Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2026.2054

Research Article

Recognition of audio source recording device using Mel frequency cepstral coefficients and recurrent neural networks

Venkata Lalitha NARLA[1], Gulivindala SURESH[1], D. P. GANGWAR[2], KRKV PRASAD[3,*], Rita Rani BHATTACHARJEE[4]

[1]*Department of Electronics and Communication Engineering, Aditya University, Surampalem, A.P, India*
[2]*Central Forensic Science Laboratory, Chandigarh, Punjab, India*
[3]*Department of Computer Science Engineering (AI & ML), Aditya University, Surampalem, A.P, India*
[4]*Centre for VIT Happiness and Well Being, Vellore Institute of Technology, Vellore, 632014, India*

ARTICLE INFO

Article history
Received: 16 February 2025
Revised: 01 June 2025
Accepted: 23 August 2025

Keywords:
Audio Source Recording Device; Constant-Q Transform; Digital Forensics; Mel Frequency Cepstral Coefficients, Recurrent Neural Networks, Long Short-Term Memory Networks

ABSTRACT

Accurate identification of audio source recording devices is paramount in digital forensic investigations, including topics like copyright protection, tamper detection, and audio source forensics. This work presented a novel method for learning feature representations using temporal audio characteristics, such as Mel Frequency Cepstral Coefficients (MFCC) and Constant-Q Transform (CQT), obtained from segmented acoustic features. Subsequently creates a structured representation learning model by combining Long Short-Term Memory Networks (LSTM) with Recurrent Neural Networks (RNN). This model efficiently condenses spatial information, resulting in accurate recognition, by utilizing temporal modelling and time-frequency representation. Audio samples were collected from four widely used mobile devices—iPhone, Realme, Vivo, and Poco—with each contributing 70 speech recordings of 10 seconds duration, totaling 280 samples. Recordings were captured in semi-controlled indoor environments using standardized speech content to simulate real-world conditions. The outcomes of the experiment show an amazing degree of accuracy, with 96% in classifying four types of recording audio source devices. This method promises improved efficacy in a variety of forensic circumstances and represents a substantial development in audio forensic analysis. The performance metrics of audio source recording using CQT-RNN and MFCC-RNN are compared and compared with state-of-the-art methods. A user interface has been developed to facilitate the recognition of the source device for test audio signals using the proposed method. The entire study represents a significant breakthrough in audio forensic analysis with a powerful, precise, and easy-to-use solution to the problem of identifying audio source recording devices and highlights its possibility of extensive use in forensic practice.

Cite this article as: Narla VL, Suresh G, Gangwar DP, Prasad KRKV, Bhattacharjee RR. Recognition of audio source recording device using Mel frequency cepstral coefficients and recurrent neural networks. Sigma J Eng Nat Sci 2026;44(3):1572–1586.

***Corresponding author.**
*E-mail address: krvprasad219@gmail.com

This paper was recommended for publication in revised form by Regional Editor Ahmet Selim Dalkilic

INTRODUCTION

Audio Source Recording (ASR) devices are important in the domain of digital forensics, and they serve as a basis for numerous investigative methods. To establish the source of audio records in court proceedings and to identify cases of interference or violation, the possibility of tracing the recordings with certain devices accurately is essential [1]. Although they are essential, current methods of recognizing ASR devices are often limited by a lack of information use, leading to worse accuracy in real-world scenarios [2].

The purpose of this work is to solve these problems by proposing a novel solution relying on learning feature representations. The methodology tries to address the shortcomings of existing methods that have limited their utility by exploiting the natural characteristics of audio signals and the modern machine-learning methods. It is based on the strength of the time-audio characteristics and mainly on Constant-Q Transform (CQT) and Mel Frequency Cepstral Coefficients (MFCC) [3], which distinguishes the audio data in detail throughout the temporal domain. The feature extraction technique can be granulized by adding time segmentation to enable a more nuanced analysis of audio data. Developing on these basic aspects, it suggests combining the Recurrent Neural Network (RNN) and the Long Short-Term Memory Network (LSTM) to form a structured representation learning model. This hybrid architecture exploits the properties of both RNNs and LSTMs, leveraging their abilities to model time to effectively store and analyze the temporal structure of audio signals. The ability to reduce spatial data with the help of time-frequency representation in order to more accurately detect devices is one of the main characteristics of the methodology. The methodology, with its ability to introduce temporal variations and frequency-related characteristics, breaks the limitations of the currently available feature extraction schemes, offering an in-depth toolset to study audio forensics.

Key Contributions of the work

- A comparative analysis of MFCC-RNN and CQT-RNN models for mobile device-based audio source recognition, highlighting the strengths of each spectral feature.
- Development and evaluation of a lightweight MFCC-RNN model that achieves a high classification accuracy of 96.49%, suitable for real-time forensic applications.
- Integration of a user-friendly Gradio interface to demonstrate the model's real-world usability for non-technical users in forensic investigations.

The rest of the article is as follows: Section II is about a survey of the state-of-the-art works. Section III proposes an ASR device identification method based on feature extraction and training of RNN. Section IV and Section V discuss results and mention conclusions, respectively.

LITERATURE SURVEY

Intricacies of the experimental framework, findings and implications of state-of-the-art works are discussed in this section. Through this detailed examination, seek to demonstrate the approach's transformative potential and implications for the field of digital audio forensics.

Multimedia forensics involves determining the source device of a signal. Cemal Hanilçi et al. [4] investigated the identification of mobile phone brands and models using speech recordings. By extracting mel-frequency cepstral coefficients (MFCCs) from the audio, the system captures device-specific characteristics. Classification was performed using vector quantization and support vector machines, achieving identification accuracies of 92.56% and 96.42%, respectively, across 14 different phones. The results demonstrate the effectiveness of speech-based device recognition techniques.

Source cell phone microphone recognition using non-voice segments of DAR is proposed by C.Hanilci and T. Kinnunen [5]. Two datasets, viz. TIMIT and Live records are considered for experimentation using classifiers, viz. SVM, GMM-ML and GMM-MMI. Features, namely MFCC and LFCC, are explored in non-voice segments. It is concluded that extracting features using non-voice portions resulted in higher recognition rates when compared with features from voiced segments. O.Eskidere explored LPCC, PLPC and MFCC to obtain features, and a Gaussian mixture model was utilized to determine source microphones [6]. This scheme is evaluated on 16 different sources for speaker-dependent and speaker-independent cases. The authors claimed that LPCC features have provided the highest recognition rate. Piczak examines the application of the use of CNNs for environmental sound classification [7]. Their study reveals that CNNs can be used to automatically categorize ambient noises such as animal calls, machinery sound, and urban sounds. With the hierarchical feature learning features of CNNs, the proposed method has a high degree of classification accuracy, establishing the foundation of the future audio analysis systems in environmental monitoring and surveillance. Salamon and Bello examine how deep CNNs and data augmentation methods are used to categorize environmental sounds [8]. Their results emphasize the power of CNNs to learn discriminative features based on the spectrogram representations of ASR. The suggested method improves the robustness of the classification model, resulting in better performance of the model on real-life audio classification tasks.

Gemmeke et al. presented the ontology-based and human-tagged collection of audio events called AudioSet [9]. Their research aims to contribute to the development of audio event detection systems by providing a large-scale dataset annotated with distinct audio events. AudioSet has proven to be a useful resource for academics working on audio event detection, enabling them to train and evaluate machine learning models on real-world audio datasets.

Authors S.Q.Z.Huang and Y.L.S.Shi [10] proposed a deep learning approach with noise as an intrinsic feature to identify the source device of DAR. The authors compared Softmax, Multilayer perceptron and Convolutional neural network classifiers and compared parameters in one classifier. G.Baldini & I.Amerini [11] presented a detection and authentication method by exciting smartphone microphones with non-speech portions at different frequencies. A broad database of 32 smartphones was utilized to assess the performance of this method. Authors reported that the CNN employed provided significant identification and authentication accuracy in different operational scenarios and the presence of Gaussian noise, babble and street noise. A copy-move forgery detection of speech recording based on the correlation between pitch and formant has been proposed [12]. Here, speech is divided into voiced and non-voiced speech segments, then pitch and formant sequences are extracted as features for voiced segments. Similarities between formant sequences as well as pitch sequences are calculated with the help of a dynamic time-warping algorithm. Authors used WSJ and TIMIT speech databases to evaluate the performance of post-processing operations, viz., white Gaussian noise, pink noise, median filter, MP3 compression, etc., and compared with the state-of-the-art works.

Gianmarco Baldini et al. [13] present a method to identify mobile phones by analyzing unique patterns in their built-in microphones using convolutional neural networks (CNNs). Microphone signals from 34 different devices were collected to train and test the model. The study found that CNNs outperformed traditional classifiers such as K-nearest neighbor and support vector machines, even when noise was added. It underscores the future possibilities of microphone-based device fingerprinting in security and digital forensics.

The performance of Gaussian Supervector (GSV) features in microphone-recognition focuses on the impact of Universal Background Model (UBM) parameters [14]. The raw GSV, containing both microphone and speech information, can be noisy. To improve GSV quality, the authors proposed a kernel-based projection method that maps the raw GSV to a new feature space, aiming to separate microphone and speech information. Experimental results show that the projected GSV consistently outperforms the raw GSV using linear Support Vector Machine (SVM) and Sparse Representation-based Classifier (SRC), demonstrating the effectiveness of the projection method. Diego Renza, Jaisson Vargas and Dora M. Ballesteros [15] proposed a scheme to identify manipulations in the audio signal through MFCC, PCA and RSA. Original and manipulated audio hash is compared with the help of a BER threshold value and measured the integrity of the content.

A deep neural network-based real-time sound source localization (SSL) model designed for low-power IoT devices using microphone arrays [16]. The SSL model processes multi-channel acoustic data through parallel convolutional neural network layers to capture unique delay patterns across frequency ranges, estimating both fine and coarse voice locations. The model achieved 91.41% accuracy in fine location estimation and a 7.43° direction of arrival error on noisy data, with a processing time of 7.811 ms per 40 ms samples on a Raspberry Pi 4B. The model is suitable for camera-based humanoid robots to enhance voice interaction in crowded environments. ASR devices, along with the environment in which audio is recorded, are identified [17]. In this work, the authors automatically extracted the environment and microphone features using a Convolutional Neural Network and Long-Short Term Memory from the speech signal. The authors conducted experiments and calculated classification accuracy by considering only voiced segments, only unvoiced segments and combined segments. In this investigation, they concluded that using unvoiced segments got good accuracy results compared to voiced segments. In this experimentation, speech datasets that are recorded in three environments and four recording devices are used with different audio quality levels. ASR device identification based on the speech recordings is presented in [18]. Three fingerprints are extracted, viz., channel response, cuccovillo, band energy difference from speech recordings and classified in this work. The performance of the three features is studied individually by observing the classification accuracy.

Vishal A. Hadoltikar et al. [19] investigated how audio format conversion affects mobile phone source identification using recorded speech signals. The study utilizes Mel Frequency Cepstral Coefficients (MFCC) and their derivatives as distinguishing features and models microphone characteristics through Gaussian Mixture Models (GMM). Experiments are conducted on the publicly available MOBIPHONE dataset, with recordings converted into eight different audio formats, including both lossy and lossless types. The results show that the accuracy of identification is consistent among the different formats showing that MFCC-based features are strong in this task. This work deals with one of the aspects of audio forensics that have not been thoroughly studied before.

Chunyan Zeng et al. [20] introduced a higher-order audio recording device recognition technique based on a new feature representation, named the Sequential Gaussian Mean Matrix (SGMM), which was based on segmented acoustic data. The method combines Convolutional Neural Networks (CNN) and Bidirectional Long Short-term Memory (BiLSTM) networks to learn and represent spatial and temporal features. The system has a high recognition rate of 98.78 with the help of this hybrid architecture. The technique is much better than current methods and it has better performance in digital forensic applications like device identification and tamper detection. Chunyan Zeng et al. [21] proposed a deep learning system to identify audio sources devices with a Squeeze-and-Excitation (SE) self-attention mechanism to enhance the performance of the recognition. The model achieves approximately 1.5%

performance improvement compared to regular CNNs and is tested on two publicly available datasets. The incorporation of SE into both residual and traditional convolutional networks demonstrates how network structure affects the effectiveness of attention. Also, transfer learning is used to improve the recognition of 141 devices, which provides 4-7 percent of improvement in several measures. The findings underscore the possibilities of self-attention and transfer learning in overcoming the limitations in data and improving performance in digital audio forensics.

The study of multimedia forensics has touched on different methods of recognizing audio source recording devices and using founrier coefficients, MFCC, LFCC, and non-voice segments to better recognition. They include machine learning classifiers like SVM, GMM, and CNNs, which prove to be accurate in a variety of situations. Recent developments focus on deep learning solutions, such as deep clustering and real-time sound source localization, improving the soundness and versatility of audio forensic analysis. Challenges remain in extracting inherent device characteristics and developing effective recognition models to improve device identification accuracy.

The two main challenges in the identification of recording devices are:

- Determining the inherent qualities of the source devices and extracting expressive information from recordings.
- Creating effective recognition models is the second step, and it will significantly increase the accuracy of recording device recognition.

The proposed methods, ASR-CQT-RNN and ASR-MFCC-RNN, will provide appropriate solutions for these problems.

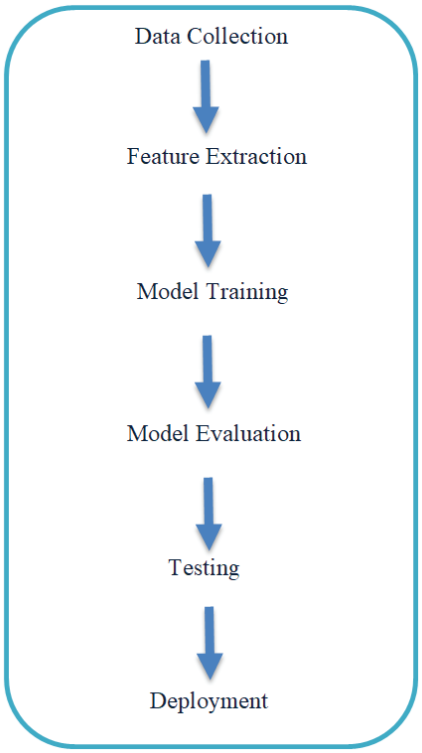


Figure 1. Generic Flow.

PROPOSED SYSTEM

The proposed system for identifying audio source recording devices involves creating a diverse dataset from various devices and extracting features using Mel Frequency Cepstral Coefficients (MFCC) and Constant-Q Transform (CQT). These features are then processed using a structured representation learning model that combines Long Short-Term Memory Networks (LSTM) with Recurrent Neural Networks (RNN) to enhance recognition accuracy. A user interface was developed to facilitate real-world applications, enabling efficient and accurate identification of audio source devices in forensic investigations. The performance of the system is confirmed by comparison with the state-of-the-art procedures that guarantee the robustness and reliability of the system.

The generic flow of the proposed work is shown in Figure 1 which involves six steps: Data Collection, Feature Extraction, Model Training, Evaluation, Testing and Deployment.

1) Data collection

Create a dataset of ASRs from various source devices, guaranteeing diversity in terms of device, recording settings, and content. The expression is shown in Eq (1).

$$S_i^n = \begin{bmatrix} a_1^1 & a_2^1 & \cdots a_i^1 \\ a_1^2 & a_2^2 & \cdots a_i^2 \\ \vdots & \vdots & \vdots \\ a_1^n & a_2^n & \cdots a_i^n \end{bmatrix} \quad (1)$$

Where i indicates the number of samples

n indicates the number of ASR devices based on the dataset

2) Feature Extraction

Extract useful features from preprocessed audio data. Popular time-frequency representations are MFCC, CQT, and others. The feature extraction step transforms raw audio data into a machine-learning algorithm-readable format but preserves important audio signal properties. Compare two salient feature extraction methods in the proposed system of enhanced source recording device recognition i.e., CQT and MFCC.

a) CQT

CQT is a time-frequency analysis method to extract features. The CQT is an alternative to the more widely used Fourier Transform that divides the signal into frequency bins separated logarithmically. This is to imply that the bins

are placed in a way that the number of bins in each octave is the same. This is helpful as it can simulate a logarithmic perception of frequency by the human auditory system.

CQT resembles the Fourier Transform, which transforms a signal in time to frequency. CQT provides a more human-friendly frequency representation, since it uses a logarithmic frequency bin spacing. CQT finds a lot of applications in audio signal processing, e.g. in music analysis, audio synthesis and in audio feature extraction used in audio classification and audio identification. It offers a frequency representation that better reflects the tonal properties of audio signals and is well adapted to applications that need these properties. Figure 2 depicts the CQT feature extraction method.

$$C_i^n = CQT(S_r^n) \tag{2}$$

$$Features_{CQT}^n = \begin{bmatrix} C_1^1 & C_2^1 & \cdots & C_i^1 \\ C_1^2 & C_2^2 & \cdots & C_i^2 \\ & \vdots & \vdots & \vdots \\ C_1^n & C_2^n & \cdots & C_i^n \end{bmatrix} \tag{3}$$

Where $Features_{CQT}^n$ are features of CQT from different ASR models and samples.

The CQT feature extraction method is provided with the audio samples in equation (2) to obtain acoustic characteristics. They are represented in equation (3).

b) Mel-Frequency Cepstral Coefficient (MFCC)

The most popular feature extraction method in speech and audio processing is called MFCC. MFCC describes the spectrum qualities of sound in a form that is suitable for a range of machine learning applications, such as music analysis and speech recognition [22]. A group of coefficients that represent the sound source's power spectrum form is known as an MFCC. The Mel-Scale is used to simulate how the human ear hears sound frequency after the raw audio input has been transformed into a frequency domain using a technique similar to the Discrete Fourier Transform (DFT) [23].

Lastly, the mel-scaled spectrum is used to compute the cepstral coefficients. Because they eliminate unnecessary information while highlighting audio signal elements essential to human speech perception, MFCCs are very advantageous. They can therefore be used for tasks including speech-to-text conversion, emotion detection, and speaker recognition. The MFCC feature extraction method is shown in Figure 3.

$$M_i^n = MFCC(S_i^n) \tag{4}$$

$$Features_{MFCC}^n = \begin{bmatrix} M_1^1 & M_2^1 & \cdots & M_i^1 \\ M_1^2 & M_2^2 & \cdots & M_i^2 \\ & \vdots & \vdots & \vdots \\ M_1^n & M_2^n & \cdots & M_i^n \end{bmatrix} \tag{5}$$

Where $Features_{MFCC}^n$ is MFCC features for different ASR devices and samples.

The threshold value is calculated by comparing the same device sample features.

3) Model Training

In speech and audio processing, characteristics are extracted using MFCC or CQT. It highlights human auditory awareness by capturing the spectral characteristics of sound. By effectively encoding audio data, MFCC enables machine learning algorithms to categorize, identify, and examine auditory patterns. They are extensively utilized in

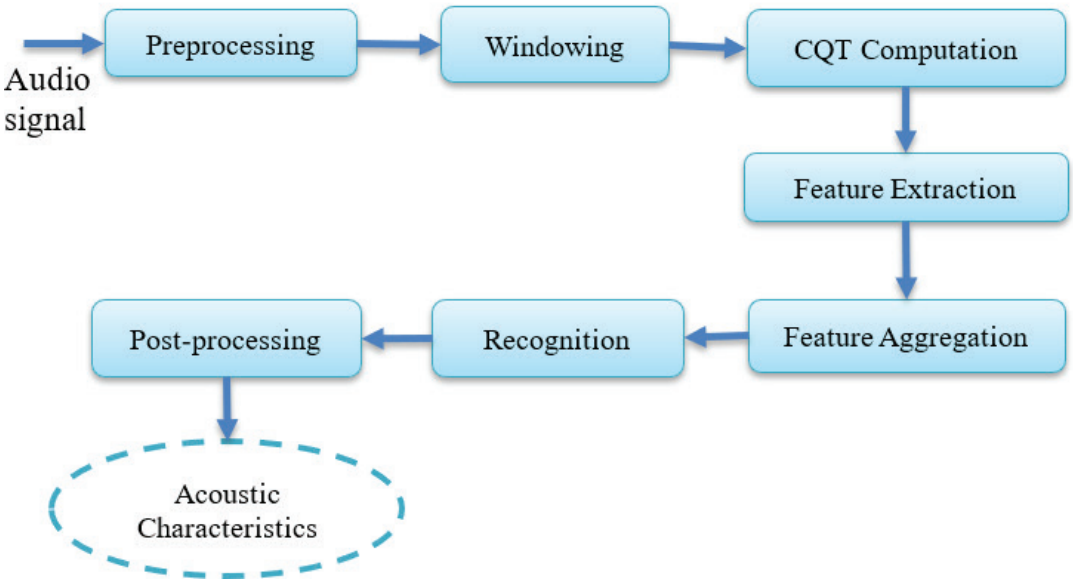


Figure 2. CQT Feature Extraction method.

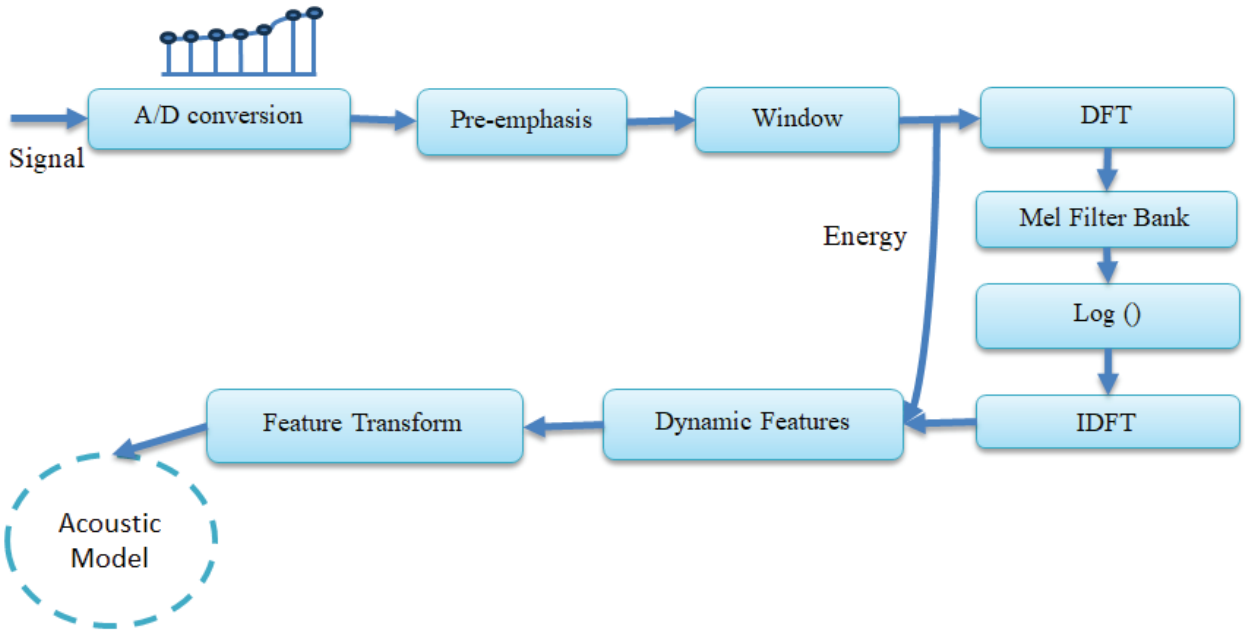


Figure 3. MFCC feature extraction method.

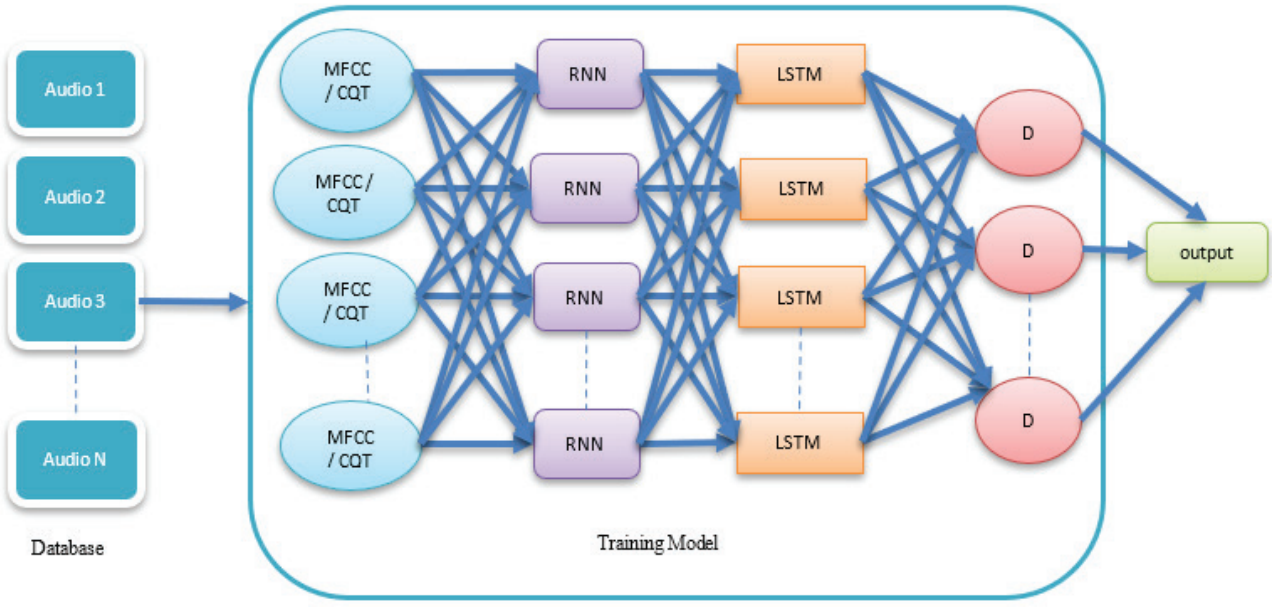


Figure 4. Model Training.

audio recognition, speech identification, and the analysis of music. The block diagram of the model training is shown in Figure 4.

a) RNN

A deep learning model designed to analyze and convert sequential data inputs into specific sequential data outputs is called an RNN. RNNs are composed of neurons, which are nodes for processing data that work together to do complex tasks. There are three levels of neurons: input, output, and hidden, shown in Figure 5 [24]. While the output layer generates the final product, the input layer receives the data to process. Prediction, analysis, and data processing happen at the hidden layer.

The information is moving in a single direction through the feed-forward neural network (FNN): through hidden

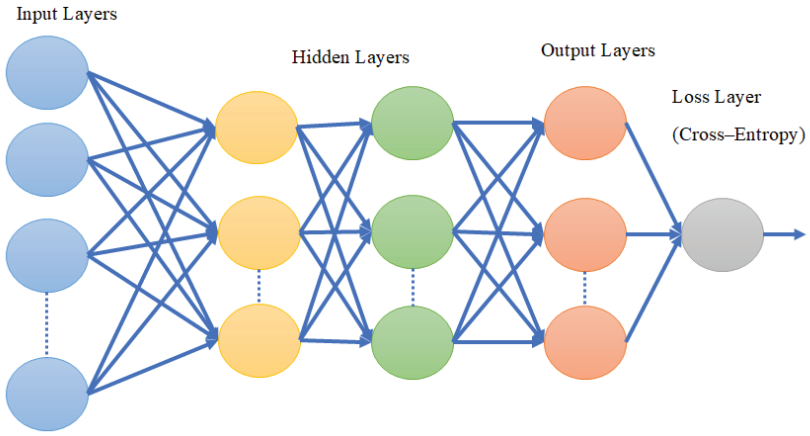


Figure 5. RNN Architecture.

layers, from the input layer to the output layer. There is a direct flow of information throughout the network. The prediction of the future is not good for FNNs because they can't remember what has been communicated to them. Since the FNN is only looking at existing inputs, it does not have a deadline. Apart from its instruction, it is incapable of remembering anything from the past. Information is looped across an RNN. It takes into consideration the information it acquires from prior inputs in addition to that of the current input when deciding.

b) LSTM

RNNs are extended by LSTM networks. The RNN's layers are constructed from LSTMs. Three gates are present in a long short-term memory cell: input, forget, and output [25]. These gates control whether to allow data to enter the system (input gate), discard data that isn't needed (forget gate), or allow data to affect the output at the current time step (output gate).

$$B_m^{CQT} = RNN(LSTM(Features_{CQT}^n)) \text{ using CQT} \quad (6)$$

$$B_m^{MFCC} = RNN(LSTM(Features_{MFCC}^n)) \text{ using MFCC} \quad (7)$$

4) Model Evaluation

Assess the trained model's performance on the test set using suitable evaluation metrics, such as accuracy, precision, recall and F1-score using formulas given below [26]. Analyze any misclassifications to identify areas for improvement. Higher values for precision, recall, and F1 score indicate better performance, with the best score possible shown in Table 1 for CQT and Table 2 for MFCC.

$$Accuracy = \frac{True_P + True_N}{True_P + True_N + False_P + False_N} \quad (8)$$

$$Precision = \frac{True_P}{True_P + False_P} \quad (9)$$

$$Recall = \frac{True_P}{True_P + False_N} \quad (10)$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (11)$$

5) Model Testing

The new audio sample is given to the trained model, then it predicts the output. It recognizes the ASR device. The block diagram of the model for testing is shown in Figure 6.

For CQT

$$PD_{CQT} = B_m^{CQT}(a_t) \quad (12)$$

where a_t is test sample

$$Features_{a_t}^{CQT} = \begin{bmatrix} C_{t_1}^1 & C_{t_2}^1 & \cdots & C_{t_i}^1 \\ C_{t_1}^2 & C_{t_2}^2 & \cdots & C_{t_i}^2 \\ \vdots & \vdots & \vdots & \\ C_{t_1}^n & C_{t_2}^n & \cdots & C_{t_i}^n \end{bmatrix} \text{using CQT} \quad (13)$$

$$if\ Features_{a_t}^{CQT} \geq Th_{CQT} : \text{Predicts the model} \quad (14)$$

Where Th_{CQT} is a threshold value for CQT

For MFCC

MFCC model testing is shown in the following equations.

$$PD_{MFCC} = B_m^{MFCC}(a_t) \quad (15)$$

where a_t is test sample

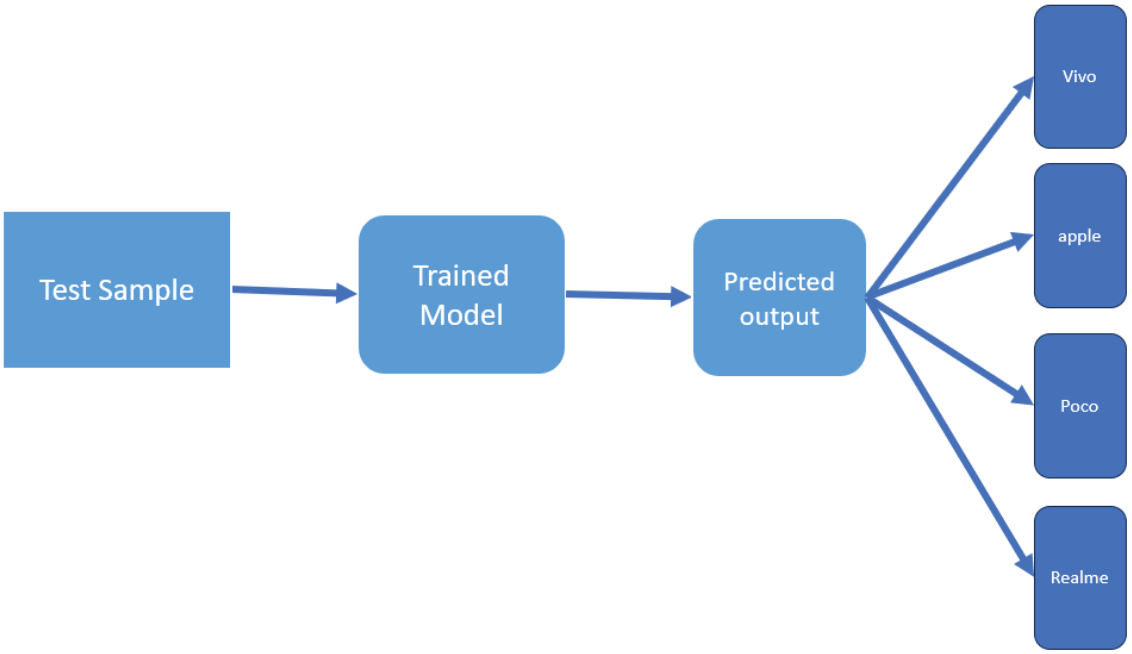


Figure 6. Model testing.

$$Features_{a_t}^{MFCC} = \begin{bmatrix} M_{t_1}^{1} & M_{t_2}^{1} & \cdots & M_{t_i}^{1} \\ M_{t_1}^{2} & M_{t_2}^{2} & \cdots & M_{t_i}^{2} \\ \vdots & \vdots & \vdots & \\ M_{t_1}^{n} & M_{t_2}^{n} & \cdots & M_{t_i}^{n} \end{bmatrix} \quad (16)$$

$$if\ Features_{a_t}^{MFCC} \geq Th_{MFCC} : \text{Predicts the model} \quad (17)$$

Where Th_{MFCC} is a threshold value for MFCC

6) Deployment

As soon as the model reaches satisfactory performance, implement it in real-life usage. It is now able to categorize the origin of new audio recordings in accordance with their characteristics.

The proposed solution develops the sphere of audio forensic analysis to a new level, taking advantage of the strength of MFCC and temporal modelling methods, which is likely to result in greater effectiveness in various forensic scenarios.

In general, the results indicate that it is important to employ appropriate feature extraction methods in ASR device detection exercises. The outstanding performance of MFCC proves that it can be used to identify source devices accurately, by their unique audio traits, which opens the way to more efficient and trusted forensic analysis in digital evidence.

RESULTS AND DISCUSSION

The main aim of this research is to determine the effectiveness of the suggested methodology through large-scale experimental evaluations. To illustrate the resilience and versatility of the methodology in a wide range of forensic cases using varied datasets, which have recordings of many different source devices. By conducting a lot of tests and validation, to provide empirical evidence of the improved performance of the model over other methodologies.

The proposed ASR-CQT-RNN and ASR-MFCC-RNN are compared and tested on four ASR device datasets. The following are steps that make up the dataset collection process. Audio samples are downloaded using ITU-T test signals from telecommunications systems. It takes around 12 minutes to merge audio samples into a large corpus of data. The recorded audio samples were obtained using different devices, where each device Split 12-minute audio into 70 samples from the given data in an efficient way. The sequence in which the samples were combined from TIMIT was such that the long speech data was separated into 560 short sample segments, each lasting around 10 seconds, based on the quantity and duration of the TIMIT corpus. The ASR devices used in this work are iPhone, Realme, Vivo, and Poco. This also emphasizes the significance of taking into account a variety of aspects, such as recording settings, background noise, and device characteristics, while developing experiments and preprocessing procedures for audio source recognition.

The performance metrics for ASR-CQT-RNN and ASR-MFCC-RNN methods are shown in Table 1 and Table 2. Furthermore, confusion matrix for both approaches were supplied to help visualize their performance in identifying audio sources, shown in Figures 7 & 8. Table 1 and Table 2 report accuracy rates of 79.94% for ASR-CQT-RNN and

Table 1. Performance Metrics for ASR-CQT-RNN

	Precision	Recall	F1 – score	support
Vivo	0.86	1.00	0.92	12
Poco	0.72	0.87	0.79	15
Apple	1.00	0.47	0.64	15
Realme	0.72	0.87	0.79	15
Accuracy	-	-	0.79	57
Macro avg	0.83	0.80	0.78	57
Weighted avg	0.82	0.79	0.78	57

96.49% for ASR-MFCC-RNN, showing that ASR-MFCC-RNN surpasses ASR-CQT-RNN in this situation. The testing results showed that ASR-MFCC-RNN regularly surpassed ASR-CQT-RNN in terms of accuracy, confirming its superiority as a feature extraction method for recording device recognition.

Figure 7 shows a confusion matrix of the ASR-CQT-RNN. It is a table indicating the performance of a classification model when it is applied to a set of the test data whose true values are known. The data in the matrix are as shown below:

Rows are the actual classes (True labels): Vivo, Poco, Apple, and Realme. The columns show the predicted classes (predicted labels) of the model; they are Vivo, Poco, Apple, and Realme. The number of correct and incorrect predictions of the model is shown in the matrix, with the diagonal showing correct predictions. These are the specific counts:

In Figure 7, Vivo had 12 correct guesses; 0 incorrect, Poco had 13 correct predictions, 2 wrong (both predicted Realme), and Apple had seven right predictions, 1 wrong prediction for Vivo, three for Realme, and four for Poco. Realme had 13 right guesses, one bad prediction as Vivo, and one as Poco.

For example, Vivo and Poco both exhibit a high number of correct predictions (dark blue), with 12 and 13, respectively, whereas Apple has a more spread-out confusion, with 7 correct but multiple misclassifications among other brands.

Figure 8 shows a confusion matrix of ASR-MFCC-RNN, in which the dark navy blue color represents true positives

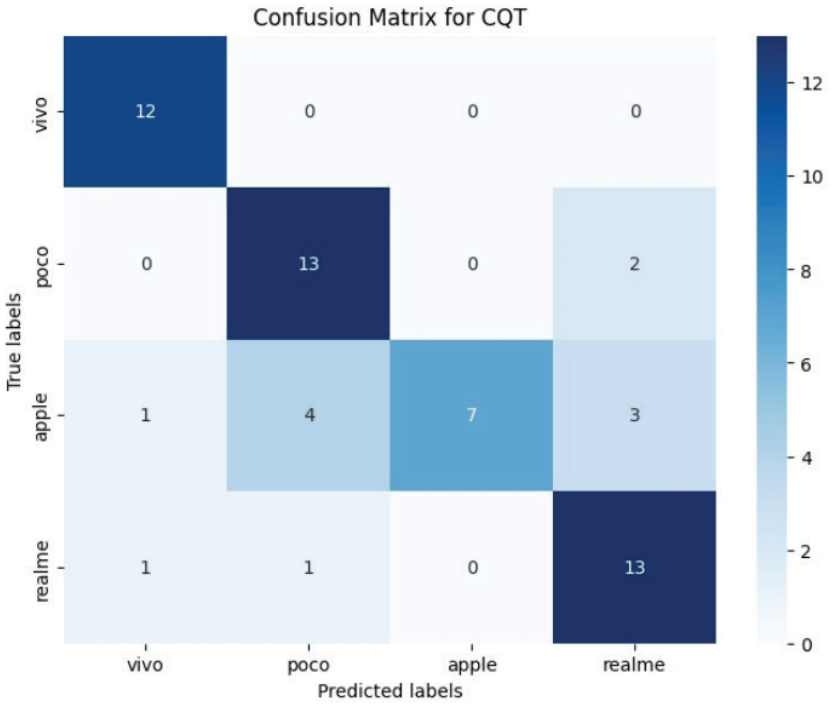


Figure 7. Confusion Matrix of ASR-CQT-RNN.

Table 2. Performance metrics of ASR-MFCC-RNN

	Precision	Recall	F1 - score	support
Vivo	1.00	1.00	1.00	12
Poco	0.88	1.00	0.94	15
Apple	1.00	0.87	0.93	15
Realme	1.00	1.00	1.00	15
accuracy			0.96	57
Macro avg	0.97	0.97	0.97	57
Weighted avg	0.97	0.96	0.96	57

and the light color represents true negatives. False predictions are not possible. For Vivo 12, correct guesses and 0 incorrect, Poco had 15 correct predictions, Apple had 13 right predictions and two wrong predictions as Poco, and Realme had 15 right predictions.

CQT, while commonly used in audio analysis, did not consistently perform well in tests. On the other hand, MFCC evolved as a more reliable and accurate feature representation method for recognizing audio source recording devices. MFCC, which is derived from the human auditory system's frequency resolution properties, collects critical information about the spectrum envelope of audio sources. This format is effective in capturing the unique characteristics of the different recording devices and hence they can be easily identified. Moreover, using MFCC together with a structured representation learning model, which consists of RNNs and LSTM, enhances device recognition accuracy.

Figure 9 displays training and validation accuracy per epoch of MFCC and CQT strategies of feature extraction. The X-axis gives the number of training epochs or iterations through the data set. Each epoch is a single repeat of the entire training set. The Y-axis displays the accuracy of the model with the training and validation datasets. Accuracy is often measured as the percentage of correctly identified examples of all the instances. The training accuracy curve shows the accuracy of the model changing during the epochs on the training dataset. The validation accuracy curve shows the accuracy of the model per epoch on the validation data. The intersection of both training and validation curves can be significant. A model can be overfitting when its validation accuracy is often less than

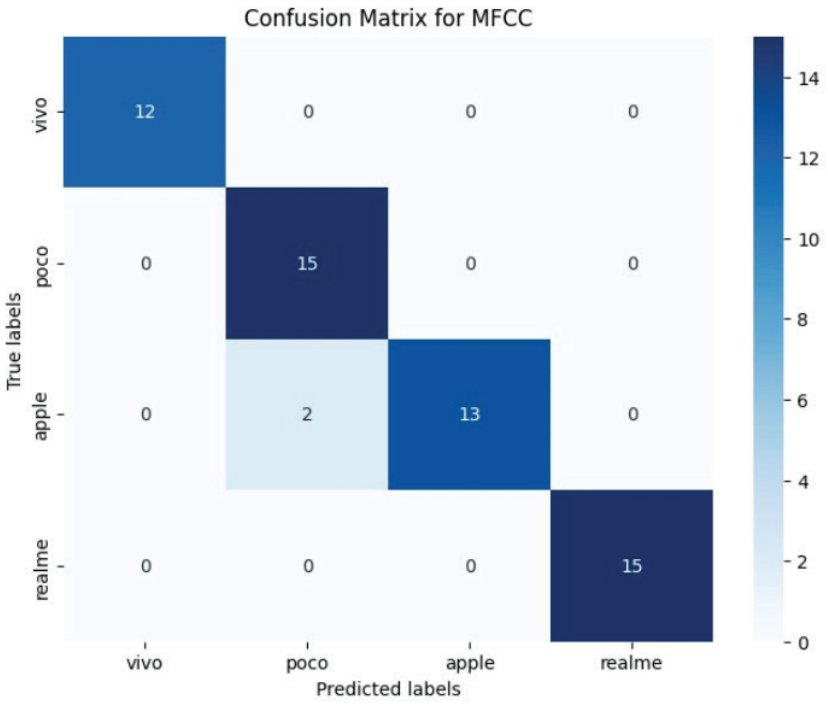


Figure 8. Confusion Matrix of ASR-MFCC-RNN.

Figure 9. Accuracy Plot.

Table 3. Comparison of the proposed work with state-of-the-art work

Previous papers	Accuracy
Yuechi Jiang et.al (2019)[14] *Scheme-1 (SRC)*	88.37%
Yuechi Jiang et.al (2019)[14] *Scheme-2 (SVM)*	89.04%
Jungbeom Ko1 et.al [16]	91.41%
Chunyan Zeng et al. [21]	95.2%
ASR-CQT-RNN (Proposed method-1)	79.94%
ASR-MFCC-RNN (Proposed method-2)	96.49%

its training accuracy. This means that the model is overfitting the training data and not generalizing the new data. Good generalization is demonstrated when the accuracy of the validation is comparable to the training accuracy. The comparison of accuracy rates, the presentation of confusion matrices, and the evaluation of various ASR devices and potential problems all help to provide a full understanding of the experimental results.

The performance of the proposed work in terms of accuracy is compared with state-of-the-art works, which are shown in Table 3. It is observed that the accuracy of this model is more when compared to previous works. This suggests that the combination of MFCC features with RNN and LSTM models was highly effective for the task.

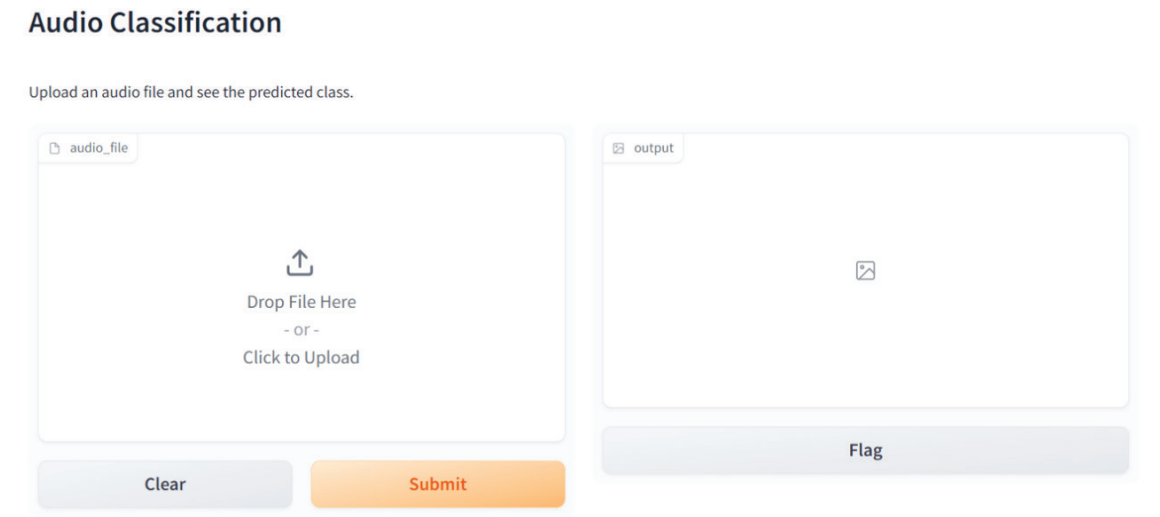
Table 4 presents the computational metrics for the proposed ASR-CQT-RNN and ASR-MFCC-RNN models. It includes training time, inference time, model size, and memory footprint to evaluate real-time applicability. The MFCC-RNN model trains faster and infers slightly quicker, making it more efficient for deployment, while both models remain lightweight and resource-friendly.

This proposed work is also implemented for the user interface, which is shown in Figure 10. The user interface screen is divided into two parts; one part of the screen is

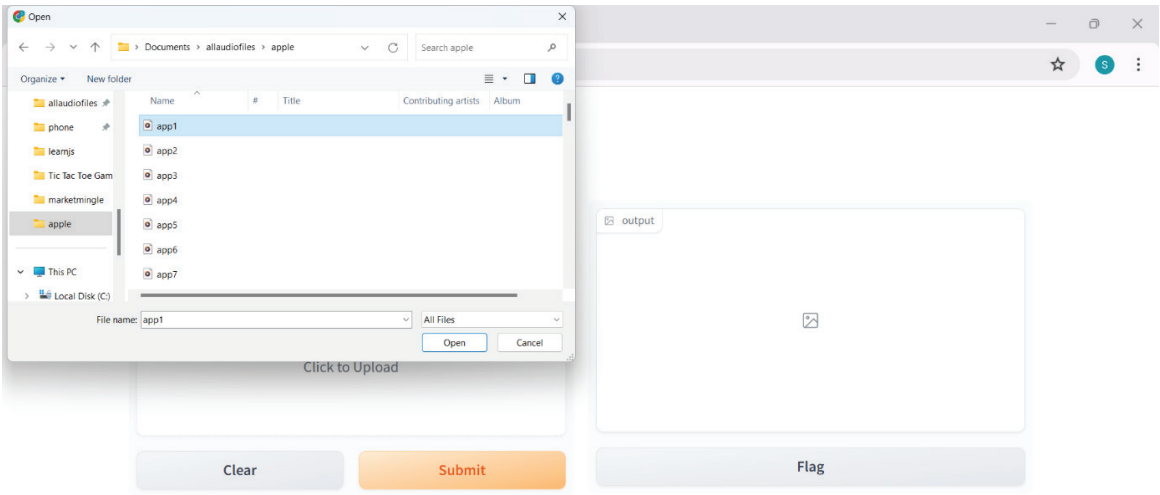
Table 4. Resource usage and computational performance

Metric	ASR-CQT-RNN Model	ASR-MFCC-RNN Model
Training Time	360 seconds	320 seconds
Model Parameters	115,268	115,268
Model Size (.h5 file)	450 KB	450 KB
Memory Footprint	0.44 MB	0.44 MB
Inference Time per Audio Sample	0.06 seconds	0.03 seconds
Batch Size Used	32	32

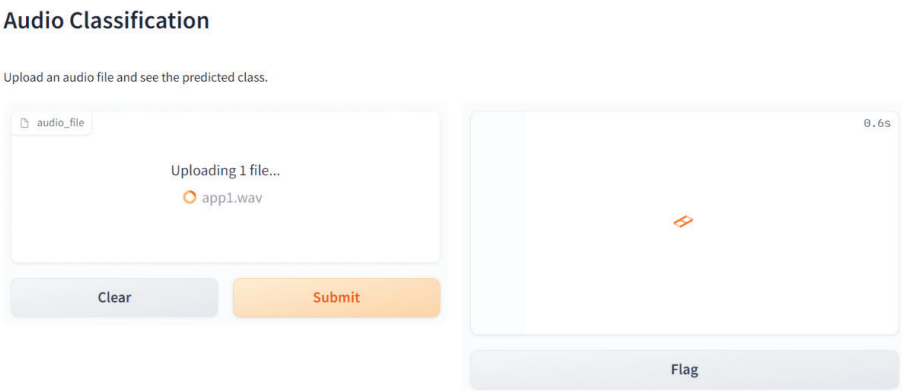
to upload the test audio signal, and the second part of the screen is to get the output ASR device shown in Figure 10(a) and (b). Whenever the test audio sample is uploaded, which is shown in Figure (c) and Figure (d), the proposed code will be run in the background, and it will produce the predicted source device, which is shown in Figure (e).



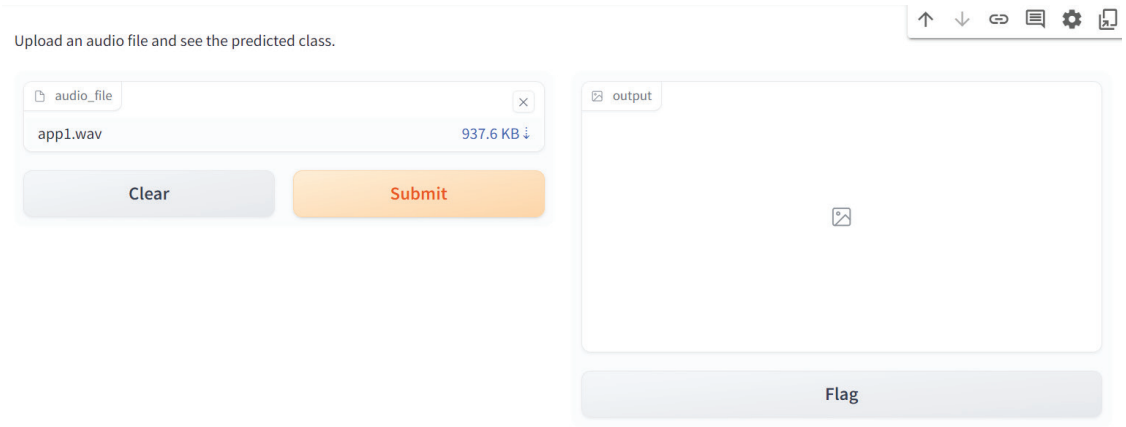
(a)



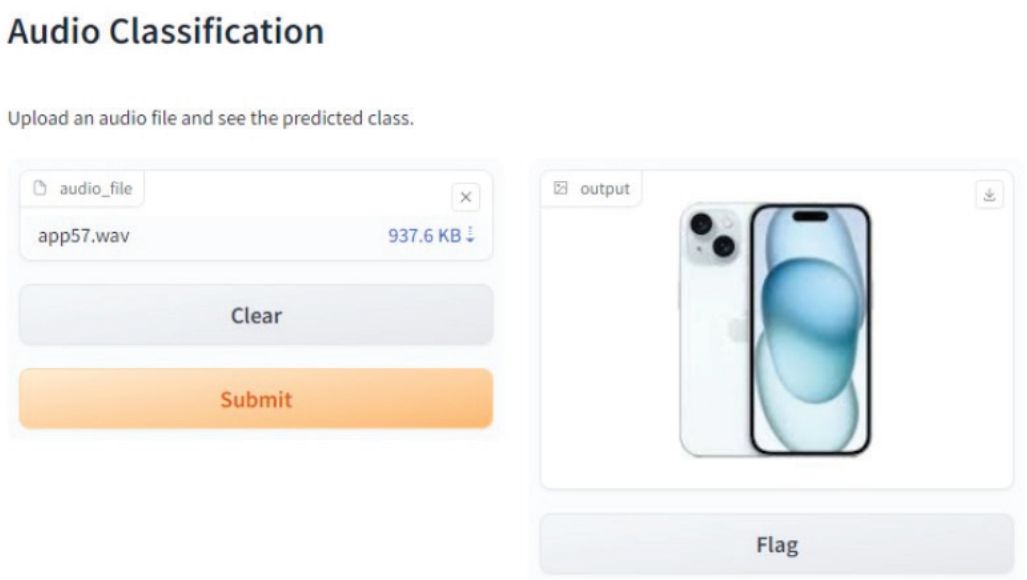
(b)



(c)



(d)



(e)

Figure 10. User Interface.

CONCLUSION

In conclusion, the findings show that integrating MFCC with RNN and LSTM improves ASR device recognition accuracy. While MFCC and CQT were compared, it was discovered that CQT did not consistently perform well on this task. The performance of this proposed work in terms of accuracy is compared with state-of-the-art works and using MFCC with RNN and LSTM resulted in the 96.49% accurate identification of source devices based on distinctive audio characteristics. This work is implemented in the frontend design for a better user interface to identify the ASR device of a given test audio signal.

In future, we intend to improve the resilience of this model by including approaches for dealing with fluctuations in the audio database, such as ambient noise and recording settings. Future research will focus on broadening the applicability of this approach to solve growing audio forensics difficulties, such as detecting deepfake audio recordings and identifying tampered or modified content. We propose extending our model by integrating GAN-detection features or spectral fingerprinting techniques to detect tampered or synthesized audio.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] Wang Z, Zhan J, Zhang G, Ouyang D, Guo H. An end-to-end transfer learning framework of source recording device identification for audio sustainable security. Sustainability 2023;15:1-22. [CrossRef]

[2] Zeng C, Kong S, Wang Z, Li K, Zhao Y. Digital audio tampering detection based on deep temporal-spatial features of electrical network frequency. Information 2023;14:1-22. [CrossRef]

[3] Narla VL, Suresh G, Singh MK, Vinod Kumar M. Speech signal splicing detection system based on MFCC and DTW. Int Res J Multidiscip Technovation 2024;6:170-181. [CrossRef]

[4] Hanilçi C, Ertaş F, Ertas T, Eskidere Ö. Recognition of brand and models of cell-phones from recorded speech signals. IEEE Trans Inf Forensics Secur 2012;7:625-634. [CrossRef]

[5] Hanilçi C, Kinnunen T. Source cell-phone recognition from recorded speech using non-speech segments. Digit Signal Process 2014;35:75-85. [CrossRef]

[6] Eskidere Ö. Source microphone identification from speech recordings based on a Gaussian mixture model. Turk J Electr Eng Comput Sci 2014;22:754-767. [CrossRef]

[7] Piczak KJ. Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP); 2015; Boston, USA. p. 1-6. [CrossRef]

[8] Salamon J, Bello JP. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process Lett 2017;24:279-283. [CrossRef]

[9] Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC, et al. Audio set: An ontology and human-labeled dataset for audio events. In: IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings; 2017. p. 776-780. [CrossRef]

[10] Qi S, Huang Z, Li Y, Shi S. Audio recording device identification based on deep learning. In: 2016 IEEE International Conference on Signal and Image Processing, ICSIP 2016; 2017. p. 426-431. [CrossRef]

[11] Baldini G, Amerini I. Smartphones identification through the built-in microphones with convolutional neural network. IEEE Access 2019;7:158685-158696. [CrossRef]

[12] Yan Q, Yang R, Huang J. Robust copy-move detection of speech recording using similarities of pitch and formant. IEEE Trans Inf Forensics Secur 2019;14:2331-2341. [CrossRef]

[13] Baldini G, Amerini I, Gentile C. Microphone identification using convolutional neural networks. IEEE Sens Lett 2019;3:1-4. [CrossRef]

[14] Jiang Y, Leung FHF. Source microphone recognition aided by a kernel-based projection method. IEEE Trans Inf Forensics Secur 2019;14:2875-2886. [CrossRef]

[15] Renza D, Vargas J, Ballesteros DM. Robust speech hashing for digital audio forensics. Appl Sci 2020;10:249. [CrossRef]

[16] Ko J, Kim H, Kim J. Real-time sound source localization for low-power IoT devices based on multi-stream CNN. Sensors 2022;22:4650. [CrossRef]

[17] Qamhan MA, Altaheri H, Meftah AH, Muhammad G, Alotaibi YA. Digital audio forensics: Microphone and environment classification using deep learning. IEEE Access 2021;9:62719-62733. [CrossRef]

[18] Giganti A, Cuccovillo L, Bestagini P, Aichroth P, Tubaro S. Speaker-independent microphone identification in noisy conditions. arXiv Preprint arXiv:2206.11640. 2022. [CrossRef]

[19] Hadoltikar VA, Ratnaparkhe V, Kumar R. Effect of format conversion on source identification from audio recordings: A study for forensic purposes. SN Comput Sci 2024;5:19. [CrossRef]

[20] Zeng C, Feng S, Wang Z, Zhao Y, Li K, Wan X. Audio source recording device recognition based on representation learning of sequential Gaussian mean matrix. Forensic Sci Int Digit Investig 2024;48:301676. [CrossRef]

[21] Zeng C, Zhao Y, Wang Z, Li K, Wan X, Liu M. Squeeze-and-excitation self-attention mechanism enhanced digital audio source recognition based on transfer learning. Circuits Syst Signal Process 2025;44:480-512. [CrossRef]

[22] Singh MK. Multimedia application for forensic automatic speaker recognition from disguised voices using MFCC feature extraction and classification techniques. Multimed Tools Appl 2024;83:1-19. [CrossRef]

[23] Mohammed Muntaz O, Osman B. Parabolic filter Mel frequency cepstral coefficient and fusion of features for speaker age classification. Sigma J Eng Nat Sci 2020;38:2177-2191.

[24] Singh MK. A text independent speaker identification system using ANN, RNN, and CNN classification technique. Multimed Tools Appl 2024;83:48105-48117. [CrossRef]

[25] Oykum Esra Y, Selcuk A, Ersoy O. Prediction of BIST price indices: A comparative study between traditional and deep learning methods. Sigma J Eng Nat Sci 2020;38:1693-1704.

[26] Sahu AK, Hassaballah M, Rao RS, Suresh G. Logistic-map based fragile image watermarking scheme for tamper detection and localization. Multimed Tools Appl 2023;82:24069-24100. [CrossRef]