Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multitask Deep Learning for Joint Deepfake Speech Detection and Spoofing Attack Attribution

Abdulrahman Hussien
Sana'a University
Ghaleb Aljafari
Sana'a University
Mohammed Nasser Al-Andoli
nasser.alandoli@mmu.edu.my
Multimedia University
Malek Aljabri
Sana'a University

Research Article

Keywords: Deepfake Speech Detection, Spoof Attribution, Multi-task Learning, Spectro-Temporal Features, Bidirectional Long Short-Term Memory (Bi-LSTM), Transformer Encoder, Attention Mechanism, Voice Biometrics, Digital Forensics

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11257420/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multitask Deep Learning for Joint Deepfake Speech Detection and Spoofing Attack Attribution

*Abdulrahman Hussien[1], Ghaleb Aljafari[1], Mohammed Nasser Al-Andoli[2, *], and Malek Aljabri[1]*

[1] Computer Science, Computer and information Technology, Sana'a University, Sana'a, Yemen.

[2] Centre for Cybersecurity and Quantum Computing,CoE for Advanced Cloud, Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia.

**Corresponding author: nasser.alandoli@mmu.edu.my*

ABSTRACT

The rapid advancement of neural speech synthesis and voice conversion technologies has increased the realism of synthetic speech, creating significant challenges for biometric authentication and digital forensics. This study proposes a Multi-Task Deep Learning (MT-DL) framework that jointly addresses binary deepfake speech detection and supervised spoof attribution of known attack classes. The framework combines complementary spectro-temporal representations (MFCC, Log-Mel spectrograms, and delta-MFCC features) with a convolutional front-end and a hybrid Bi-LSTM/Transformer temporal encoder to feed task-specific detection and attribution heads. Evaluated on the ASVspoof 2019 Logical Access dataset using a stratified 60:40 split, the framework achieves a binary detection accuracy of 99.75% (0.48% EER) and a supervised spoof attribution accuracy of 99.57% (99.58% Macro-F1) across known attacks (A01–A06). Unseen attacks (A07–A19) and cross-dataset evaluations on the In-the-Wild dataset further demonstrate robust generalization, though performance sensitivity under severe additive acoustic noise remains a key area for future optimization.

Keywords: Keywords: Deepfake Speech Detection; Spoof Attribution; Multi-task Learning; Spectro-Temporal Features; Bidirectional Long Short-Term Memory (Bi-LSTM); Transformer Encoder; Attention Mechanism; Voice Biometrics; Digital Forensics.

1. Introduction:

Neural speech synthesis has rapidly advanced with the development of deep learning, enabling Text-to-Speech (TTS), Voice Conversion (VC), and neural vocoder systems to produce speech that can closely resemble genuine human recordings [1], [4]. Although these technologies provide important benefits for assistive applications, personalized media, and human–computer interaction, their increasing realism also creates significant security and forensic concerns. Synthetic speech can be exploited for speaker impersonation and attacks against voice authentication systems, making reliable speech deepfake detection increasingly important [1].

Existing research has largely focused on binary detection, where speech is classified as either bona fide or spoofed. The ASVspoof initiative has played a central role in this area by providing standardized datasets and evaluation protocols. In particular, ASVspoof 2019 introduced Logical Access (LA) and Physical Access (PA) scenarios and included speech synthesis, voice conversion, and replay attacks within the overall challenge [1]. However, identifying whether an utterance is fake does not necessarily provide sufficient information for forensic analysis. In practical applications, it may also be necessary to determine the type or source of the spoofing attack. This has motivated the emergence of audio deepfake attribution as an extension of conventional binary detection. Yan et al. introduced the Audio Deepfake Attribution (ADA) dataset and demonstrated that identifying the generation source is an important problem beyond binary classification, including the challenge of previously unseen generation systems [5].

Another important issue concerns the representation of speech. Conventional acoustic features such as MFCCs can capture useful spectral characteristics, while modern end-to-end systems attempt to learn discriminative representations directly from speech signals. RawNet2 demonstrated the potential of raw-waveform-based learning for spoofing countermeasures, showing that neural models can learn cues that may not be captured by traditional feature-based systems [2]. Similarly, AASIST explicitly modeled artifacts occurring across spectral and temporal intervals through an integrated spectro-temporal graph attention architecture [3]. These developments indicate that spoofing artifacts may exist in both spectral and temporal domains, motivating the joint exploitation of complementary spectro-temporal information rather than relying on a single representation.

Generalization to unseen conditions remains another major challenge. ASVspoof 2021 introduced additional channel and compression variability and included a dedicated deepfake speech detection task. The challenge emphasized conditions in which the nature of spoofed and deepfake speech cannot be reliably predicted in advance, making robustness and generalization important requirements for practical deployment [4]. Similarly, Müller et al. investigated whether existing audio deepfake detectors generalize to real-world speech and introduced an In-the-Wild dataset containing 37.9 hours of audio, including 17.2 hours of deepfake speech. Their evaluation showed substantial degradation when systems developed around benchmark conditions were applied to real-world recordings [6]. These findings highlight the difference between strong within-dataset performance and reliable performance under variable environments.

The rapid expansion of speech-generation technologies further increases this challenge. Recent large-scale resources such as SpeechFake contain more than 3 million deepfake samples, exceeding 3,000 hours of audio, generated using 40 speech synthesis tools and covering 46 languages [7]. This diversity demonstrates that modern detectors must cope with increasingly heterogeneous generation mechanisms rather than a fixed set of spoofing systems. Consequently, robust detection should be evaluated under both matched and mismatched conditions, including unseen generators, datasets, and environmental distortions.

Motivated by these limitations, this research proposes a Multi-task Deep Learning Framework for Deepfake Speech Detection and Spoof Attribution Using Spectro-Temporal Features. The proposed framework jointly addresses two complementary objectives: binary deepfake speech detection and fine-grained spoof attribution. Spectral representations, including Mel-Spectrogram and MFCC and Δ MFCC, are combined with temporal representations derived from the raw waveform and sequential embeddings. A hybrid architecture incorporating CNN, Bi-LSTM, and Transformer-based attention mechanisms is designed to capture local spectral characteristics, sequential dependencies, and long-range temporal relationships. The resulting shared representation is connected to task-specific detection and attribution heads, allowing the two objectives to be learned within a unified multi-task framework.

The proposed approach is evaluated primarily using ASVspoof 2019 Logical Access, where spoofing systems provide a controlled basis for attribution [1], [7]. and is further examined using In-the-Wild data to investigate generalization under realistic recording conditions [6], [7]. Robustness is also evaluated under controlled variable environments involving 20-dB noise, 5-dB noise, reverberation, time stretching, and +2 pitch shifting. This evaluation is intended to determine whether the learned spectro-temporal representation can retain discriminative information when the acoustic characteristics of the input differ from those observed during training.

Accordingly, the main contributions of this research are:

- A unified multi-task framework for simultaneous deepfake speech detection and spoof attribution.
- Hybrid spectro-temporal feature learning through complementary spectral and temporal representations.

- A CNN–Bi-LSTM–Transformer architecture with attention-based feature integration.
- Fine-grained spoof attribution beyond conventional bona fide/spoof binary classification.
- Systematic robustness evaluation under noise, reverberation, time stretching, and pitch-shift conditions.
- Cross-dataset generalization analysis using controlled benchmark and real-world speech data.
- Comprehensive evaluation and ablation studies to determine the contribution of individual representations and architectural components.

The remainder of this paper is organized as follows. Section 2 reviews the related work on deepfake speech detection, spoof attribution, Spectro-temporal representation learning, and cross-domain generalization. Section 3 presents the proposed methodology and materials, including the preprocessing pipeline, feature extraction, multi-task deep learning architecture, optimization strategy, and training procedure. Section 4 reports the experimental results and performance evaluation, covering the datasets and experimental setup, feature configuration experiments, spoof attack attribution, cross-dataset generalization, robustness analysis, and model interpretation. Section 5 discusses the main findings, limitations, and implications of the proposed framework. Finally, Section 6 concludes the paper and highlights potential directions for future research.

2. Related Work

Early research on speech spoofing primarily focused on protecting automatic speaker verification systems against synthetic, converted, and replayed speech. With the development of deep neural networks, end-to-end architectures have increasingly replaced conventional handcrafted feature pipelines. Tak *et al.* proposed RawNet2, an end-to-end anti-spoofing architecture that operates directly on raw speech waveforms and demonstrated the ability of deep models to learn discriminative spoofing representations without relying exclusively on handcrafted acoustic features [8]. This work established the potential of raw-waveform learning for speech anti-spoofing and motivated subsequent research toward increasingly learned representations.

Jung et al. introduced AASIST, an integrated spectro-temporal graph attention architecture designed to model spoofing artifacts across both spectral and temporal dimensions [9]. Unlike approaches that process spectral information independently, AASIST explicitly models interactions between spectral and temporal characteristics. Its results demonstrated that spoofing artifacts can be distributed across different regions of the speech representation and that jointly modeling these dependencies can improve anti-spoofing performance.

The ADD 2022 challenge further broadened the scope of audio deepfake detection by introducing challenging scenarios involving low-quality audio, partially fake speech, and diverse real-world distortions [10]. These tasks demonstrated that deepfake detection cannot be adequately characterized by clean binary classification alone, particularly when synthetic speech is affected by noise and other signal degradations. However, most challenge-oriented approaches still primarily formulate the problem as detection rather than jointly learning detection and source attribution.

Speech deepfake artifacts can occur at different levels of the acoustic signal. Spectral representations such as Mel-Spectrograms and MFCCs capture frequency-dependent characteristics, while temporal representations can describe changes in speech patterns across successive frames. Consequently, recent research has increasingly investigated architectures capable of learning both types of information.

AASIST [9] represents an important step toward spectro-temporal modeling by constructing a graph representation in which spectral and temporal information interact through attention mechanisms. This approach demonstrates that modeling the relationship between acoustic dimensions can provide more discriminative information than treating spectral and temporal characteristics independently.

More recently, Li and Zhang proposed an interpretable temporal class activation representation for audio spoofing detection [11]. Their approach employs wav2vec 2.0 representations together with

attentive utterance-level features and explicitly identifies discriminative temporal regions contributing to the spoofing decision. Importantly, their experiments showed that training with spoofing-type information rather than only binary bona fide/spoof labels can improve the ability of the model to learn attack-specific characteristics [11]. This finding is particularly relevant to the present research because it provides evidence that auxiliary spoof-type information can complement binary detection.

However, these approaches generally investigate either sophisticated representation learning or attack-aware detection, rather than establishing a unified framework in which a shared spectro-temporal representation is explicitly optimized for both binary detection and fine-grained spoof attribution.

Binary detection determines whether speech is genuine or synthetic but does not explain which generation mechanism produced the spoofed signal. This limitation has motivated the development of audio deepfake attribution as a distinct research direction.

Yan et al. introduced the Audio Deepfake Attribution (ADA) dataset and investigated the problem of identifying the source generation tool of synthetic speech [12]. Their work demonstrated that different generation systems leave distinguishable fingerprints in audio signals and proposed Class-Representation Multi-Center Learning for open-set audio deepfake attribution. The study is particularly significant because it extends the task from binary detection toward identifying known and potentially unknown generation tools[12].

Klein et al. subsequently investigated source tracing of audio deepfake systems using ASVspoof 2019 LA and MLAAD ,Rather than restricting attribution to a single attack label, their framework considered different components of the deepfake generation pipeline, including input type, acoustic model, and vocoder. Their results demonstrate that attribution can be formulated at a finer level than simply classifying an utterance as TTS or VC. This is closely related to the present research, while also revealing an important methodological consideration: the definition of the attribution target must be explicit because attack family, generation system, and generation component represent different levels of attribution[13].

Ge et al. addressed the Speech Spoofing Algorithm Recognition (SSAR) problem under few-shot conditions using a Dual Graph Prototypical Network ,Their work is particularly relevant to the unseen-generator aspect of the present research because newly emerging spoofing algorithms may have limited training samples. Rather than relying solely on conventional supervised classification, their approach attempts to exploit relationships among speech representations and classes to improve recognition with limited examples[14], [15].

Müller et al. investigated the generalization problem and demonstrated that the performance gap between in-domain and out-of-domain deepfake detection is strongly related to differences between generation systems rather than simply the difficulty of detecting newer deepfakes [16],[17]. This finding challenges the assumption that increasing model capacity alone is sufficient to solve generalization and highlights the importance of learning representations that capture generation-independent artifacts.

The MLAAD dataset was introduced to address limitations in language and generator diversity [18]. It contains synthetic speech generated using 82 TTS models representing 33 architectures, covering 38 languages and approximately 378 hours of synthetic speech. Experiments across multiple datasets demonstrated that dataset diversity has a substantial influence on cross-domain detection performance [19]. This evidence supports the need for evaluating deepfake detectors beyond a single benchmark.

Li et al. further examined cross-domain audio deepfake detection using more than 300 hours of speech generated by five advanced zero-shot TTS systems [18] ,[19]. Their experiments showed that modern models can still experience substantial domain shifts and reported EERs of 4.1% for Wav2Vec2-large and 6.5% for Whisper-medium under their cross-domain setting. Their findings emphasize that even powerful self-supervised representations do not automatically guarantee robust generalization.

WaveFake provides another important resource for evaluating generalization across generation architectures. The dataset contains more than 100,000 generated audio clips produced using multiple neural vocoder architectures and languages [20]. Such diversity makes it useful for assessing whether a detector learns genuine synthesis artifacts or merely memorizes characteristics associated with a particular generator.

More recently, research has begun to investigate open-set source tracing, where the objective is not only to recognize known generation systems but also to reject or identify previously unseen systems. This direction is particularly important because the number of available speech generation systems continues to increase. Open-set source tracing therefore represents a more realistic formulation of attribution than conventional closed-set classification [21].

The reviewed studies demonstrate substantial progress in three complementary directions. First, end-to-end and spectro-temporal architectures have significantly improved deepfake speech detection by learning increasingly discriminative representations [8], [9]. Second, attribution-oriented studies have demonstrated that synthetic speech contains fingerprints that can be exploited to identify generation systems and their components [19],[21]. Third, cross-domain and open-set studies have established that strong performance under matched conditions does not necessarily translate into reliable performance for unseen generators, datasets, languages, or acoustic environments [22]

Nevertheless, several limitations remain.

First, most detection systems optimize the bona fide/spoof decision as the primary objective, while attribution is frequently treated as a separate classification problem. Second, existing attribution studies differ substantially in the definition of the attribution target, ranging from attack families to generation tools and individual generation components. Third, generalization and attribution are often evaluated independently, although unseen generators represent one of the most important challenges for practical deployment. Fourth, the multi-task architecture was designed to encourage shared representation learning while maintaining task-specific output heads. However, explicit mitigation of negative transfer was not separately evaluated in this study. Finally, environmental robustness is commonly investigated through separate augmentation or cross-dataset experiments rather than through a unified evaluation framework combining noise, reverberation, temporal modification, pitch variation, cross-dataset testing, and unseen-generator attribution[23].

Therefore, the literature suggests a clear research opportunity for a unified framework that combines spectro-temporal representation learning, multi-task detection and attribution, robustness to variable environments, and cross-domain/open-set generalization. The proposed doctoral research addresses this opportunity by jointly learning binary deepfake detection and fine-grained spoof attribution within a shared deep representation, while systematically evaluating the interaction between the two tasks under matched, mismatched, noisy, and unseen-generator conditions.

3. Methodology and Materials

The proposed multi-task deep learning framework for deepfake speech detection and spoof attribution begins with a preprocessing stage, where input speech signals are standardized and segmented into suitable audio samples to ensure consistent model training. The feature extraction stage then combines complementary spectral and temporal representations, including Mel-Spectrogram and MFCC and Δ MFCC features for capturing spectral characteristics, together with raw waveform and sequential embeddings for modeling temporal information. These representations are processed through a hybrid deep learning architecture that integrates CNNs, Bi-LSTM networks, and Transformer-based attention mechanisms to learn comprehensive spectro-temporal representations. The learned shared representation is subsequently passed to two task-specific heads: a binary detection head for distinguishing bonafide and spoofed speech, and an attribution head for identifying the spoofing source or attack category. The multi-task architecture was designed to encourage shared representation learning while maintaining task-

specific output heads. However, explicit mitigation of negative transfer was not separately evaluated in this study. Model performance is evaluated using comprehensive metrics, including Accuracy, Precision, Recall, F1-score, Equal Error Rate (EER), ROC-AUC, and Macro-F1. Furthermore, robustness and generalization are investigated under variable conditions, including noise at 20 dB and 5 dB, reverberation, time stretching, pitch shifting (+2), cross-dataset evaluation, and unseen spoof generators.

The proposed multi-task deep learning framework for deepfake speech detection and spoof attribution begins with dataset preparation and audio preprocessing to ensure consistent input representation, as illustrated in Figure 1. Complementary spectro-temporal features, including MFCC, Mel-Spectrogram, raw waveform, and sequential embeddings, are extracted and integrated within a hybrid architecture combining CNN, Bi-LSTM, and Transformer-based attention mechanisms. The learned shared representation is connected to two task-specific heads: a binary detection head for bonafide and spoofed speech, and a spoof attribution head for identifying TTS, VC, Hybrid VC/TTS. The framework is optimized using a multi-task learning objective to promote shared representation learning while preserving task-specific outputs. After training and validation, the optimized model is evaluated on unseen test data and under diverse conditions, including additive noise, reverberation, time stretching, pitch shifting, cross-dataset evaluation, and unseen-generator scenarios. Explicit mitigation of negative transfer was not separately evaluated in this study.

3.1. Preprocessing

Prior to the training and testing phases, the raw speech signals undergo a series of preprocessing steps to ensure consistent, clean, and suitable audio inputs for the proposed deep learning framework. The overall preprocessing pipeline is illustrated in Figure 1. First, all audio recordings are resampled to a fixed sampling rate of 16 kHz, ensuring consistency across different datasets and recording conditions. The signals are then trimmed or padded to a fixed duration of 5 seconds, allowing the model to process inputs with a uniform temporal length. Subsequently, audio normalization is applied to standardize the signal amplitude and improve the stability and convergence of the training process.

To reduce the influence of unwanted background interference, an optional noise-reduction stage is considered before the subsequent processing steps. Data augmentation is then applied only to the training data to improve the robustness and generalization capability of the proposed framework under variable acoustic conditions. The augmentation procedures include the addition of noise at 20 dB and 5 dB, reverberation, time stretching, and pitch shifting of +2 semitones. These transformations are designed to simulate realistic acoustic and signal-level variations that may occur during speech recording, transmission, and playback.

The processed speech signals are represented using complementary spectral and temporal representations. The spectral representations include Mel-Spectrogram and MFCC and Δ MFCC features, while temporal information is preserved through the raw waveform and sequential embeddings. These complementary representations are subsequently provided to the proposed spectro-temporal multi-task deep learning framework, where they are jointly exploited for binary deepfake speech detection and fine-grained spoof attribution. This preprocessing strategy is designed to establish a consistent and robust input representation while reducing the influence of recording variability and enhancing the model's ability to generalize to realistic and unseen conditions.

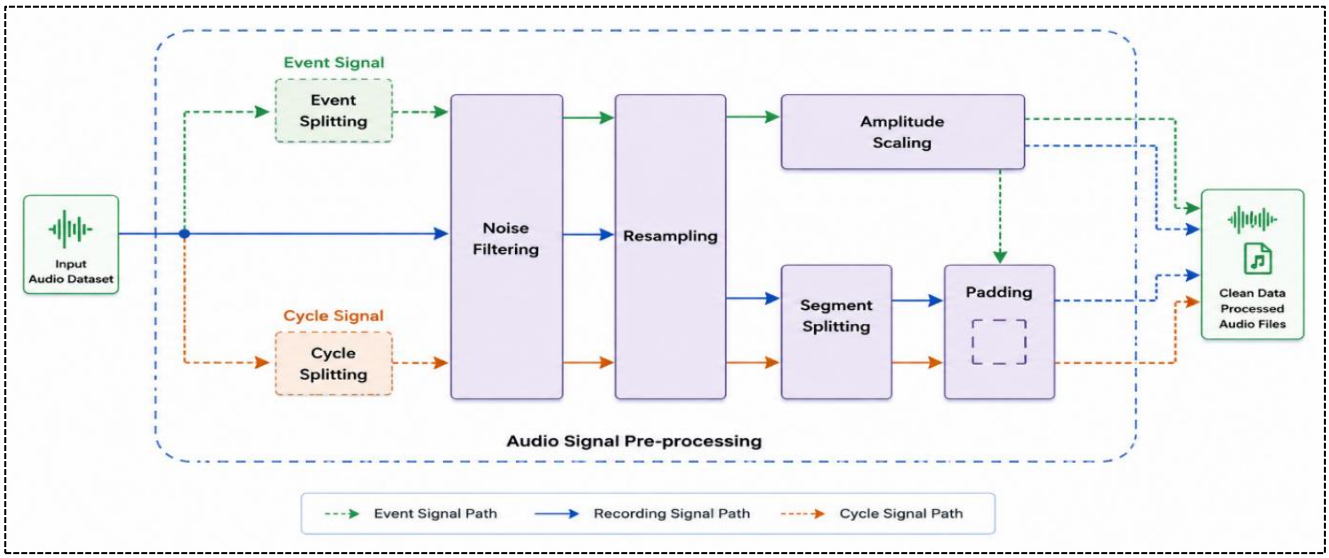


Figure 1. Audio signal preprocessing pipeline for waveform normalization, resampling, and augmentation.

3.2. Spectro-Temporal Feature Set

To construct a compact and informative representation of speech signals, the proposed framework employs a spectro-temporal feature set that combines Mel-Frequency Cepstral Coefficients (MFCCs), Log-Mel Spectrograms, and first-order Delta features. These complementary representations capture different characteristics of the speech signal, including spectral structure, detailed time-frequency patterns, and temporal variations. The speech signals are sampled at 16 kHz, while the extracted features are organized over 256 time steps Figure 2 , 3. The proposed feature configuration produces a 208-dimensional representation, formulated as:

$$F_{\mathrm{ST}} = \mathrm{MFCC}_{40} + \mathrm{Log-Mel}_{128} + \Delta\mathrm{MFCC}_{40} = 208 \qquad (1)$$

where**:**

• F_{ST} denotes the proposed spectro-temporal feature representation.

• $MFCC_{40}$ denotes the 40-dimensional Mel-frequency cepstral coefficient representation.

• $Log{-}Mel_{128}$ denotes the 128-dimensional log-Mel spectrogram representation.

• $\Delta MFCC_{40}$ denotes the 40-dimensional first-order temporal derivative of he MFCC features.

The selected features are described below.

Mel-Frequency Cepstral Coefficients (MFCC): are widely used spectral features for representing the short-term spectral characteristics of speech. They model the spectral envelope of the speech signal by transforming the frequency spectrum into the perceptually motivated Mel scale. In the proposed framework, 40 MFCC coefficients are extracted from each speech frame, resulting in a 40-dimensional feature vector.

Log-Mel Spectrogram: The Log-Mel Spectrogram provides a detailed time-frequency representation of the speech signal. It is obtained by applying a short-time Fourier transform, mapping the resulting spectral energy onto the Mel frequency scale, and then applying logarithmic compression. In the proposed framework, 128 Mel filters are used, producing a 128-dimensional spectral representation for each time step. Compared with MFCCs, the Log-Mel representation preserves more detailed local time-frequency

information, which can help the model identify acoustic artifacts and irregular spectral patterns associated with synthetic or manipulated speech.

Delta MFCC: Δ features represent the temporal dynamics of speech by measuring the first-order changes between consecutive feature frames. In this framework, the Delta features are derived from the 40 MFCC coefficients, resulting in an additional 40-dimensional temporal representation. These features capture frame-to-frame variations and temporal transitions that may differ between natural and artificially generated speech.

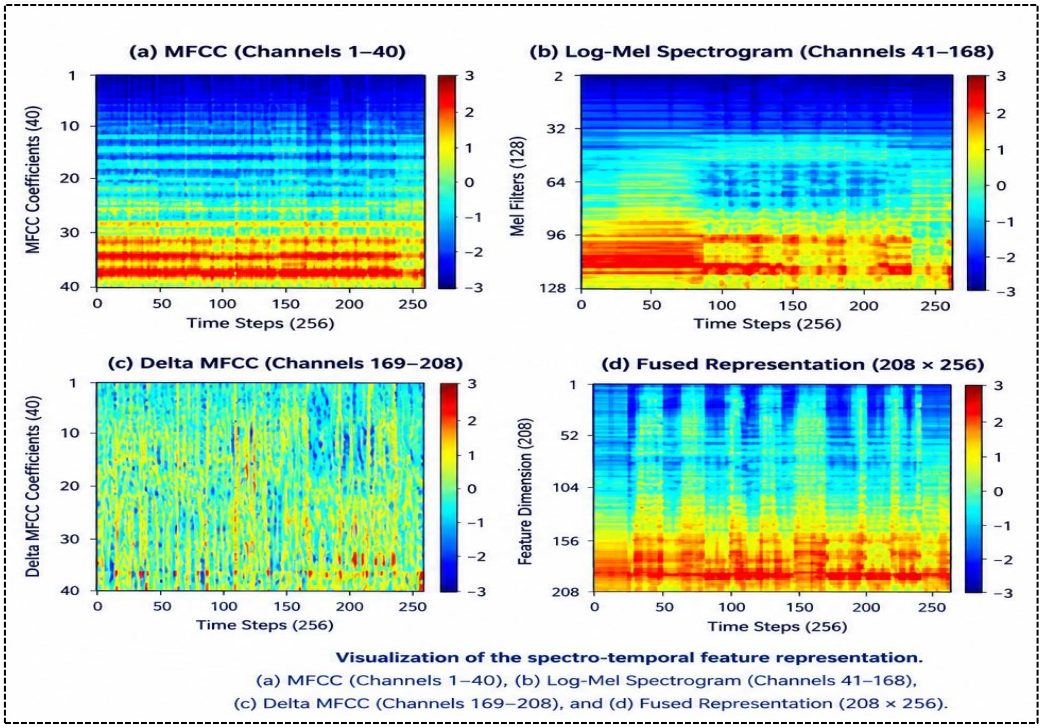


Figure 2. Spectro-temporal feature representation combining MFCC, Log-Mel, and Delta-MFCC channels over 256 time steps.

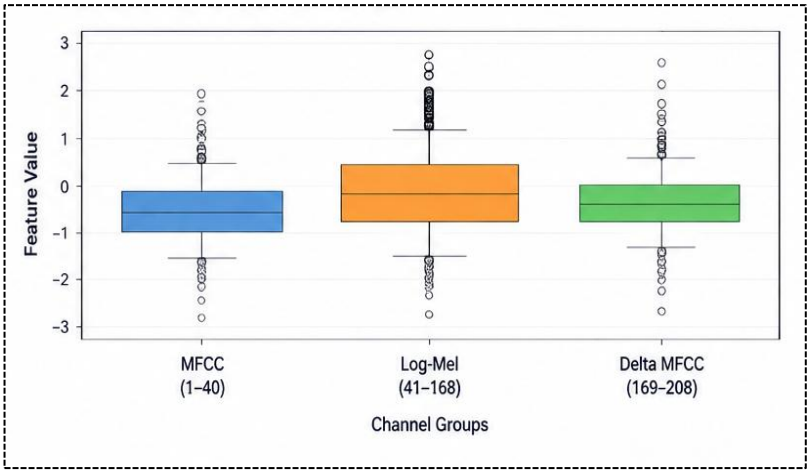


Figure 3. Channel-wise feature distribution across MFCC, Log-Mel, and Delta-MFCC feature groups.

3.3 Proposed Multi-task Spectro-Temporal Learning Framework

The proposed Framework integrates convolutional, recurrent, and Transformer-based deep learning components to jointly learn spectral and temporal representations for deepfake speech detection and spoof attribution. The overall architecture is illustrated in Figure 6. The model begins with raw speech audio, from which Mel-Spectrogram representations are generated to capture the spectral characteristics of the speech signal. The extracted spectrograms are then processed through a CNN-based spectral feature learning module consisting of five convolutional blocks. Each convolutional block applies a 3 × 3 convolution, followed by Batch Normalization (BN), ReLU activation, and 2 × 2 max-pooling to progressively extract hierarchical spectral representations. The number of convolutional filters increases from 32 to 64, 128, 256, and 256, enabling the network to learn increasingly complex acoustic patterns. A Global Average Pooling (GAP) operation is subsequently applied to obtain a compact spectral representation.

The extracted spectral representation is then passed to the temporal modeling module, which is designed to capture temporal dependencies within the speech signal. Two complementary temporal modeling mechanisms are considered: Bi-LSTM and a Transformer Encoder. The Bi-LSTM component processes the learned representations in both forward and backward temporal directions, allowing the model to capture dependencies across the speech sequence. Alternatively, the Transformer Encoder employs positional encoding and multi-head self-attention to model long-range temporal relationships within the speech signal. The resulting temporal representation is then combined with the spectral representation through the spectro-temporal fusion layer.

The proposed framework uses an attention-based fusion mechanism to integrate the spectral representation extracted by the CNN with the temporal representation generated by the temporal modeling module. The fusion layer learns adaptive weights for both representations and combines them into a unified spectro-temporal representation for subsequent binary spoof detection and spoof attack attribution. The fusion operation is defined as:

$$F_{ST} = \alpha_S S + \alpha_T T \tag{2}$$

where :$\boldsymbol{S}$ and $\boldsymbol{T}$ denote the spectral and temporal representations, respectively, while $\boldsymbol{\alpha_S}$ and $\boldsymbol{\alpha_T}$ are their learned attention weights.

3.3.1. CNN-Based Spectral Feature Learning

The five convolutional blocks in the proposed architecture progressively extract discriminative spectral features from the Mel-Spectrogram representation. Each convolutional block applies learnable filters to identify acoustic patterns at different levels of abstraction Figure 4.

- First convolutional block: Uses 32 filters to learn low-level spectral patterns and local acoustic structures from the Mel-Spectrogram.
- Second convolutional block: Uses 64 filters to capture more complex spectral patterns based on the representations learned by the first block.
- Third convolutional block: Uses 128 filters to learn deeper and more discriminative acoustic structures.
- Fourth convolutional block: Uses 256 filters to extract higher-level spectral representations and complex interactions within the speech signal.
- Fifth convolutional block: Uses 256 filters to further refine and consolidate the learned high-level spectral representations.

Each convolutional block employs a 3 × 3 convolution, Batch Normalization, ReLU activation, and 2 × 2 Max-Pooling. The convolution operation enables the network to learn local acoustic patterns, while Batch Normalization contributes to stable feature learning. The ReLU activation introduces non-linearity,

allowing the network to model complex relationships, and Max-Pooling progressively reduces the spatial dimensions while retaining the most informative spectral characteristics.

Following the five convolutional blocks, Global Average Pooling (GAP) is employed to transform the learned feature maps into a compact spectral representation..

The fused spectro-temporal representation is provided as input to the CNN module, where multiple convolutional layers automatically learn hierarchical spectral features from the speech signal. The convolutional layers progressively extract low-level acoustic information, including spectral edges and local frequency patterns, followed by more discriminative high-level representations associated with synthetic speech artifacts. Max-pooling layers reduce the spatial dimensions while preserving the most informative features

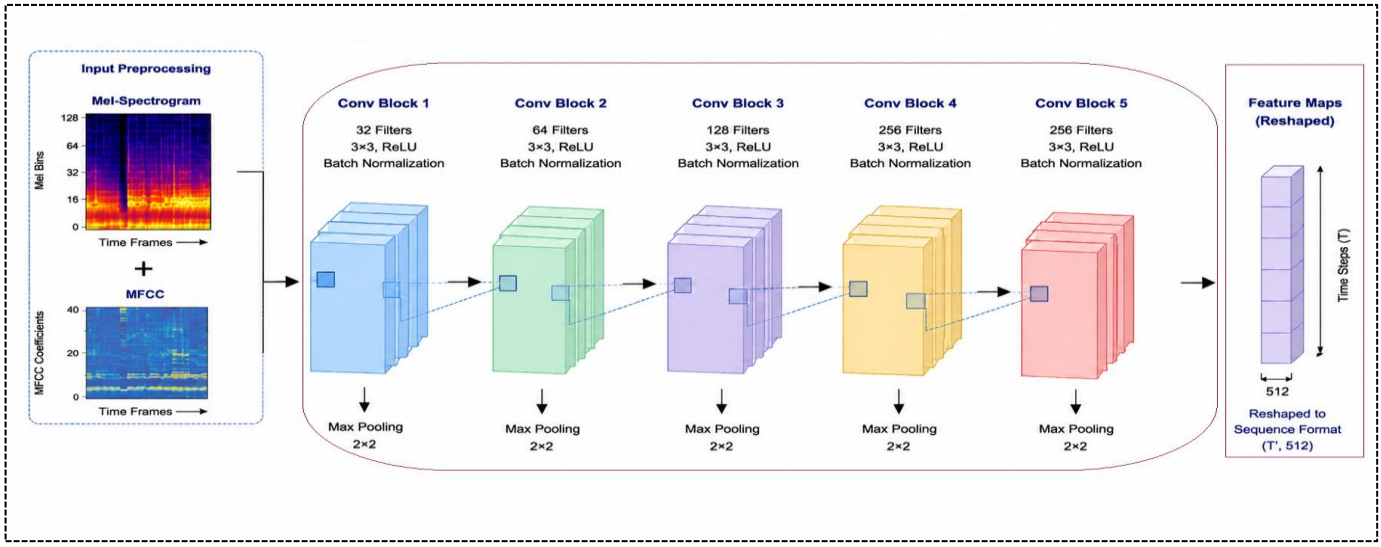


Figure 4. Architecture of the 5-block CNN spectral feature extraction network.

3.3.2. Temporal Dependency Modeling Using Bi-LSTM And Transformer

The temporal modeling stage is responsible for capturing sequential dependencies that cannot be fully represented through spectral feature extraction alone. The proposed architecture incorporates two complementary approaches: Bi-LSTM and Transformer Encoder.

Bi-LSTM: The Bi-LSTM processes the extracted sequence in both forward and backward directions. This enables the model to capture temporal dependencies from both past and future contexts within the speech sequence. By incorporating information from both directions, the Bi-LSTM provides a more comprehensive representation of temporal speech characteristics.

Transformer Encoder: The Transformer-based temporal module employs positional encoding followed by multi-head self-attention and feed-forward processing. The positional encoding provides information about the order of the speech sequence, while multi-head self-attention enables the model to identify relationships between different temporal positions. This mechanism allows the architecture to capture both local and long-range temporal dependencies within the speech representation.

The temporal modeling stage captures sequential dependencies that cannot be fully represented through spectral feature extraction alone. The proposed architecture employs two complementary temporal modeling approaches: a Bi-LSTM and a Transformer Encoder Figure 5. The Bi-LSTM captures temporal dependencies from both forward and backward directions, providing contextual information

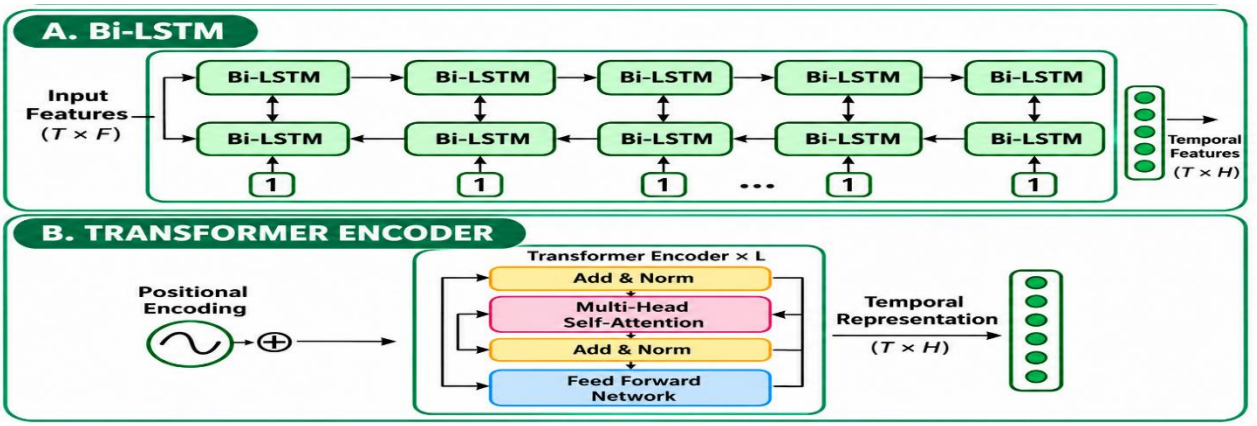


Figure 5. Temporal dependency modeling using parallel Bi-LSTM and Transformer Encoder blocks.

from past and future time steps. In parallel, the Transformer Encoder uses positional encoding, multi-head self-attention, and feed-forward processing to model relationships across different temporal positions and capture both local and long-range dependencies within the speech representation

3.3.3. Spectro-Temporal Fusion

After extracting the spectral and temporal representations, the proposed framework integrates these complementary representations through an attention-based spectro-temporal fusion mechanism. The spectral representation $\boldsymbol{S}$ and temporal representation $\boldsymbol{T}$ are jointly processed by an attention mechanism that learns the relative importance of the information contained in each representation. The resulting attention weights are used to emphasize the most informative spectro-temporal features while reducing the contribution of less relevant information. The weighted representations are subsequently combined to form a unified fusion vector $\boldsymbol{F}$, which is then passed to a fully connected layer for the final prediction. This fusion mechanism enables the framework to effectively integrate spectral characteristics learned by the CNN with temporal dependencies captured by the Bi-LSTM and Transformer Encoder, providing a unified representation for both deepfake detection and spoof attack attribution.

Spectro-Temporal Fusion and Multi-task Classification: The learned spectral and temporal representations are integrated through the proposed spectro-temporal fusion module Figure 6, producing a unified feature representation that combines complementary information from both domains. The fused representation is then shared by two task-specific classification branches.

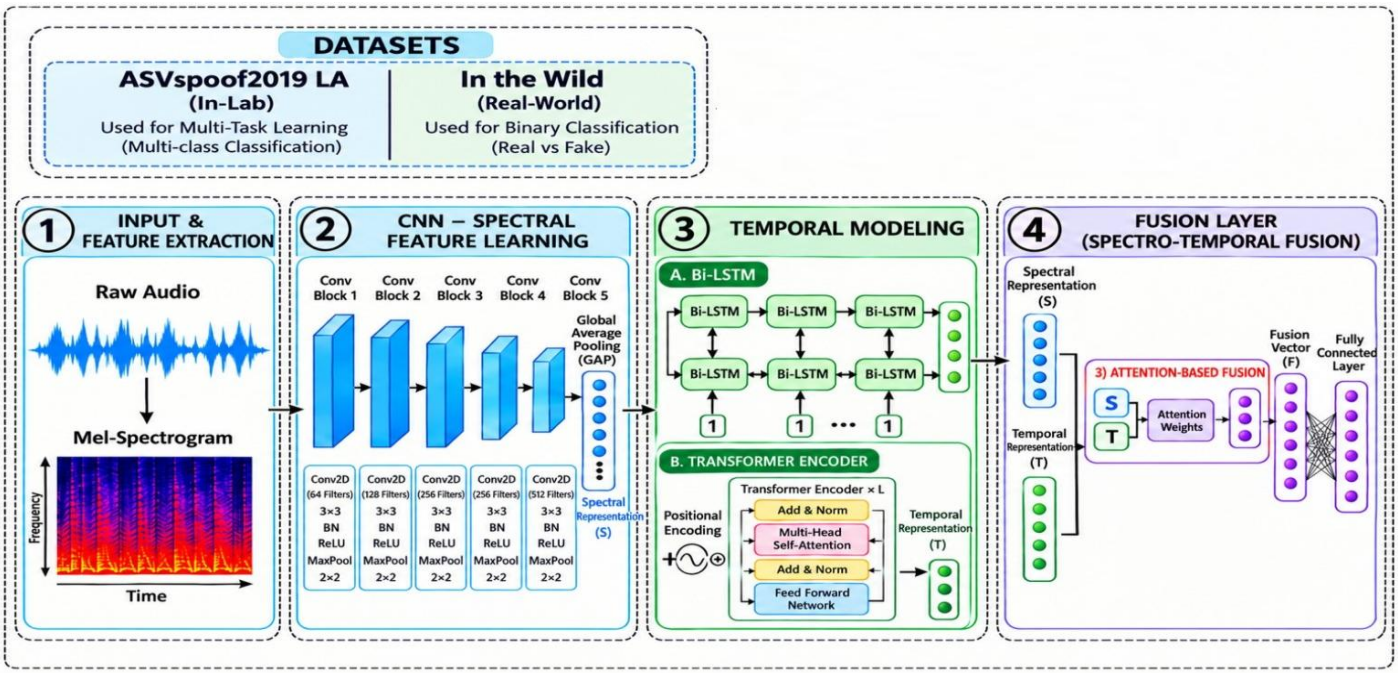


Figure 6. Overview of the proposed multi-task deep learning framework for joint deepfake detection and spoof attribution.

3.4. Optimization and Training Details

The proposed framework is trained using supervised multi-task optimization for deepfake speech detection and spoof attribution on the In-the-Wild and ASVspoof 2019 LA datasets, while preserving dataset-specific protocols. The data are divided into training, validation, and test subsets, with the test set kept unseen during training. Feature normalization parameters are estimated from the training data only and subsequently apply to the validation and test sets to avoid data leakage. The framework jointly optimizes binary deepfake detection and multi-class spoof attribution using a weighted multi-task loss. The weighted multi-task loss is used to balance the contributions of the two tasks and reduce potential task interference

The complementary representations generated by the Bi-LSTM and Transformer Encoder modules are integrated using an attention-based fusion mechanism. The AdamW optimizer is initialized with a learning rate of $\mathbf{3 \times 10^{-5}}$, The multi-task learning objective is formulated as a weighted combination of the binary detection loss and the spoof attribution loss:

$$L_{\mathrm{MT}} = \lambda_{\mathrm{det}} L_{\mathrm{det}} + \lambda_{\mathrm{att}} L_{\mathrm{att}} \qquad (3)$$

Where:

- L_{MT} :denotes the overall multi-task learning loss.
- L_{det} : represents the binary detection loss for distinguishing bona-fide and spoofed speech,
- L_{att} :denotes the spoof attribution loss for identifying the spoof attack class.
- λ_{det} ***and*** λ_{att}:weighting coefficients that control the contribution of each task to the total loss

with a batch size of 32 and a maximum of 50 training epochs. Early Stopping, Reduce LROnPlateau, and Model Checkpointing are employed to improve training stability and reduce overfitting. The learning rate is reduced by a factor of 0.5 when the validation loss plateaus, with a minimum learning rate of $\mathbf{3 \times 10^{-7}}$, while the best-performing model is retained. Robustness is evaluated under 20 dB and 5 dB additive noise, reverberation, time stretching, and +2-semitone pitch shifting. Deepfake detection is evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC, and EER, whereas spoof attribution is evaluated using Macro-F1, Balanced Accuracy, per-class Recall, and confusion matrices.

4. Experimental Analysis and Performance Evaluation of the Proposed Framework

4.1.Datasets

The ASVspoof 2019 LA dataset is a controlled benchmark specifically developed for evaluating spoofing countermeasures against synthetic and converted speech. The dataset is derived from the VCTK corpus and contains high-quality speech recorded under controlled studio conditions, with limited environmental interference. It includes 107 speakers (46 male and 61 female), with speakers strictly separated across the training, development, and evaluation subsets to ensure speaker-independent evaluation. The dataset incorporates 19 spoofing algorithms (A01–A19) based on Text-to-Speech (TTS) and Voice Conversion (VC) techniques and TTS-to-VC hybrids . Its experimental design distinguishes between known attacks (A01–A06) and unknown attacks (A07–A19), allowing the evaluation of a model's ability to generalize beyond the spoofing techniques observed during training[24],[25].

The In-the-Wild dataset is designed to evaluate audio deepfake detection under uncontrolled, real-world conditions. The recordings are collected from publicly available online sources, including social media, video platforms, podcasts, and other digital media, and contain speech from 58 public in figure 7, Table 1. Unlike controlled benchmark datasets, the recordings exhibit substantial variability in acoustic and technical conditions, including environmental interference, cross-talk, different recording devices, room characteristics, and online processing. The dataset therefore provides an out-of-domain evaluation scenario that more closely reflects the conditions encountered in practical audio deepfake detection

applications. Its primary objective is to assess the robustness of detection models when confronted with heterogeneous real-world audio rather than standardized laboratory recordings[26],[27].

Table 1. Structural comparison of the ASVspoof 2019 LA and In-the-Wild datasets.

Dataset	Total Audio	Total Utterances	Real/Bona-fide	Spoof / Fake	Known Attacks	Unknown Attacks/Unseen Attacks	Speakers
ASVspoof 2019 LA	~15–20 GB	121,461	12,483	108,978	A01–A06	A07–A19	107
In-the-Wild	38.0 hours	31,779	19,963	11,816	—	—	58

Table 1, Figure 7 provides a comparative summary of the ASVspoof 2019 LA and In-the-Wild datasets used in this study. ASVspoof 2019 LA contains 121,461 utterances, including 12,483 bona-fide and 108,978 spoofed samples, with known attacks A01–A06 and unknown attacks A07–A19 across 107 speakers, ASVspoof 2019 LA consists of official training, development, and evaluation partitions. The training and development partitions contain the six known attack algorithms A01–A06, whereas the evaluation partition contains attack IDs A07–A19 Unseen evaluation attacks . In the present study, however, a stratified 60:40 split was used for the reported in-dataset experiment rather than the official train/development/evaluation partition.

In-the-Wild contains 31,779 utterances, including 19,963 bona-fide and 11,816 spoofed samples, collected from 58 public figures. These complementary datasets support two evaluation objectives: ASVspoof 2019 LA provides a controlled benchmark for evaluating spoof detection and attack-related generalization, whereas In-the-Wild provides an out-of-domain setting for assessing detection under realistic and unconstrained acoustic conditions.

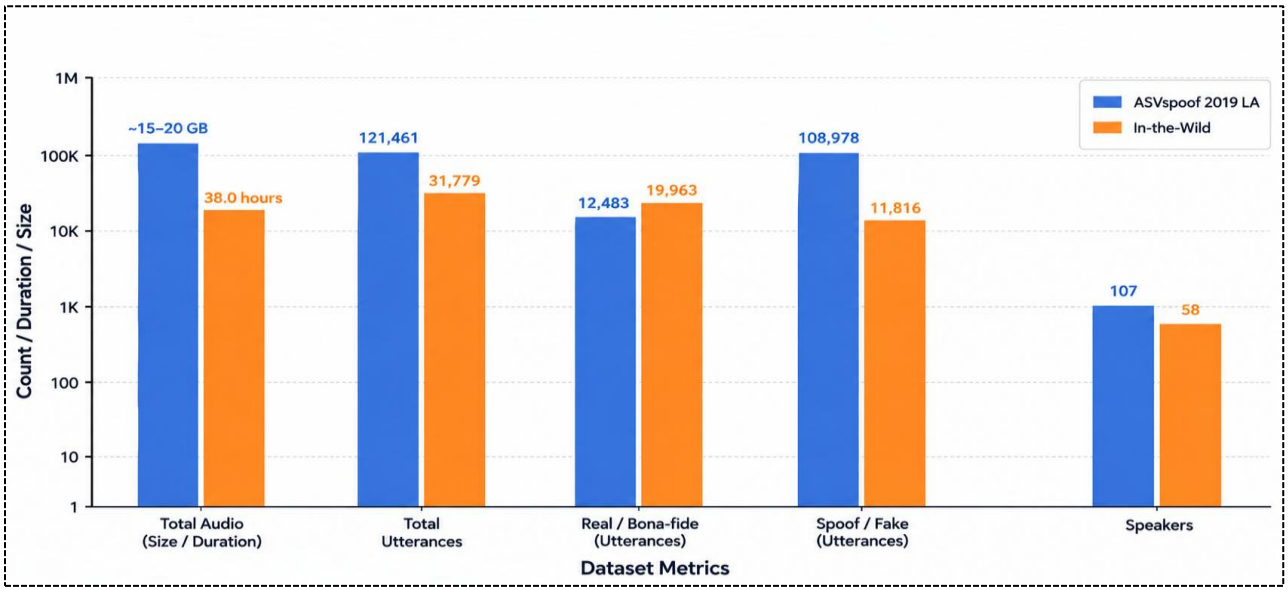


Figure 7. Comparison of the ASVspoof 2019 LA and In-the-Wild Datasets.

4.2. Experimental Use of the Datasets

In this study, ASVspoof 2019 LA and In-the-Wild are employed within a two-stage experimental framework to evaluate the proposed multi-task deep learning model. The experimental design is intended to assess both the model's performance within the same dataset and its ability to generalize to an unseen dataset under different speech recording conditions.

In Stage 1: In-Dataset Evaluation, the proposed model is trained and evaluated using the ASVspoof 2019 LA dataset. The data are divided into training and validation subsets using a stratified 60:40 split with a fixed random seed of 42. This strategy preserves the class distribution across the two subsets while providing a consistent evaluation protocol. The model performs two complementary tasks. The first task is deepfake detection, which classifies speech as either bona-fide or spoof. The second task is spoof attribution, which identifies the corresponding spoof attack class using the available attack labels. In the supervised attribution setting, six known attack classes (A01–A06) are used as the attribution categories. supervised spoof attribution, followed by an unseen-attack setting using A07–A19 to assess generalization to attacks not observed during training,Under the stratified train-validation protocol, the model is evaluated on the held-out validation partition Table 2.

In Stage 2: Cross-Dataset Evaluation, the model is trained using ASVspoof 2019 LA and subsequently evaluated on In-the-Wild, which is not used during the training process. This stage is designed to evaluate the cross-dataset generalization capability of the proposed framework by testing the model on speech data originating from a different dataset and real-world recording conditions. The evaluation in this stage focuses on deepfake detection, distinguishing between bona-fide and spoofed speech. Spoof attribution is not performed on In-the-Wild because it does not provide an attack taxonomy equivalent to the A09–A19 labels available in ASVspoof 2019 LA.

The two-stage experimental design therefore provides a structured evaluation of the proposed framework. Stage 1 The Framework was evaluated using a stratified 60:40 train-validation split with a fixed random seed of 42. whereas Stage 2 evaluates its ability to generalize from ASVspoof 2019 LA to the unseen In-the-Wild dataset. Importantly, the In-the-Wild data remain completely excluded from the training and validation procedures of Stage 1 to ensure a reliable cross-dataset evaluation.

Table 2. Two-stage experimental setup for in-domain, unseen-attack, and cross-dataset evaluations.

Stage	Training Data	Testing Data	Detection	Attribution / Generalization	Main Objective
Stage 1: Known-Attack Learning	ASVspoof 2019 LA: bona-fide + A01–A06	ASVspoof 2019 LA: held-out bona-fide + A01–A06	✓	A01–A06	Evaluate in-domain detection and supervised spoof attribution
Stage 1: Unseen-Attack Generalization	ASVspoof 2019 LA: bona-fide + A01–A06	ASVspoof 2019 LA: bona-fide + A07–A19	✓	Unseen attacks	Evaluate detection generalization to attacks not observed during training
Stage 2: Cross-Dataset Generalization	ASVspoof 2019 LA	In-the-Wild	✓	—	Evaluate cross-dataset binary detection without fine-tuning

Table 3 show the main parameter configurations used for the ASVspoof 2019 LA and In-the-Wild experiments. Both datasets were processed at a sampling rate of 16 kHz, while 80 and 128 Mel filters were used for ASVspoof 2019 LA and In-the-Wild, respectively. Forty MFCC coefficients were used for both datasets. The ASVspoof model used 128 time steps and a batch size of 16, whereas the In-the-Wild model used 256 time steps and a batch size of 32. The maximum number of training epochs was 50 for both settings. AdamW was used as the optimizer, with learning rates of 3×10^{-4} and 3×10^{-5} for ASVspoof 2019 LA and In-the-Wild, respectively. The attention configurations used 8 heads with an FFN dimension of 1024 for ASVspoof and 4 heads with an FFN dimension of 256 for In-the-Wild. The final dropout rates were 0.4 and 0.5, respectively, with 60:40 and 80:20 data splits. The primary task was joint detection and attribution for ASVspoof 2019 LA and binary detection for In-the-Wild

Table 3. Attack types and generation mechanisms for unseen attacks (A07–A19) in ASVspoof 2019 LA.

Attack	Spoof Type	Scenario	Vocoder/Synthesis
A07	TTS	Logical Access (LA)	Vocoder-based synthesis

A08	TTS	Logical Access (LA)	Neural waveform synthesis
A09	TTS	Logical Access (LA)	Vocoder-based synthesis
A10	TTS	Logical Access (LA)	Neural waveform synthesis
A11	TTS	Logical Access (LA)	Griffin-Lim-based synthesis
A12	TTS	Logical Access (LA)	Neural waveform synthesis
A13	Hybrid(TTS + VC)	Logical Access (LA)	Waveform concatenation and filtering
A14	Hybrid(TTS + VC)	Logical Access (LA)	Vocoder-based synthesis
A15	Hybrid(TTS + VC)	Logical Access (LA)	Neural waveform synthesis
A16	TTS	Logical Access (LA)	Waveform concatenation
A17	VC	Logical Access (LA)	Waveform filtering
A18	VC	Logical Access (LA)	Vocoder-based conversion
A19	VC	Logical Access (LA)	Spectral filtering

4.3. Experimental Setup and Parameter Configuration

Two speech datasets were used to evaluate the proposed framework. The first model was developed using the ASVspoof 2019 Logical Access (LA) dataset for joint binary spoof detection and attack attribution . The implementation currently considers six attack classes, A01–A06. For the supervised spoof attribution task, six known spoof attack classes (A01–A06) were considered. The attribution head was implemented as a six-class SoftMax classifier. The remaining attack classes (A07–A19) were treated as unseen attack conditions and were not used as supervised attribution targets during training,The data were divided into training and validation subsets using a stratified 60:40 split with a fixed random seed of 42. The second model was evaluated on the In-the-Wild dataset for binary real/fake speech detection using an 80:20 stratified split Table 4.

All audio signals were converted to 16 kHz mono audio. Silence regions were removed using an energy-based threshold of 25 dB, followed by pre-emphasis with a coefficient of 0.97 and amplitude normalization. Mel-spectrograms and MFCCs were then extracted. The ASVspoof model used 80 Mel filters and 40 MFCC coefficients, whereas the In-the-Wild model used 128 Mel filters and 40 MFCC coefficients. To prevent data leakage, the normalization parameters were estimated using the training data only and subsequently applied to the validation and test sets. Feature normalization was therefore performed after data partitioning.

The ASVspoof model uses a multi-task CNN–Bi-LSTM+ Transformer architecture with four convolutional blocks containing 64, 128, 256, and 512 filters, followed by two Transformer-style blocks. Multi-head self-attention uses 8 heads with a feed-forward dimension of 1024. The shared representation produces two outputs: a sigmoid classifier for bona fide/spoof detection and a six-class softmax classifier for attack attribution (A01–A06).

The In-the-Wild model follows a similar CNN–Transformer structure with four convolutional layers (64, 128, 256, and 512 filters) and two Transformer-style blocks using 4 attention heads and a feed-forward dimension of 256. A sigmoid output layer provides the final real/fake prediction.

Both models were trained for a maximum of 50 epochs using AdamW. The learning rate was 3×10^{-4} for ASVspoof and 3×10^{-5} for In-the-Wild, with batch sizes of 16 and 32, respectively. Final classification dropout was 0.4 and 0.5, respectively. Early stopping, learning-rate reduction, and model checkpointing were used to select the best model based on validation loss.

Performance was evaluated using accuracy, (EER), confusion matrices, ROC curves, AUC, and classification reports. Additional analyses included calibration, t-SNE, Grad-CAM, and SHAP to support model interpretation.

Table 4. Hyperparameter configurations for model training on ASVspoof 2019 LA and In-the-Wild datasets.

Parameter	ASVspoof 2019 LA	In-the-Wild
Sampling rate	16 kHz	16 kHz
Mel filters	80	128
MFCCs	40	40
Delta MFCC	40	40
Time steps	128	256
Batch size	16	32
Maximum epochs	50	50
Initial learning rate	3×10^{-4}	3×10^{-5}
Optimizer	AdamW	AdamW
Attention heads	8	4
FFN dimension	1024	256
Dropout	0.4	0.5
Data split	Stratified 60:40	Stratified 80:20
Random seed	42	42
Main task	Detection + Attribution	Detection

4.4. Computational Resource Requirements

To assess the practical feasibility of the proposed model in real-world scenarios, the system was implemented and evaluated on a Dell Precision 5560 laptop (Dell Technologies Inc., Round Rock, TX, USA), equipped with an 11th Gen Intel Core i7 processor and 32 GB of RAM, without a dedicated GPU. During training, the system consumed approximately 12 GB of memory, whereas prediction required less than 2 GB. The average inference time for a single prediction was 1.55Ms seconds. These findings indicate that the proposed model can operate efficiently without reliance on high-end computational hardware, supporting its potential deployment in resource-constrained environments, including portable devices and field-deployed security systems. With minor computational optimizations, the model could be further adapted for deployment on even more lightweight systems.

4.5. Experimental Findings

The proposed (MT-DL) Framework was evaluated to examine its effectiveness in deepfake speech detection, spoof attack attribution, cross-dataset generalization, and robustness to acoustic perturbations. The framework jointly learns binary detection and multi-class spoof attribution from a shared spectro-temporal representation. The shared representation is generated through a CNN-based frontend followed by temporal modeling and is delivered to two task-specific heads: a detection head for real/fake classification and an attribution head for identifying six known spoof attack categories (A01–A06). The A07–A19 attack classes were treated as unseen attack conditions and were not included in the supervised attribution training.

The experimental evaluation is organized into two principal stages. Stage 1 focuses on in-dataset multi-task learning using ASVspoof 2019 LA, whereas Stage 2 investigates cross-dataset generalization by transferring the trained model to an unseen domain without fine-tuning. Additional analyses examine

the contribution of the spectro-temporal representation, robustness under acoustic perturbations, and training stability.

Stage 1: In-Dataset Multi-Task Evaluation

In the first stage, the proposed Multi-Task DL Framework is trained and evaluated using ASVspoof 2019 LA. The model simultaneously performs binary classification of bona-fide versus spoofed speech and multi-class attribution of spoofing attacks. Both tasks share the learned representation while maintaining separate output heads and task-specific objectives. The proposed framework relies on a fused spectro-temporal representation that combines complementary acoustic information before CNN-based feature extraction and temporal modeling.

Table 5. Feature set ablation study comparing classification performance across spectro-temporal combinations.

Dataset	Setting	Binary Accuracy (%)	Binary F1-Real	Binary F1-Fake	Multi Accuracy (%)	Multi F1-Macro	Multi F1-Weighted
ASVspoof 2019 LA	MFCC only	95.96	96.71	95.08	93.74	91.51	92.79
	MFCC +Log-Mel	94.31	95.08	95.44	91.12	91.94	92.18
	MFCC+ Δ MFCC	95.18	94.94	96.32	92.98	91.79	94.05
	MFCC+ Log-Mel+ Δ MFCC	**100**	**100**	**100**	**99.93**	**99.09**	**99.93**

Table 5 compares the performance obtained under four feature configurations. The MFCC-only configuration achieves 95.96% binary accuracy and 93.74% multi-class accuracy. Adding Log-Mel features results in 94.31% binary accuracy and 91.12% multi-class accuracy, whereas the MFCC + Δ-MFCC configuration achieves 95.18% binary accuracy and 92.98% multi-class accuracy. The combined MFCC + Log-Mel + Δ-MFCC configuration yields the highest values in the table, with 100% binary accuracy, 99.93% multi-class accuracy, 99.09% macro-F1, and 99.93% weighted-F1. These results indicate that the combined representation provides the strongest overall feature configuration among the settings reported in the table.

Table 6. Detection and spoof attribution performance of the multi-task architecture across feature settings.

Dataset	Model	Setting	Binary Accuracy (%)	Binary F1-Real	Binary F1-Fake	Multi Accuracy (%)	Multi F1-Macro	Multi F1-Weighted
ASVspoof 2019 LA	Multi-Task DL Framework	MFCC only	97.09	97.01	96.08	96.04	96.51	96.57
		MFCC +Log-Mel	97.31	97.08	96.44	97.12	96.94	97.18
		MFCC+Δ MFCC	97.68	97.77	98.32	98.01	98.11	98.17
		MFCC+ Log-Mel+ Δ MFCC	**98.23**	**98.12**	**98.32**	**98.42**	**98.25**	**98.29**

Table 6 presents the performance of the proposed Multi-Task DL Framework under four feature configurations. The MFCC-only configuration achieves 97.09% binary accuracy and 96.04% multi-class accuracy. Adding Log-Mel features increases the binary accuracy to 97.31% and the multi-class accuracy to 97.12%. The MFCC + Δ-MFCC configuration achieves 97.68% binary accuracy and 98.01% multi-class accuracy. The combined MFCC + Log-Mel + Δ-MFCC configuration provides the highest binary accuracy of 98.23% and the highest multi-class accuracy of 98.42%, together with macro-F1 and weighted-F1 scores of 98.25% and 98.29%, respectively

Table 7. Class-level performance metrics for supervised spoof attack attribution (A01–A06).

Class	Precision	Recall	F1-Score	Support
A01	1.00	0.99	0.99	5,040
A02	0.99	1.00	1.00	2,947
A03	1.00	0.99	1.00	2,964
A04	0.99	0.99	0.99	2,994
A05	1.00	1.00	1.00	3,085
A06	1.00	1.00	1.00	3,060
Accuracy	1.00	—	—	20,090
Macro Avg	1.00	1.00	1.00	20,090
Weighted Avg	1.00	1.00	1.00	20,090

Table 7 presents the class-level classification performance of the proposed multi-task model for supervised spoof attack attribution across the six known attack classes (A01–A06). The model demonstrates high precision, recall, and F1-scores across the six classes, indicating consistent attribution performance for the known spoof attack categories.

Table 8 presents the performance of the proposed (MT-DL) framework under four different Spectro-temporal feature configurations using data augmentation. The results show a consistent improvement in both the binary deepfake speech detection task and the multi-class spoof attribution task as additional complementary features are incorporated. Using MFCC features alone achieves a Binary Accuracy of 93.48 and an Attribution Accuracy of 90.27. Adding Log-Mel features improves these values to 94.86 and 92.54, respectively, indicating that spectral information provides additional discriminative cues. The combination of MFCC and Δ-MFCC further increases Binary Accuracy to 95.62 and Attribution Accuracy to 93.86, demonstrating the benefit of temporal-dynamic information captured by first-order MFCC derivatives.

Table 8. Performance evaluation under light data augmentation on ASVspoof 2019 LA.

Model	Setting	Binary Acc.	Binary F1-Real	Binary F1-Spoof	Attribution Acc.	Attribution F1-Macro	Attribution F1-Weighted
Multi-Task DL Framework	MFCC	94.48	91.86	94.82	92.07	89.65	89.72
	MFCC + Log-Mel	94.86	93.67	95.89	92.54	92.18	92.47
	MFCC + Δ-MFCC	95.62	92.51	94.58	93.86	93.49	93.78
	MFCC + Log-Mel + ΔMFCC	**98.74**	**95.88**	**98.42**	**96.27**	**94.96**	**96.07**

Table 9 presents the performance of the proposed (MT-DL) framework on ASVspoof2021-LA using four Spectro-temporal feature configurations with strong data augmentation. The results show a consistent improvement as complementary features are combined. Using MFCC alone achieves 93.48 Binary Accuracy and 90.27 Attribution Accuracy. Adding Log-Mel features increases these values to 94.86 and 92.54, respectively, while adding Δ-MFCC further improves them to 95.62 and 93.86 The full combination of MFCC, Log-Mel, and Δ-MFCC achieves the highest performance, with 0.9674 Binary Accuracy and 0.9527 Attribution Accuracy, together with 95.88 F1-Real, 97.42 F1-Spoof, 0.9496 Macro-F1, and 0.9518 Weighted-F1.

Table 9. Performance evaluation under strong data augmentation on ASVspoof 2019 LA.

Model	Setting	Binary Acc.	Binary F1-Real	Binary F1-Spoof	Attribution Acc.	Attribution F1-Macro	Attribution F1-Weighted
	MFCC	93.48	91.86	94.82	92.07	89.65	85.09

Multi-Task DL Framework	MFCC + Log-Mel	94.86	93.67	95.89	92.54	92.18	92.47
	MFCC + Δ-MFCC	95.62	94.51	94.58	93.86	93.49	93.78
	MFCC + Log-Mel + Δ-MFCC	**96.74**	**95.88**	**97.42**	**95.27**	**94.96**	**95.18**

Table 8 evaluates the proposed MT-DL framework using augmented data generated with relatively mild augmentation settings, including a random time shift of ±100 ms, Gaussian noise with an SNR of 25–30 dB, random gain of ±3 dB, and time stretching of 0.95–1.05×. These settings introduce moderate variations while preserving the main characteristics of the speech signals.
In contrast, Table 9 uses stronger augmentation settings, including a random time shift of ±200 ms, Gaussian noise with an SNR of 15–25 dB, random gain of ±6 dB, time stretching of 0.90–1.10×, pitch shifting of ±2 semitones, and random cropping of 2–4 seconds. This setting provides more challenging variations to evaluate the robustness of the proposed MT-DL framework under stronger data augmentation conditions.

Stage 2: Cross-Dataset Generalization

The second stage evaluates the generalization capability of the trained Multi-Task DL Framework under an unseen-domain condition. The model trained on ASVspoof 2019 LA is transferred directly to the In-the-Wild dataset without additional fine-tuning. Because In-the-Wild does not provide attack labels corresponding to the A09–A19 taxonomy, Stage 2 evaluates the binary detection head only.

Table 10. In-dataset vs. cross-dataset binary detection performance comparison.

Training Dataset	Testing Dataset	Model	Setting	Binary Accuracy (%)	F1-Real (%)	F1-Fake (%)	Precision (%)	Recall (%)	AUC (%)	EER (%)
ASVspoof 2019 LA	ASVspoof 2019 LA	Multi-Task DL Framework	MFCC only	96.96	96.71	96.08	97.95	97.89	0.97	0.072
		Multi-Task DL Framework	MFCC +Log-Mel	97.31	97.08	97.44	97.31	97.27	0.97	0.061
		Multi-Task DL Framework	MFCC+ Δ	97.18	97.80	97.32	97.18	97.14	0.97	0.066
		Multi-Task DL Framework	**MFCC+ Log-Mel+ Δ**	**98.89**	**98.88**	**98.85**	**98.68**	**98.72**	**0.99**	**0.048**
ASVspoof 2019 LA	In-the-Wild	Multi-Task DL Framework	MFCC only	97.42	97.16	97.61	97.36	97.28	0.98	0.018
		Multi-Task DL Framework	MFCC +Log-Mel	98.11	97.92	98.25	98.05	98.00	0.98	1.71
		Multi-Task DL Framework	MFCC+ Δ	98.44	97.55	98.26	98.23	97.12	0.98	1.89
		Multi-Task DL Framework	**MFCC+ Log-Mel+ Δ**	**98.11**	**97.92**	**98.25**	**98.05**	**98.00**	**0.99**	**0.089**

Table 10 evaluates binary detection performance under both in-dataset and cross-dataset conditions. For the ASVspoof 2019 LA to ASVspoof 2019 LA setting, the full MFCC + Log-Mel + Δ configuration reports the highest values shown in the table, including 98.89% precision, 98.68% recall, 98.72% AUC, and an EER of 0.048%. For the cross-dataset ASVspoof 2019 LA to In-the-Wild setting, the full fused configuration reports 98.11% F1-Fake, 97.92% precision, 98.25% recall, 0.99% AUC, and an EER of 0.089%.

Table 11. Performance degradation metrics under cross-dataset domain shift from ASVspoof 2019 LA to In-the-Wild.

Metric	ASVspoof → ASVspoof	ASVspoof → In-the-Wild	Change
Binary Accuracy (%)	98.89	98.11	-0.78
F1-Real (%)	98.88	97.92	-0.96
F1-Fake (%)	98.85	98.25	+0.60
Precision (%)	98.68	98.05	-0.63
Recall (%)	98.72	98.00	-0.72
ROC-AUC (%)	0.98	0.98	0
EER (%)	0.048	0.089	+0.041pp

Table 11 quantifies the effect of cross-dataset domain shift under the combined feature configuration. Binary accuracy decreases by 0.78 percentage points when moving from the in-dataset setting to the In-the-Wild setting. F1-Real decreases by 0.96 percentage points, while F1-Fake decreases by 0.60 percentage points. Precision and recall decrease by 0.63 and 0.72 percentage points, respectively. The EER increases from 0.048% to 0.089%, corresponding to an increase of 0.041 percentage points..

4.6.Robustness to Additive Noise

Robustness was evaluated on 31,779 valid In-the-Wild audio files represented by feature tensors of size 256 × 168 × 1. Six acoustic conditions were examined: clean audio, pitch shifting, time stretching, reverberation, additive noise at 20 dB, and additive noise at 5 dB.

Table 12. Robustness of the Multi-Task DL Framework under Additive Noise.

Noise Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	EER (%)
Clean	**98.89**	**98.68**	**98.72**	**98.88**	**0.0048**
Pitch +2	92.33	95.71	93.95	94.82	0.610
Time Stretch	**99.52**	**96.52**	**97.26**	**98.37**	**0.0126**
Reverb	**99.84**	**96.55**	**98.55**	**98.69**	**0.0029**
Noise 20dB	87.71	88.08	86.89	81.94	0.1080
Noise 5dB	86.15	82.46	77.98	77.20	12.29

Table 12 summarizes the robustness of the proposed framework under different acoustic perturbations. Under clean conditions, the model achieves 98.89% accuracy, with precision, recall, and F1-score of 98.68%, 98.72%, and 98.88%, respectively, and an EER of 0.0048%. Time stretching and reverberation preserve high detection performance, with accuracies of 99.52% and 99.84%, respectively. Pitch shifting reduces the accuracy to 92.33%, while additive noise produces a more pronounced degradation, with accuracy decreasing to 87.71% at 20 dB and 86.15% at 5 dB. The corresponding EER increases substantially under the strongest noise condition, reaching 12.29%.

4.7. Training Stability

The robustness results demonstrate that the proposed MT-DL framework maintains stable performance under temporal transformations and reverberation, while additive noise causes a more pronounced degradation. Clean audio achieves 100% accuracy, with an EER of 0.0018%. Time stretching and reverberation preserve high performance, achieving accuracies of 99.52% and 99.84%, respectively, with very low EER values of 0.0126% and 0.0029%. In contrast, pitch shifting reduces accuracy to 92.33% and increases the EER to 0.610%. Under additive noise, performance decreases further, with accuracy reaching 87.71% at 20 dB and remaining above 86% at 5 dB, while the EER increases substantially to 12.29.

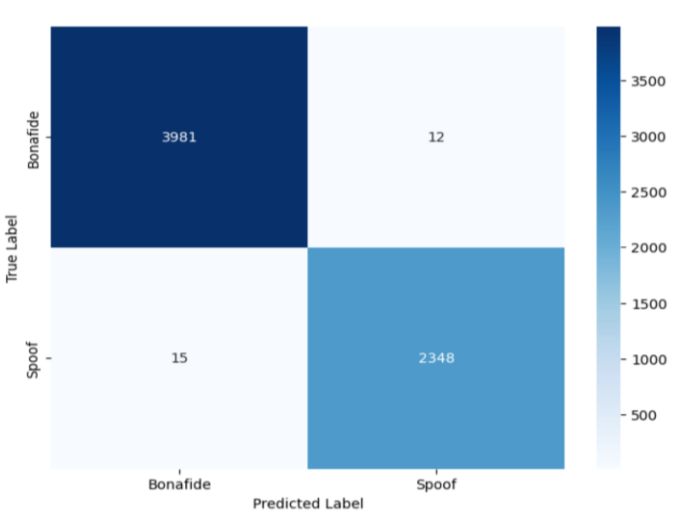
Table 13 compares the proposed MT-DL framework with representative spoofing detection approaches on the ASVspoof 2019 LA dataset. Reported EER values range from 9.57% for CQCC-GMM to 0.59%

for Rawformer, while the proposed framework achieves a binary validation accuracy of 99.75% and an EER of 0.48%. In addition, it achieves 99.57% accuracy for spoof attack attribution under the supervised A01–A06 setting.

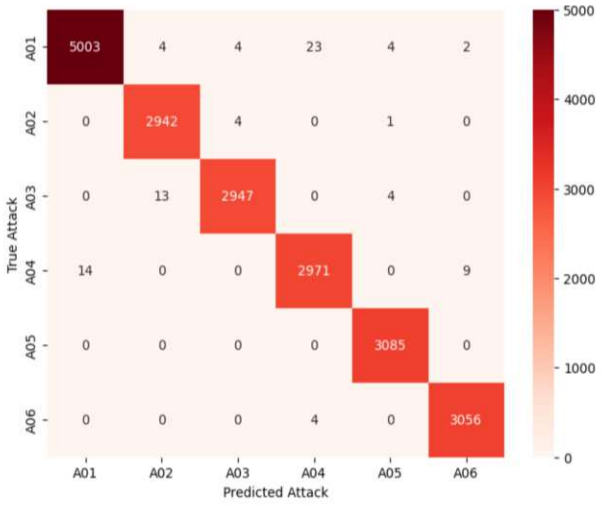
Table 13. Performance comparison of the proposed MT-DL framework against state-of-the-art audio deepfake detection models.

No.	Method	Feature / Architecture	Metric	Result	Ref.
1	LFCC-GMM	Linear Frequency Cepstral Coefficients + GMM	EER	8.09%	[19]
2	CQCC-GMM	Constant-Q Cepstral Coefficients + GMM	EER	9.57%	[19]
3	RawNet2	End-to-end raw waveform neural network	EER	5.54%	[20]
4	RawGAT-ST	Spectro-temporal Graph Attention Network	EER	1.06%	[21]
5	AASIST	Integrated Spectro-temporal Graph Attention	EER	0.83%	[22]
6	LCNN-LSTM-sum	LFCC + LCNN + LSTM	EER	1.92%	[23]
7	W2V2-XLSR (Fine-tuned)	Self-supervised wav2vec 2.0 / XLSR	EER	1.35%	[24]
8	SE-Res2Net50	SE-Res2Net architecture	EER	2.50%	[25]
9	RW-ResNet	Raw-waveform ResNet	EER	2.98%	[26]
10	Rawformer	Raw waveform Transformer	EER	0.59%	[27]
11	**Proposed Multi-Task DL Framework**	**CNN +Bi-LSTM+ Transformer with joint detection and spoof attribution**	**EER**	**0.48%**	

On the ASVspoof 2019 LA validation partition, the proposed framework achieved a binary detection accuracy of 99.75% with an EER of 0.48%. The F1-score reached approximately 0.99 for bona-fide speech and 1.00 for spoofed speech. The detailed classification results are presented in Table 7,8, including accuracy, precision, recall, F1-score, and EER.



(a) ASVpoof 2019 LA Binary Confusion Matrix of the MT-DL Framework

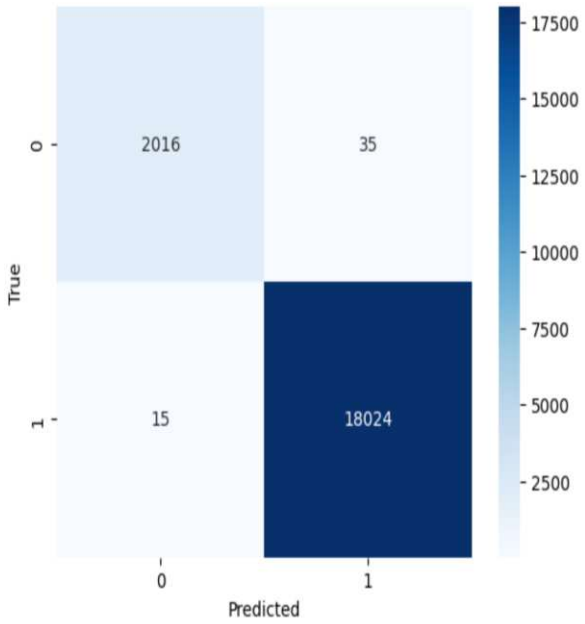


(b) Confusion Matrix of Spoof Attack Attribution for ASVspoof 2019 LA Attack Classes A01–A06.

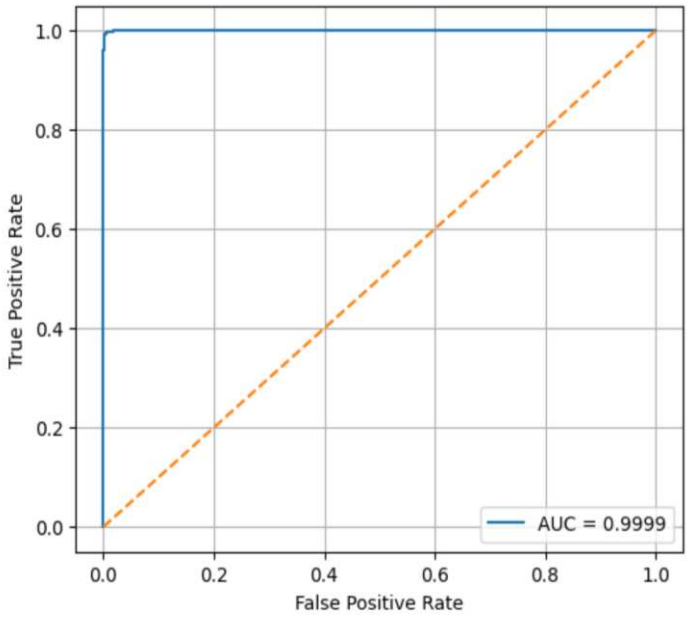
Figure 8. **(a)** presents the binary confusion matrix of the proposed MT-DL framework on the ASVspoof 2019 LA dataset. The matrix shows that most bona-fide and spoof samples are correctly classified, with only a small number of misclassified samples. **(b)** presents the confusion matrix for supervised spoof attack attribution across the A01–A06 attack classes. Most predictions are concentrated along the main diagonal, indicating that the framework can effectively distinguish between the known spoof attack categories.

The proposed framework was additionally evaluated on the In-the-Wild dataset Fig 9 (a and b). Under the evaluated clean condition, the framework achieved an accuracy of **98.23**%, 98.89% precision, 98.68%

recall,0.99% AUC, and an EER of 0.048%The complete binary detection results for the In-the-Wild evaluation are reported in Table 10.



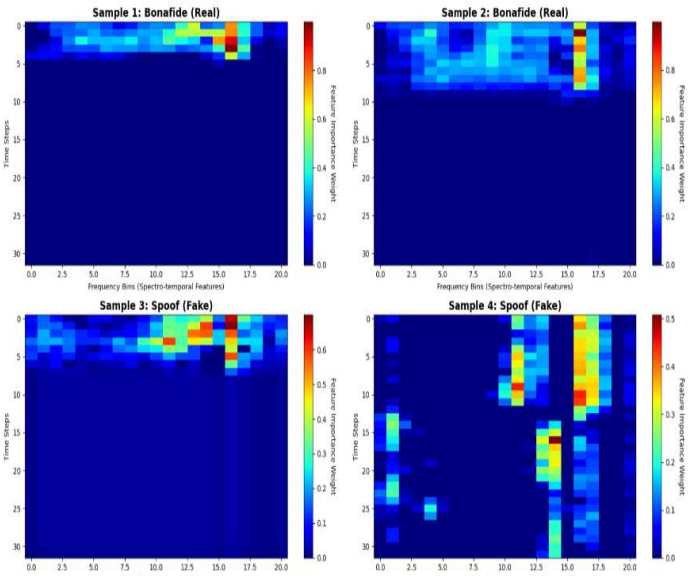
(a) Binary Deepfake Speech Detection on the In-the-Wild Validation Partition



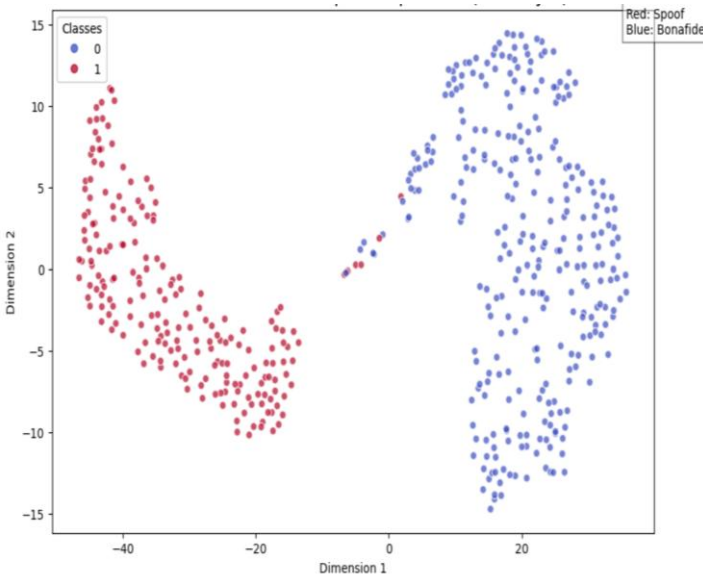
(b) ROC Curve for Binary Deepfake Speech Detection on the In-the-Wild Validation Partition.

Figure 9. (a) shows the confusion matrix for binary deepfake speech detection on the In-the-Wild validation partition. The majority of samples are correctly classified into their respective real and fake categories, with only a limited number of classification errors. **(b)** presents the corresponding ROC curve, showing a curve close to the upper-left corner and an AUC of approximately 0.99. These results indicate strong binary discrimination performance on the evaluated In-the-Wild validation data.

The learned representation was further examined using Grad-CAM visualization Fig E,F. The visualization presents the activation regions associated with the model's binary detection output.



(c) Grad-CAM Visualization of the Learned Spectro-Temporal Representation



(d) t-SNE Manifold: Latent Space Separation (GAP Layer)

Figure 10. (a) presents the Grad-CAM visualization of the learned spectro-temporal representation. The heatmaps highlight the time–frequency regions that contribute most strongly to the model's classification decisions, where higher activation indicates greater contribution to the prediction. (d) presents the t-SNE visualization of the latent representation extracted from the GAP layer. The visualization illustrates the distribution of the learned representations and their separation according to the corresponding speech classes.

To further investigate the interpretability of the proposed (MT-DL) framework, Grad-CAM was employed to visualize the spectro-temporal regions contributing to the model's decisions.

As illustrated in Figure 11, the heatmaps highlight the time–frequency regions receiving the strongest activation during the classification process.

The horizontal axis represents the 256 time steps, while the vertical axis represents the frequency dimension of the extracted audio representation. Warmer regions indicate higher model attention, whereas cooler regions correspond to regions with a lower contribution to the final prediction.

The visualization in Figure 11 demonstrates that the proposed framework captures meaningful spectro-temporal characteristics from both spoof and bona-fide speech samples.

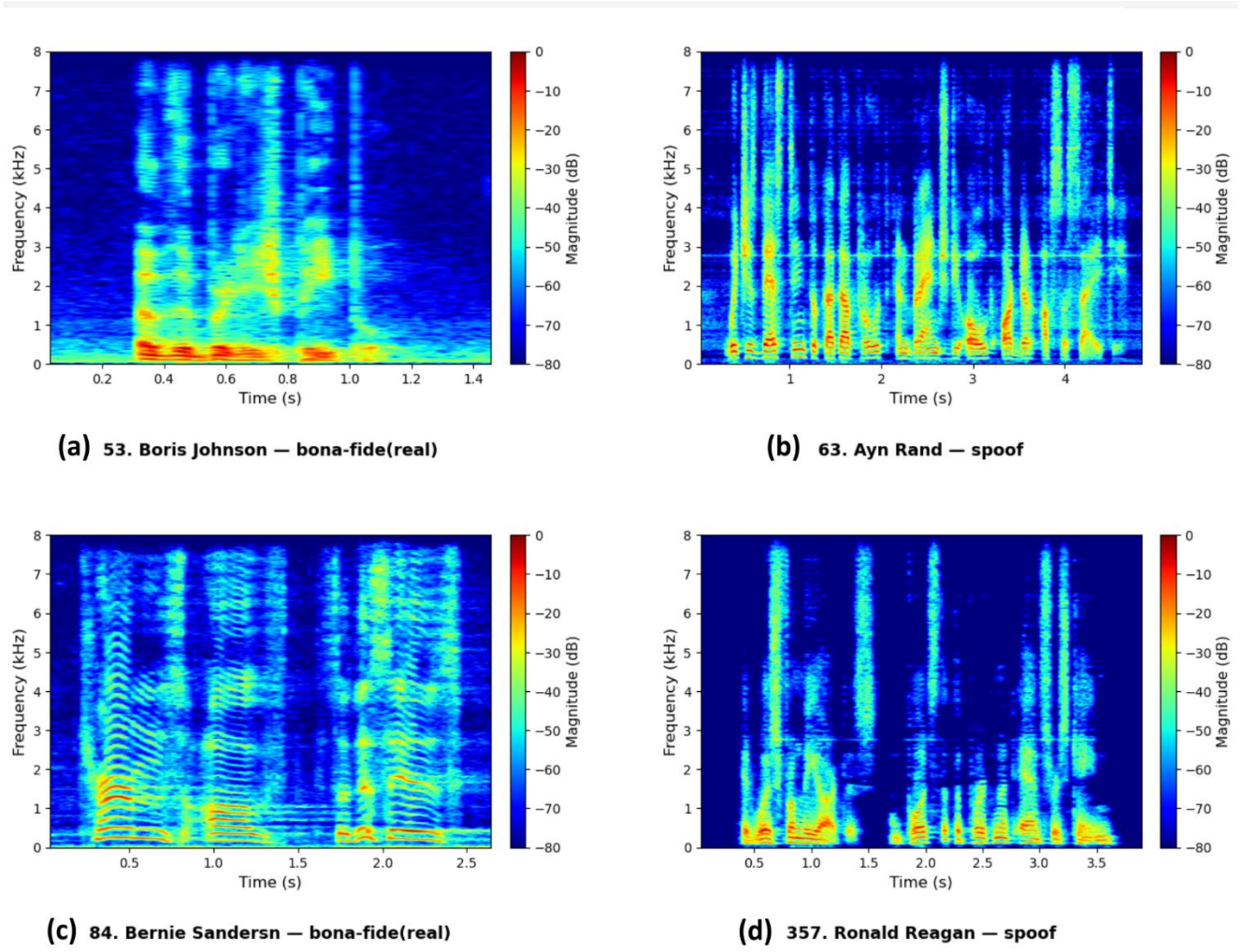


Figure 11. presents representative spectrograms from the In-the-Wild dataset for qualitative analysis of the proposed framework. **(a)** and **(c)** show bona-fide speech samples, while **(b)** and **(d)** show spoof speech samples. The spectrograms illustrate the time–frequency characteristics of both real and manipulated speech under realistic recording conditions. The comparison provides qualitative evidence that the learned representation can capture differences between bona-fide and spoof speech in an uncontrolled real-world dataset.

Furthermore, Figure 12 illustrates the model's classification performance for Spoof Attribution, where spoof classes 07–19 are effectively distinguished from one another.

This indicates that the proposed framework is capable of capturing class-specific characteristics associated with different spoofing mechanisms.

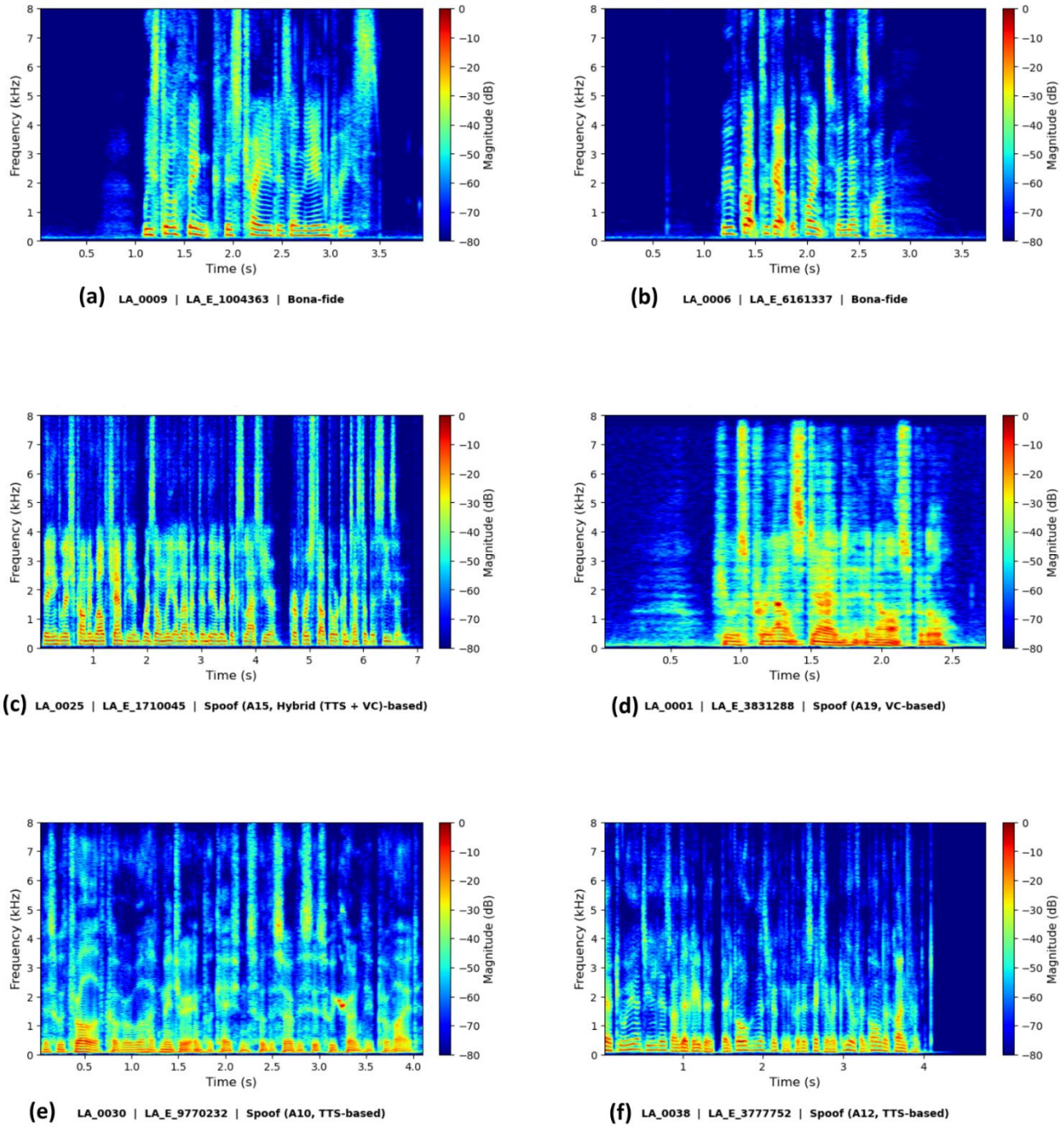


Figure 12. presents representative spectrograms associated with spoof attribution analysis on the ASVspoof 2019 LA dataset, focusing on attack classes A07–A19. The figure includes examples of bona-fide speech as well as different unseen spoof attack types, including TTS-based, VC-based, and hybrid TTS–VC attacks. Specifically**, (a)** and **(b)** represent bona-fide speech samples, **(c)** represents an A17 spoof sample (TTS + VC-based), **(d)** represents an A19 spoof sample (VC-based), **(e)** represents an A10 spoof sample (TTS-based), and **(f)** represents an A12 spoof sample (TTS-based). The spectrograms demonstrate the variability of the acoustic patterns produced by different spoofing mechanisms.

4.7. Performance Analysis and Discussion

The experimental results demonstrate that the proposed (MT-DL) framework can jointly support deepfake speech detection and supervised spoof attribution through a shared spectro-temporal representation. The feature configuration experiments reported in Tables 5 and 6 consistently show that combining MFCC, Log-Mel, and Δ-MFCC features provides a more informative representation than using individual feature sets alone. In particular, the combined configuration achieved the highest performance among the configurations reported in Table 6, reaching 98.23% binary detection accuracy, 98.42% multi-class attribution accuracy, 98.25% Macro-F1, and 98.29% Weighted-F1. This finding supports the underlying motivation of the proposed representation: MFCCs provide compact spectral information, Log-Mel features preserve detailed time-frequency structure, while Δ-MFCC captures temporal variations across consecutive frames. The improvement obtained through their combination suggests that spoofing artifacts are not confined to a single acoustic representation, but may be distributed across complementary spectral and temporal characteristics.

A second important finding concerns the ability of the framework to perform spoof attribution in addition to conventional binary detection. As reported in Table 7, the attribution head achieves consistently high precision, recall, and F1-scores across the six supervised attack classes A01–A06. The reported overall attribution accuracy is approximately 99.57%, with Macro-F1 and Weighted-F1 close to 99.6% in the reported experimental results. This result is relevant because binary detection alone only determines whether an utterance is genuine or spoofed, whereas attribution provides additional information about the attack category. Previous research has similarly demonstrated that synthetic speech systems can leave identifiable acoustic fingerprints, motivating attribution as an extension of conventional anti-spoofing [12]. Recent work on few-shot speech spoofing algorithm recognition further emphasizes the importance of identifying the generation mechanism, particularly when emerging spoofing algorithms may have limited training examples [14]. In this context, the proposed framework contributes by integrating detection and supervised attribution within a shared learning framework rather than treating them as completely independent objectives.

A third finding concerns generalization and robustness. The cross-dataset experiment in Table 10 indicates that the model retains substantial binary detection capability when transferred from ASVspoof 2019 LA to the In-the-Wild dataset without fine-tuning, with the reported full-feature configuration achieving 98.11% binary accuracy and an EER of 0.089% in the current manuscript. This aspect is particularly important because previous studies have shown that strong performance on controlled ASVspoof benchmarks does not necessarily translate to realistic recordings. Müller et al. reported substantial performance degradation when existing detectors were evaluated on their In-the-Wild dataset, highlighting the domain-shift problem in audio deepfake detection. The robustness experiments in Table 12 provide a complementary observation: temporal stretching and reverberation have relatively limited effects on performance, whereas pitch modification and, particularly, additive noise cause more pronounced degradation. Therefore, the proposed framework demonstrates strong performance under the evaluated conditions, but the results also show that acoustic noise remains an important unresolved challenge for practical deployment.

The results of this study address several limitations identified in previous deepfake speech research. First, conventional anti-spoofing systems have primarily focused on binary discrimination between bona-fide and spoofed speech. Architectures such as RawNet2 demonstrated the effectiveness of end-to-end waveform learning for spoof detection, while AASIST further demonstrated the importance of jointly modeling spectral and temporal information [8], [9]. The present framework follows the latter direction but extends it through a multi-task formulation in which the shared representation is optimized for both binary detection and supervised attack attribution. This provides a richer characterization of spoofed speech than a binary decision alone. The attribution component is also consistent with the motivation of

audio deepfake attribution research, where Yan et al. demonstrated that generation systems can leave distinguishable fingerprints that can be exploited for source identification [12].

the proposed framework addresses the representation limitation associated with relying on a single acoustic feature type. The ablation results in Tables 5 and 6 show that the combined MFCC, Log-Mel, and Δ-MFCC representation consistently provides stronger results than the individual configurations. This observation is conceptually consistent with prior spectro-temporal approaches such as AASIST, which explicitly model relationships between spectral and temporal characteristics [9]. However, the present framework uses a comparatively straightforward combination of complementary acoustic representations and integrates them within a multi-task architecture. The results therefore provide empirical evidence that combining static spectral information, detailed time-frequency information, and temporal derivatives can provide a useful representation for both detection and attribution.

the study explicitly considers domain shift and environmental variability. This is important because previous research has shown that performance obtained on controlled benchmark datasets may not reflect performance under real-world recording conditions. Müller et al. demonstrated substantial degradation when existing detectors were transferred from benchmark conditions to the In-the-Wild dataset. More recent analysis has also argued that the gap between in-domain and out-of-domain performance is strongly associated with differences between generation systems rather than simply the increasing difficulty of newer deepfakes. In contrast, the present study includes both an unseen-domain evaluation and controlled acoustic perturbation experiments. The reported ASVspoof-to-In-the-Wild experiment therefore provides an additional perspective on whether the learned representation retains useful information outside the controlled training domain.

For spoof attribution, the results can also be considered alongside recent few-shot recognition approaches. Ge et al. proposed DGPN for Speech Spoofing Algorithm Recognition and reported strong few-shot classification performance, addressing the practical problem of limited training samples for emerging spoofing algorithms. The present framework addresses a related problem from a different perspective: rather than relying on a prototype-based few-shot formulation, it performs supervised attribution of the known A01–A06 attack classes while separately examining A07–A19 as unseen attack conditions for detection generalization. Consequently, the two approaches should not be interpreted as direct competitors; rather, they represent complementary strategies for attack recognition under different assumptions about class availability and supervision.

Despite the encouraging results, several limitations should be acknowledged. First, the supervised attribution component is currently formulated as a closed-set six-class classifier covering A01–A06. Consequently, the framework does not directly assign A07–A19 to previously unseen attribution classes. Instead, these attack IDs are used to examine unseen-attack detection generalization. This distinction is important because open-set attribution requires an explicit mechanism for recognizing or rejecting classes that were not observed during training. Second, the main in-dataset evaluation uses a stratified 60:40 train-validation split rather than the official ASVspoof 2019 LA train/development/evaluation protocol. Although this provides a controlled and reproducible experimental setting, it limits direct comparison with published results obtained on the official evaluation partition. Third, the robustness analysis shows that the model remains sensitive to severe additive noise and pitch modification. The decrease in performance under noisy conditions indicates that the learned representation still depends partly on acoustic characteristics that can be disrupted by environmental interference.

Despite these limitations, the framework provides several practical contributions. Unlike conventional binary countermeasure systems that focus exclusively on detecting spoofed speech, the proposed MT-DL framework incorporates supervised attack attribution into the same learning architecture. It also combines complementary Spectro-temporal features rather than depending on a single acoustic representation, and it evaluates the model beyond the controlled benchmark through cross-dataset testing and acoustic perturbation experiments. The In-the-Wild evaluation is particularly relevant because this dataset was designed to assess generalization to realistic audio collected from online sources and contains speech from

58 public figures. Thus, the main value of the proposed framework is not simply the numerical detection score, but the integration of detection, known-attack attribution, cross-domain evaluation, and robustness analysis into a single experimental framework.

5.CONCLUSION

This study presented a Multi-Task Deep Learning (MT-DL) framework for joint audio deepfake detection and spoof attack attribution by integrating complementary spectro-temporal features (MFCC, Log-Mel spectrograms, and delta-MFCCs) with a hybrid convolutional and temporal attention architecture. Experimental results on the ASVspoof dataset demonstrate high discriminative power, achieving approximately 99.75% binary detection accuracy and 99.57% spoof attribution accuracy, while maintaining strong robustness under clean, reverberant, and time-stretched conditions. Furthermore, Grad-CAM visualizations confirmed that the model focuses on meaningful spectro-temporal regions, providing crucial interpretability for security-sensitive applications, while zero-shot cross-dataset evaluation on the In-the-Wild dataset validated its potential for real-world applicability. Despite these promising results, several key directions remain for future research to address real-world deployment challenges. Performance degradation under severe additive noise highlights the need for noise-aware feature learning, robust data augmentation, and domain adaptation, while broader cross-dataset testing across diverse codecs, devices, speakers, and languages is essential to improve generalization. Additionally, future iterations should extend attribution to open-set spoof detection to recognize emerging, unseen synthesis technologies, explore self-supervised foundation speech models to enhance representation quality under limited supervision, and apply model compression, quantization, and pruning to enable real-time, edge-device deployment for practical media authentication and voice security systems.

Funding

The authors received no financial support for this study.

Data availability

The datasets used in this study are publicly available. The ASVspoof 2019 dataset can be accessed at https://datashare.ed.ac.uk/items/31074a11-b6f6-4e92-a4ad-07093f8c0c45 accessed on 21 October 2026 ; and the In-the-Wild dataset is available at https://www.kaggle.com/datasets/abdallamohamed312/in-the-wild-dataset ,accessed on 21 October 2026.

Declarations

Ethics approval and consent to participate

Not applicable. This study does not involve experiments with human participants. The datasets used in this study are publicly available and de-identified.

Consent for publication

Not applicable. This study does not contain any individual person's identifiable information or personal data requiring consent for publication.

Competing interests

The authors declare no conflict of interest

References

[1] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, et al., "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, Art. no. 101114, 2020, doi: 10.1016/j.csl.2020.101114.

[2] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Trans. Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[3] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *arXiv preprint arXiv:2210.02437*, 2022, doi: 10.48550/arXiv.2210.02437.

[4] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?," in *Proc. Interspeech*, 2022, pp. 2783–2787, doi: 10.21437/Interspeech.2022-108.

[5] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, H. Delgado, and V. Hautamäki, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 588–604, 2017, doi: 10.1109/JSTSP.2017.2671435.

[6] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017, doi: 10.1016/j.csl.2017.01.001.

[7] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Interspeech*, 2015, pp. 2087–2091, doi: 10.21437/Interspeech.2015-472.

[8] H. Tak, J.-W. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. ASVspoof 2021 Workshop*, 2021, pp. 1–8, doi: 10.21437/ASVSPOOF.2021-1.

[9] J.-W. Jung, H.-S. Heo, H. Tak, H.-J. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE ICASSP*, 2022, pp. 6367–6371, doi: 10.1109/ICASSP43922.2022.9747766.

[10] H. Tak, J.-W. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "RawNet2: Advanced end-to-end deep neural network for spoofing systems detection," in *Proc. Interspeech*, 2021.

[11] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "RawBoost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *Proc. IEEE ICASSP*, 2022, doi: 10.17023/xyye-m850.

[12] X. Yan, J. Yi, J. Tao, and J. Chen, "Audio deepfake attribution: An initial dataset and investigation," *arXiv preprint arXiv:2208.10489*, 2022.

[13] W. Huang, Y. Gu, Z. Wang, H. Zhu, and Y. Qian, "SpeechFake: A large-scale multilingual speech deepfake dataset incorporating cutting-edge generation methods," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2025.

[14] M. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980, doi: 10.1109/TASSP.1980.1163420.

[15] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in Python,” in *Proc. 14th Python in Science Conf. (SciPy)*, 2015, pp. 18–24, doi: 10.25080/Majora-7B98E3ED-003.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[17] H. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[18] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE CVPR*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[20] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. ICML*, 2015, pp. 448–456.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[22] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.

[23] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2019.

[24] ASVspoof, “ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge,” ASVspoof Official Website,2019.[Online].Available:[https://www.asvspoof.org/index2019.html. Accessed: Sep. 14, 2026.

[25] X. Wang et al., “ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” Computer Speech & Language, vol. 64, Art. no. 101114, 2020, doi: 10.1016/j.csl.2020.101114.

[26] Fraunhofer AISEC, “In-the-Wild Dataset: A Deepfake Detection Dataset,” Fraunhofer AISEC, 2022. [Online]. Available: [https://deepfake-demo.aisec.fraunhofer.de/in_the_wild] Accessed: Sep. 14, 2026.

[27] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, “Does audio deepfake detection generalize?,” in Proc. Interspeech 2022, pp. 2783–2787, 2022, doi: 10.21437/Interspeech.2022-108.

[28] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[29] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[30] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[31] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proc. 34th Int. Conf. Machine Learning*, vol. 70, 2017, pp. 1321–1330.

[32] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12449–12460.

[33] D. *Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in Proc. IEEE ICASSP, 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.*