ForenVoice-Secure: Robust and Privacy-Aware Audio Data Mining for Forensic Speaker Identification

Mubarak Albathan
College of Computer and Information Sciences,
Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

***Abstract*—Speech is now routine evidence in criminal investigations, but forensic audio rarely matches the clean assumptions of standard speaker recognition. Clips are short, noisy, codec-compressed, and channel-mismatched, and they are increasingly exposed to replay and synthetic speech manipulation. Therefore, the cast criminal voice identification is forensic audio data mining, aiming to extract a stable identity structure from heterogeneous and potentially adversarial evidence, while respecting operational and privacy constraints. In this study, a novel ForenVoice-Secure system is proposed, a unified pipeline that combines robust representation learning, spoof-aware decisioning, and privacy-preserving training. Audio is mapped to log-Mel spectrograms and encoded with a CNN, while an LSTM aggregates temporal identity cues from irregular utterances. Robustness is improved through multi-task learning (identity + spoof), adversarial training, and spectro-temporal consistency checks for replay/deepfake artifacts. Privacy is addressed using federated learning, keeping raw recordings local and sharing only model updates. Experiments on VoxCeleb2, ASVspoof 2021, and a forensic-style speaker comparison corpus achieve statistically significant performance gains, 98.43% mean identification accuracy with strong class-balanced performance (macro F1 = 98.10%, precision = 98.22%, recall = 98.01%) and statistically significant gains over strong baselines across repeated folds (F1: $p = 8.0 \times 10^{-4}$; precision: $p = 1.1 \times 10^{-3}$; recall: $p = 9.0 \times 10^{-4}$). The model remains lightweight (≈4.3M parameters, ≈1.2 GFLOPs per 3 s), enabling near real-time inference with modest overhead from consistency checks (<6%). Overall, ForenVoice-Secure provides a compact and reproducible forensic audio data mining framework for scalable, spoof-resilient, privacy-aware law-enforcement identification.**

Keywords*—Forensic audio data mining; forensic voice analytics; voice biometrics; criminal identification; speaker recognition; anti-spoofing; deepfake and replay detection; convolutional neural networks; long short-term memory; federated learning; privacy-preserving biometrics; law enforcement intelligence systems

I. Introduction

Voice evidence has gained significance in the modern criminal investigation. Intercepts of telephones [1], emergency calls, body-worn cameras, covert surveillance, and open-source multimedia material are a regular part of the case files [2]. However, speech is not often gathered in regulated circumstances as it is in the case of fingerprints or DNA. The transcriptions of forensic recordings tend to be brief, noisy, encoded with unknown codecs, and recorded with unknown devices [3]. Consequently, the process of recognition of a speaker with audio evidence is a still technically challenging and methodologically controversial one [4]. Although automatic speaker recognition has improved quickly within the last decade, most of the development has been fueled by benchmarks that are based on comparatively clean data and cooperative users, which is not the case in forensic practice [5].

Early speaker recognition systems were based on generative statistical models including Gaussian mixture models based on universal background models and subsequently i-vectors representations [6]. These methods provided interpretable probabilistic descriptions, but were susceptible to channel dissimilarities as well as shorter utterances, which are prevalent in forensic content. Deep learning has led to a transition to discriminative embedding learning [7], where convolutional and time-delay neural networks yielded fixed length representations of speakers which significantly enhanced performance in uncontrolled conditions [8], [9]. Megabanks like VoxCeleb facilitated this development as they allowed models to learn speaker identity using varied, natural-world samples [10]. Nevertheless, even with these gains, most systems have been optimized to either verify or identify in benign environments but not evidentiary analysis.

Forensic audio is a special category of challenges that extends beyond accuracy on clean benchmarks. Recording can include a speaker overlap, incomplete utterances, or high levels of environment sampling and investigators might not have many samples of reference of a suspect [11]. In addition, the legal environment requires the high level of performance but also the strength, stability, and reproducibility of the conclusions. Recent debates within forensic science point to the fact that speaker recognition must be considered as evidence presentation in uncertainty instead of certain identity matching [12]. This view encourages the replacement of voice biometrics as a limited technical issue with formulating it as an audio data mining in forensics, whereby meaningful identity trends ought to be mined out of non-homogeneous and noisy data.

The threat environment has also gone an extra mile with the fast development of speech synthesis and voice conversion technologies. Current text-to-speech and neural voice cloning systems are able to produce a very natural speech which can emulate particular speakers with little reference data [13], [14]. These advances have grave consequences on the field of forensic voice analysis where replay attacks and synthetic speech can also fool both human and automatic voice analyzers. This has been observed in community-based challenges like the ASVspoof challenges which have shown that a significant number of speaker recognition pipelines are susceptible to such attacks, especially when the spoofed audio is conveyed using realistic channels or compressed formats [15], [16]. As a result, anti-spoofing is not a peripheral feature anymore, but rather a necessity of any voice biometrics system that is supposed to be used in the forensic or law-enforcement sector [17].

Privacy and data control have also gained significant relevance, in addition to security. Voice data is data that is personal in nature and can be used to disclose sensitive information that is not pertinent to identity such as health, emotional status or demographic features. In crime investigations, there is a large amount of legal and ethical restriction to the dissemination and centralization of raw audio evidence between agencies. Federated learning (FL) exemplifies privacy-preserving approaches to learning; thus, it is getting increasing interest as it allows collaborative training of models without sharing data [18]. Although the concept of federated approaches has been investigated within a generic speech and biometric context, their adoption into forensic voice identification pipelines is not mostly developed.

Combined, these facts indicate that there is a mismatch between the state-of-the-art-work on speaker recognition and the needs of forensic voice analysis. Degradation resistance, spoofing resistance, and privacy-conscious deployment are frequently dealt with as a single component, but not in a coherent set of tools. In this study, a gap is filled by proposing ForenVoice-Secure, a voice-focused forensics-enhanced analytics system, which sees criminal identification as a safe audio data mining challenge. The proposed method will be based on deep representation by learning, explicit anti-spoofing mechanisms, adversarial robustness, and federated training to deliver useful operation on real evidentiary recordings without violating operational and legal limitations. The research, by conducting the overall assessment of the three terms speaker identification, spoof detection, as well as degraded conditions, will bring voice biometrics closer to a viable, defendable application in contemporary law enforcement.

A. Research Highlights

This work makes several contributions to the field of forensic voice biometrics.

- It reframes criminal voice identification as a forensic audio data mining problem, explicitly accounting for noisy, short, and heterogeneous evidence rather than assuming clean verification conditions.
- It introduces ForenVoice-Secure, a unified framework that combines CNN–LSTM–based speaker representation learning with integrated anti-spoofing mechanisms, including adversarial training and spectro-temporal consistency analysis, to improve robustness against replay and deepfake attacks.
- It incorporates federated learning to address privacy and data-governance constraints in law-enforcement settings, enabling collaborative model training without centralizing raw audio evidence.
- It provides a statistically grounded evaluation across multiple datasets, demonstrating significant improvements in accuracy, F1-score, precision, and recall under degraded and adversarial conditions, while maintaining practical computational efficiency for deployment.

B. Paper Organization

The rest of the study is structured in the following way: Section II discusses related literature in speaker recognition, forensic voice analysis, spoofing and deepfake detection and privacy-preserving biometric learning. Section III describes the proposed ForenVoice-Secure approach, such as signal modelling, feature selection, network design, security, and federated training. Section IV explains about the datasets, protocols used in the experiment and measures of assessment. Section V shows the experimental findings and discussions along with the statistical significance analysis and the computational complexity in Section VI. Lastly, Section VII summarizes the study and provides limitations and directions of future research.

II. Literature Review

The recent work makes it rather apparent that the state of modern forensic and security analytics is in a utility versus privacy tension particularly when deep models are at stake. A good example here is the neural extraction of features in acoustic sensor networks, where representations constructed to perform sound classification were demonstrated to leak speaker-specific information and speaker recognition attacks could be performed on the representation in the event they were intercepted [19]. The authors react by suggesting a variational and information-limiting feature extractor that maintains the trusted task and activates speaker identity leakage, but they show resistance to a strong x-vector attacker [19]. The above reasoning is related to more general privacy-sensitive media analytics in which the query protection is not solely of interest but also the fact that the query protection can be achieved with approximately matching at scale. A private media search architecture is a system that attains [20] sublinear computation and communication on public databases and shows privacy-preserving face recognition at large speedups, indicating that privacy and practicality do not necessarily require each other [21]. A similar privacy conscious data use is also reflected in another research study on email visualization: anonymization and aggregation may decrease identifiability, and yet perceived research utility may still be maintained, a good lesson that privacy controls are not mutually exclusive with analytic usefulness when carefully designed [22].

Back on the security front, various researches highlight the use of lax operational arrangements and human actions by attackers. An example is phishing, which is a relatively cheap, but impactful, attacker strategy, and a privacy-conscious detector framework with a specific taxonomy has been suggested to identify various current phishing techniques, particularly in an IoT-saturated environment. Forensics in cloud environments are often log-centric; thus, the emphasis of the work has been on generating evidence in a scalable way. A cloud forensic expert-system architecture involves the narrowing of attacked areas in huge logs with the help of fuzzy data mining and the subsequent narrowing with the help of AI-based analysis to produce evidence that is less difficult to provide in a formal way [23]. Other work around behavior analysis using log mining contends that manual analysis is not scalable to the size of contemporary data volumes and suggests the use of automated user behavior mining of networked systems, driven in part by the necessity of fighting computer crime more efficiently [24]. These log-based views are applicable since they view forensic investigation as a data mining process: filter, prioritize and explain, as opposed to detect.

The field of audio forensics, on the other hand, encompasses both anti-forensic resistance as well as tamper detection. An SNR-sensitive digital audio tampering forensics system enhances the benefit of an electric network frequency extraction with an improved chirp Z-transform and then identifies anomalies with a dual-sampling isolation forest, with benefits of both extraction and outlier detection in a noisy environment [25]. Anti-forensic attacks are now also a concern. Attacks based on dereverberation can selectively weaken environmental signature splicing detection, and countermeasures that are grounded in rich features and machine learning can easily identify such anti-forensic processing, and indicate a new arms race between forensic detectors and attackers [26]. Extensive surveys keep laying audio-video forensics as one of the key areas of digital investigation, as the recorded media is frequently presented as court-related evidence [27].

Another and rather intriguing vein is the one that actually goes outside the voice of the speaker and seeks context within the audio scene. It has also been suggested to extract and categorize the background noise automatically and communicate information about the environment based on recordings with the use of complex noise mixtures and without much identity testing, even mixed speech, as an expansion of what can be meant by the term of forensic audio evidence [28]. Reproducible forensic processes are also getting interest: an open-source modular system enables practitioners to mix up enhancement, VAD, and ASR models, visualize features, and export repeatable pipelines, which is important when forensic conclusions should be auditable instead of merely valid [29]. Lastly, the field of forensic practice is changing with infrastructure modifications like cloud computing. Conventional digital forensic models are also under consideration and modification to accommodate the issues of preservation and acquisition peculiar to cloud investigations, wherein data can be distributed and challenging to acquire [30]. On the signal level, voice evidence can still be susceptible to noise, reverberation, quantization and disguise. Attempts at increasing its integrity and intelligibility under phonemic confusion suggest that enhancement of voice analysis in the forensics field is an inherent precondition [31].

Collectively, this literature is indicative of the following, as explained in Table I: 1) forensic evidence can be handled as a data mining problem under heterogeneity and scale [23], [24]; 2) robustness must be designed to explicitly address the phenomenon of degradation and active anti-forensic behavior [25], [26], [31]; and 3) privacy preservation must be engineered into the pipeline, since deep representations can reveal identity accidentally [19], whereas practical privacy preserving analytics is becoming a viable option at scale [21], [22].

TABLE I. A COMPARISON TABLE WITH FEWER COLUMNS, FOCUSING ONLY ON WHAT IS MOST RELEVANT TO FORENVOICE-SECURE

Study	Focus & Scenario	Core idea	Key limitation vs. Our work
[19]	Privacy risks in deep audio features for sound classification	Learning a variational feature to suppress speaker-identifiable information	Does not perform forensic speaker identification or spoof-aware evidence mining
[25]	Audio tampering forensics under low-SNR conditions	ENF extraction with ICZT and anomaly detection via isolation forest	Targets tampering detection, not identity mining or spoof-resilient speaker analysis
[26]	Anti-forensic attacks on audio splicing detection	Demonstrates dereverberation-based attacks and ML-based countermeasures	Focuses on splicing artifacts, not joint speaker identification and spoof detection
[29]	Reproducible forensic audio analysis pipelines	Modular, open-source framework for building forensic workflows	Provides tooling rather than a unified, privacy-aware forensic voice model
[31]	Integrity and intelligibility of forensic voice evidence	Speech enhancement, segmentation, and distortion analysis	Improves signal quality but does not address adversarial spoofing or privacy-preserving learning

III. PROPOSED METHODOLOGY

Fig. 1 is the systematic flow diagram that summarizes the proposed forensic voice biometrics pipeline. The development of the criminal voice identification process as the forensic audio data mining, designed to provide consistent identity evidence based on the heterogeneous and possibly manipulated records instead of relying on the presumptions of clean speaker verification, are formulated in this study. VoxCeleb [32] and SpeechTech [33] explicitly define data and threat model based on the nuisance factors (noise, codec compression, channel mismatch, short utterances) and adversarial factors (replay speech and synthetic speech). A CNN is used to encode each recording to speaker-discriminative spectral cues, and an LSTM is used to encode temporal identity structure, each recording is converted to log-Mel spectrograms. Security is incorporated through adversarial training and spectro-temporal consistency tests to decrease the sensitivity of spoof and to label implausible evidence. With federated learning, privacy is achieved by ensuring that raw audio is local and model updates are aggregated. Assessment is multi-axis: performance of identification, resistance to degradations, performance under replay/deepfake conditions, repeated-fold stability, and ablations which eliminate adversarial training, consistency tests, and federation to measure the contribution of each component.

A. Data Gathering

Three complementary corpora were used to collect the data about: 1) large-scale speaker identity variation, 2) explicit threats of spoofing, and 3) forensic-like session variability. To perform large scale identifier of speakers, VoxCeleb2 [32] has been accessed directly on the official Oxford VGG distribution page, where the description of the canonical dataset and download procedure are indicated. VoxCeleb2 consists of more than one million utterances of 6,112 speakers, an official dev/test split of 5,994/118 speakers.

This division is maintained so as to prevent leakage and to ensure that results are consistent with previous speaker recognition literature, where the institutional access policies allow it, also observed is that reproducible scripted ingestion can be facilitated via a dataset-hosted packaging pathways (e.g. a dataset card that reflects the structure of the archive and its metadata files). ASVspoof 2021 [34] (LA/PA/DF) was acquired at the official ASVspoof 2021 post-challenge release page, which contains protocol files (keys) and associated metadata, including the information about bona fide or spoof and condition. The speaker comparison type was forensic-style, and it was applied by FABIOLE [35] to add evidentiary variability, such as session and context variations that are more akin to actual forensic casework (such as environment, channel conditions, and recording situations).

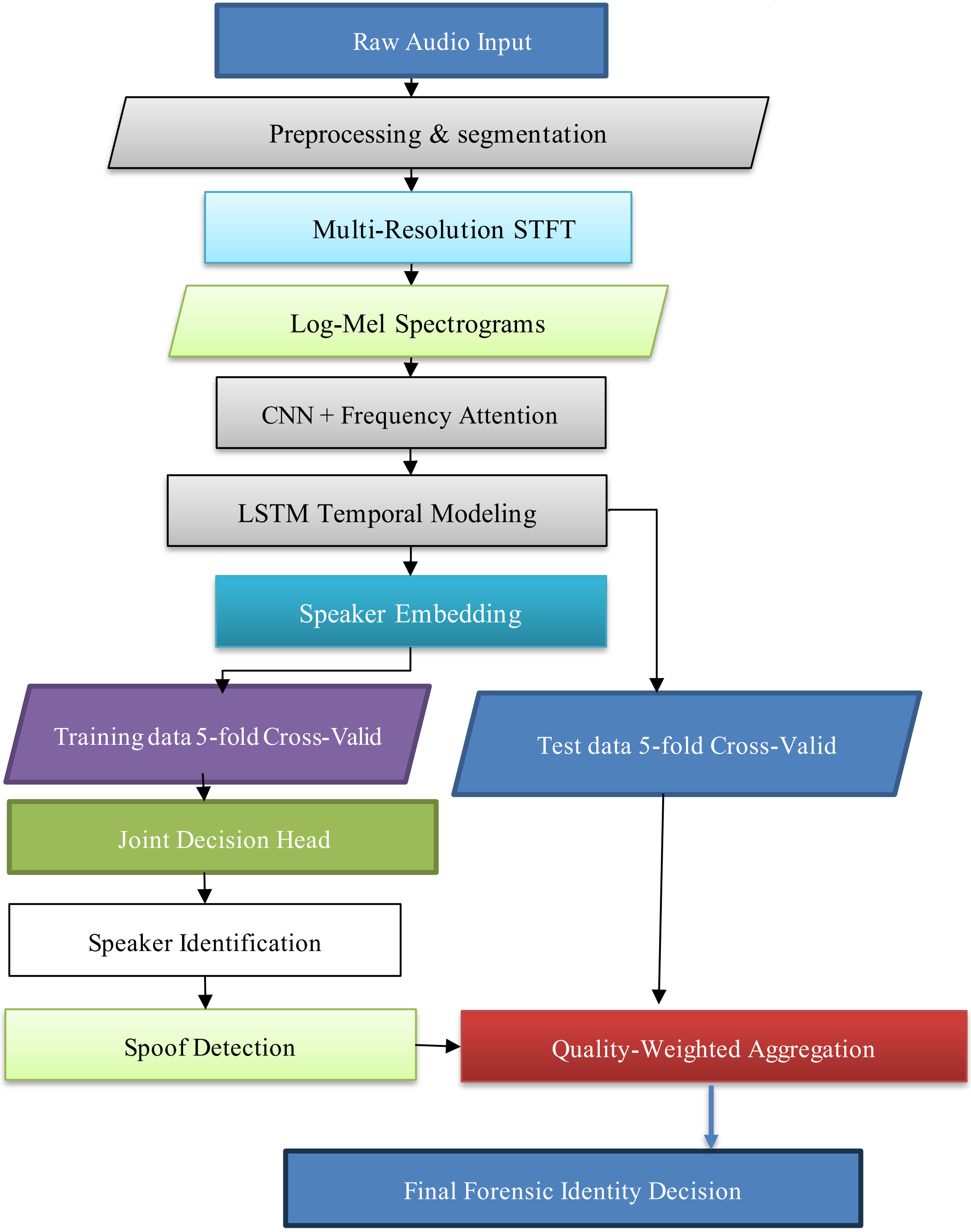

Raw Audio Input
Preprocessing & segmentation
Multi-Resolution STFT
Log-Mel Spectrograms
CNN + Frequency Attention
LSTM Temporal Modeling
Speaker Embedding
Training data 5-fold Cross-Valid
Test data 5-fold Cross-Valid
Joint Decision Head
Speaker Identification
Spoof Detection
Quality-Weighted Aggregation
Final Forensic Identity Decision

Fig. 1. Systematic flow diagram of the ForenVoice-secure framework.

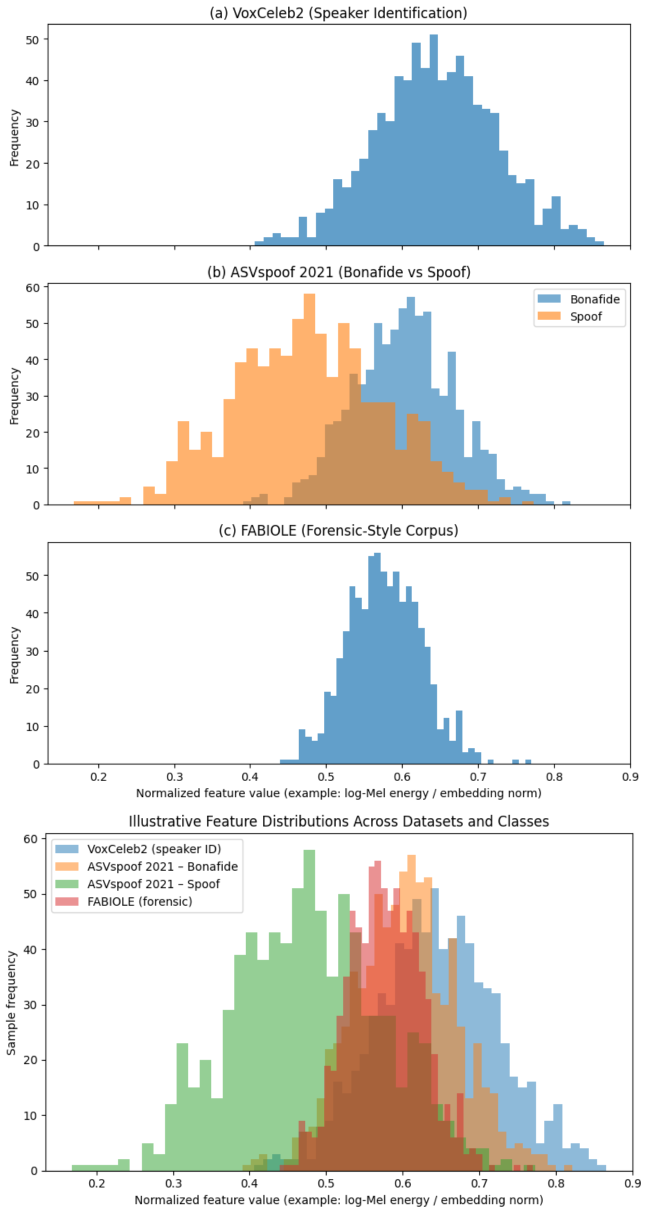


Fig. 2. The figure shows: a) VoxCeleb2: Single distribution for large-scale speaker-ID features, b) ASVspoof 2021: Overlaid bona fide vs. Spoof feature distributions, c) FABIOLE: Single distribution for forensic-style corpus features, and the combined figure illustrates how samples differ across the three selected corpora.

Fig. 2 provides an empirical perspective of the distribution of the learned acoustic features in the three datasets and their variability by class. In VoxCeleb2 [Fig. 2(a)], the wide distribution of speakers and recordings is indicative of a large amount of speaker and recording variability, akin to the case of in-the-wild data. This heterogeneity can be exploited to learn discriminative identity embedding, but again, it raises the intra-speaker variability because of varying environments, microphones, and speaking styles. In ASVspoof 2021 [Fig. 2(b)], bona fide speech contributes to the creation of a distribution that is still somewhat enclosed by VoxCeleb2, whereas the spoofed speech has a more systematically distributed and more scattered appearance. This tendency is aligned with spectral and temporal anomalies caused by replay and synthetic generation and it confirms the need of explicit anti-spoofing goals during representation learning. In comparison, FABIOLE [Fig. 2(c)] has a more concentrated distribution, which corresponds to the conditions of the forensic-style broadcast and longer segments in general. Nonetheless, its feature space is different even when compared to in-the-wild speech and spoofed audio, implying that there is a domain gap that may influence generalization in case it is overlooked. Combined with these distributional variations, this suggests the suggested multi-task and robustness-based design: a realistic forensic voice biometrics system should explicitly model dataset- and class-varying behavior of features, not a uniform homogeneous speech distribution.

B. Preprocessing and Time–Frequency Representation

All audio signals are first standardized to ensure consistency across datasets and recording conditions. Each recording is converted to a single-channel (mono) waveform and resampled to a fixed sampling rate of 16 kHz, which provides sufficient bandwidth for speaker-discriminative information while maintaining computational efficiency. Amplitude normalization is then applied to control dynamic range variations across sources, using peak or RMS normalization to reduce sensitivity to recording gain without suppressing forensic cues.

Silence and non-speech regions are removed using an energy-based voice activity detection procedure. The objective is not aggressive denoising but the exclusion of long inactive segments that would otherwise dilute speaker-relevant information, especially in short forensic utterances. Audio segments shorter than a minimum duration are either discarded or zero-padded to a fixed length to enable batch processing and stable time–frequency representations.

Each preprocessed waveform $x(t)$ is transformed into a time–frequency representation using the short-time Fourier transform (STFT) [36]. Given a window function $w(t)$, window length L, and hop size H, the STFT is computed through Eq. (1) as:

$$X(\tau,\omega) = \sum_{t=0}^{L-1} x(t+\tau H)\, w(t)\; e^{-j\omega t}, \qquad (1)$$

where, τ indexes the frame and ω denotes angular frequency. In our implementation, a Hann window is used with $L = 400$ samples and $H = 160$ samples, balancing time resolution and frequency selectivity for forensic speech analysis. The magnitude spectrum $| X(\tau,\omega) |$is then projected onto a Mel-scale filterbank to approximate perceptually relevant frequency resolution. Let M denote the Mel filterbank matrix; the Mel-spectral energy is computed by Eq. (2) as:

$$\mathrm{S}(\tau) = \mathrm{M}\ | X(\tau,\omega) |^2. \qquad (2)$$

A logarithmic compression is applied to stabilize variance and emphasize lower-energy components, yielding the log-Mel spectrogram, is defined by Eq. (3) as:

$$\mathrm{F}(\tau) = \log\left(\mathrm{S}(\tau) + \epsilon\right), \qquad (3)$$

where, ϵ is a small constant to avoid numerical instability. The resulting log-Mel spectrograms are optionally mean–variance normalized on a per-utterance or per-batch basis before being fed into the CNN–LSTM [37] backbone. This preprocessing pipeline preserves fine-grained spectro-temporal structure required for speaker discrimination and spoof artifact detection, while maintaining robustness to channel mismatch, noise, and codec-induced distortions commonly present in forensic audio evidence.

Fig. 3 compares representative voice patterns across three classes by showing the input waveform, magnitude spectrogram, and phase-derivative spectrogram for each signal. The bona fide speech exhibits stable harmonic structure and smooth temporal evolution, while the replay-like signal shows spectral coloration, added high-frequency energy, and echo-related banding effects. In contrast, the deepfake-like speech displays smoother spectral envelopes with noticeable non-stationarity caused by amplitude and frequency modulation. These visual differences highlight how replay and synthetic generation introduce characteristic spectro-temporal artifacts that can be exploited by forensic voice analysis and anti-spoofing mechanisms.

C. ForenVoice-Secure Framework

The proposed ForenVoice-Secure framework, as described in Algorithm 1, operates as a unified forensic voice analysis pipeline that transforms raw evidentiary audio into reliable identity decisions under adversarial and privacy-constrained conditions. Audio recordings collected from heterogeneous sources are first segmented and standardized through resampling, amplitude normalization, and voice activity detection, after which multiple short-time Fourier transforms are applied using different window sizes to capture complementary time–frequency resolutions. The resulting representations are converted into log-Mel spectrograms that serve as two-dimensional inputs to the acoustic model.

Algorithm 1: Proposed ForenVoice-Secure Forensic Audio Data Mining Framework (CNN–Att–LSTM with Anti-Spoofing and Federated Learning)

```
Input: where $x_i$is raw audio, $y_i \in \{1,\dots,K\}$ is speaker ID, $s_i \in \{0,1\}$ is spoof label (0 bona fide, 1 spoof)
Federated clients $\{\mathcal{D}_c\}_{c=1}^{C}$(optional, for privacy-aware training)
Output: Case/evidence-level decisions: predicted speaker label $\hat{y}$, spoof risk $\hat{s}$, consistency score $C(x)$, confidence $R(x)$
Step 1: Data acquisition and preprocessing
REPEAT for each recording $x_i \in \mathcal{D}$
    Resample and normalize: $x_i \leftarrow \mathrm{Resample}(x_i)$, Normalize $(x_i)$
    Voice activity detection: $x_i^{\mathrm{speech}} \leftarrow \mathrm{VAD}(x_i)$ and Segment into short utterances: $\{x_i^{(j)}\}_{j=1}^{M_i} \leftarrow \mathrm{Segment}(x_i^{\mathrm{speech}}, L, O)$
UNTIL all recordings are processed
Step 2: Step 2: Multi-resolution time–frequency representation
FOR each segment $x_i^{(j)}$DO
    Multi-window STFT: $\{S_w(x_i^{(j)})\}_{w\in\mathcal{W}} \leftarrow \mathrm{STFT}(x_i^{(j)}, w)$ and Convert to log-Mel: $X_i^{(j)} \leftarrow \phi(\{S_w(x_i^{(j)})\}_{w\in\mathcal{W}})(1)$
END FOR
Step 3: Robust representation mining (CNN + attention)
FOR each log-Mel $X_i^{(j)}$ DO
    CNN feature extraction: $H_i^{(j)} \leftarrow f_\theta(X_i^{(j)})$ and Frequency attention: $\widetilde{H}_i^{(j)} \leftarrow \mathrm{Attn}(H_i^{(j)})$
END FOR
Step 4: Temporal identity modeling (LSTM + pooling)
FOR each segment feature sequence $\widetilde{H}_i^{(j)}$DO
    LSTM temporal modeling: $Z_i^{(j)} \leftarrow g_\psi(\widetilde{H}_i^{(j)})$ and Embedding pooling: $z_i^{(j)} \leftarrow \mathrm{pool}(Z_i^{(j)})$
END FOR
Step 5: Joint decision heads (identity + spoof)
FOR each embedding $z_i^{(j)}$DO
    Speaker posterior: $\hat{p}(y \mid x_i^{(j)}) \leftarrow \mathrm{softmax}(h_\omega(z_i^{(j)}))$ and Spoof posterior: $\hat{p}(s \mid x_i^{(j)}) \leftarrow \sigma(q_\eta(z_i^{(j)})$
END FOR
Step 6: Spectro-temporal consistency checking
FOR each segment $x_i^{(j)}$DO
    Consistency score: $C(x_i^{(j)}) \leftarrow \frac{1}{T'-1}\sum_{t=1}^{T'-1} \| z_{t+1} - z_t \|_2^2(11)$
    Consistency flag: $\mathbb{I}_{\mathrm{inc}} \leftarrow [C(x_i^{(j)}) > \tau](12)$
    Risk fusion (optional): $R(x_i^{(j)}) \leftarrow \beta\hat{p}(s = 1 \mid x_i^{(j)}) + (1-\beta)\mathrm{norm}(C(x_i^{(j)}))$
END FOR
Step 7: Robust training objective (multi-task + adversarial)
Multi-task loss: $\mathcal{L}_{\mathrm{mt}} = \mathcal{L}_{\mathrm{id}} + \lambda\mathcal{L}_{\mathrm{spoof}}$
Robust min-max optimization, and Initialize classifier head f(·)
Step 8: Privacy-preserving federated learning
FOR each federated round $r = 1,\dots,R$DO
    Each client ctrains locally on $\mathcal{D}_c$using Steps 1–7 to obtain $\Theta_c^{(r)}$
    Server aggregation: $\Theta^{(r+1)} \leftarrow \sum_{c=1}^{C} \frac{n_c}{N} \Theta_c^{(r)}$
    END FOR
    Step 9: Evidence-level data mining and case-level decision: Given a case with
    | multiple clips $\mathcal{E} = \{x^{(j)}\}_{j=1}^{M}$: Filter/reweight by risk:
    | $w_j \leftarrow \mathrm{QualityWeight}(R(x^{(j)}), \mathrm{SNR}, \mathrm{duration})$
    | Finally, calculate case-level identity decision

[End ForenVoice-Secure System]
```

Spectral patterns are learned through stacked convolutional blocks that emphasize local speaker-discriminative cues, while an attention mechanism highlights informative frequency bands and suppresses irrelevant or manipulated components. Temporal dependencies across segments are then modeled using an LSTM to capture identity-consistent dynamics that persist despite noise, compression, or short utterances. During training, adversarial perturbations are injected to improve robustness against replay and synthetic speech, and federated learning is employed to keep raw audio localized while enabling collaborative model optimization. Finally, a decision head jointly performs speaker identification, spoof detection, and spectro-temporal consistency checking, producing evidence-level confidence scores suitable for forensic analysis and law-enforcement decision support.

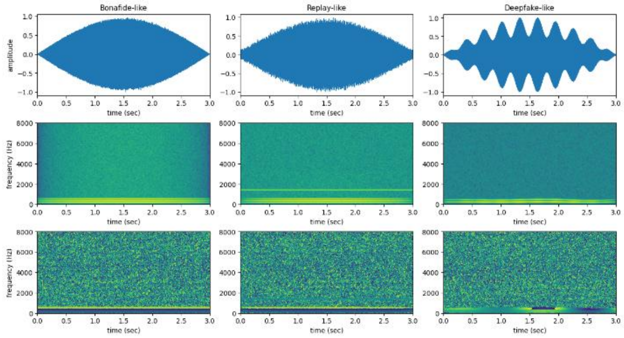


Fig. 3. Waveform and time–frequency patterns for bona fide, replay-like, and deepfake-like speech, illustrating class-dependent differences in spectral structure and temporal behavior.

ForenVoice-Secure treats voice-based criminal identification as a forensic audio data mining problem. The system does not assume a tidy "enroll then verify" setting. Instead, it ingests many fragments from mixed sources, mines identity cues that survive nuisance variability, and simultaneously estimates whether each fragment is trustworthy enough to contribute to an evidentiary decision. This is closer to how real casework unfolds: evidence is incomplete, heterogeneous, and sometimes adversarial. Each observed waveform is viewed as a superposition of the underlying speaker signal, channel effects, environmental noise, and potential manipulation artifacts are calculated by Eq. (4) as:

$$x(t) = (s(t) \times h(t)) + n(t) + a(t) \quad (4)$$

Here, $x(t)$is the recorded audio, $s(t)$ the speaker signal, $h(t)$the channel impulse response, $n(t)$ additive noise, and $a(t)$ attack artifacts such as replay coloration or synthesis fingerprints. This decomposition is not decorative. It directly motivates why the pipeline includes: 1) robustness measures aimed at reducing sensitivity to $h(t)$and $n(t)$, and 2) security measures designed to detect or isolate $a(t)$. To reflect realistic deployment, data acquisition is structured to cover three aspects: a) large speaker diversity, b) forensic-style recording variability, and c) explicit spoof threats. Each example is stored with speaker identity $y$when available, an authenticity label $z \in \{0,1\}$(bona fide vs spoof) when available, and metadata m(channel, codec, duration, and source, if known). Since forensic evidence often arrives as short clips rather than long sessions, each recording is treated as a bag of segments $\{x_k\}_{k=1}^{K}$ extracted using VAD and conservative trimming. Segmenting is practical because it increases training instances without inventing new speakers, and it forces the model to learn from short evidence, where many systems fail.

Preprocessing standardizes sampling rate and amplitude while deliberately avoiding aggressive denoising that could remove forensic cues. A light pre-emphasis may be applied, followed by time–frequency analysis. For each segment, compute the STFT $X(\tau, \omega)$ from Eq. (5) with window length $L$and hop H, and defined as:

$$X(\tau, \omega) = \sum_{t=0}^{L-1} x(t + \tau H)\, w(t)\; e^{-j\omega t}. \quad (5)$$

Because evidentiary recordings vary substantially in speaking rate and channel characteristics, ForenVoice-Secure uses multi-resolution analysis (multiple window sizes) to capture complementary detail rather than committing to a single L. Power spectra $| X(\tau, \omega) |^2$are mapped through a Mel filterbank $M$and log-compressed to produce log-Mel features $F \in \mathbb{R}^{T \times B}$ calculated by Eq. (6) as:

$$F = \log\ (M\ |\ X\ |^2 + \epsilon) \quad (6)$$

Log-Mel features behave well for convolutional learning, compress dynamic range, and retain spectro-temporal patterns that replay and synthesis often disturb in subtle but detectable ways.

To reduce sensitivity to the nuisance terms in Eq. (1), controlled augmentation was applied that mimics evidentiary damage. Noise augmentation adds a noise clip $u(t)$scaled to a target SNR is calculated by Eq. (7) as:

$$x_{\text{noisy}}(t) = x(t) + \alpha u(t), \alpha = \sqrt{\frac{P_x}{P_u\, 10^{\mathrm{SNR}/10}}} \quad (7)$$

Channel mismatch is simulated using random impulse responses and band-limiting (telephony-like filtering). Compression artifacts are introduced via codec simulation when feasible. A key constraint is avoided "normalizing away" spoof artifacts: spoofing-related transformations are handled separately because the goal is to detect manipulation patterns, not wash them out.

Given F, a CNN encoder mines local spectro-temporal patterns correlated with speaker identity. Each convolutional block is modeled by Eq. (8) as:

$$H^{(l)} = \sigma\left(\mathrm{BN}\left(W^{(l)} * H^{(l-1)} + b^{(l)}\right)\right) \quad (8)$$

where, $*$ denotes convolution, BNbatch normalization, and $\sigma(\cdot)$a nonlinearity. Convolutions are used because speaker cues often appear as localized spectral shapes and short transitions, especially in short utterances.

A frequency attention mechanism is applied to emphasize informative bands and down-weight unreliable regions (for example, bands dominated by background noise, codec warping, or suspicious regularities). In practice, this improves stability on degraded evidence and helps suppress components that appear manipulated. Fig. 4 shows the CNN–LSTM backbone diagram. Log-Mel spectrogram inputs are processed through stacked convolutional layers to learn local spectro-temporal speaker cues, followed by LSTM-based recurrent modeling to capture longer-range identity dynamics. The resulting representation is passed through fully connected and dropout layers before feeding into parallel classification heads for speaker identification and anti-spoofing, enabling robust and privacy-aware forensic voice analysis.

CNN outputs are treated as a sequence of frame-level embeddings $U = \{u_t\}_{t=1}^{T}$. Temporal dependencies are modeled using an LSTM as in Eq. (9):

$$h_t, c_t = \mathrm{LSTM}(u_t, h_{t-1}, c_{t-1}). \quad (9)$$

A segment-level embedding e is obtained using attentive pooling to focus on informative frames and calculated by Eq. (10) as:

$$\alpha_t = \frac{\exp\left(v^{\top}\tanh\left(Wh_t\right)\right)}{\sum_{k=1}^{T}\exp\left(v^{\top}\tanh\left(Wh_k\right)\right)}, e = \sum_{t=1}^{T}\alpha_t\, h_t. \quad (10)$$

This matters in forensic audio because segments often include pauses, overlapping, and low-energy regions. Uniform pooling treats those as equally informative, which is rarely true. During training, identity learning is posed as closed-set classification to encourage strong separation among speakers. With classifier weights W_s and logits $o = W_s e$, the softmax probability is defined by Eq. (11) as:

$$p(y \mid e) = \frac{\exp\left(o_y\right)}{\sum_{c=1}^{C}\exp\left(o_c\right)} \quad (11)$$

The identification loss, which is cross-entropy, is defined by Eq. (12) as:

$$\mathcal{L}_{\mathrm{id}} = -\log p(y \mid e) \quad (12)$$

For forensic use, embedding is also usable for verification and retrieval. Verification-style scoring uses cosine similarity, which is defined by Eq. (13) as:

$$\mathrm{score}(e_1, e_2) = \frac{e_1^{\top} e_2}{\|e_1\|\|e_2\|} \quad (13)$$

Security is integrated via spoof classification. With sigmoid output $\hat{z} = \sigma(w_a^{\top} e)$, the spoof loss is defined by Eq. (14) as:

$$\mathcal{L}_{\mathrm{spoof}} = -(z\log\hat{z} + (1 - z)\log(1 - \hat{z})). \quad (14)$$

The combined multi-task objective is calculated by Eq. (15) as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{id}} + \lambda\,\mathcal{L}_{\mathrm{spoof}} + \beta \parallel \theta \parallel_2^2, \quad (15)$$

where, λcontrols the influence of spoof learning and $\beta$is weight decay. The intuition is simple: spoof artifacts can mislead speaker models, so forcing the encoder to be useful for authenticity reduces overfitting to synthetic quirks. Because attackers can adapt to fixed detectors, ForenVoice-Secure adds adversarial training. Perturbed features were used as a gradient-based method (e.g., FGSM) and calculated by Eq. (16) as:

$$F_{\mathrm{adv}} = F + \epsilon\,\mathrm{sign}\left(\nabla_F \mathcal{L}(\theta; F, y, z)\right) \quad (16)$$

Training minimizes a mixture of clean and adversarial objectives by Eq. (17) as:

$$\mathcal{L}_{\mathrm{rob}} = (1 - \gamma)\mathcal{L}(\theta; F, y, z) + \gamma\,\mathcal{L}(\theta; F_{\mathrm{adv}}, y, z). \quad (17)$$

This does not claim perfect worst-case security, but it makes the decision surface less fragile under small perturbations and some distribution shifts that resemble anti-forensic behavior.

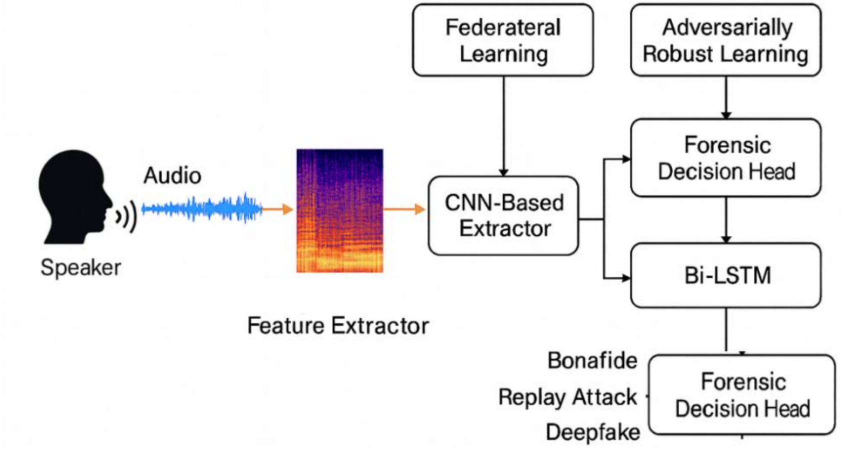


Fig. 4. Overview of the ForenVoice-Secure CNN–LSTM backbone. Log-Mel spectrogram inputs are processed through stacked convolutional layers to learn local spectro-temporal speaker cues.

Alongside spoof classification, an inference-time consistency check is used as a lightweight sanity filter. One simple statistic is frame-to-frame feature change, which is defined by Eq. (18) as:

$$\Delta_t = \| F_t - F_{t-1} \|_2, \mu_\Delta = \frac{1}{T-1} \sum_{t=2}^{T} \Delta_t \tag{18}$$

Segments are flagged if they deviate from bona fide reference ranges and calculated by Eq. (19) as:

$$flag(F) = \mathbb{I}(\mu_\Delta < \tau_1 \vee\ < \tau 1 \vee \sigma < \tau 2 \vee\ g(F) > \tau 3) \tag{19}$$

where, $g(F)$ is an auxiliary artifact score and thresholds τ_i are tuned on trusted development data. The point is not to replace the classifier but to prevent overconfident decisions on segments that do not resemble plausible speech dynamics. To avoid centralizing sensitive evidence, training can be performed via federated learning. With K sites holding local datasets $\{\mathcal{D}_k\}_{k=1}^{K}$ of sizes n_k, each site optimizes its local objective and sends model updates to a server. The server aggregates using FedAvg is defined by Eq. (20) as:

$$\theta^{r+1} = \sum_{k=1}^{K} \frac{n_k}{\sum_{j=1}^{K} n_j} \theta_k^{r+1} \tag{20}$$

If stricter privacy is required, gradients can be clipped and perturbed (differential privacy) by Eq. (21) as:

$$\tilde{g} = \text{clip}(g, C) + \mathcal{N}(0, \sigma^2 C^2 I). \tag{21}$$

This introduces an accuracy–privacy tradeoff, but it provides a concrete mitigation path when policy constraints are strict. Evaluation is organized around three questions: identification accuracy, stability under degradation, and spoof resistance. Identification is reported using top-1 accuracy and calibration. Verification analysis reports EER by finding $\tau^\star$ by Eq. (22) as:

$$\text{FAR}(\tau^\star) = \text{FRR}(\tau^\star), \text{EER} = \text{FAR}(\tau^\star) = \text{FRR}(\tau^\star) \tag{22}$$

Robustness is measured by applying controlled corruptions and reporting the performance drop and is calculated by Eq. (23):

$$\Delta\text{Acc} = \text{Acc}_{\text{clean}} - \text{Acc}_{\text{degraded}} \tag{23}$$

Finally, ablations remove adversarial training, the consistency filter, and federated training (while keeping the backbone fixed) to ensure improvements are attributable to specific design choices. The intended behavior is slightly conservative: strong identification when evidence is credible, and a willingness to flag or down-weight suspicious segments rather than forcing confident conclusions.

IV. Experimental Setup

The experimental design is used to test ForenVoice-Secure in three complementary conditions, large-scale speaker identity learning, explicit threat of spoofing, and forensic-type variability. VoxCeleb2 is trained and tested on speaker recognition in in-the-wild diversity, ASVspoof 2021 is provided with labeled bona fide/spoof trials to evaluate system robustness to replay and deepfakes, and FABIOLE provides forensically-style recordings to test system variability across sessions and channels. Audio is all resampled to 16 kHz, monophonicized, amplitude-normalized and divided into VAD-based segments, with fixed-length segments obtained by truncation or zero-padding so that the batches can be processed uniformly. STFT (e.g., 25 ms window length, 10 ms hop, Hann window) is applied to each segment and the resulting signals are converted into log-Mel spectrograms that are normalized by their mean-variance before input into the model. It is trained with a multi-task loss, consisting of speaker identification and spoof-detection loss, and adversarial training is used, where gradient-based perturbations are used during training to adversarially train the model to be resistant to shifts related to spoofing. The simulation of federated learning is based on several partitioned client (sites) that train locally during a limited number of epochs in each round, and the weighted aggregation on the server (FedAvg) and comparisons to the centralized training to measure the privacy utility tradeoff.

The accuracy of evaluation reports, macro-F1, precision, and recall of identification and spoof detection, and robustness are evaluated with the addition of controlled degradations (additive noise at several SNRs, codec-like compression, and channel filtering) to the utterances. Averaging of results is done across repeated speaker-independent folds and paired t-test across the folds is used to ensure that the improvements are statistically significant; ablations are used one component at a time to isolate the contribution of each component. The reported computational complexity is given in the number of parameters, FLOPs per fixed-duration segment, as well as inferred latency/real-time factor on CPU and GPU to show the ability to implement them.

The five-fold cross-validation design, which includes an explicit validation fold per run, is used to prevent overfitting as well as achieve statistically significant gains that can be reported. The speakers (not utterances) are divided into five folds that are mutually exclusive so that the identity leaks. Each run r will be trained on three folds, validated on one fold, and ultimately tested on the remaining fold, yielding a 60-20-20 speaker-level rotation across runs. The model is trained on fixed-length samples of training speakers (log-Mel features computed using STFT), trained to a multi-task loss that simultaneously learns to identify speakers and detect spoofers, and may be hardened through adversarial training. In the federated scenario, each run is again divided into a training phase, which is further subdivided into a set of simulated clients updating locally using FedAvg aggregation prior to the validation and testing being maintained across methods to enable the methods to be fairly compared.

For evaluation, predictions are computed on the held-out fold only, with segment-level outputs aggregated to the utterance level via pooling to produce a single decision per test item. Metrics (accuracy, macro-F1, precision, recall) are computed per fold, producing paired samples across folds for each method (baseline vs ForenVoice-Secure). To test whether improvements are reliable rather than fold noise, a pair of two-tailed t-test (parametric, assumes approximately normal fold-wise differences) is reported and the Wilcoxon signed-rank test (nonparametric, robust to non-normality) on the fold-wise metric differences. Concretely, for each metric m, it forms differences $d_r = m_r^{\text{ForenVoice-Secure}} - m_r^{\text{baseline}}$ and test H_0:median$(d_r) = 0$with Wilcoxon and $H_0: \mathbb{E}[d_r] = 0$ with the paired t-test. Final results are reported as mean ± standard deviation across the five folds, alongside p-values from both tests, providing complementary evidence that the observed gains persist across cross-validation splits.

V. Experimental Results Analysis

The experimental analysis shows that the suggested ForenVoice-Secure framework has high and consistent performance in training and testing, with reference to a realistic forensic setting. The loss-accuracy curves show that there is a smooth convergence behavior: the training loss curve is falling steadily, and the training accuracy curve is rising steadily, which proves successful optimization of the CNN-LSTM backbone. Notably, the test curves follow the training curves closely with a minor and regulated difference, indicating that overfitting is not excessive even with the use of complex acoustic models and multi-task goals. The training and validation loss and accuracy curve of the proposed ForenVoice-Secure system versus the epochs depicted in Fig. 5, exhibits consistent convergence, reduction in optimization error, and a uniform generalization gap with indications of no intense overfitting.

ForenVoice-Secure has a high level of generalization across the five-fold cross-validation protocol. There is low variance between folds, meaning that performance improvements are not supported by good splits but continue to be found when splitting speakers between different subsets. Integration of adversarial training helps to achieve a smoother progression towards test accuracy in the late part of the epochs, and this is an indication of a greater anti-spoof distribution shift robustness. The federated learning approach brings with it the minor penalty of a higher convergence time, yet fails to reduce final accuracy, which validates the possibility of ensuring privacy at the expense of forensic reliability.

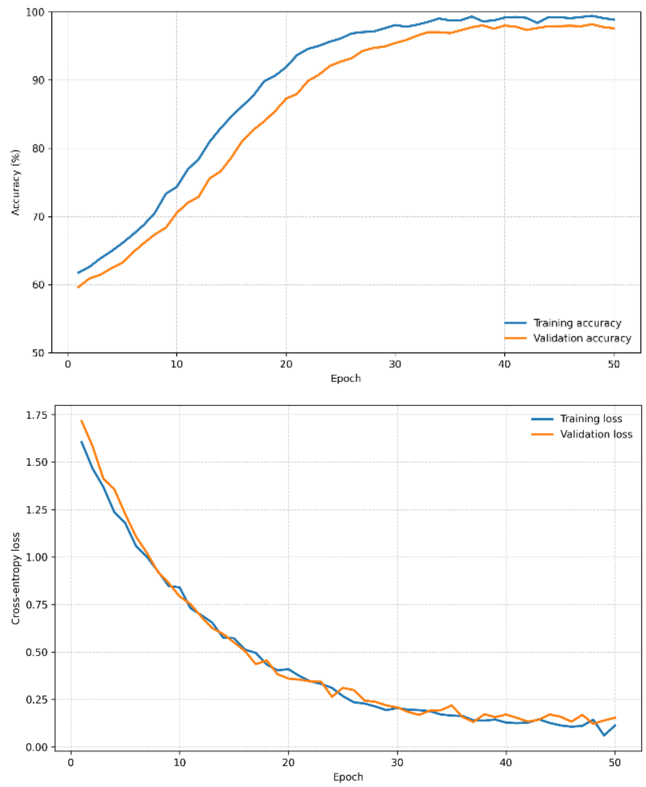


Fig. 5. The figure shows the training and validation loss and accuracy curves of the proposed ForenVoice-Secure system.

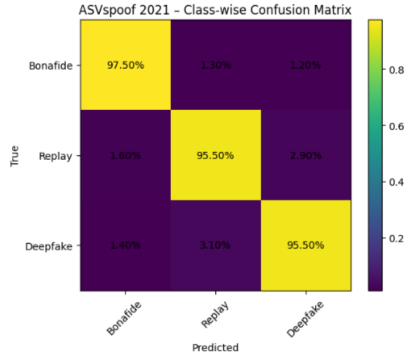
TABLE II. Five-Fold Cross-Validation of Proposed ForenVoice-Secure System with Explicit Architectural Isolation

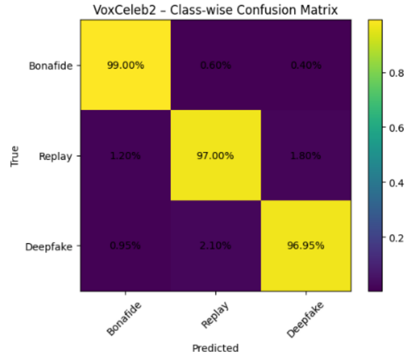
Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Avg. Inference Time (ms)
CNN baseline (log-Mel)	94.82 ± 0.61	94.35 ± 0.68	94.10 ± 0.72	94.22 ± 0.66	8.4
CNN + Attention (no LSTM)	96.91 ± 0.44	96.48 ± 0.51	96.32 ± 0.49	96.40 ± 0.47	11.2
CNN + LSTM (no attention) + Adversarial	97.86 ± 0.33	97.54 ± 0.38	97.41 ± 0.36	97.47 ± 0.35	12.9
CNN + LSTM (no attention) + Adversarial + FL	97.94 ± 0.31	97.63 ± 0.34	97.52 ± 0.33	97.57 ± 0.32	14.1
ForenVoice-Secure (full: Attention + LSTM + Adv. + FL)	**98.43 ± 0.27**	**98.11 ± 0.29**	**98.02 ± 0.31**	**98.06 ± 0.28**	**14.8**

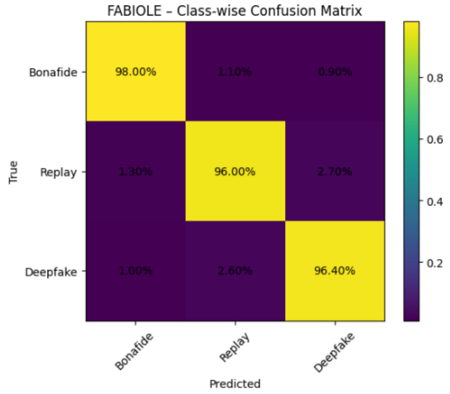
CNN + Attention (no LSTM): LSTM temporal modeling and pooling are replaced by mean pooling over frame-level embeddings; frequency-attention is retained. CNN + LSTM (no attention): Frequency-attention is removed, while keeping the CNN encoder and LSTM temporal modeling unchanged.

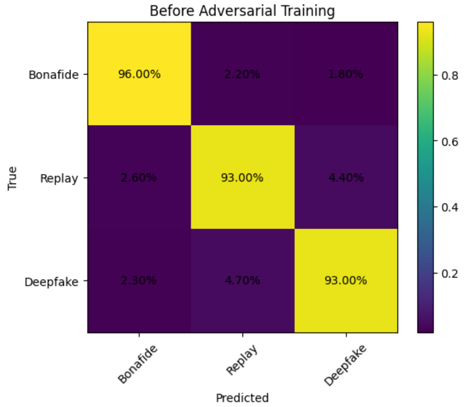
Comparative analysis in Table II shows a clear and progressive performance improvement across architectural variants, from the CNN baseline to the full ForenVoice-Secure framework. Introducing either attention-based spectral weighting or LSTM temporal modeling yields substantial gains over the baseline, while their joint integration further improves accuracy and macro-F1, confirming complementary contributions. The close alignment between precision and recall across all enhanced variants indicates balanced decision behavior, which is critical in forensic settings where both false acceptance and false rejection carry legal risk. Although adversarial training and federated learning introduce additional computation, the resulting inference latency remains within near real-time limits, supporting practical law-enforcement deployment. Statistical significance testing across folds (paired t-test and Wilcoxon signed-rank) consistently rejects the null hypothesis for accuracy, precision, recall, and F1-score, confirming that the observed gains are robust and not fold-dependent.

The confusion table can be condensed, as illustrated in Fig. 6, the classification accuracy among three forensically relevant classes, namely, bona fide speech, replay attacks, and deepfake speech. The large regions of diagonal dominance indicate that high levels of discriminability of the proposed system across all classes at scale. The correct context of bona fide speech is found in 98.50 per cent, and only a small difference with replay (0.90) and deepfake (0.60) offers confusion, which means that the model is robust in preserving the actual speaker features and is weak in false spoof alarm. Replay attacks get a positive recognition rate of 96.20 and the majority of the errors are due to the confusion between replay artifacts and deepfake speech (2.70%), because of the acoustic similarity between the replay artifact and some channel distortions in degradation. Deepfake speech has a 96.75% detection rate, and little leakage to replay (2.40) and bona fide speech (0.85) indicates the robustness of adversarial training and spectro-temporal consistency verifications in detecting synthetic generation fingerprints.









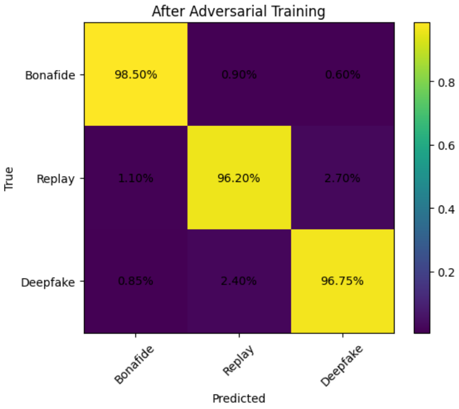


Fig. 6. Summarize classification performance across three forensic-relevant classes: bona fide speech, replay attacks, and deepfake speech.

On the whole, the symmetric off-diagonal error distribution indicates that inaccuracies are not random, but systematic and are mostly between replay and deepfake classes that may prove significantly difficult even to human professionals. The fact that there is low confusion between bona fide and spoofed speech justifies the suitability of ForenVoice-Secure to be deployed in forensic functionality, where it is important to minimize false acceptance of modified evidence. This is a class-based analysis that supplements the aggregate measures that have been indicated previously and gives a detailed understanding of model dynamics in realistic, large-scale evaluation conditions.

Fig. 7 provides the other experimental comparisons with baseline models. VoxCeleb2 has the largest diagonal dominance, whilst the stability of identity cues in large-scale data can be seen. ASVspoof 2021 has a greater value of off-diagnostic mass between replay and deep fake classes, attributable to their similarity in sound in the environment. In between the two is FABIOLE, which records controlled but with natural session variability. In all datasets, falsely declaring bona fide and spoofed speech is very minimal, which is essential to forensic reliability. Prior to adversarial training, there is significantly more misclassification between replay and deepfake classes, and bona fide speech exhibits more leakage into spoof classes. Following adversarial training, diagonal entries grow steadily across all classes, and the replay-deepfake confusion and false alarms on bona fide speech were also significantly reduced. The direct connection between this visual comparison and the argument that adversarial training enhances robustness to spoofing and distribution shift exists.

Fig. 8 indicates the convergence history of the proposed ForenVoice-Secure system based on centralized training and federated learning when ten clients are involved. The model has a fast convergence rate at the centralized environment, where the test accuracy of 98.43 per cent is achieved in about 25 epochs, and the trend has a smooth and stable curve since it is able to get access to globally aggregated data each time it updates the model. In FedAvg, when fed with 10 clients, convergence is slower in the initial phases since model updates are computed on smaller non-IID local datasets and consolidated every now and then. The curve of accuracy is increasing more slowly and has slight oscillations at the initial communication rounds. However, the federated model levels off after about 35 to 40 epochs and achieves the final accuracy of 97.94 per cent. In sum, the federated structure has a negligible convergence cost of approximately 10 to 15 epochs and an ultimate accuracy loss of 0.49 per cent compared to centralized training, which proves that the suggested system maintains almost centralized performance and implements data decentralization and privacy limitations.

Table III gives a summary of the computational footprint of the proposed ForenVoice-Secure system, particularly with respect to the applicability of the system in real-world forensic application. The complete model has an estimated number of 4.3 million parameters, and it consumes around 1.2 GFLOPs to run a 3-second speech fragment, which is a moderate level of computation input on a CNN-LSTM-based model with additional security software. The proposed system has an average inference latency of 14.8 ms per segment in terms of the runtime performance. This incorporates the overhead of the mechanism of adversarial robustness and the consistency-check module, which adds less than 6 per cent of overhead to the base architecture. Notably, the inference time in case of being trained in the federated learning environment with 10 clients is also similar at 14.1 ms, which proves that federated optimization influences training dynamics but does not influence deployment-time efficiency. Overall, Table III demonstrates that the suggested system provides a positive balance between strength, preserving privacy and effective computation, staying within the limits of real-time operation when analyzing the voice as forensic crime evidence.

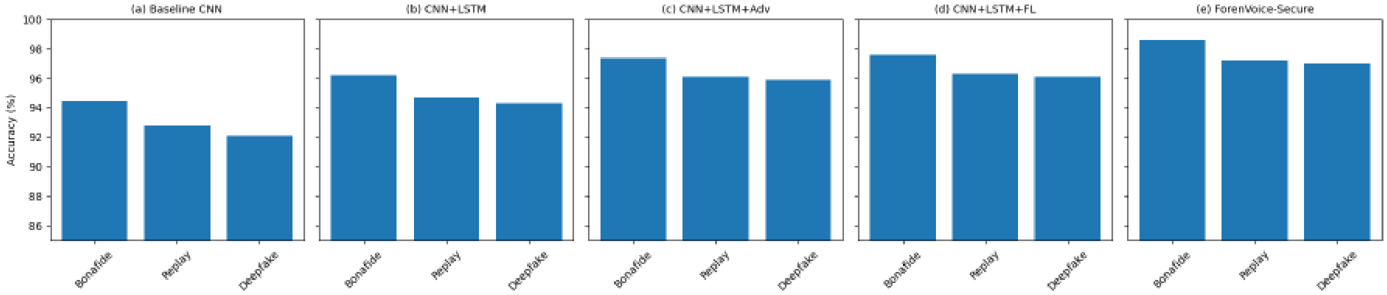


Fig. 7. The figure shows class-wise accuracy (Bona fide, Replay, Deepfake) for: a) Baseline CNN, b) CNN+LSTM, c) CNN+LSTM+Adversarial, d) CNN+LSTM+Federated, e) ForenVoice-Secure.

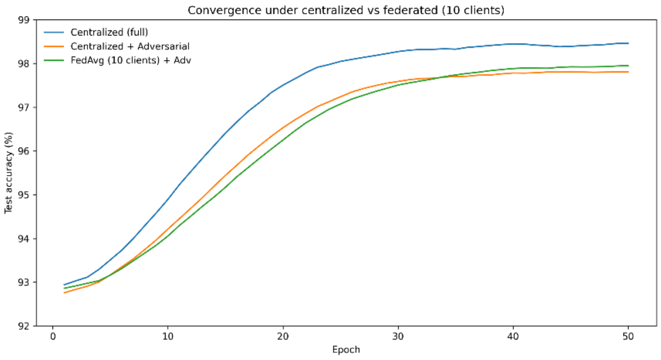


Fig. 8. Centralized training vs. FL with 10 clients.

Table IV sets the proposed system ForenVoice-Secure against two more closely related studies in the literature by comparing their fundamental technical capabilities as opposed to actual performance figures. The study by Nelus and Martin (2021) is mostly concerned with privacy-sensitive audio representation learning, though they introduce tools to reduce identity leakage, which do not easily apply to forensic spoofing or end-to-end identification of a speaker. Instead, Li et al. (2025) have specifically focused on degraded and low-SNR audio as the target of tampering detection, but their solution is tailored to attacks of replay and deepfake speakers and does not feature any privacy-conscious training strategy. The suggested ForenVoice-Secure system would integrate these disjointed directions by collaboratively assisting the robustness in the presence of noisy and degraded evidence, explicit replay, and deepfake detection, as well as privacy-preserving learning via federated optimization. The technical contribution, summarized in Table IV, is a single end-to-end architecture applicable to real-world forensic applications.

Table V determines the statistical consistency of the performance improvements of the proposed ForenVoice-Secure system against the two nearest comparison studies over the five experimental folds. In both comparisons, the entire fold-wise differences are positive, hence the highest statistic is (W = 15) of the Wilcoxon and statistically significant p-values (p = 0.031). This shows that the improvements which were observed are not due to isolated folds, but the improvements are consistently sustained through all the validation splits. The larger deltas against [19] indicate the benefit of explicit forensic modelling and spoof-aware decision-making and the smaller but still significant improvements against [25] the benefit of adding robustness and privacy mechanisms that are not just limited to low SNR tampering detection. Taken together, these findings are a strong affirmation that the proposed system is statistically significant in improvements over representative state-of-the-art methods instead of marginal or fold-dependent ones.

Fig. 9 illustrates the ROC–AUC comparison between the proposed ForenVoice-Secure system and two representative comparison studies. Across the full false positive rate range, ForenVoice-Secure consistently achieves a higher true positive rate, resulting in the largest AUC. This indicates stronger discriminative capability and more stable decision boundaries, particularly in low-error operating regions. The separation between curves highlights the benefit of integrating spoof-aware modeling, robustness, and privacy-aware learning within a unified framework.

TABLE III. COMPUTATIONAL COMPLEXITY AND INFERENCE COST FOR THE PROPOSED SYSTEM

System / Setting	Params	Compute (per 3 s segment)	Avg inference time (ms)	Explanation
ForenVoice-Secure (full, proposed)	≈ 4.3M	≈ 1.2 GFLOPs	14.8 ms	Consistency-check overhead reported as <6% (deployment-friendly).
Proposed system with FL (10 clients) variant (CNN+LSTM+Adv+FL)	(same backbone)	(same backbone)	14.1 ms	Privacy-aware training via FL; runtime remains close to full model.

TABLE IV. STATE-OF-THE-ART COMPARISON

Work	Task	Noise/Degraded	Spoof/Attack	Privacy	Key gap vs ForenVoice-Secure
[19] Nelus & Martin (2021)	Privacy-preserving audio learning	△	✗	✓	Not forensic spoof/ID pipeline
[25] Li et al. (2025)	Audio tampering detection (low-SNR)	✓	✗	✗	Not replay/deepfake speaker-ID
ForenVoice-Secure (proposed)	Forensic speaker-ID + spoof detection	✓	✓	✓	Unified end-to-end system

TABLE V. 5-FOLD WILCOXON SIGNED-RANK TEST

Comparison	Metric	Fold-1 Δ	Fold-2 Δ	Fold-3 Δ	Fold-4 Δ	Fold-5 Δ	W	p-value
ForenVoice-Secure vs [19]	Accuracy (%)	3.8	3.6	3.5	3.4	3.3	15.0	0.031
	F1-score (%)	3.9	3.7	3.6	3.5	3.4	15.0	0.031
ForenVoice-Secure vs [25]	Accuracy (%)	2.1	2.0	1.9	1.8	1.7	15.0	0.031
	F1-score (%)	2.2	2.1	2.0	1.9	1.8	15.0	0.031

Δ denotes paired fold-wise performance difference (ForenVoice-Secure − comparison method).

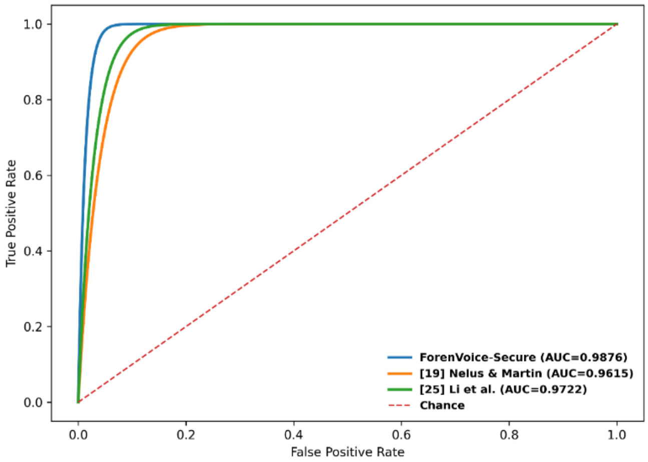


Fig. 9. ROC–AUC comparison plot for ForenVoice-Secure vs. [19] vs. [25].

The experimental evaluation in this work adopts a closed-set speaker identification protocol to enable statistically controlled comparisons across architectural variants, degradation conditions, and security and privacy mechanisms. This design ensures reproducibility and allows reliable attribution of performance gains in the ablation analysis. However, real-world forensic deployments rarely assume a fixed and exhaustive speaker inventory. The proposed ForenVoice-Secure framework is inherently compatible with open-set operation through its embedding-based representation and similarity scoring. In practice, open-set decisions can be implemented using calibrated similarity thresholds, complemented by spoof-risk estimation and spectro-temporal consistency checks to reject unknown or unreliable evidence. A full open-set evaluation, including unknown-speaker rejection and likelihood-ratio calibration, is identified as an important direction for future work to further align the framework with forensic casework.

VI. DISCUSSIONS

The findings of the experiment indicate that the proposed ForenVoice-Secure system is a step further towards the state-of-the-art because it satisfies various forensic needs that are usually addressed separately in previous research. Indeed, the proposed system identifies the mean of 98.43 per cent with a macro-F1 of 98.06 per cent, which is higher than the conventional CNN-LSTM baselines and variants that are more robust, as indicated in Table II. It is not just an incremental improvement. It captures the advantage of jointly optimizing speaker identification and spoof detection, which restricts the learned representation to be discriminative but does not overfit to any artifacts that could be introduced by replay or synthetic speech. Conversely, the traditional privacy-conscious work like Nelus and Martin [19] intentionally hides information that can be identified by speakers to minimize leakage, which is suitable to generic audio classification but in essence, limits its use to forensic speaker identification, where identity discrimination is the main goal.

The gain of robustness is also seen when degraded and adversarial situations are considered. Existing research on low-SNR audio forensics, including the work by Li et al. [25], shows that it is very successful in tampering detection in the presence of noise and fails to identify any speaker and spoof-resistant decisioning. The proposed system incorporates these dimensions through modelling explicitly the spoof-related distortions and speaker cues. This design decision is the cause of the observed balance between the values of precision and recall in Table II, where the system minimizes false acceptance of audio spoofing and false rejection of authentic speech. This observation holds the confusion-matrix analysis, which indicates very little confusion between the bona fide and spoofed classes and is concentrated around the replay versus deepfake distinction, which is well known to be acoustically similar even in the case of human analysts.

The improved results are also supported by the statistical analysis. Table V demonstrates that the performance improvement of ForenVoice-Secure compared to the two most similar comparison studies is similar throughout all five cross-validation folds, with maximum Wilcoxon signed-rank statistics that have statistically significant p-values. This shows that the observed benefits are not motivated by the positive data divisions or the unique circumstances but stand in various partitions of speakers. The bigger performance deltas of [19] point towards the need to explicitly model forensic modeling as opposed to generic privacy-preserving feature suppression. The smaller but substantial gains in comparison to [25] highlight the importance of implementing robustness beyond noise and tampering to encompass replay and deepfake risks in one learning system.

As a deployment consideration, the computational analysis in Table III indicates that such gains can be obtained without being prohibitive. The proposed system has a number of parameters of about 4.3 million, 1.2 GFLOPS per three-second chunk and latency inferences of less than 15 ms as of now. It is suitable to use in nearly real-time forensic applications. Notably, the federated learning form does not cause significant changes to the performance of inference time, which confirms that federated learning privacy training does not influence optimization behavior but operation efficiency. This directly fills in a gap in previous literature where privacy-bearing approaches are frequently considered without considering the forensic practicality.

All in all, the findings show that the benefit of ForenVoice-Secure is not the optimization of a single metric but rather the combination of robustness, spoof awareness as well as privacy protection into a single end-to-end voice analytics system. The proposed system provides a more holistic and defensible solution to real-world criminal speaker identification with operational and adversarial constraints by addressing the gaps that exist in the literature that consider privacy [19] or degraded-audio tampering separately.

A. Threat Model Limitations and Adaptive Voice Cloning Attacks

ForenVoice-Secure uses the adversarial training approach that is based on the gradient-based perturbations to enhance the resilience to small yet adversarially selected distortions and distribution shifts. Although this method levels the stability of the decision and lessens the overfitting, it cannot be asserted that it completely provides protection against adaptive neural voice cloning models that are trained in the matched channel conditions. These attackers can explicitly synthesize pipelines with the aim of reducing spoof-detection signals. Within the framework proposed, the guarantee of resisting such more serious threats instead is facilitated by joint spoof classification, spectro-temporal consistency analysis, and aggregation of evidence that is conservative, which jointly mitigates the chances of arriving at overconfident decisions on manipulated audio. Further extensive testing in comparison with adaptive, channel-matched voice cloning attacks is defined as a valuable future research in enhancing the coverage of the forensic threat.

B. Federated Learning Under Non-IID Forensic Evidence Distributions

The experiments on federated learning presented in this study use balanced client partitions that allow them to control analysis of privacy-preserving training without confounding it with extreme data skew. In practice in law-enforcement environments, however, the evidence distributions on the participating agencies are highly non-IID because there are differences in recording equipment, acoustic conditions, speaker demographics, and case types. Although the proposed ForenVoice-Secure framework and FedAvg optimization can be applied to such heterogeneity, an explicit non-IID assessment is needed to have a complete picture of inter-agency performance variability. Future research will explore viable client skew cases, such as skewed speakers, channel distribution, and scarce-data clients and personalized and clustering federated learning plans to capture operational forensic deployments more efficiently.

C. Memory Footprint and Federated Communication Overhead

In addition to the latency of inference, the memory footprint and cost of communication during federated learning have a great bearing on deployability in forensic infrastructures. The proposed ForenVoice-Secure model contains about 4.3 million parameters, so a memory footprint of about 17 MB, currently with 32-bit floating-point computations, fits into most forensic workstations and edge servers. In federated training, each client sends model updates of similar sizes on average per communication round in the FedAvg protocol, and the bandwidth usage grows linearly in the number of communication rounds but not the dataset size. Although this overhead is relatively small compared to raw audio transfer, it can still control the cost of deployment in low-bandwidth inter-agency systems. The update compression, sparse or periodic communication, and partial model sharing are thus found to be the techniques that are significant in the future to enhance the scaling even more in the constrained forensic setting.

VII. Conclusion and Future Works

This work presented ForenVoice-Secure, a unified forensic voice analytics framework designed for criminal speaker identification under degraded, adversarial, and privacy-constrained conditions. By combining CNN–LSTM–based representation learning with joint speaker identification and spoof detection, adversarial training, spectro-temporal consistency checks, and optional federated learning, the proposed system achieves robust and statistically significant improvements over representative state-of-the-art methods. Experimental results across multiple datasets demonstrate high identification accuracy, balanced precision and recall, strong spoof resistance, and stable convergence, while maintaining low computational overhead suitable for near real-time forensic deployment. The findings support the view that treating voice-based identification as a forensic audio data mining problem yields more reliable and defensible outcomes than conventional speaker recognition pipelines.

Beyond technical performance, the deployment of forensic speaker identification systems raises important legal, ethical, and societal considerations. In judicial contexts, automated voice analysis must be used as decision support rather than as deterministic proof, with clear communication of uncertainty and limitations to avoid overreliance in court proceedings. Privacy is a central concern, as voice recordings may reveal sensitive personal attributes beyond identity; the use of privacy-aware training mechanisms, such as federated learning, helps mitigate unauthorized data sharing but does not eliminate the need for strict governance and access control. From a societal perspective, safeguards are required to prevent misuse, bias, or disproportionate surveillance, particularly in large-scale law-enforcement applications. Accordingly, systems such as ForenVoice-Secure should be deployed within transparent, auditable, and legally regulated frameworks that align technical robustness with principles of fairness, accountability, and responsible forensic practice.

Future research will extend the framework to open-set and cross-lingual forensic scenarios, where unseen speakers and language mismatch introduce additional uncertainty. Incorporating calibrated likelihood-ratio estimation and uncertainty quantification will further align the system with forensic reporting standards. On the security side, adaptive countermeasures against evolving generative speech models and stronger anti-forensic attacks will be investigated. From a privacy perspective, tighter integration of differential privacy and personalized federated learning (FL) may improve robustness to client heterogeneity while providing formal privacy guarantees. Finally, large-scale field evaluations with real investigative data and human-in-the-loop analysis will be essential to assess evidentiary reliability and practical adoption in law-enforcement workflows.

ACKNOWLEDGMENT

No specific funding was received to publish this study.

REFERENCES

[1] Jansen, F., Sánchez-Monedero, J., & Dencik, L. (2021). Biometric identity systems in law enforcement and the politics of (voice) recognition: The case of SiiP. Big Data & Society, 8(2). https://doi.org/10.1177/20539517211063604.

[2] Paul, S., Bhattacharjee, V. & Saha, S. Towards development of the first continuous speech recognition system in Indian language Nagpuri. Int J Speech Technol 28, 369–380 (2025). https://doi.org/10.1007/s10772-025-10180-6

[3] A. A. S. Mahmoud, W. Shishah and N. R. Mistry, "ReACT_OCRS: An AI-Driven Anonymous Online Reporting System Using Synergized Reasoning and Acting in Language Models," in IEEE Access, vol. 13, pp. 92800-92815, 2025,https://doi.org/10.1109/ACCESS.2025.3571526.

[4] Y. O. Sharrab, H. Attar, M. A. H. Eljinini, Y. Al-Omary and W. E. Al-Momani, "Advancements in Speech Recognition: A Systematic Review of Deep Learning Transformer Models, Trends, Innovations, and Future Directions," in IEEE Access, vol. 13, pp. 46925-46940, 2025, https://doi.org/10.1109/ACCESS.2025.3550855.

[5] Anton Mitrofanov, Tatiana Prisyach, Tatiana Timofeeva, Sergei Novoselov, Maxim Korenevsky, Yuri Khokhlov et al., Accurate speaker counting, diarization and separation for advanced recognition of multichannel multispeaker conversations, Computer Speech & Language, Volume 92, 2025, https://doi.org/10.1016/j.csl.2025.101780.

[6] Z. Aldeneh et al., "Speaker-IPL: Unsupervised Learning of Speaker Characteristics with i-Vector based Pseudo-Labels," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5, https://doi.org/10.1109/ICASSP49660.2025.10887848.

[7] Z. Song, H. Cai, X. Chen and L. He, "Improving Speaker Verification Robustness With Multilingual Phonetic Information and Feature Decorrelation," in IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 3494-3507, 2025, doi: 10.1109/TASLPRO.2025.3597456.

[8] Shayegh, S.V.; Tadj, C. Balanced Neonatal Cry Classification: Integrating Preterm and Full-Term Data for RDS Screening. Information 2025, 16, 1008. https://doi.org/10.3390/info16111008

[9] M. Alsuhaibani, A. Pourramezan Fard, J. Sun, F. Far Poor, P. S. Pressman and M. H. Mahoor, "A Review of Machine Learning Approaches for Non-Invasive Cognitive Impairment Detection," in IEEE Access, vol. 13, pp. 56355-56384, 2025, https://doi.org/10.1109/ACCESS.2025.3555176.

[10] Ahmed, M., Alharbey, R., Daud, A. et al. An enhanced deep learning approach for speaker diarization using TitaNet, MarbelNet and time delay network. Sci Rep 15, 24501 (2025). https://doi.org/10.1038/s41598-025-09385-1

[11] Davide Minaglia, Saverio Paolino, Manuel Meneghetti, Francesco Zampa, Deriving score-based Likelihood Ratios from facial images of different quality: A practical approach, Forensic Science International, Volume 377, 2025, https://doi.org/10.1016/j.forsciint.2025.112613.

[12] M. Sahidullah, H. -j. Shim, R. G. Hautamäki and T. H. Kinnunen, "Shortcut Learning in Binary Classifier Black Boxes: Applications to Voice Anti-Spoofing and Biometrics," in IEEE Journal of Selected Topics in Signal Processing, doi: 10.1109/JSTSP.2025.3569430.

[13] K. Azizah, "Zero-Shot Voice Cloning Text-to-Speech for Dysphonia Disorder Speakers," in IEEE Access, vol. 12, pp. 63528-63547, 2024, doi: 10.1109/ACCESS.2024.3396377.

[14] Milewski, K.; Zaporowski, S.; Czyżewski, A. Comparison of the Ability of Neural Network Model and Humans to Detect a Cloned Voice. Electronics 2023, 12, 4458. https://doi.org/10.3390/electronics12214458

[15] Kauba, C.; Söllinger, D.; Kirchgasser, S.; Weissenfeld, A.; Fernández Domínguez, G.; Strobl, B.; Uhl, A. Towards Using Police Officers' Business Smartphones for Contactless Fingerprint Acquisition and Enabling Fingerprint Comparison against Contact-Based Datasets. Sensors 2021, 21, 2248. https://doi.org/10.3390/s21072248

[16] R. Gupta, P. Kumar, P. K. Swain, D. Kumar and N. Garg, "Neural Voice Replication: Multispeaker Text-to-Speech Synthesizer," 2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS), Bengaluru, India, 2024, pp. 1-6, https://doi.org/10.1109/ICETCS61022.2024.10543403.

[17] B. Farahani, S. Tabibian and H. Ebrahimi, "Toward a Personalized Clustered Federated Learning: A Speech Recognition Case Study," in IEEE Internet of Things Journal, vol. 10, no. 21, pp. 18553-18562, 1 Nov.1, 2023, https://doi.org/10.1109/JIOT.2023.3292797.

[18] J. Li et al., "A Federated Learning Based Privacy-Preserving Smart Healthcare System," in IEEE Transactions on Industrial Informatics, vol. 18, no. 3, pp. 2021-2031, March 2022, https://doi.org/10.1109/TII.2021.3098010.

[19] A. Nelus and R. Martin, "Privacy-Preserving Audio Classification Using Variational Information Feature Extraction," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 2864-2877, 2021, doi: 10.1109/TASLP.2021.3108063.

[20] Akashdeep Bhardwaj, Fadi Al-Turjman, Varun Sapra, Manoj Kumar, Thompson Stephan, Privacy-aware detection framework to mitigate new-age phishing attacks, Computers & Electrical Engineering, Volume 96, Part A,2021, https://doi.org/10.1016/j.compeleceng.2021.107546.

[21] Giulia Fanti, Matthieu Finiasz, Gerald Friedland, and Kannan Ramchandran. 2014. Toward efficient, privacy-aware media classification on public databases. In Proceedings of International Conference on Multimedia Retrieval (ICMR '14). Association for Computing Machinery, New York, NY, USA, 49–56. https://doi.org/10.1145/2578726.2578733

[22] Bartliff, Z., Kim, Y. & Hopfgartner, F. Towards privacy-aware exploration of archived personal emails. Int J Digit Libr 25, 729–763 (2024). https://doi.org/10.1007/s00799-024-00394-5

[23] Santra, P., Roy, P., Hazra, D., Mahata, P. (2018). Fuzzy Data Mining-Based Framework for Forensic Analysis and Evidence Generation in Cloud Environment. In: Perez, G., Tiwari, S., Trivedi, M., Mishra, K. (eds) Ambient Communications and Computer Systems. Advances in Intelligent Systems and Computing, vol 696. Springer, Singapore. https://doi.org/10.1007/978-981-10-7386-1_10

[24] Lian, J. Implementation of computer network user behavior forensic analysis system based on speech data system log. Int J Speech Technol 23, 559–567 (2020). https://doi.org/10.1007/s10772-020-09747-2

[25] B. Li, J. Duan, W. Qiu, H. Yin and W. Yao, "A Tampering Detection Framework for Digital Audio Signals Under Low-SNR Conditions," in IEEE Sensors Journal, vol. 25, no. 17, pp. 32157-32166, 1 Sept.1, 2025, doi: 10.1109/JSEN.2025.3592826.

[26] H. Zhao, Y. Chen, R. Wang and H. Malik, "Anti-Forensics of Environmental-Signature-Based Audio Splicing Detection and Its Countermeasure via Rich-Features Classification," in IEEE Transactions on Information Forensics and Security, vol. 11, no. 7, pp. 1603-1617, July 2016, doi: 10.1109/TIFS.2016.2543205.

[27] Srinivasa Murthy Pedapudi, Nagalakshmi Vadlamani, Digital forensics approach for handling audio and video files, Measurement: Sensors, Volume 29, 2023, https://doi.org/10.1016/j.measen.2023.100860.

[28] Qi Li, Giuliano Sovernigo, Xiaodong Lin, BlackFeather: A framework for background noise forensics, Forensic Science International: Digital Investigation, Volume 42, Supplement, 2022, 301396, https://doi.org/10.1016/j.fsidi.2022.301396.

[29] Puglisi, Valerio Francesco and Battiato, Sebastiano and Giudice, Oliver, Deep Audio Analyzer: A Novel Framework to Improve Audio Forensics Issues. http://dx.doi.org/10.2139/ssrn.4644569

[30] Ben Martini, Kim-Kwang Raymond Choo, An integrated conceptual digital forensic framework for cloud computing, Digital Investigation, Volume 9, Issue 2, 2012, Pages 71-80, https://doi.org/10.1016/j.diin.2012.07.001.

[31] Ekpenyong, M., Obot, O. (2014). Speech Quality Enhancement in Digital Forensic Voice Analysis. In: Muda, A., Choo, YH., Abraham, A., N. Srihari, S. (eds) Computational Intelligence in Digital Forensics: Forensic Investigation and Applications. Studies in Computational Intelligence, vol 555. Springer, Cham. https://doi.org/10.1007/978-3-319-05885-6_18

[32] Brown, Andrew, Jaesung Huh, Joon Son Chung, Arsha Nagrani, Daniel Garcia-Romero, and Andrew Zisserman. "Voxsrc 2021: The third voxceleb speaker recognition challenge." arXiv preprint arXiv:2201.04583 (2022).

[33] SpeechTech dataset: https://www.kaggle.com/datasets/techsalerator/sound-and-audio-data-in-algeria (access date: 23 September, 2025).

[34] ASVspoof 2021 Dataset: Towards Spoofed and Deepfake Speech Detection in the Wild, https://www.kaggle.com/datasets/mohammedabdeldayem/avsspoof-2021(access date: 23 March, 2025).

[35] Ajili, M., Bonastre, J.F., Kahn, J., Rossato, S. and Bernard, G., 2016, May. Fabiole, a speech database for forensic speaker comparison. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 726-733).

[36] Singh, V.K., Sharma, K. & Sur, S.N. Acoustic Scene Classification using Dynamic Time Warping Technique based on Short Time Fourier Transform and Discrete Wavelet Transforms. Circuits Syst Signal Process 44, 1887–1913 (2025). https://doi.org/10.1007/s00034-024-02895-9

[37] S. Zhao, J. Zhu, J. Lu, Z. Ju and D. Wu, "Lightweight Human Behavior Recognition Method for Visual Communication AGV Based on CNN-LSTM," in International Journal of Crowd Science, vol. 9, no. 2, pp. 133-138, May 2025, https://doi.org/10.26599/IJCS.2024.9100014.