-set scenarios, we introduce a KNN module that performs instance-based reasoning in the embedding space. After

training f_{cnn} , we collect all training embeddings $\{\mathbf{z}_j\}$ and fit a distance-weighted KNN classifier. At test time, for a query \mathbf{z} , the KNN output is computed as:

$$\mathbf{p}_{\text{KNN}} = \frac{\sum_{k=1}^K w_k \cdot \mathbf{y}_k}{\sum_{k=1}^K w_k}, \quad w_k = \frac{1}{\|\mathbf{z} - \mathbf{z}_k\|_2^2 + \epsilon}$$

where \mathbf{z}_k are the K nearest neighbors with corresponding labels \mathbf{y}_k , and ϵ is a small constant to ensure numerical stability.

Ensemble Fusion and Open-Set Detection: We fuse the GNN and KNN predictions via convex combination:

$$\mathbf{p}_{\text{ens}} = \alpha \cdot \mathbf{p}_{\text{GNN}} + (1 - \alpha) \cdot \mathbf{p}_{\text{KNN}}, \quad \alpha \in [0, 1]$$

To determine whether a sample belongs to a known or unknown generator, we use confidence-based routing with threshold τ and optionally an entropy-based uncertainty signal:

- **Confidence thresholding (main):** if $\max(\mathbf{p}_{\text{ens}}) < \tau$, the sample is labeled as *unseen*.
- **Entropy thresholding (optional):** if $\mathcal{H}_{\text{attn}} > \tau_e$, the sample is labeled as *unseen*.

Unless stated otherwise, we use confidence thresholding as our default decision rule and ablate τ in Sec. 4.4 (Fig. 3); the entropy criterion is included as an auxiliary uncertainty signal and not used for primary model selection.

Training Objective: The GNN head is trained using cross-entropy loss over the seen classes:

$$\mathcal{L}_{\text{CE}} = - \sum_{i=1}^N y_i \log(\mathbf{p}_{\text{GNN},i})$$

The KNN module is non-parametric and requires no training.

4 Experiments

4.1 Benchmark Dataset

We conduct our study on the Diffusion-Based Synthetic Speech Dataset (DiffSSD) (Bhagtani et al., 2025), which contains high-quality synthetic and real speech samples. This dataset is specifically designed to evaluate models in both source tracing and unseen generator detection tasks for synthetic speech. In total, DiffSSD includes around

200 hours of labeled audio, with 70,000 synthetic and 24,226 real speech samples. The real speech in DiffSSD comes from LibriSpeech, which contributes 11,126 samples, and LJ Speech, with 13,100 samples. The synthetic portion is produced by ten TTS systems—eight open-source generators (GradTTS, OpenVoiceV2, ProDiff, WaveGrad2, Xttsv2, YourTTS, DiffGAN-TTS, and UnitSpeech) and two commercial tools (ElevenLabs and PlayHT). Each synthetic sample is created using a set of 5,000 English text lines covering everyday topics such as conversations, weather, quotes, and general descriptions. These text lines were generated using ChatGPT-3.5 and were carefully filtered to avoid repetition. The dataset is divided into three parts: the training set with 31,690 samples, the validation set with 7,423 samples, and the test set with 54,613 samples. This split supports two types of evaluation—closed-set, where all generator types are included during training, and open-set, where the test set includes synthetic speech from generators not seen during training, such as the commercial tools PlayHT and ElevenLabs.

Training details: We follow the predefined train/dev/test splits provided by DiffSSD. Models are trained on the training split and selected using the dev split; all final numbers are reported on the held-out test split (including the ID/OOD setting). We train for up to 50 epochs using Adam with a learning rate of 1×10^{-3} and batch size 32, and apply early stopping based on dev performance.

4.2 Evaluation Metrics

To evaluate the performance of our proposed framework, we employ widely adopted metrics: Accuracy (ACC), F1-score (F1), and Equal Error Rate (EER). Accuracy provides a straightforward measure of correct classifications. F1-score, as the harmonic mean of precision and recall, offers a balanced view of performance in class-imbalanced scenarios. EER, represents the point at which false acceptance and false rejection rates are equal, making it particularly insightful for open-set evaluation. These metrics have been used in Phukan et al. (2025) on deepfake detection and source attribution, ensuring methodological consistency across domains and reinforcing the generalizability of our evaluation framework.

PTM	DEV			TEST			ID/OOD		
	ACC \uparrow	F1 \uparrow	EER \downarrow	ACC \uparrow	F1 \uparrow	EER \downarrow	ACC \uparrow	F1 \uparrow	EER \downarrow
KNN									
Whisper	92.59	90.15	5.52	85.41	84.50	7.83	60.36	58.15	25.56
Unispeech	86.14	85.06	16.05	75.13	73.32	19.99	43.14	41.37	30.14
X-vector	82.63	80.27	14.32	76.39	74.54	19.47	49.22	47.61	32.91
wav2vec 2.0	88.41	86.38	7.18	87.76	85.91	8.22	50.37	48.04	27.33
WavLM	90.43	88.26	8.94	82.43	80.83	10.52	49.19	47.21	30.36
ECAPA	78.37	76.15	12.39	72.88	70.43	18.19	38.01	36.95	33.57
Mamba-T	89.32	87.19	11.05	80.91	79.30	10.45	67.91	65.36	28.39
Mamba-S	91.26	89.07	7.11	80.66	78.81	8.44	69.95	68.09	23.68
Mamba-B	94.05	93.28	5.32	89.41	87.19	7.62	72.56	70.24	22.83
GNN									
Whisper	97.29	96.11	4.37	93.64	93.21	5.78	64.13	62.88	22.46
Unispeech	91.26	89.26	15.89	81.38	75.18	17.54	47.28	45.86	28.85
X-vector	86.95	82.94	12.33	80.23	79.25	17.48	52.34	50.58	27.61
wav2vec 2.0	92.05	85.20	5.74	93.40	90.79	6.92	54.01	52.06	24.38
WavLM	93.62	91.67	6.86	89.37	86.33	8.25	53.27	51.52	28.46
ECAPA	82.36	78.25	10.39	75.03	72.71	17.09	42.69	40.54	30.57
Mamba-T	94.29	91.22	8.96	85.51	82.64	9.23	71.92	69.66	26.29
Mamba-S	96.41	92.64	5.18	87.26	85.42	7.34	72.75	70.12	21.41
Mamba-B	96.35	96.15	5.29	93.40	93.32	5.62	71.22	71.02	19.28
KNN + GNN									
Whisper	97.56	96.29	3.38	94.19	92.22	4.33	76.67	74.14	17.10
Unispeech	92.31	91.25	9.03	82.63	80.36	13.72	63.15	61.90	22.91
X-vector	88.19	86.99	7.79	81.59	80.41	13.20	64.66	62.35	20.96
wav2vec 2.0	93.46	90.34	3.95	94.38	92.37	5.39	72.05	70.26	19.65
WavLM	95.09	93.34	5.15	91.21	89.84	6.71	74.28	71.47	21.35
ECAPA	83.27	81.06	6.55	77.61	75.29	15.07	62.36	60.32	23.91
Mamba-T	95.64	93.12	4.16	87.37	85.28	7.94	80.17	78.66	20.31
Mamba-S	96.10	94.37	3.91	89.24	87.32	5.90	86.54	84.12	17.36
Mamba-B	98.11	96.86	2.33	95.52	94.21	4.32	88.91	86.53	14.78

Table 2: Performance comparison across different Pretrained Models (PTMs) using GNN, KNN, and their combination. The Blue gradient indicates performance from highest to lowest. Table 4 uses the same scheme.

4.3 Experimental Results

Table 1 presents performance using FCN and CNN classifiers on individual pre-trained representations. Among the baselines, Mamba-B consistently outperforms other PTMs across all splits. Under the CNN setup, it achieves ACC: 83.27%, F1: 82.63%, and EER: 4.26% on the DEV set, while maintaining reasonable generalization to in-domain/out-of-domain (ID/OOD) settings (ACC: 69.01%, EER: 17.97%). Other representations like Whisper and wav2vec 2.0 show competitive performance in seen scenarios but degrade substantially in the presence of unseen generators. ECAPA and UniSpeech yield relatively lower results, reflecting the limitations of speaker-centric and monolingual embeddings in generalizing to diverse generative artifacts. These findings indicate that while CNN-based classifiers can leverage strong pre-trained representations for attribution, they are still limited in open-set scenarios.

Baseline clarification: We evaluate three categories of methods: (i) Attribution-only models, which assume all test samples originate from known generators and output a closed-set class prediction (e.g., GNN-only); (ii) Open-set-only mod-

els, which focus on detecting unseen generators without fine-grained attribution among seen classes (e.g., KNN-only); and (iii) *Unified* models, which jointly perform attribution and open-set detection. **SIGNAL** belongs to the third category and explicitly combines GNN-based attribution with KNN-based open-set reasoning. Table 2 reports all three settings, enabling direct comparison of the trade-offs between attribution accuracy, open-set detection, and their joint optimization. The standalone KNN branch improves over CNN/FCN baselines on ID/OOD detection—for example, Mamba-B achieves an ID/OOD EER of 22.83%, compared to 26.27% from CNN. The GNN head further enhances attribution performance by modeling class-level relations, and improves F1 and EER for several PTMs including Whisper and Wav2Vec 2.0. Notably, Whisper’s EER drops from 5.52% (KNN) to 4.37% (GNN) on the DEV set. The combined GNN+KNN configuration—our proposed **SIGNAL** framework—achieves the best performance across all settings. With Mamba-B, **SIGNAL** attains ACC: 98.11%, F1: 96.86%, and EER: 2.33% on the DEV set, while demonstrating strong generalization to unseen generators with ID/OOD ACC: 88.91% and

EER: 14.78%. Even smaller models like Mamba-T show noticeable gains under this hybrid setup, validating the complementary strengths of graph reasoning and local neighborhood-based inference.

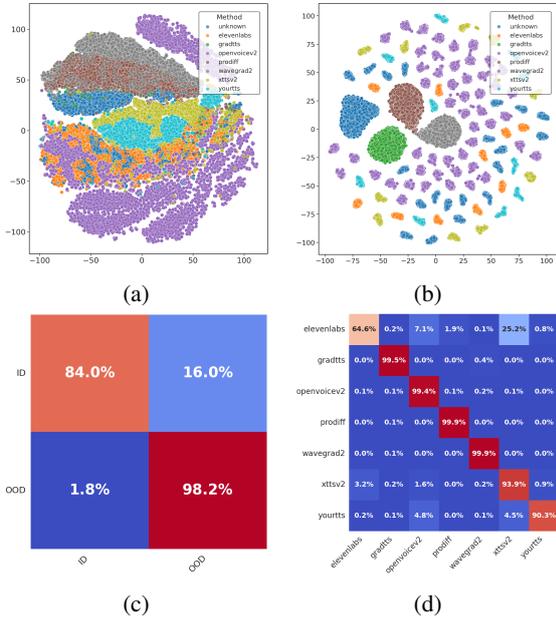


Figure 2: Subfigure a and b depict t-SNE : (a) shows the raw embedding space on the DiffSSD test set, while (b) illustrates enhanced class separation after GNN-based refinement. While subfigure c and d present confusion matrices (c) shows ID/OOD separation on DiffSSD using the GNN+KNN ensemble, and (d) shows full test set attribution performance.

PTM	DEV			ID/OOD		
	ACC	F1	EER	ACC	F1	EER
FCN						
Whisper	69.62	67.42	10.64	58.16	56.67	26.79
Unispeech	48.59	47.11	20.35	56.31	54.08	33.89
x-vector	71.12	69.73	12.16	<u>64.26</u>	<u>62.36</u>	32.61
wav2vec 2.0	66.25	64.56	11.28	57.45	54.34	28.50
wavLM	42.81	41.19	23.30	59.25	57.13	34.11
ECAPA	70.85	68.34	13.98	62.58	60.24	36.22
MAMBA-T	70.92	69.46	9.04	60.62	59.37	30.52
MAMBA-S	79.56	78.02	6.22	63.76	61.52	25.22
MAMBA-B	80.39	79.14	5.49	65.77	63.28	23.14
CNN						
Whisper	71.64	68.91	8.66	61.03	60.17	21.26
Unispeech	50.05	49.88	18.05	59.16	58.28	27.91
x-vector	73.04	71.33	10.23	<u>67.39</u>	<u>66.58</u>	25.80
wav2vec 2.0	67.71	66.17	9.38	60.74	59.95	24.52
wavLM	45.67	44.05	20.67	62.65	62.02	26.13
ECAPA	72.34	70.12	11.28	65.89	64.90	29.89
MAMBA-T	72.55	70.99	8.29	63.74	62.97	25.12
MAMBA-S	<u>81.64</u>	<u>80.16</u>	<u>5.01</u>	67.05	66.12	21.34
MAMBA-B	83.27	82.63	4.26	69.01	68.12	17.97

Table 3: Performance of various Pretrained Models (PTMs) on the SingFake dataset using FCN and CNN classifiers. Models are trained on DiffSSD and directly evaluated on SingFake (zero-shot setting).

These outcomes reaffirm that success in synthetic

speech detection depends on more than just model size or architecture—it requires the right pairing of representations and reasoning. **SIGNAL** reflects this principle by effectively combining structured graph learning and local similarity-based inference. Furthermore, Figure 2 provides a comprehensive visual analysis of **SIGNAL** behavior. Subfigures 2a and 2b present t-SNE projections of the learned embedding space. In 2a, the raw DiffSSD test embeddings show overlapping regions across generator classes, indicating limited separability. After GNN-based refinement in 2b, the clusters become more compact and well-separated, highlighting the GNN’s effectiveness in modeling inter-class relationships for better attribution. Subfigures 2c and 2d show confusion matrices under ID/OOD and full test settings, respectively.

Additional Experiments: To evaluate the generalization capability of our framework beyond synthetic speech, we experiment on SingFake (Zang et al., 2024), a benchmark dataset for singing voice deepfake detection (SVDD). The corpus comprises 28.93 hours of bonafide and 29.40 hours of deepfake clips across five languages and 40 singers, including diverse generative models and musical contexts. We evaluate **SIGNAL** in a zero-shot setting—models trained on DiffSSD are directly tested on SingFake without any fine-tuning. From Table 3, we observe that baseline CNN classifiers achieve moderate performance, with Mamba-B attaining the best results among them (F1: **82.63%**, EER: **17.97%** on ID/OOD). However, Table 4 shows that our hybrid **SIGNAL** architecture significantly enhances performance across all PTMs. Mamba-B again leads with the highest F1-score of **90.99%** and the lowest EER of **7.92%** under the ID/OOD split, demonstrating strong generalization to out-of-domain singing data. Even smaller models such as Mamba-S benefit considerably, reducing EER to **10.38%** with the hybrid setup. These findings affirm the robustness and transferability of **SIGNAL** to domains with different acoustic structures, such as musical and multilingual content.

4.4 Ablation Study

To evaluate the impact of threshold selection in our GNN+KNN hybrid architecture, we perform an ablation analysis on both the DiffSSD and SingFake datasets using the best-performing representation, Mamba-B. As described earlier, our model applies a confidence-based routing mechanism, where samples with prediction confidence above a threshold

PTM	DEV			ID/OOD			DEV			ID/OOD			DEV			ID/OOD		
	ACC ↑	F1 ↑	EER ↓	ACC ↑	F1 ↑	EER ↓	ACC ↑	F1 ↑	EER ↓	ACC ↑	F1 ↑	EER ↓	ACC ↑	F1 ↑	EER ↓	ACC ↑	F1 ↑	EER ↓
	KNN						GNN						KNN + GNN					
Whisper	75.51	74.31	7.57	62.97	61.04	18.26	77.45	76.44	6.17	64.98	64.07	16.19	85.52	84.41	4.92	70.46	69.61	14.13
Unispeech	70.27	69.76	17.71	61.19	60.26	19.17	72.42	71.34	15.48	63.01	62.26	17.08	80.38	79.17	14.73	68.32	67.36	15.67
X-vector	76.39	75.76	9.54	67.31	66.34	16.35	78.85	77.91	8.73	71.77	70.26	15.36	87.34	86.29	7.18	77.64	76.86	14.88
wav2vec 2.0	71.77	70.13	9.01	61.72	60.19	15.61	73.45	71.97	8.31	64.65	63.81	14.34	80.89	79.76	6.64	70.06	69.15	13.83
wavLM	74.57	73.01	17.83	63.44	61.52	18.87	76.54	75.07	15.66	66.79	66.12	16.27	85.17	84.32	14.39	72.41	71.33	16.02
ECAPA	73.89	72.11	10.32	68.28	65.22	21.36	78.22	77.18	9.74	70.05	69.27	19.56	86.61	85.48	8.75	76.04	75.20	19.07
MAMBA-T	76.65	75.32	7.99	64.15	62.78	15.25	78.22	77.44	7.38	67.97	67.15	14.28	87.14	86.19	6.76	73.96	72.29	13.92
MAMBA-S	84.11	83.28	4.87	67.64	66.91	13.74	86.12	85.16	4.58	71.42	70.42	11.96	89.29	87.95	4.29	77.26	76.49	10.38
MAMBA-B	86.79	84.78	4.19	70.14	69.07	10.71	87.20	86.04	4.18	73.47	72.52	9.52	92.28	90.99	3.81	79.66	78.55	7.92

Table 4: Performance of different PTMs on the SingFake dataset using KNN, GNN, and the proposed GNN+KNN hybrid **SIGNAL**.

0.1	6.50	18.50	5.10	12.40
0.2	5.80	17.20	4.60	10.60
0.3	5.10	16.10	4.20	9.30
0.4	4.50	15.00	3.70	8.40
0.5	4.32	14.78	3.81	7.90
0.6	4.29	14.79	3.73	8.10
0.7	5.40	16.30	4.60	8.90
0.8	6.20	17.90	5.30	10.10
0.9	7.00	19.60	6.00	11.50
	DIFSSD ID	DIFSSD ID/OOD	SINGFAKE IO	SINGFAKE IO/OOD

Figure 3: Threshold τ sensitivity of **SIGNAL** (KNN+GNN) with Mamba-B.

τ are classified as originating from seen generators, while those below τ are flagged as unseen. Figure 3 presents the Equal Error Rate (EER) across a range of threshold values ($\tau \in [0.1, 0.9]$), evaluated on four evaluation splits: DiffSSD ID, DiffSSD ID/OOD, SingFake ID, and SingFake ID/OOD. The results clearly indicate that $\tau = 0.5$ offers a balanced trade-off—minimizing EER on both in-distribution and open-set splits. This analysis validates our design choice of setting $\tau = 0.5$ as the open-set decision boundary in **SIGNAL**.

5 Conclusion

This work address the dual challenge of synthetic speech attribution and detection of unseen generators. We show that Mamba-based representations are highly effective in capturing generator-specific traits, including prosodic patterns and synthesis artifacts, owing to their advanced temporal modeling. Building on this insight, we introduce a hybrid framework **SIGNAL** that combines graph-

based relational modeling with instance-level KNN inference to exploit both global structure and local similarity. Our study highlights the overlooked potential of graph-enhanced modeling for robust synthetic speech detection, providing a strong foundation for advancing attribution and generalization in generative speech forensics. This work sets a solid baseline for future research in structured and generalizable synthetic speech analysis.

6 Limitations and Future Work

First, our evaluation is conducted on two publicly available benchmarks (DiffSSD and SingFake), which, while diverse, do not cover all possible synthesis conditions or generators. Second, **SIGNAL** relies on a decision threshold for open-set detection; although empirically stable, threshold selection remains a challenge in fully unconstrained deployment scenarios. Third, our current framework

identifies whether a sample originates from an unseen generator but does not perform fine-grained attribution among multiple unseen sources. Extending SIGNAL toward hierarchical or cluster-based modeling of unseen generators is a promising direction for future work.

Interpretability and Explainability: Although the framework achieves strong performance, its decisions remain somewhat opaque. The current framework lacks mechanisms for explaining why a particular speech sample is attributed to a specific generator or flagged as unseen. This limits its usability in high-stakes scenarios where transparency is essential. Enhancing interpretability—through attention visualization or prototype-level explanations—is an important direction for future work.

Threshold sensitivity: SIGNAL relies on a decision threshold to balance attribution and open-set detection. While Figure 3 shows stable performance across a wide range of values, we acknowledge that no fixed threshold is universally optimal under all real-world deployment conditions. Adaptive or data-driven threshold selection remains an important direction for future work.

7 Ethical Statement

This study addresses the growing concern around synthetic speech and singing voice generation by proposing methods to detect and attribute artificially generated content. We use only publicly available datasets (DiffSSD and SingFake), and no personal or sensitive data is involved. Our framework is developed purely for detection and research purposes—not for generating or misusing synthetic content.

References

- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.
- Kratika Bhagtani, Amit Kumar Singh Yadav, Paolo Bestagini, and Edward J Delp. 2025. Diffssd: A diffusion-based dataset for speech forensics. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, and Xiangzhan Yu. 2021. Unispeech-sat: Universal speech representation learning with speaker aware pre-training. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6152–6156.
- Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi. 2020. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6184–6188. IEEE.
- Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. *Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification*. In *Interspeech 2020*, pages 3830–3834.
- Falih Gozi Febrinanto, Kristen Moore, Chandra Thapa, Jiangang Ma, Vidya Saikrishna, and Feng Xia. 2025. Vision graph non-contrastive learning for audio deepfake detection with limited labels. *arXiv preprint arXiv:2501.04942*.
- Albert Gu and Tri Dao. 2024. *Mamba: Linear-time sequence modeling with selective state spaces*. In *First Conference on Language Modeling*.
- Yinlin Guo, Haofan Huang, Xi Chen, He Zhao, and Yuehai Wang. 2024. Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12702–12706. IEEE.
- Lian Huang and Chi-Man Pun. 2024. Self-attention and hybrid features for replay and deep-fake audio detection. *arXiv preprint arXiv:2401.05614*.
- Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. 2024. Natural-speech 3: zero-shot speech synthesis with factorized codec and diffusion models. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. *Diffwave: A versatile diffusion model for audio synthesis*. In *International Conference on Learning Representations*.
- Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and 1 others.

2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034.
- Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6706–6713.
- Jagabandhu Mishra, Manasi Chhibber, Hye-jin Shim, and Tomi H Kinnunen. 2025. Towards explainable spoofed speech attribution and detection: A probabilistic approach for characterizing speech synthesizer components. *Computer Speech & Language*, page 101840.
- Orchid Chetia Phukan, Girish, Mohd Mujtaba Akhtar, Swarup Ranjan Behera, Priyabrata Mallick, Pailla Balakrishna Reddy, Arun Balaji Buduru, and Rajesh Sharma. 2025. [Towards Source Attribution of Singing Voice Deepfake with Multimodal Foundation Models](#). In *Interspeech 2025*, pages 1673–1677.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. [Fastspeech 2: Fast and high-quality end-to-end text to speech](#). In *International Conference on Learning Representations*.
- Hyun-seo Shin, Jungwoo Heo, Ju-ho Kim, Chan-yeong Lim, Wonbin Kim, and Ha-Jin Yu. 2024. Hm-conformer: A conformer-based audio deepfake detection system with hierarchical pooling and multi-level classification token aggregation methods. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10581–10585. IEEE.
- David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.
- Adriana Stan, David Combei, Dan Oneata, and Horia Cucu. 2025. [TADA: Training-free Attribution and Out-of-Domain Detection of Audio Deepfakes](#). In *Interspeech 2025*, pages 1543–1547.
- Yun Tan, Xiaoqian Weng, and Jiangzhang Zhu. 2025. Dual-channel spoofed speech detection based on graph attention networks. *Symmetry*, 17(5):641.
- Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi H. Kinnunen, and Kong Aik Lee. 2019. [Asvspoof 2019: Future horizons in spoofed and fake audio detection](#). In *Interspeech 2019*, pages 1008–1012.
- Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Haniççi, Md Sahidullah, and Aleksandr Sizov. 2015. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association*, pages 2037–2041. International Speech Communication Association.
- Sarthak Yadav and Zheng-Hua Tan. 2024. [Audio mamba: Selective state spaces for self-supervised audio representations](#). In *Interspeech 2024*, pages 552–556.
- Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023. [Audio deepfake detection: A survey](#). *ArXiv*, abs/2308.14970.
- Yongyi Zang, You Zhang, Mojtaba Heydari, and Zhiyao Duan. 2024. Singfake: Singing voice deepfake detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12156–12160. IEEE.