

Article

Frame-Level Audio Forgery Localization Using Handcrafted and Neural Features

Mostafa Moallim ¹, Taqwa A. Alhaj ^{2,*}, Fatin A. Elhaj ³, Inshirah Idris ⁴ and Tasneem Darwish ⁵

¹ Center of Artificial Intelligence for Security, Naif Arab University for Security Sciences, Riyadh 14812, Saudi Arabia; mmoallim@nauss.edu.sa

² Department of Cybersecurity and Digital Forensics, Center of Artificial Intelligence for Security, Naif Arab University for Security Sciences, Riyadh 14812, Saudi Arabia

³ College of Computing & Intelligent Systems, University of Khorfakkan, Sharjah 18119, United Arab Emirates; fatin.mohamed@ukf.ac.ae

⁴ Computer Science and Information Technology College, Wadmedani Ahlia University, Wad Medani P.O. Box 402, Sudan; inshirah.idris@gmail.com

⁵ Department of Computer Science, St. Francis Xavier University, Antigonish, NS B2G 2W5, Canada; tdarwish@stfx.ca

* Correspondence: talhaj@nauss.edu.sa

Abstract

Audio forgery has emerged as a significant security and forensic challenge, driven by rapid advances in generative artificial intelligence and the widespread availability of audio editing tools, which enable the creation of highly realistic manipulated speech with minimal technical expertise. Existing approaches predominantly operate at the file level, providing only coarse binary decisions without identifying when or where manipulation occurs. This study addresses fine-grained temporal localization through a unified frame-level localization framework. We introduce a controlled forgery generation framework derived from the TIMIT speech corpus, applying atomic, localized manipulations under strict temporal constraints and producing precise frame-level annotations across diverse manipulation types. Building on this dataset, we then propose a transform-agnostic localization-driven detection approach using temporal inconsistency modeling, enabling unified analysis across heterogeneous manipulations at frame-level resolution. To analyze forensic evidence, we present an evidence-stratified modeling paradigm comparing three complementary strategies: a handcrafted anomaly-based method, a deep localization model leveraging pretrained wav2vec 2.0 representations, and a hybrid approach combining both through confidence-aware fusion and temporal consistency reinforcement. A systematic experimental analysis evaluates the effects of representation adaptation, hybrid fusion, and manipulation type on detection and localization performance. Results show that handcrafted features are insufficient for reliable frame-level localization, while task-adapted wav2vec 2.0 achieves strong and consistent performance. The hybrid approach does not consistently improve frame-level accuracy but yields substantial gains in segment-level localization by enforcing temporal coherence. Per-transform analysis confirms robust performance across most manipulations, with deletion-based operations remaining the most challenging.

Keywords: audio forgery localization; frame-level analysis; speech forensics; handcrafted features; neural localization; hybrid modeling



Academic Editors: Xiao-Rong Ding, Ying-Ren Chien and Pavel Loskot

Received: 27 February 2026

Revised: 16 April 2026

Accepted: 24 April 2026

Published: 7 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Audio signals are essential components of contemporary communication systems. Recent developments in audio editing software have made it possible to create extremely realistic altered recordings. Such modifications are typically confined to brief temporal regions while leaving the surrounding acoustic environment unchanged. It is challenging to identify the tiny temporal discrepancies introduced by this partial alteration. This emphasizes the necessity of efficient methods for detection and authentication [1,2].

Recent deep learning approaches, particularly those based on self-supervised speech representations, such as wav2vec 2.0 [3], HuBERT, and WavLM, have demonstrated strong generalization capabilities for audio analysis tasks. These representations capture rich acoustic and linguistic information and have been successfully applied to detect spoofing and manipulation [4–6]. However, their effectiveness for temporal localization remains insufficiently explored. In particular, the extent to which pretrained representations encode manipulation-relevant temporal cues, how representation adaptation affects localization performance, and how neural models compare to traditional handcrafted signal descriptors when operating at fine temporal granularity remain unclear [5].

Despite growing interest in audio forgery detection, existing work exhibits several limitations when precise temporal localization is required. First, most approaches focus on file-level classification, providing only a global decision without identifying when manipulation occurs [6–9]. Second, localization is often treated indirectly through weakly supervised mechanisms such as attention maps or class activation, which are not trained using frame-level ground truth and therefore yield imprecise temporal boundaries. Third, many publicly available datasets rely on loosely constrained synthetic editing pipelines that introduce unrealistic artifacts or ambiguous boundary patterns, limiting their forensic relevance [10]. Finally, prior benchmarks rarely isolate different manipulation types under a unified and controlled protocol, making it difficult to analyze how specific operations such as insertion, deletion, or temporal warping affect localization difficulty [1].

In this study, we address these issues by introducing a controlled framework for analyzing localized audio manipulations. Our dataset generation process applies small, time-limited edits to real speech recordings, keeping the surrounding audio unchanged and enforcing strict rules on segment separation, minimum length, and timing gaps. We used a dual-regime approach to handle both short and normal-length recordings, allowing consistent supervision across different audio lengths. Each sample included detailed pair-level annotations describing the manipulated segments, type of transformation, and exact timing.

We present a localization-driven approach for detecting audio manipulation based on transform-agnostic temporal inconsistency modeling built on this dataset. The technique works at the frame level and learns to recognize abnormal local deviations in prosodic, temporal, and spectral properties that result from various editing operations such as shifting, time warping, copy, pitch adjustment, and deletion. Even in cases where altered sections are short-lived or sparsely dispersed, the framework facilitates the detection of partial and multi-segment modifications by aggregating the localized evidence over time.

Furthermore, we introduce an evidence-stratified modeling paradigm that explicitly distinguishes between three complementary sources of information: handcrafted acoustic cues, representation-driven neural embeddings, and their hybrid integration methods. By enforcing identical temporal supervision and evaluation protocols across these strata, the proposed formulation enables a systematic analysis of the agreement, disagreement, and failure patterns among different signal representations, offering insights into the strengths and limitations of each modeling approach for temporal audio analysis.

From a forensic perspective, this work belongs to the class of passive, content-based audio analysis methods and focuses on frame-level localization rather than global detection.

While the individual components employed in this study, handcrafted acoustic anomaly detection, self-supervised speech representations, and confidence-aware fusion have been explored separately in prior work, their systematic integration for fine-grained frame-level localization of signal-level audio editing operations represents a distinct contribution. The novelty of this work lies not in proposing new model architectures, but in three underexplored directions: (1) the design of a controlled forensic benchmark with precise frame-level annotations across diverse manipulation types, which, to the best of our knowledge, is not currently available in the literature; (2) a principled comparative evaluation of complementary forensic evidence sources under identical supervision and evaluation conditions, enabling systematic analysis of their individual and combined strengths; and (3) a confidence-gated hybrid localization framework that operates at inference time without retraining, providing forensically interpretable and temporally coherent outputs. Together, these contributions establish a reproducible foundation for future research in fine-grained audio forgery localization.

The main contributions of this study are summarized as follows:

1. A controlled dataset generation framework for localized audio manipulations, featuring strict temporal constraints, dual-regime handling of short and normal recordings, and structured pair-level annotations that support detection and localization within a unified format.
2. A transform-agnostic localization-driven detection approach based on temporal inconsistency modeling, enabling unified analysis of heterogeneous manipulation types at frame-level resolution.
3. An evidence-stratified signal modeling paradigm that systematically compares handcrafted features, self-supervised neural representations, and their hybrid integration under identical localization settings.
4. A comprehensive experimental analysis examining the effect of representation adaptation, hybrid cue fusion, and manipulation type on both detection and localization performance.

Together, these contributions provide a unified signal-processing framework for studying and addressing the challenges of detecting and localizing partial audio manipulations and establishing principled guidelines for dataset design and temporal modeling in future audio analysis systems.

2. Related Work

The rapid advancement of audio generation and editing technologies has shifted the landscape of audio manipulation, necessitating a parallel evolution in detection methodologies, from traditional signal processing techniques to advanced deep learning models. This section reviews key developments in audio forgery analysis, organized into handcrafted signal-level methods, deep-learning-based detection, and dataset design considerations, highlighting the progression of the field and the specific gaps addressed by this work.

Audio forgery analysis methods can be broadly categorized into active and passive authentication approaches. Active methods rely on embedded watermarks or prior integrity mechanisms introduced during recording, whereas passive methods infer authenticity from intrinsic properties of the audio signal without requiring prior information.

Within passive audio forensics, approaches can be further divided into container-based and content-based methods. Container-based approaches analyze metadata, compression artifacts, or file structure, while content-based approaches focus on signal-level, spectral, or environmental inconsistencies introduced during manipulation.

The present study falls within passive, content-based audio forensics and specifically addresses the problem of fine-grained, frame-level localization of partial manipulations. Unlike many existing detection methods that provide primarily file-level decisions, this work focuses on identifying the precise temporal regions affected by localized editing operations. Accordingly, the related work in this section is organized into handcrafted signal-level methods, deep-learning-based approaches, and dataset design considerations, which together represent the most relevant branches for frame-level audio forgery localization.

2.1. Handcrafted Signal-Level Methods for Audio Forgery Analysis

Early research in audio forensics relied on handcrafted acoustic features to identify disruptions in signal properties and detect audio manipulations. Based on the domain from which features are extracted, these approaches can be organized into four primary categories: Electric Network Frequency (ENF)-based, audio-signal-based, image-based, and environment-based. Each category leverages distinct traces left by recording and manipulation processes, offering complementary strengths and facing unique limitations.

Electric Network Frequency (ENF)-based methods: These detect tampering by analyzing the ENF signal embedded in audio recordings captured using mains-powered devices. These methods compare the extracted ENF signature with reference databases, which can also provide information about the time and location of the recording [2]. Researchers have used ENF analysis to spot splicing forgeries because combining audio segments can noticeably change the ENF signal [11]. Some studies use only single ENF features and simple machine learning models, which often leads to weaker results [12]. For example, Reis et al. [13] used kurtosis features of ENF with a Support Vector Machine (SVM) classifier, while Esquef et al. [14] looked at ENF variation to find splice points. Lin et al. used spectral phase analysis, which is more resistant to noise than direct ENF trace analysis [11]. More recent approaches combine multiple ENF features with deep learning. For instance, The ENFformer [12], a deep learning model for ENF-based tampering detection, combines zero-order phase, first-order phase, and frequency features. Mao et al. [15] calculated multiple ENF features (MUSIC, Hilbert linear prediction, and Welch) for input to convolutional neural networks. However, ENF-based methods only work well with audio that has a high enough Signal-to-Noise Ratio (SNR) [12]. These techniques depend heavily on recording conditions and do not work if the ENF trace is weak or hidden by noise [11,15]. Still, they can check for edits with low computational cost [15].

Audio-Signal-Based Methods: These methods detect tampering by analyzing inherent characteristics of the audio signal itself, extracting features directly from audio content encompassing structure, spectral characteristics, and internal consistency [2,11,12]. Farid et al. [16] employed bicoherence analysis to quantify third-order correlation intensity. Lin et al. [17] utilized short-time Fourier transform on sub-bands for spectral phase analysis, constructing feature vectors from high-order statistical residuals of spectral phases—an approach that reduces noise but entails significant computational complexity. Yan et al. [18] introduced a detection scheme using features such as standard deviation of the difference signal and its high-frequency components, analyzed with an SVM classifier, though relying exclusively on basic SVM may yield suboptimal accuracy. Other approaches examine deviations from standard statistical characteristics of digital audio, including mathematical distributions and bandwidth fluctuations [12]. Yang et al. [19] employed the count of small-value Modified Discrete Cosine Transform (MDCT) coefficients to differentiate authentic from counterfeit MP3 files. Luo et al. [20] utilized both MDCT and Mel-frequency cepstral coefficients (MFCC) to examine prior audio compression methods, and developed an Amplitude Cooccurrence Vector (ACV) for detecting post-processing artifacts in high-frequency components. Chen et al. [21] employed Wavelet Packet Singular Value Decomposition to leverage correlations

in adjacent sampled sequences and continuity across frequency bands for detection and localization. Zhou et al. [22] investigated quantum properties of digital signals through quantum tunneling theory and an improved quantum genetic algorithm for detecting phase changes. Challenges in this domain encompass significant computational requirements, long processing times, and the possibility of increased false detection rates [12], while these techniques can be effective, they often make distinct assumptions about compression codecs, recording instruments, and ambient conditions—assumptions that may not hold in real-world scenarios [10].

Image-Based Methods: This category involves converting audio signals into image representations, typically spectrograms, which are then analyzed using image processing techniques and deep learning models [8,23]. These approaches leverage the strengths of computer vision to identify visual patterns indicative of tampering. Recent studies emphasize the power of spectrogram analysis for forgery detection. Wang et al. [8] proposed robust copy-move detection and localization using Constant-Q Filterbank Cepstral Coefficient (CFCC) features, demonstrating improved performance under compression and noise. Zhao et al. [23] designed a CNN-based model that analyzes spectrogram patches to localize audio copy-move forgeries with high accuracy. Recent spectrogram-based approaches have further advanced copy-move forgery detection through keypoint matching strategies. Ustubioglu et al. [7] proposed a SIFT keypoint-based feature matching approach on Mel spectrograms, demonstrating robust file-level detection performance against common post-processing operations including compression, filtering, and noise using the TIMIT and Arabic Speech corpora. Building on this direction, Ozgen and Altay [24] introduced a hybrid keypoint strategy combining SIFT, FAST, and FREAK descriptors on Mel spectrograms, achieving file-level F1-scores exceeding 0.93 under clean conditions with improved robustness against post-processing attacks. However, both methods address copy-move forgery exclusively and provide only file-level or block-level decisions, without frame-level temporal annotations or support for diverse manipulation types. These methods exploit the rich time-frequency representations of spectrograms, which often reveal hidden artifacts introduced during tampering. The strength of image-based approaches lies in their ability to transfer mature image forensics and CNN architectures to audio forensics. However, their performance may be sensitive to spectrogram resolution choices and susceptible to adversarial post-processing such as noise filtering. Despite these challenges, image-based methods are increasingly effective for splicing, copy-move, and even synthetic audio detection when combined with deep neural networks [8,23].

Environment-Based Methods: This domain focuses on detecting inconsistencies or changes in the acoustic environment, background noise, or device-specific traces within an audio recording [11]. Zhao et al. [25] evaluated the similarity of Environmental Signature (ES) characteristics by analyzing the magnitude of acoustic channel impulse response and ES features. Cuccovillo et al. [26] conducted audio splicing detection through microphone classification, utilizing distinctive signatures from various source devices. Yu et al. [27] performed splicing forgery detection by extracting characteristics from background noise using MFCC and Pearson correlation coefficients. Pan et al. [28] identified fake audio merges by analyzing variations in noise levels across questionable audio files. A fundamental limitation of environment-based methods is that they can be circumvented when recordings are captured under controlled conditions or with the same device [11]. Additionally, like other handcrafted approaches, they rely on assumptions about recording conditions that may not generalize.

In summary, handcrafted signal-level methods provide interpretable, computationally efficient tools for audio forgery analysis, each exploiting different physical or statistical traces of manipulation. However, they share common limitations: dependence on specific

recording conditions, sensitivity to post-processing, and an inability to provide the fine-grained temporal localization required for detecting partial manipulations where only short segments are altered. These constraints motivate the exploration of data-driven approaches that can learn more robust and generalizable representations directly from data.

2.2. Deep Forgery Detection

Recent progress in audio forgery detection has increasingly shifted toward deep learning (DL) methods, which automatically learn discriminative features directly from raw waveforms or time–frequency representations instead of relying on handcrafted descriptors. Unlike traditional approaches based on ENF analysis, spectral features, or noise characteristics, deep learning models are capable of modeling complex nonlinear relationships and jointly integrating diverse forensic cues within a unified framework. One prominent research direction transforms audio signals into image-like spectrogram representations, enabling the use of convolutional neural networks (CNNs) and established image forensics techniques [8,23]. Spectrogram-based analysis exposes subtle tampering artifacts embedded in the time-frequency structure and has demonstrated strong effectiveness in detecting copy-move, splicing, and compression-related manipulations. For example, integrating Constant-Q cepstral representations with CNN-based processing enhances robustness against compression artifacts and background noise, while patch-based CNN localization strategies allow precise identification of manipulated regions within the spectrogram domain. Beyond spectrogram-based approaches, several end-to-end deep architectures have been introduced for audio forgery detection. Siamese CNN models learn similarity-aware feature embeddings that generalize effectively to noisy environments and previously unseen forgery types [6]. Multi-scale residual networks augmented with attention mechanisms improve sensitivity to signal discontinuities and other manipulation-related artifacts, thereby enhancing detection performance under challenging acoustic conditions [9]. In addition, decreasing-kernel convolutional architectures support adaptive feature extraction from fused spectrogram representations, facilitating robust copy-move detection within fully end-to-end learning frameworks [29]. Overall, deep learning techniques offer strong generalization capabilities and enable the integration of multiple forensic evidence sources, including spectral, ENF, and noise-related features. However, most existing approaches remain detection-oriented, providing utterance-level or patch-level predictions rather than precise frame-level temporal localization of manipulations. Bridging this gap between accurate detection and temporally consistent localization remains a key open challenge in audio forgery analysis.

2.3. Dataset Design for Localization Tasks

The development and rigorous evaluation of audio forgery detection methods are fundamentally constrained by the availability, quality, and forensic realism of benchmark datasets. Existing resources span multiple domains, including clean speech corpora such as TIMIT [30], WSJ [31], and KSUD [32]; spoofing-focused datasets such as ASVspoof [33] and Fake-or-Real [34]; and robustness-oriented collections that account for environmental noise and device variability, including NOIZEUS and MOBIPHONE. Together, these datasets provide diversity in speakers, languages, recording environments, and manipulation conditions, supporting extensive experimentation in speech processing and audio authenticity research. However, despite this breadth, most available datasets are not specifically designed for fine-grained audio forgery localization. Clean speech corpora primarily target automatic speech recognition or speaker identification tasks, while spoofing datasets concentrate on detecting synthetic or converted speech rather than edits applied to genuine recordings. Likewise, noise- and device-robustness datasets emphasize performance

under acoustic variability but rarely include precise temporal annotations of manipulated segments or standardized editing protocols. The lack of temporally consistent ground truth, controlled manipulation procedures, and transform-balanced editing scenarios significantly restricts systematic comparison among forensic algorithms and hinders progress toward frame-level localization. Consequently, many current detection models remain optimized for clip-level or utterance-level classification, leaving the accurate identification of manipulation boundaries largely unresolved. These limitations underscore the necessity of forensically structured datasets that integrate:

- Controlled and reproducible manipulation generation;
- Precise frame-level temporal annotations;
- Standardized evaluation across diverse forgery categories.

These requirements align with established forensic guidelines for digital audio authentication, which emphasize reproducibility, chain of custody, and methodological rigor in forensic analysis [35].

To address this gap, the present work introduces a TIMIT-based controlled forgery dataset explicitly designed for frame-level audio forgery localization. This dataset facilitates systematic investigation of temporal inconsistency modeling while supporting interpretable integration of multiple forensic evidence sources.

2.4. Contextual Comparison with Related Work

Table 1 summarizes the key methodological distinctions between representative prior work and the proposed framework across task formulation, dataset design, and evaluation granularity, highlighting differences that preclude direct numerical comparison.

Table 1. Contextual comparison of representative audio forgery analysis approaches and the proposed framework. Differences in task formulation and evaluation granularity preclude direct numerical comparison. [†] Block-level spectrogram localization. [‡] Proposal-based segment localization evaluated via mAP. [§] Approximate region indication without dense temporal annotation.

Method	Forgery Type	Approach	Dataset	Primary Metric	Granularity	Frame GT
Reis et al. [13]	Splicing	ENF + SVM	Custom	Accuracy	File	No
Chen et al. [21]	Splicing	Wavelet SVD	Custom	Detection rate	Region [§]	No
Ustubioglu et al. [7]	Copy-move	SIFT on Mel spectrogram	TIMIT, Arabic	F1-score	Block [†]	No
Zhao et al. [23]	Copy-move	CNN spectrogram patches	Custom	Accuracy	Region	No
Wang et al. [8]	Copy-move	CFCC similarity matching	Custom	Precision / Recall	Segment (matching-based)	No
Ozgen and Altay [24]	Copy-move	SIFT + FAST + FREAK on Mel spectrogram	Arabic, Turkish	F1-score	File	No
Shaaban et al. [6]	Deepfake	Siamese CNN	Custom	Accuracy	File	No
Das et al. [9]	Deepfake	ResNet + attention	Custom	Accuracy	File	No
Shi et al. [29]	Copy-move	Decreasing-kernel CNN	Custom	Accuracy	File	No
Wu et al. [5]	Synthetic splicing	wav2vec2 + proposal refinement	LAV-DE, ASVS2019PS	mAP	Segment (proposal-based) [‡]	No
Proposed	Multi-transform (copy, pitch, warp, shift, deletion)	Handcrafted anomaly + wav2vec 2.0 + Hybrid	TIMIT-derived controlled benchmark	Frame F1, IoU	Frame (10 ms)	Yes

3. Methodology

3.1. Overview of the Proposed Framework

Figure 1 illustrates the proposed end-to-end framework for localized audio forgery analysis. The pipeline is organized into four sequential stages: data preparation, controlled forgery generation, evidence modeling, and evaluation.

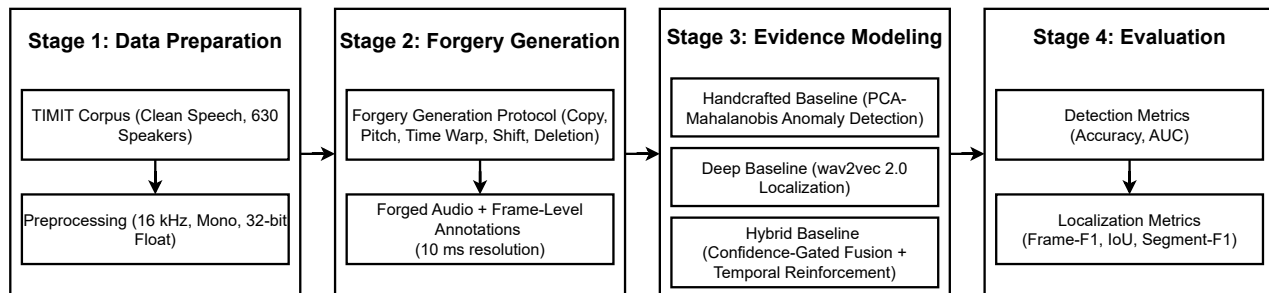


Figure 1. Overview of the proposed audio forgery framework.

Clean speech recordings are first standardized through a unified preprocessing step, producing 16 kHz mono waveforms used consistently across all experiments. A deterministic forgery generation protocol then introduces localized atomic manipulations within short temporal regions, producing forged audio together with structured, sample-accurate annotations.

These annotated recordings are subsequently analyzed using three complementary evidence branches. A handcrafted feature baseline captures low-level signal irregularities through anomaly-based statistics, a deep learning baseline based on wav2vec 2.0 learns task-adapted frame-level representations, and a hybrid baseline fuses neural predictions with handcrafted anomaly cues. All methods operate under identical temporal supervision and data splits to ensure controlled comparison.

Finally, predictions from each branch are evaluated using unified metrics for binary detection and frame-level localization, where detection serves as an auxiliary measure of global separability and localization constitutes the primary forensic task.

3.2. Dataset Construction and Statistics

This study employs a single controlled dataset derived from the TIMIT corpus, which serves as the sole source for all training and evaluation experiments. Rather than introducing a new speech corpus, the benchmark is specifically designed for fine-grained audio forgery localization. TIMIT provides clean, carefully recorded, and phonetically balanced speech, making it a well-established foundation for controlled speech analysis. Building on these properties, we construct a derived dataset by applying strictly constrained, time-limited manipulations to real speech signals and pairing each forged sample with sample-accurate temporal annotations. Table 2 summarizes the resulting dataset statistics and split configuration.

3.3. Preprocessing

All audio recordings are subjected to a unified and minimal preprocessing procedure. Each file is resampled to 16 kHz, converted to single-channel (mono) format, and represented using 32-bit floating-point precision. No denoising, dereverberation, dynamic range compression, or enhancement techniques are applied. A minimum duration constraint of approximately 2 s is enforced to ensure feasibility of localized manipulation under the imposed segment length and temporal gap constraints. Audio recordings near this threshold are handled through a regime-aware generation strategy rather than padding,

repetition, or exclusion. The same preprocessing procedure is applied consistently to both clean and forged audio across training and evaluation splits.

Table 2. Summary of the localized audio forgery benchmark derived from TIMIT.

Property	Description
Source corpus	TIMIT (English read speech)
Language	English
Sampling rate	16 kHz
Training files (clean/forged)	4261/4261
Test files (clean/forged)	1548/1548
Clean-to-forged ratio	1:1
Number of speakers	630
Speaker overlap	None (following the official TIMIT train/test split)
Avg. duration per file	~3 s
Total duration	~5.5 h (clean + forged)
Manipulation types	Copy, pitch, time warp, shift, deletion
Manipulated segments per file	1–3 (variable)
Annotation resolution	10 ms frame-level masks

3.4. Forgery Generation Protocol

We adopt a controlled and fully reproducible forgery generation protocol designed to simulate localized audio editing through atomic manipulations, including copy–paste, pitch modification, time warping, local temporal shift, and partial deletion. Rather than applying global transformations or synthesizing entire utterances, the generator modifies short temporal regions while leaving the remainder of the signal unchanged, thereby confining forensic cues to localized temporal inconsistencies.

Given a clean waveform $\mathbf{x} \in \mathbb{R}^L$ of duration D seconds, the generator produces a forged waveform $\tilde{\mathbf{x}}$ together with structured ground truth annotations describing all applied manipulations. Generation is deterministic at the file level: for each audio file indexed by i , a fixed random seed $s = s_0 + i$ is used, ensuring full reproducibility across dataset splits.

Let $\mathcal{G}$ denote the set of supported atomic transformations, including copy–paste, pitch modification, time warping, local temporal shift, and partial deletion.

For each audio file, the generator samples K non-overlapping source–target segment pairs

$$\{(S_k, T_k)\}_{k=1}^K$$

where each source segment $S_k = [s_k^s, s_k^e]$ and target segment $T_k = [t_k^s, t_k^e]$ is defined in continuous time. Each pair is modified using exactly one atomic transformation $g_k \in \mathcal{G}$. Composite or chained operations are explicitly disallowed.

Segment sampling follows a constrained non-overlapping strategy that enforces: (i) no overlap between source segments, (ii) no overlap between source and target segments, (iii) a minimum temporal gap $\Delta_{\min}$ between all segments, and (iv) segment duration sampled uniformly within predefined bounds. Segment boundaries are sampled in continuous time and converted to sample-accurate indices using the sampling rate. Target boundaries are clamped to the valid signal duration after transformation to account for operations that alter segment length.

To accommodate variability in audio duration, the generator employs a regime-aware strategy. Let $d_{\min}$ denote the minimum segment duration. Audio files shorter than

$$2d_{\min} + \Delta_{\min} \quad (1)$$

are assigned to a short-duration regime and receive exactly one localized forgery using a restricted transformation subset, denoted by $\mathcal{G}_{\text{short}}$, which includes copy–paste, pitch modification, and time warping.

Segment length is constrained to at most 60% of the audio duration to prevent dominance of manipulated content.

Longer recordings follow the normal regime, where multiple forgeries may be applied. The number of forgeries K is determined automatically based on signal duration and structural constraints, subject to predefined minimum and maximum limits. If valid non-overlapping segments cannot be sampled under the requested configuration, the generator falls back to a guaranteed single-forgery setup.

Each applied manipulation produces a structured annotation record containing a unique pair identifier, source and target temporal boundaries, transformation type, transformation parameters, and the assigned duration regime. All temporal annotations are stored in seconds with sample-level precision. This pair-level representation supports binary detection, frame-level localization, and future attribution analyses without modification to the dataset format.

Algorithm 1 summarizes the localized forgery generation procedure.

Algorithm 1: Localized Forgery Generation

Input: Clean audio $\mathbf{x}$, sampling rate f_s , configuration $\mathcal{C}$

Output: Forged audio $\tilde{\mathbf{x}}$, annotations $\mathcal{A}$

Initialize deterministic random seed based on file index;

Compute audio duration D ;

if $D < 2d_{\min} + \Delta_{\min}$ **then**

Sample one valid source–target segment pair under $\mathcal{C}$;
 Select transformation $g \in \mathcal{G}_{\text{short}}$;
 Apply localized transformation to target segment;
 Record annotation;

else

Determine maximum feasible number of edits K under structural constraints;
 Sample K non-overlapping source–target segment pairs;
for $k = 1$ **to** K **do**
 Sample transformation $g_k \in \mathcal{G}$;
 Sample transformation parameters from $\mathcal{C}$;
 Apply localized transformation to target segment;
 Record annotation;

Return $\tilde{\mathbf{x}}$ and $\mathcal{A}$;

3.5. Frame-Level Forgery Localization

The primary forensic task addressed in this work is frame-level localization of audio forgeries. Given an input audio recording, the objective is to identify the temporal regions that have been subjected to localized manipulation. Localization provides fine-grained temporal evidence of forgery and is treated as the primary task, while binary detection is derived through temporal aggregation and described separately.

Let $\mathbf{x} \in \mathbb{R}^L$ denote an input audio waveform of duration D seconds sampled at 16 kHz. Localization is formulated as the prediction of a binary temporal mask $\mathbf{m} = \{m_1, m_2, \dots, m_N\}$, where each element m_t indicates whether the corresponding time frame contains forged content. The number of frames is determined by a fixed hop size $h = 10$ ms as $N = \left\lfloor \frac{D}{h} \right\rfloor$. This temporal resolution balances boundary precision with

annotation stability; coarser resolutions reduce boundary accuracy, while finer resolutions provide diminishing returns given encoder stride and windowing effects. Multiple disjoint forged regions may be present within a single recording, and no assumptions are made regarding their number, duration, or temporal arrangement.

Frame-level ground truth supervision is derived deterministically from the structured pair-level annotations produced during forgery generation. Each annotated target segment is converted into a binary mask by marking all frames whose temporal centers fall within the corresponding target boundaries. Formally, a frame centered at time th is labeled as forged if

$$m_t = \begin{cases} 1, & \text{if } th \in [t_k^s, t_k^e] \text{ for any target segment } k, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

This conversion is transform-agnostic and identical across all forgery types and duration regimes. Frames outside all annotated target segments are labeled as pristine.

To enable efficient training and uniform batching, localization models operate on fixed-duration audio segments of 4.0 s. This duration provides sufficient temporal context to capture local manipulation artifacts while limiting the inclusion of unrelated speech content. Empirically, we observed that modest variations in crop duration did not alter the relative performance trends between handcrafted, neural, and hybrid methods. During training, random temporal crops are sampled from each audio file, and the corresponding segments of the ground truth mask are extracted using the same temporal offsets. During evaluation, center cropping is applied to ensure deterministic alignment. Audio recordings shorter than the required duration are zero-padded symmetrically, with mask alignment preserved using the same hop size.

Precise temporal boundaries of audio edits are inherently ambiguous due to windowing effects, encoder stride, and gradual signal transitions. Accordingly, all ground truth masks remain strictly binary during training, and boundary ambiguity is addressed at evaluation time through boundary-tolerant localization metrics.

The localization task encompasses all supported forgery operations, including copy–paste, pitch modification, time warping, local temporal shift, and partial deletion. The formulation is intentionally transform-agnostic: models are not required to identify the specific manipulation applied and operate without access to oracle detection labels or prior knowledge of the number or placement of forged regions.

3.6. Handcrafted Feature Baseline

To provide a classical and interpretable reference point, we employ an unsupervised PCA–Mahalanobis distance–based anomaly localization baseline built on handcrafted acoustic features. This model operates as a statistical deviation detector that learns the distribution of clean speech and identifies localized forgeries as frames that deviate from this learned reference. Unlike neural baselines, it does not rely on task-specific supervision or learned semantic representations, serving instead as a principled signal-level baseline.

Handcrafted features are extracted directly from the waveform using a 25 ms analysis window and a 10 ms hop size, matching the temporal resolution used throughout the localization framework. The feature set is designed to capture both spectral structure and short-term temporal dynamics, and includes mel-frequency cepstral coefficients (MFCCs) with first- and second-order derivatives, spectral flux, root-mean-square (RMS) energy, and zero-crossing rate. All features are computed on raw audio without denoising, dereverberation, or enhancement in order to avoid suppressing potential forensic artifacts.

For each recording, this process yields a temporally aligned feature matrix $\mathbf{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_T\}$, where T denotes the number of frames and D the feature dimensionality.

Rather than performing per-utterance normalization or heuristic aggregation, anomaly detection is formulated using a global statistical model trained exclusively on clean speech. Let $\mathbf{X} \in \mathbb{R}^{N \times D}$ denote the collection of all frame-level feature vectors extracted from clean training recordings. A principal component analysis (PCA) transform is first learned to decorrelate features and reduce dimensionality $\mathbf{y}_t = \mathbf{W}^\top (\mathbf{f}_t - \boldsymbol{\mu})$, where $\mathbf{W} \in \mathbb{R}^{D \times r}$ contains the top r principal components and $\boldsymbol{\mu}$ is the global feature mean estimated from clean data.

On the PCA-projected space, a robust covariance estimator is fit using the Ledoit-Wolf shrinkage method. Given the estimated mean $\boldsymbol{\mu}_y$ and precision matrix $\boldsymbol{\Sigma}^{-1}$, a frame-level anomaly score is computed as the squared Mahalanobis distance:

$$s_t = (\mathbf{y}_t - \boldsymbol{\mu}_y)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_t - \boldsymbol{\mu}_y) \quad (3)$$

This formulation explicitly models deviations from the global distribution of clean speech and is sensitive to localized spectral and temporal inconsistencies introduced by partial manipulations.

To suppress spurious frame-level fluctuations, anomaly scores are optionally smoothed using a short median filter. A fixed threshold τ , selected on the development split by maximizing Frame F1, is then applied to obtain a binary localization mask:

$$\hat{m}_t = \begin{cases} 1, & s_t \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

No parameter tuning or threshold adjustment is performed on the test set.

This handcrafted localization method is fully unsupervised: ground-truth annotations are never used during model fitting and are employed only for threshold selection and evaluation. As such, it serves as a principled lower-bound reference that highlights the limitations of purely signal-level anomaly detection. At the same time, it provides complementary boundary-sensitive cues that are later exploited in the proposed hybrid framework to stabilize neural predictions.

Although the handcrafted model is primarily designed for frame-level anomaly localization, it is also evaluated in a file-level detection setting to provide a classical baseline for global forgery discrimination. For detection, frame-level anomaly scores $\{s_t\}$ are aggregated into a fixed-dimensional utterance representation using summary statistics, including the mean, maximum, and variance. These aggregated features are passed to a logistic regression classifier trained with file-level supervision. All classifier hyperparameters are selected on the development split, and no test-time tuning is performed.

3.7. Neural Representation Baseline (wav2vec 2.0)

To evaluate the role of learned representations in localized audio forgery analysis, we employ deep learning baselines built on pretrained self-supervised speech encoders. The same encoder backbone is used consistently across both detection and localization tasks to ensure a fair comparison. Any observed performance differences therefore arise from task formulation and supervision rather than architectural variation.

All deep baselines are based on the wav2vec 2.0 base model (wav2vec2-base), pretrained on the LibriSpeech 960-h corpus of clean English speech. The encoder produces 768-dimensional contextualized frame-level representations. For task adaptation, only the final two transformer layers are unfrozen, while earlier layers remain frozen. Given a raw waveform input $\mathbf{x} \in \mathbb{R}^L$, the encoder produces a sequence of contextualized frame-level embeddings $\mathbf{H} = \{\mathbf{h}_1, \dots, \mathbf{h}_{T'}\}$, where $\mathbf{h}_t \in \mathbb{R}^D$, $D = 768$ for the base model, and T' depends on the encoder stride. The pretrained encoder is not exposed to any forgery-related supervision during pretraining.

For binary detection, the encoder is partially adapted by unfreezing the final two transformer layers, while all earlier layers remain frozen. A fixed-length utterance-level representation is obtained via temporal average pooling, $\mathbf{z} = \frac{1}{T'} \sum_{t=1}^{T'} \mathbf{h}_t$, and passed to a lightweight classification head consisting of two fully connected layers with ReLU activation. The model is trained using cross-entropy loss with file-level supervision. Temporal structure is intentionally discarded to reflect the global nature of the detection task.

For localization, temporal structure is preserved. Frame-level embeddings are processed directly to estimate a forgery likelihood at each time step. As in the detection setting, only the final two transformer layers of the encoder are unfrozen. Let $\mathbf{H} \in \mathbb{R}^{T' \times D}$ denote the encoder output. A shallow temporal prediction head, implemented as a one-dimensional convolutional network, operates on this sequence: $\mathbf{p}_t = \sigma(f_\theta(\mathbf{h}_t))$, where f_θ denotes the convolutional head and σ is the sigmoid function. The head consists of a Conv1D layer with kernel size 3 and ReLU activation, followed by a pointwise Conv1D projection to a single output channel.

Frame-level ground truth masks are downsampled when necessary to match the encoder temporal resolution. Training is performed using binary cross-entropy loss with logits. Localization is learned without access to oracle detection labels or prior knowledge of the number or type of manipulations present in the audio.

By sharing a common pretrained backbone across detection and localization tasks, these baselines provide a controlled assessment of how learned speech representations support both global and fine-grained forensic inference. Unfreezing only the final transformer layers yields a favorable balance between task adaptation and overfitting risk. Deeper adaptation did not change qualitative performance trends and resulted in less stable training behavior.

3.8. Hybrid Confidence-Gated Localization with Temporal Reinforcement

Handcrafted and neural representations capture complementary forensic cues. Handcrafted features are sensitive to low-level signal irregularities, while deep neural representations encode higher-level contextual and phonetic consistency. The neural evidence branch is instantiated using the adapted wav2vec 2.0 localization model, whose pretrained contextual representations and calibrated frame-level posteriors make it well suited for confidence-aware fusion. To leverage these complementary properties without introducing additional supervision, joint training, or feature-level coupling, we propose a hybrid localization framework that integrates both evidence sources through confidence-aware fusion and explicit temporal consistency enforcement.

The proposed hybrid method operates entirely at inference time and does not involve end-to-end retraining, feature concatenation, or task-specific tuning. Fusion is performed deterministically at the frame level, enabling controlled analysis under identical evaluation conditions.

Let $\mathbf{p} \in [0, 1]^T$ and $\mathbf{s} \in \mathbb{R}^T$ denote the frame-level posterior probabilities produced by the neural localization model and the anomaly scores produced by the handcrafted detector, respectively. Here, p_t and s_t denote the corresponding scalar values at frame t , and T denotes the number of frames at the neural encoder resolution, consistent with T' defined in Section 3.7.

Because the two pipelines may operate at different native temporal resolutions, handcrafted scores are temporally aligned to the neural timeline using linear interpolation. Rather than directly combining binary masks, fusion is governed by the confidence of the neural prediction. Three confidence regimes are defined using fixed thresholds τ_{high} and τ_{low} :

- If $p_t \geq \tau_{\text{high}}$, the frame is accepted as forged based on neural evidence alone.
- If $p_t \leq \tau_{\text{low}}$, the frame is rejected as genuine.
- If $\tau_{\text{low}} < p_t < \tau_{\text{high}}$, the neural prediction is considered uncertain, and the handcrafted anomaly score is consulted. The frame is marked as forged only if $s_t \geq \tau_{\text{hand}}$.

This confidence-gated strategy ensures that handcrafted evidence is used exclusively to resolve ambiguous neural predictions, preventing it from overriding high-confidence neural decisions or introducing spurious detections in clearly genuine regions.

Although confidence-gated fusion improves the reliability of frame-level decisions, forged regions are inherently temporally contiguous. To enforce this structural prior, a lightweight temporal reinforcement step is applied to the fused binary mask. This step consists of short-gap filling, minimum segment length enforcement, and limited local dilation over a small temporal radius.

Importantly, temporal reinforcement does not introduce new evidence or modify confidence thresholds. Instead, it stabilizes contiguous detections, reduces temporal fragmentation, and improves segment-level coherence while preserving the conservative nature of the fusion logic.

3.9. Evaluation Metrics

The proposed framework is evaluated on two tasks: binary forgery detection and frame-level forgery localization. All metrics are computed on held-out test data, with all thresholds and hyperparameters fixed using the development split.

3.9.1. Detection Metrics

Binary forgery detection is evaluated at the file level using classification accuracy and the area under the receiver operating characteristic curve (AUC). Accuracy measures the proportion of correctly classified recordings,

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{y}_i = y_i) \quad (5)$$

where $y_i \in \{0, 1\}$ denotes the ground-truth label and $\hat{y}_i$ is obtained by thresholding the forged-class probability at 0.5. AUC provides a threshold-independent measure of discriminative performance by assessing the ranking quality between forged and clean samples.

3.9.2. Localization Metrics

Forgery localization is evaluated at a fixed temporal resolution of 10 ms using binary frame-level masks. Predicted probabilities are thresholded using a fixed value selected on the development split.

Frame-level precision, recall, and F1 score are computed as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (6)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (7)$$

$$\text{Frame-F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively, computed by comparing predicted and ground-truth frame labels. Precision (6) and Recall (7) are combined to compute Frame F1 (8), which is treated as the primary localization metric.

To quantify overall overlap consistency between predicted and ground-truth manipulated regions, frame-level intersection over union (IoU, (9)) is additionally reported:

$$\text{IoU} = \frac{|\mathbf{m}^{\text{pred}} \cap \mathbf{m}^{\text{gt}}|}{|\mathbf{m}^{\text{pred}} \cup \mathbf{m}^{\text{gt}}|}. \quad (9)$$

To account for boundary uncertainty introduced by windowing effects, encoder stride, and gradual signal transitions, a boundary-tolerant Frame F1 is computed by allowing a temporal tolerance of ± 50 ms around ground-truth boundaries. Under this tolerance, predicted frames within the specified temporal window are considered correctly matched for the purpose of computing TP, FP, and FN.

Under this formulation, tolerance-adjusted counts TP, FP, and FN are computed by considering predicted frames within the specified temporal window as correct matches. These adjusted counts are then used to compute boundary-tolerant Precision, Recall, and Frame F1 in the same manner as their standard definitions.

Finally, a segment-level F1 score is reported as a coarse-grained diagnostic metric. Binary frame masks are converted into contiguous temporal segments, and a predicted segment is considered correct if it overlaps a ground-truth segment within a temporal tolerance and exceeds a minimum IoU threshold. Segment-level F1 is reported as a complementary forensic indicator that captures temporal coherence and structural consistency of predicted regions, providing a measure of practical forensic utility that is distinct from but complementary to frame-level boundary precision.

4. Results and Discussion

4.1. Experimental Protocol

All results are reported on the held-out test split. Binary detection and frame-level localization are evaluated using the data partitions defined in Section 3. Localization performance is assessed at a fixed temporal resolution of 10 ms. For all methods requiring thresholding, thresholds are selected on the development split and fixed for test evaluation.

Binary detection is evaluated at the file level, while localization models operate on fixed-duration audio segments as described in the methodology. No parameter tuning or threshold adjustment is performed on the test set.

4.2. Binary Forgery Detection Performance

Binary forgery detection is evaluated as a secondary task to provide a coarse file-level indication of whether an audio recording contains any manipulation. Detection does not localize edits in time, but instead serves as a global reference for separability between clean and forged signals. Performance is reported for handcrafted and deep representations. The hybrid method is not considered here, as it is designed specifically for frame-level localization rather than global classification. The deep model substantially outperforms the handcrafted baseline across both metrics, while handcrafted statistics capture certain global spectral and energy deviations introduced by manipulation, they rely on coarse distributional differences and therefore exhibit limited discriminative capacity. In contrast, the pretrained wav2vec 2.0 encoder provides high-capacity contextual representations that model phonetic structure, prosody, and long-range temporal consistency. Fine-tuning these representations for forgery detection yields near-perfect ranking performance and significantly higher classification accuracy, as shown in Table 3.

Table 3. Binary forgery detection performance on the test set. The deep model is based on a wav2vec 2.0 encoder with the final two transformer layers fine-tuned. Results are reported from the best development-selected checkpoint.

Method	Accuracy	ROC-AUC
Handcrafted (Logistic Regression)	0.856	0.926
Deep (wav2vec 2.0)	0.988	0.998

Figures 2 and 3 further illustrate this performance gap. Both visualizations clearly demonstrate the superior discriminative capability of the deep model compared to the handcrafted baseline, consistently showing stronger separation between clean and forged recordings across evaluation settings.

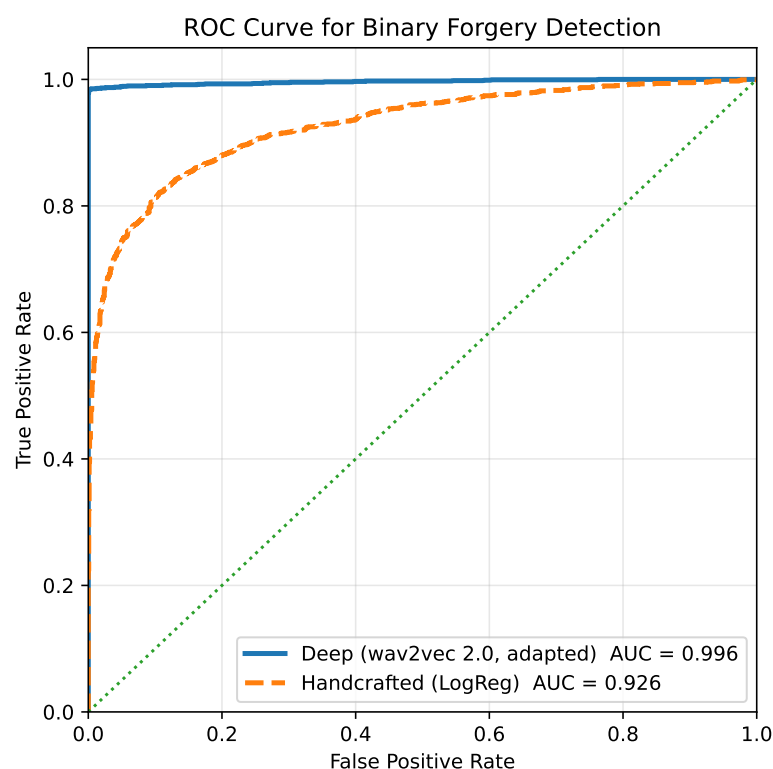


Figure 2. Receiver operating characteristic (ROC) curves for binary forgery detection. The dotted line represents the random classifier baseline.

Although detection confirms that manipulated recordings can be reliably distinguished at the file level, it provides only a binary decision and offers no insight into the temporal extent or location of edits. For forensic analysis, such coarse predictions are insufficient. Consequently, we next focus on frame-level localization, which provides fine-grained temporal evidence of manipulation.

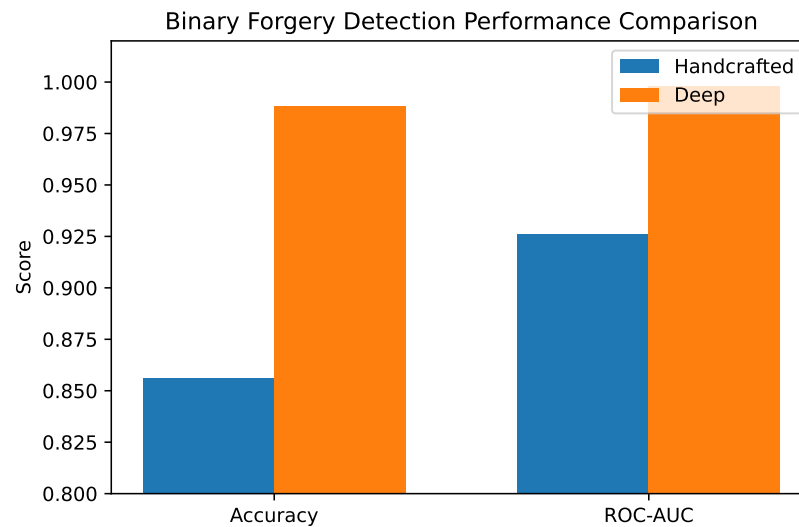


Figure 3. Bar-chart comparison of binary forgery detection performance.

4.3. Frame-Level Forgery Localization Performance

Frame-level forgery localization constitutes the primary task of this study and requires models to identify the precise temporal regions affected by manipulation. Localization performance is evaluated using Frame F1 (8) as the principal metric. Segment F1 is additionally reported as a complementary indicator of forensic utility, capturing the coherence and structural consistency of predicted tampered regions. Boundary-tolerant Frame F1 and frame-level intersection-over-union (IoU, (9)) are reported as complementary measures capturing boundary accuracy and overall overlap consistency, respectively, as shown in Table 4.

Table 4. Frame-level forgery localization performance on the test set.

Method	Frame F1	Boundary F1	Segment F1	IoU
Handcrafted (Anomaly-based)	0.4491	0.4495	0.4088	0.2940
Deep (wav2vec 2.0)	0.7519	0.7542	0.7064	0.6339
Hybrid (Deep + Handcrafted)	0.7471	0.7475	0.7938	0.6171

Several important observations emerge from these results. First, the handcrafted anomaly-based baseline performs weakly across all localization metrics, while it achieves moderate frame-level accuracy, its low IoU and segment-level F1 indicate fragmented detections and poor temporal coherence. This behavior is expected, as the method is entirely unsupervised and relies solely on low-level statistical deviations from clean speech rather than task-specific learning. Handcrafted descriptors are sensitive to local spectral and energy fluctuations, many of which arise naturally in speech and are not indicative of manipulation, leading to frequent false positives and unstable temporal predictions.

Second, the deep wav2vec 2.0-based localization model achieves strong and consistent performance across all metrics. Leveraging pretrained contextual representations that encode phonetic structure, prosody, and long-range temporal dependencies, the model effectively captures subtle inconsistencies introduced by localized manipulations such as segment insertion, pitch modification, or temporal warping. This is reflected not only in high Frame F1 and Boundary F1 scores, but also in the highest IoU value among the evaluated methods, indicating superior overlap consistency between predicted and ground-truth manipulated regions.

Finally, the proposed hybrid framework exhibits a distinct and complementary performance profile, while its Frame F1 and Boundary F1 scores are marginally lower than those

of the deep model, it achieves a substantial improvement in Segment F1, outperforming the neural baseline by a wide margin. Although the hybrid model yields slightly lower IoU than the deep-only approach, this reduction reflects more conservative frame-level decisions rather than degraded localization quality. The confidence-gated fusion restricts the influence of handcrafted anomaly cues to regions of neural uncertainty, while temporal reinforcement suppresses spurious detections and stabilizes contiguous segments. As a result, the hybrid framework produces more temporally coherent and structurally consistent forged regions, even when frame-wise overlap is slightly reduced.

Overall, these findings highlight a clear trade-off between fine-grained frame-level accuracy and segment-level coherence, while learned contextual representations are essential for accurate and stable overlap with manipulated regions, integrating handcrafted anomaly cues through principled confidence gating and temporal reinforcement yields more coherent and forensically meaningful segment-level localization without aggressive threshold tuning or transform-specific heuristics.

While Frame F1 is treated as the primary localization metric in this study—reflecting the finest achievable temporal resolution—segment-level coherence carries distinct and complementary forensic significance. In practical audio forensic analysis, the objective extends beyond correctly labeling individual frames to identifying manipulated regions as temporally contiguous and well-bounded segments. Models that produce fragmented or temporally inconsistent predictions, even with strong frame-level accuracy, may be less actionable in forensic workflows.

The hybrid framework is designed with this consideration in mind, incorporating confidence-gated fusion and temporal reinforcement to promote structural consistency in predicted regions. As a result, it demonstrates improved segment-level coherence while maintaining competitive frame-level performance. Accordingly, Frame F1 and Segment F1 should be interpreted as complementary indicators: the former quantifies boundary precision, while the latter reflects the practical usability of the predictions in forensic analysis.

4.4. Per-Transform Localization Performance

Results are presented for the adapted deep neural localization model (Table 5) and the hybrid fusion model (Table 6). The handcrafted anomaly-based baseline is omitted from this analysis, as its overall localization performance was uniformly low and did not exhibit meaningful variation across manipulation types.

4.4.1. Deep (wav2vec 2.0): Per-Transform Localization Performance

The results show that the adapted wav2vec 2.0 localization model achieves consistently strong performance across most manipulation types, as shown in Table 5. Pitch modification, time warping, copy-paste, and local temporal shift operations all yield Frame F1 scores in the range of approximately 0.75 to 0.81, indicating that the model generalizes well across a diverse set of additive and modification-based edits. These transformations typically introduce either boundary artifacts or contextual inconsistencies that are effectively captured by the model's learned temporal representations.

Table 5. Per-transform localization performance of the Deep (wav2vec 2.0) model on the test set.

Transform	Frame F1	Boundary F1	Segment F1	<i>n</i>
Time warp	0.787	0.789	0.761	315
Local shift	0.757	0.759	0.685	294
Copy	0.783	0.784	0.731	349
Pitch	0.811	0.813	0.792	350
Partial deletion	0.568	0.574	0.500	240

Among the evaluated manipulations, partial deletion is clearly the most challenging, exhibiting a noticeable drop in performance across all metrics. Unlike insertion or replacement operations, deletion removes content without introducing foreign acoustic material, resulting in weaker boundary cues and fewer explicit signal discontinuities. As a result, localization must rely primarily on contextual inconsistencies, which are inherently more difficult to detect precisely at the frame level.

Despite this challenge, the model still achieves a Segment F1 of 0.50 for deletion-based forgeries, suggesting that it can identify manipulated regions at a coarse temporal scale even when fine-grained boundary localization is more ambiguous. Overall, the relatively stable performance across most transforms highlights the robustness of the adapted wav2vec 2.0 encoder for localized audio forgery detection, with sensitivity primarily observed for deletion-based edits that lack strong acoustic artifacts.

4.4.2. Hybrid Localization Performance

To assess whether handcrafted anomaly cues provide complementary evidence beyond neural predictions, we report per-transform localization performance of the hybrid fusion model with confidence-gated fusion and temporal reinforcement. Results are summarized in Table 6.

Table 6. Per-transform localization performance of the hybrid model.

Transform	Frame F1	Boundary F1	Segment F1	<i>n</i>
Time warp	0.782	0.784	0.685	315
Local shift	0.744	0.745	0.583	294
Copy	0.775	0.776	0.640	349
Pitch	0.794	0.795	0.696	350
Partial deletion	0.587	0.590	0.495	240

Across all manipulation types, the hybrid model achieves frame-level and boundary-level performance comparable to the adapted neural baseline, indicating that confidence-gated fusion preserves neural precision without introducing excessive false positives. Frame F1 values remain stable across time warping, copy-paste, pitch modification, and local temporal shifts, while partial deletion remains the most challenging case.

At the segment level, the hybrid model exhibits improved temporal coherence relative to frame-wise fusion alone, though gains vary across manipulation types. Operations that introduce explicit boundary artifacts, such as copy-paste, pitch modification, and time warping, benefit more from temporal reinforcement than deletion-based edits. In contrast, partial deletion yields lower segment-level performance, reflecting the absence of strong acoustic discontinuities or foreign material.

Overall, these results indicate that the hybrid framework primarily improves localization consistency rather than raw detection sensitivity, while it does not uniformly outperform the neural model at the frame level, it provides more temporally stable segment predictions across most transformation types, supporting its use as a complementary forensic localizer rather than a replacement for learned representations.

5. Limitations and Future Work

The proposed framework is designed to isolate and analyze the intrinsic difficulty of fine-grained audio forgery localization under controlled conditions, while this design enables precise evaluation and reproducibility, it does not explicitly model real-world distribution shift. In practical settings, forged recordings may undergo additional processing such as lossy compression, re-encoding, background noise, or transmission artifacts, all of which can alter the acoustic characteristics of both genuine and manipulated regions.

Moreover, all experiments are conducted on English read speech derived from the TIMIT corpus, and the benchmark focuses on single, localized manipulation operations, while TIMIT provides a clean, phonetically balanced, and reproducible foundation that is well-suited for controlled forensic analysis, it does not capture the diversity of real-world recording conditions, languages, or acoustic environments. The relatively small scale of the benchmark and its focus on read speech may limit generalization to conversational, multilingual, or noise-degraded audio. Evaluating the proposed framework on larger and more diverse datasets such as ASVspoof and PartialSpoof [4], as well as in-the-wild recordings, remains an important direction for future work and is explicitly identified as a priority.

Future work will extend the proposed framework toward more realistic forensic scenarios by evaluating localization robustness under distribution shift and on partially manipulated recordings from public benchmarks such as ASVspoof. Additional directions include analyzing user-edited speech involving manual cut-paste operations or silence removal, expanding the evaluation to multilingual and conversational speech, and studying compound or overlapping manipulations. From a modeling perspective, further improvements may be achieved through stronger temporal modeling and uncertainty-aware boundary refinement, provided that such extensions preserve the interpretability, stability, and conservative decision-making principles emphasized in this work.

6. Conclusions

This work presents a structured and reproducible framework for localized audio forgery analysis, with a primary focus on fine-grained temporal localization of partial manipulations rather than coarse file-level classification. By introducing a controlled forgery generation protocol with precise frame-level annotations, the proposed framework enables systematic and fair evaluation of localization performance across diverse manipulation types. Handcrafted signal-level methods, supervised neural representations, and hybrid fusion strategies are examined within a unified experimental setting to disentangle the contribution of different forensic evidence sources.

Experimental results demonstrate that purely anomaly-based handcrafted features are insufficient for reliable frame-level localization, while such features capture low-level spectral and energy irregularities, their lack of task-specific supervision and contextual modeling leads to high false-positive rates, fragmented detections, and weak overlap consistency. In contrast, task-adapted pretrained speech representations based on wav2vec 2.0 consistently achieve strong localization performance across all metrics, including high frame-level F1 and IoU scores, indicating accurate and stable alignment with manipulated regions.

The proposed hybrid localization framework, which combines confidence-gated fusion with explicit temporal reinforcement, does not uniformly outperform the neural model at the frame level. However, it provides a clear and consistent advantage in segment-level localization. By enforcing temporal coherence without altering confidence thresholds or introducing additional supervision, the hybrid approach substantially improves Segment F1, yielding more contiguous and forensically meaningful localization outputs. These

results show that handcrafted anomaly cues, while weak in isolation, can play a valuable supporting role when selectively integrated to stabilize neural predictions over time.

Per-transform analysis further confirms that the adapted neural model offers robust and consistent localization across a wide range of manipulation types, including copy–paste, pitch modification, time warping, and temporal shifts. Deletion-based edits remain the most challenging scenario for all methods, reflecting the absence of strong acoustic discontinuities when content is removed rather than replaced. Nevertheless, performance degradation remains moderate, suggesting that contextual modeling remains effective even under subtle manipulation conditions.

Overall, the findings highlight a clear division of strengths. Supervised neural representations provide the most accurate frame-level detection and overlap consistency for localized audio forgeries, while the proposed hybrid framework improves temporal coherence and segment-level interpretability. The presented benchmark, evaluation protocol, and analysis provide a reproducible foundation for future research in fine-grained audio forensics, including stronger temporal modeling, improved boundary refinement, and defenses against increasingly realistic and localized generative editing techniques.

Author Contributions: Conceptualization, M.M. and T.A.A.; Methodology, M.M. and T.A.A.; Software, M.M.; Validation, M.M.; Formal analysis, M.M. and T.A.A.; Writing—original draft, M.M., T.A.A. and I.I.; Writing—review and editing, F.A.E. and T.D.; Visualization, M.M. and T.A.A.; Supervision, F.A.E. and I.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Naif Arab University for Security Sciences under grant No. NAUSS-25-R02.

Data Availability Statement: The data presented in this study are openly available in TIMIT-corpus at <https://www.kaggle.com/datasets/nltkdata/timitcorpus> (accessed on 29 December 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nasr, M.A.; El-Shafai, W.; Abdel-Salam, N.; El-Rabaie, E.S.M.; El-Fishawy, A.S.; Abd El-Samie, F.E. A comprehensive survey of audio forgery detection: Challenges and novel trends. *J. Electr. Syst. Inf. Technol.* **2025**, *12*, 30. [\[CrossRef\]](#)
2. Zakariah, M.; Khan, M.K.; Malik, H. Digital multimedia audio forensics: Past, present and future. *Multimed. Tools Appl.* **2018**, *77*, 1009–1040. [\[CrossRef\]](#)
3. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
4. Zhang, L.; Wang, X.; Cooper, E.; Evans, N.; Yamagishi, J. The partialspooof database and countermeasures for the detection of short fake speech segments embedded in an utterance. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *31*, 813–825. [\[CrossRef\]](#)
5. Wu, J.; Lu, W.; Luo, X.; Yang, R.; Wang, Q.; Cao, X. Coarse-to-fine proposal refinement framework for audio temporal forgery detection and localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 7395–7403.
6. Shaaban, O.A.; Yildirim, R. Audio deepfake detection using deep learning. *Eng. Rep.* **2025**, *7*, e70087. [\[CrossRef\]](#)
7. Ustubioglu, B.; Tahaoglu, G.; Ulutas, G. Detection of audio copy-move-forgery with novel feature matching on Mel spectrogram. *Expert Syst. Appl.* **2023**, *213*, 118963. [\[CrossRef\]](#)
8. Wang, D.; Li, X.; Shi, C.; Niu, X.; Xiong, L.; Wu, H.; Qian, Q.; Qi, C. Robust copy-move detection and localization of digital audio based CFCC feature. *Multimed. Tools Appl.* **2025**, *84*, 9573–9589. [\[CrossRef\]](#)
9. Das, D.; Dixit, R. Detecting audio forgery using deep learning techniques with attention mechanisms on ResNet++. *Signal Image Video Process.* **2025**, *19*, 886.
10. Moussa, D.; Hirsch, G.; Riess, C. Towards unconstrained audio splicing detection and localization with neural networks. In *Proceedings of the International Conference on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 264–280.
11. Gulsoy, T.; Gulsoy, E.K.; Ustubioglu, A.; Ustubioglu, B.; Kablan, E.B.; Ayas, S.; Ulutas, G.; Tahaoglu, G.; Elhoseny, M. Detecting audio splicing forgery: A noise-robust approach with Swin Transformer and cochleagram. *J. Inf. Secur. Appl.* **2025**, *93*, 104130. [\[CrossRef\]](#)

12. Zeng, C.; Li, K.; Wang, Z. Enformer: Long-short term representation of electric network frequency for digital audio tampering detection. *Knowl.-Based Syst.* **2024**, *297*, 111938. [CrossRef]
13. Reis, P.M.G.I.; da Costa, J.P.C.L.; Miranda, R.K.; Del Galdo, G. ESPRIT-Hilbert-based audio tampering detection with SVM classifier for forensic analysis via electrical network frequency. *IEEE Trans. Inf. Forensics Secur.* **2016**, *12*, 853–864. [CrossRef]
14. Esquef, P.A.A.; Apolinário, J.A.; Biscainho, L.W. Edit detection in speech recordings via instantaneous electric network frequency variations. *IEEE Trans. Inf. Forensics Secur.* **2014**, *9*, 2314–2326. [CrossRef]
15. Mao, M.; Xiao, Z.; Kang, X.; Li, X.; Xiao, L. Electric network frequency based audio forensics using convolutional neural networks. In *Proceedings of the IFIP International Conference on Digital Forensics*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 253–270.
16. Farid, H. *Detecting Digital Forgeries Using Bispectral Analysis*; Massachusetts Institute of Technology: Cambridge, MA, USA, 1999.
17. Lin, X.; Kang, X. Exposing speech tampering via spectral phase analysis. *Digit. Signal Process.* **2017**, *60*, 63–74. [CrossRef]
18. Yan, Q.; Yang, R.; Huang, J. Detection of speech smoothing on very short clips. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 2441–2453. [CrossRef]
19. Yang, R.; Shi, Y.Q.; Huang, J. Defeating fake-quality MP3. In *Proceedings of the 11th ACM Workshop on Multimedia and Security*; Association for Computing Machinery: New York, NY, USA, 2009; pp. 117–124.
20. Luo, D.; Sun, M.; Huang, J. Audio postprocessing detection based on amplitude cooccurrence vector feature. *IEEE Signal Process. Lett.* **2016**, *23*, 688–692. [CrossRef]
21. Chen, J.; Xiang, S.; Huang, H.; Liu, W. Detecting and locating digital audio forgeries based on singularity analysis with wavelet packet. *Multimed. Tools Appl.* **2016**, *75*, 2303–2325. [CrossRef]
22. Zhou, C.; Pan, P.; Huang, L. Authenticity identification of speaker digital recording data based on quantum genetic algorithm. *Multimed. Tools Appl.* **2018**, *77*, 19399–19413. [CrossRef]
23. Zhao, W.; Zhang, Y.; Wang, Y.; Zhang, S. An Audio Copy-Move Forgery Localization Model by CNN-Based Spectral Analysis. *Appl. Sci.* **2024**, *14*, 4882. [CrossRef]
24. Ozgen, E.; Altay, S. Detecting Audio Copy-Move Forgeries on Mel Spectrograms via Hybrid Keypoint Features. *Appl. Sci.* **2025**, *15*, 11845. [CrossRef]
25. Zhao, H.; Chen, Y.; Wang, R.; Malik, H. Audio splicing detection and localization using environmental signature. *Multimed. Tools Appl.* **2017**, *76*, 13897–13927. [CrossRef]
26. Cuccovillo, L.; Mann, S.; Tagliasacchi, M.; Aichroth, P. Audio tampering detection via microphone classification. In *Proceedings of the 2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*; IEEE: Piscataway, NJ, USA, 2013; pp. 177–182.
27. Yu, Y. Research on Digital Audio Copy-Move and Splicing Detection. Master's Thesis, South China University of Technology, Guangzhou, China, 2018.
28. Pan, X.; Zhang, X.; Lyu, S. Detecting splicing in digital audios using local noise level estimation. In *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2012; pp. 1841–1844.
29. Shi, C.; Qiu, X.; Wu, M.; Niu, X.; Li, X.; Abdullahi, S.M. Audio copy-move forgery detection with decreasing convolutional kernel neural network and spectrogram fusion. *Complex Intell. Syst.* **2025**, *11*, 382. [CrossRef]
30. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Tech. Rep. N* **1993**, *93*, 27403.
31. Paul, D.B.; Baker, J. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the Speech and Natural Language: Proceedings of a Workshop*, Harriman, NY, USA, 23–26 February 1992.
32. Alsulaiman, M.; Muhammad, G.; Abdelkader, B.; Mahmood, A.; Ali, Z. *King Saud University Arabic Speech Database*; Hard Drive, catalog LDC2014S02; Linguistic Data Consortium: Philadelphia, PA, USA, 2014.
33. Wang, X.; Yamagishi, J.; Todisco, M.; Delgado, H.; Nautsch, A.; Evans, N.; Sahidullah, M.; Vestman, V.; Kinnunen, T.; Lee, K. A.; et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Comput. Speech Lang.* **2020**, *64*, 101114. [CrossRef]
34. Reimao, R.; Tzerpos, V. FoR: A Dataset for Synthetic Speech Detection. In *Proceedings of the SpeD*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–10.
35. SWGDE. Best Practices for Digital Audio Authentication. 2023. Available online: <https://www.swgde.org/documents/published-complete-listing/> (accessed on 20 March 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.