computers

MDPI

Article

Audio Deep Fake Detection with Sonic Sleuth Model

Anfal Alshehri [1,†], Danah Almalki [1,†], Eaman Alharbi [1,2,*] and Somayah Albaradei [1,2,*]

[1] Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah P.O. Box 80221, Saudi Arabia; alshehrianfal@gmail.com (A.A.); danaalmalkexvii@gmail.com (D.A.)
[2] Center of Research Excellence in Artificial Intelligence and Data Science, King Abdulaziz University, Jeddah, P.O. Box 80221, Saudi Arabia
* Correspondence: ealraddadi@kau.edu.sa (E.A.); salbaradei@kau.edu.sa (S.A.)
† These authors contributed equally to this work.

Abstract: Information dissemination and preservation are crucial for societal progress, especially in the technological age. While technology fosters knowledge sharing, it also risks spreading misinformation. Audio deepfakes—convincingly fabricated audio created using artificial intelligence (AI)—exacerbate this issue. We present Sonic Sleuth, a novel AI model designed specifically for detecting audio deepfakes. Our approach utilizes advanced deep learning (DL) techniques, including a custom CNN model, to enhance detection accuracy in audio misinformation, with practical applications in journalism and social media. Through meticulous data preprocessing and rigorous experimentation, we achieved a remarkable 98.27% accuracy and a 0.016 equal error rate (EER) on a substantial dataset of real and synthetic audio. Additionally, Sonic Sleuth demonstrated 84.92% accuracy and a 0.085 EER on an external dataset. The **novelty** of this research lies in its integration of datasets that closely simulate real-world conditions, including noise and linguistic diversity, enabling the model to generalize across a wide array of audio inputs. These results underscore Sonic Sleuth's potential as a powerful tool for combating misinformation and enhancing integrity in digital communications.

Keywords: artificial intelligence; deepfake detection; machine learning; deep learning; deepfake audio; audio forensics

check for updates

Citation: Alshehri, A.; Almalki, D.; Alharbi, E.; Albaradei, S. Audio Deep Fake Detection with Sonic Sleuth Model. *Computers* **2024**, *13*, 256. https://doi.org/10.3390/computers13100256

Academic Editors: Aditya Kumar Sahu, Amine Khaldi and Jatindra Kumar Dash

Received: 5 September 2024
Revised: 26 September 2024
Accepted: 1 October 2024
Published: 8 October 2024

Copyright: © 2024 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The cybersecurity landscape is constantly evolving, and a novel and highly deceptive threat has emerged: deepfakes. Unlike traditional cyberattacks that often necessitate significant technical prowess, deepfakes leverage the power of artificial intelligence (AI) to create hyper-realistic and convincing fabricated audio, video, or text content. This intrinsic characteristic makes them particularly perilous, as they can effortlessly deceive unsuspecting victims, potentially leading to devastating consequences.

The threat posed by deepfakes is significant and multifaceted, impacting individuals, organizations, and societies. Developing robust deepfake detection methods is crucial for mitigating these risks. By harnessing the power of deep learning, we can create effective tools to identify and combat deepfakes, safeguarding the integrity of digital content and maintaining public trust.

The use of generative AI enables users to produce new content quickly based on a variety of inputs. These models can be fed with text, images, sounds, animations, 3D models, or other types of data. The process is accomplished by using machine learning models, particularly generative models, which are capable of learning and replicating patterns from large datasets [1]. Simultaneously, the accessibility of audio deepfake or audio manipulation technology has increased. This technology enables individuals to generate audio content where the speaker utters statements they never actually said.

Moreover, it allows synchronization of muted videos with the correct voice. Additionally, this technology can even decipher the speaker's lip movements without the need to hear the speech.

In this study, we present Sonic Sleuth, a novel deep learning model designed specifically for detecting audio deepfakes, focusing on human speech. Our model uses a custom convolutional neural network (CNN) architecture, which significantly improves detection accuracy, outperforming traditional methods. Evaluated on both primary and external datasets, Sonic Sleuth achieves 98.27% accuracy and a 0.016 equal error rate (EER) on the primary dataset. Its generalization capabilities across diverse audio manipulation techniques demonstrate its potential for real-world applications in journalism, social media monitoring, and information verification. This research contributes to broader efforts in cybersecurity, enhancing information authenticity and mitigating the risks associated with deepfake technologies.

2. Background

2.1. Deepfake Generation

Deepfake, a blend of the terms "deep learning" and "fake", refers to images, videos, or audio that have been manipulated or created using artificial intelligence. These media can portray both real and fictional individuals and are categorized as a form of synthetic media [2]. Generating a deepfake is accomplished through a generative adversarial network (GAN), which consists of two parts, convolutional neural networks (CNNs) and deconvolutional neural networks (DNNs), as shown in Figure 1. They aim to produce data that look just like real data. The CNN is the generator model, while the DNN is the discriminator model. The generator is trained to produce data, and the discriminator is trained to distinguish between data generated by the generator and real data. When the discriminator successfully distinguishes between real and generated data, the generator generates an enhanced version of the data until the discriminator cannot differentiate between the generated and real data.

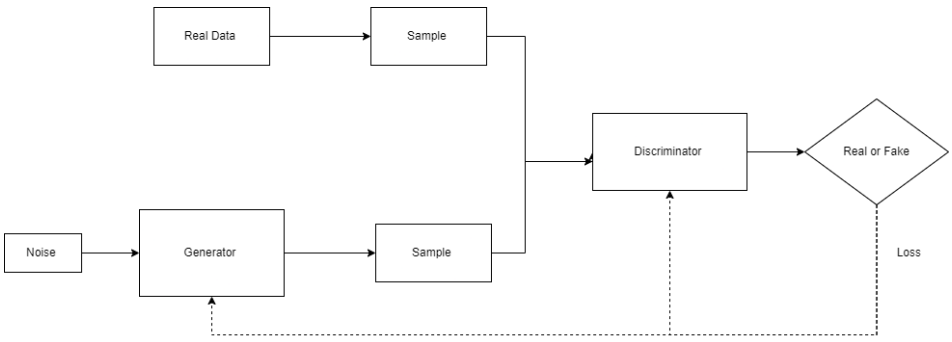

Figure 1. Generative AI structure [3].

2.2. Time Domain and Frequency Domain

The time domain represents a sound wave using a function in time where a graph of such a domain shows the change in the signal value (amplitude) over time [4]. The time domain gives us information about the wave frequency, wavelength, and amplitude. Various calculations can be made to analyze and understand a signal, including time-domain filtering, feature extraction, and signal visualization [4]. Another representation of a signal is the frequency domain, where instead of showing amplitude change over time, it shows how much of a signal lies in a specific frequency over a range of frequencies [5]. Two types of frequency-domain graphs are the spectrum and the spectrogram. A spectrum captures the signal's amplitude vs. frequency distribution at a specific point in time, while a spectrogram shows the same information across a period of time; it plots the frequency of a wave over time and uses colors to indicate the amplitude of each frequency: the brighter the color, the higher the amplitude [6].

The Fourier transform, a powerful mathematical tool, is used to convert signals from the time domain to the frequency domain [5]. This transformation breaks down a signal into the different waves of a constant frequency that make up the whole signal. It provides us with the spectrum of an acoustic wave [5]. The spectrum can then be used to obtain a spectrogram by combining the spectra of multiple overlapping audio segments into a single plot of frequency against time [6].

2.3. Literature Review

Research in the field of deepfake audio detection is a continuous effort to defend against the evolving techniques of generative AI. And, just like any AI task, robustness is a main goal of deepfake research, which requires AI to move to the next level of deep understanding [7].

One way to achieve that depends on the training data of AI models. In the case of deepfake audio detection, a common approach to achieve better generalization of detection is to combine multiple datasets, not only to train a model on both fake and real audio but also to train on data of different natures, such as a dataset that was meant for spoofing attacks or in-the-wild datasets that evaluate the model on realistic audio [8].

The dataset WaveFake [8] is one example, where six different architectures and a text-to-speech pipeline were used to generate fake audio, addressing the problem of detection models generalizing to different deepfake audio generation methods.

A similar approach was that of [9], where the "Attack Agnostic Dataset" was proposed to evaluate the generalization of models. The dataset combined two deepfake datasets and one audio spoofing dataset: WaveFake, FakeAVCeleb, and ASVspoof2019(LA subset), respectively. Spoofing attacks have different elements compared to deepfake audio since the aim is to deceive automatic speech verification (ASV) systems instead of humans [9].

Research Gap and Main Contributions

The detection of audio deepfakes presents several significant challenges that remain unresolved. A primary challenge lies in the generalization of detection models, as many current methods struggle to adapt effectively to the wide variations in input data encountered in real-world scenarios. This issue is further compounded by the expensive data processing typically required to prepare datasets for training, which limits the accessibility and scalability of deepfake detection systems [10]. Ensuring model robustness across different environments, including those with background noise and varying acoustic conditions, remains a critical challenge for researchers. Additionally, the existing research has paid limited attention to non-English languages, which restricts the global applicability of current audio deepfake detection systems [11]. Current methods also lack consideration for accent variations, an essential factor in developing a truly comprehensive detection system that can operate reliably across diverse linguistic and cultural contexts. Another important gap lies in the impact of noise during recordings, which can obscure critical audio features, making it increasingly difficult to distinguish between real and manipulated audio signals [11]. This study seeks to bridge these gaps by focusing on datasets that more accurately reflect real-world conditions, such as the "In-the-Wild" dataset, which includes background noise and a diverse range of accents. By training deep learning models on these more realistic datasets , the research aims to improve the generalization and robustness of audio deepfake detection systems, enabling them to adapt to real-world audio variations more effectively. In this way, this work contributes to developing a more generalized and scalable approach to detecting audio deepfakes in practical settings. The evolution of deep learning (DL) in the field of audio deepfake detection has led to significant advancements, particularly in terms of model architectures and training methodologies. Earlier approaches, which relied on traditional machine learning techniques, often struggled to capture the inherent complexity of audio signals. The shift to deep learning, especially through models like convolutional neural networks (CNNs), has substantially improved detection accuracy by

enabling the identification of subtle audio anomalies [11]. The preference for deep learning over traditional methods is grounded in several key advantages:

1. Feature learning: Unlike traditional techniques, DL models can autonomously learn features from raw audio data, improving their adaptability to new and unseen variations in audio signals.
2. Scalability: Deep learning models demonstrate strong scalability, maintaining or even enhancing performance as they are trained on increasingly large datasets.

The novelty of this research lies in its integration of datasets that closely simulate real-world conditions, including noise and linguistic diversity. By leveraging such datasets, this study seeks to build a robust model capable of generalizing across a wide array of audio inputs, a crucial step toward advancing the field of audio deepfake detection. The ultimate goal is to push the boundaries of current detection capabilities, making them more reliable, scalable, and adaptable to the complexities of real-world audio environments.

Spoofing audio also contains distortion and background noises that aim to confuse ASV systems. These subtle differences can be useful for increasing a model's accuracy and ability to generalize to new real-life data [9].

Additionally, ref. [9] investigated the effect of different front-ends (features) on deepfake audio detection results. This came from the idea that deepfake audio generation models pay attention to the details and features of the audio that are within the hearing scope of humans; therefore, features that are not audible to humans, i.e., of the higher frequency range, might be a giveaway that distinguishes synthesized audio from genuine audio [9]. Architectures trained on a linear frequency cepstral coefficient (LFCC) front-end performed better than those trained on features that are within the human hearing scope, such as mel-frequency cepstral coefficients (MFCCs) and spectrogram-based features [9].

Another approach related to features is taking advantage of vocoders, as in [12], a neural network that takes the features of an acoustic wave, e.g., a mel spectrogram, as an input and outputs a waveform [13]. Since vocoders are commonly the last stage in speech synthesis frameworks, identifying traces of vocoder's artifacts can help detect deepfake audio [12]. This is possible by training a model to identify the vocoder's artifacts in audio before going through a second model that detects audio deepfakes [12]. Table 1 provides a summary of the main ideas of the related research.

Table 1. Paper information.

Paper Title	Main Idea	Model	Language	Fakeness Type	Dataset
Arabic Audio Clips: Identification and Discrimination of Authentic Cantillations from Imitations	1—Compare classifiers' performance. 2—Compare recognition with human experts. 3—Authentic reciter identification.	1—Classic 2—Deep learning	Arabic	Imitation	Arabic Diversified Audio
A Compressed Synthetic Speech Detection Method with Compression Feature Embedding	1—Detect synthetic speech in compressed formats. 2—Multi-branch residual network. 3— Evaluate method on ASVspoof. 4—Compare with state of the art.	Deep learning (DNN)	Not specified	Synthetic	ASVspoof
Learning Efficient Representations for Fake Speech Detection	1—Develop efficient fake speech detection models. 2—Ensure accuracy. 3—Adapt to new forms and sources of fake speech.	1—Machine learning (CNN, RNN, SVM, KNN) 2—Deep learning (CNNs, RNNs)	English	Imitation, Synthetic	ASVSpoof, RTVCSpoof

Table 1. *Cont.*

Paper Title	Main Idea	Model	Language	Fakeness Type	Dataset
WaveFake: A Dataset for Audio Deepfake Detection	1—Create a dataset of fake audio. 2—Provide baseline models for detection. 3—Test generalization on data from different techniques. 4—Test on real-life scenarios.	Fake audio synthesizing: MelGAN, Parallel WaveGAN, Multi-band MelGAN, Full-band MelGAN, HiFi-GAN, WaveGlow Baseline detection models: Gaussian Mixture Model (GMM), RawNet2	English, Japanese	Synthetic	WaveFake
Attack Agnostic Dataset: Generalization and Stabilization of Audio Deepfake Detection	1—Analyze and compare model generalization. 2—Combine multiple datasets. 3—Train on inaudible features. 4—Monitor training for stability.	LCNN, XceptionNet, MesoInception, RawNet2, and GMM	English	Synthetic, Imitation	WaveFake, FakeAVCeleb, ASVspoof
Exposing AI-Synthesized Human Voices Using Neural Vocoder Artifacts	1—Utilize the identification of vocoders' artifacts in audio to detect deepfakes. 2—Create a dataset of fake audio using six vocoders.	Detection model: RawNet2 Dataset construction: WaveNet, WaveRNN, Mel-GAN, Parallel WaveGAN, WaveGrad, and DiffWave	English	Synthetic	LibriVoc, WaveFake, and ASVspoof2019

3. Approach

This section outlines our approach to developing the audio deepfake detection models, covering data acquisition, preprocessing, feature extraction, and training.

Existing methods often use limited or specialized datasets, which may not fully capture the diversity of real-world audio manipulations. Our approach, however, utilizes a broad range of datasets to encompass various audio types and manipulation techniques, enhancing our model's detection capabilities.

For feature extraction, we employ techniques like short-time power spectrum and constant-Q transform (CQT). These methods effectively analyze different frequency ranges, ensuring both high and low frequencies are thoroughly examined, which improves the detection of subtle audio manipulations. Figure 2 illustrates the overall approach for detecting deepfakes.

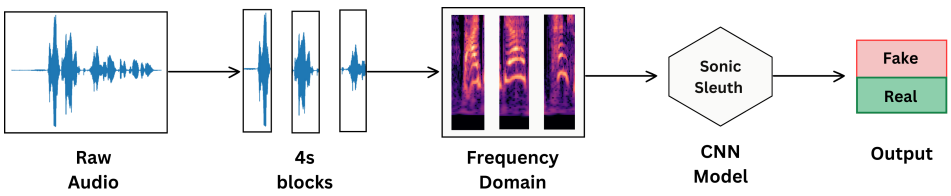

Figure 2. Deepfake detection approach.

3.1. Datasets

The dataset is a crucial element in deep learning, necessitating a thoughtful and informed selection process that considers the specific task and model requirements. For deepfake audio detection, two main types of data are required: generated audio (fake) and human voice audio (real). The selection of the dataset is a critical factor in deep learning for audio deepfake detection. The model's effectiveness depends not only on its architecture but also on the quality, diversity, and relevance of the training data. For this project,

we utilized four datasets that offer a wide range of real and synthetic (deepfake) audio: ASVspoof2019, In-the-Wild, FakeAVCeleb, and Fake-or-Real. These datasets were selected to address generalization issues, ensuring that the model can detect deepfake audio across various contexts and spoofing methods, which reflects real-world scenarios. We trained our CNN model on three datasets: ASVspoof2019, In-the-Wild, and FakeAVCeleb, which combine both types and a total of 70,000 samples. And to ensure the generalization of our model, we further tested it on the Fake-Or-Real dataset.

- **ASVspoof2019** ASVspoof (Automatic Speaker Verification Spoofing and Countermeasures) is an international challenge focusing on spoofing detection in automatic speaker verification systems [14]. The ASVspoof2019 dataset contains three sub-datasets: Logical Access, Physical Access, and Speech Deepfake. Each of these datasets was created using different techniques depending on the task, like text-to-speech (TTS) and voice conversion (VC) algorithms [14]. We utilize the train subset of the Logical Access dataset in our experiment.
  - The Logical Access (LA) subset, which we used, includes synthetic speech created using state-of-the-art TTS technologies. This dataset ensures that our model is exposed to sophisticated spoofing attacks.
  - ASVspoof2019 provides 2580 real audio files and 22,800 fake audio files, with an average length of 3 s per file.
- **In-the-Wild** is a dataset consisting of real and deepfake audio sourced from publicly available recordings of 58 politicians and celebrities. This dataset spans a total of 20.8 h of real audio and 17.5 h of fake audio, ensuring a broad representation of different speaking styles, tones, and environments [10].
  - Each speaker has an average of 23 min of real audio and 18 min of fake audio, with an average length of 4.3 s per clip. The audio was sourced from diverse environments, including media interviews, public speeches, and social media clips, which introduce realistic variances such as background noise, variable quality, and different accents.
  - The fake audio in this dataset was generated using various techniques collected from social media, mimicking public figures in both scripted and spontaneous settings. This dataset simulates real-world challenges by providing audio in uncontrolled conditions, thus testing the model's ability to generalize across diverse and noisy environments.
- **FakeAVCeleb** contains both deepfake and real audio and videos of celebrities. The fake audio was created using a text-to-speech service followed by manipulation with a voice cloning tool to mimic the celebrity's voice [15]. We extracted the audio as WAV files from each video.
  - The dataset includes 10,209 real audio files and 11,357 fake audio files, extracted from videos with an average length of 5 s per file.
  - FakeAVCeleb mimics the challenges of detecting AI-generated content from popular sources, such as social media, where fake media can spread rapidly. This dataset contributes to testing the model's robustness in detecting manipulated audio in entertainment and digital media contexts.
- **The Fake-or-Real (FoR)** dataset comprises 111,000 files of real speech and 87,000 files of fake speech. It encompasses both MP3 and WAV file formats, offering four distinct versions to suit various needs. The 'for-original' files are in their original state, while the 'for-norm' version has been subjected to normalization. 'For-2sec' is shortened to 2 s, while 'for-rerec' simulates re-recorded data, depicting deepfake from a phone call scenario [16].
  - The dataset offers four distinct versions, including a normalized version and a re-recorded version, which simulates a phone call scenario. The diversity in data

formats and audio lengths makes this dataset ideal for testing model performance in a variety of real-world applications.
- The FoR data closely **simulate common deepfake usage,** such as telephone fraud or manipulated voice recordings in conversational settings, adding another layer of complexity to the evaluation process.

The datasets used in this project are designed to represent real-world deepfake audio scenarios by capturing the following critical aspects:

1. Diversity in generation techniques: The datasets include audio generated using various deepfake technologies such as TTS, VC, and cloning techniques. This diversity ensures the model is exposed to the range of methods used to generate synthetic audio, making it more effective in real-world applications.
2. Generalization across environments: Datasets like In-the-Wild simulate real-world conditions, including background noise, different recording devices, and varied speaking environments, which are common in public audio recordings. This enhances the model's ability to detect deepfakes in uncontrolled settings, where real-world challenges such as poor audio quality or overlapping voices exist.
3. Realistic data representation: The inclusion of datasets such as ASVspoof2019 and Fake-or-Real ensures that our model is exposed to both high-quality and noisy audio, from both controlled experiments and re-recorded settings, thus improving its robustness. This combination allows the model to detect deepfakes in situations like social media voice messages, fraudulent phone calls, and doctored recordings.
4. Variety of speakers and content: With a wide range of speakers, accents, languages, and contexts (e.g., political speeches, interviews), the datasets ensure the model is not biased towards a specific type of speaker or content, but can generalize across different contexts, which is essential for real-world detection.

By combining the first three diverse datasets for training and reserving the last one, which has properties to mimic real attacks such as phishing through phone calls, for testing, we aim to achieve a comprehensive and varied dataset for our task. Further details are illustrated in Table 2.

Table 2. Datasets for model development.

Name	Size		Length	Sample Rate	File Format	URL
	Real	Fake				
ASVspoof2019	2580 files	22,800 files	Avg. 3 s	-	flac	Link
'In-the-Wild'	19,963 files	11,816 files	Avg. 4.s	16 kHz	WAV	Link
FakeAVCeleb	10,209 files	11,357 files	Avg. 5 s	-	MP4	Link
Fake-or-Real	111,000 files	87,000 files	1–20 s	-	WAV/MP3	Link

3.2. Data Preprocessing

Several essential preprocessing steps were applied to our datasets to ensure uniformity and compatibility of the audio data with our machine learning models, including:

- First, any silence in the audio was trimmed to enable the model to focus on speech parts of the audio.
- Second, all audio clips were trimmed to a standardized length of 4 s or padded by repeating the audio to avoid silence that could potentially bias the model's learning.
- Third, the audio data were downsampled to a rate of 16 kHz. Downsampling reduces the data size and computational load without significantly compromising the audio signal's quality or essential characteristics.
- Lastly, the audio was converted to a mono channel using channel averaging. Stereo audio, which uses two channels (left and right), is averaged into a single channel, resulting in a mono audio signal.

These conversions simplify the data and ensure a uniform input format for the machine learning models, which is particularly important when dealing with diverse audio sources.

3.3. Feature Extraction

In audio processing, rather than working with raw audio files, a common practice in AI is to convert audio into a visual representation, such as a spectrogram. This conversion enables the use of image classification models, like convolutional neural networks (CNNs), for audio classification tasks. To accurately capture the essential characteristics of audio signals, we explore two main types of feature extraction techniques: short-time power spectrum and constant-Q transform (CQT). These techniques are crucial for analyzing different frequency ranges effectively, ensuring that both high and low frequencies are captured for comprehensive analysis.

3.3.1. Short-Time Power Spectrum

Short-time power spectrum refers to the representation of the power distribution of a signal over short time intervals. It is a technique commonly used in signal processing and analysis to examine the frequency content and changes in a signal over time [17].

Our chosen feature extraction techniques, MFCCs (mel-frequency cepstral coefficients) and LFCCs (linear frequency cepstral coefficients), are derived from a signal's short-time power spectrum and provide compact representations of its spectral characteristics.

MFCCs is a widely used feature set in automatic speech recognition systems [18]. It is based on the human auditory system's response to sound and is particularly effective in capturing the characteristics of speech signals. LFCCs is a similar technique that has been shown to outperform MFCCs in specific applications, such as deepfake detection [8]. Both techniques are used to extract relevant information from audio signals for further analysis and processing.

3.3.2. Constant-Q Transform

The constant-Q transform (CQT) is another feature extraction technique that provides a time–frequency representation of a signal. Unlike the short-time power spectrum, which has a fixed frequency resolution, CQT offers a logarithmic frequency scale. This means that CQT provides higher-frequency resolution at lower frequencies and better time resolution at higher frequencies, which is particularly useful in analyzing music and other signals with a wide frequency range.

CQT is particularly advantageous when dealing with signals where pitch perception is crucial, as it mirrors the human auditory system's logarithmic frequency response. This makes CQT an effective technique in applications like music analysis, instrument identification, and speech processing, where capturing the harmonic structure of audio signals is important [19].

3.4. Sonic Sleuth Structure and Training Details

We employ a convolutional neural network (CNN) to classify the processed audio samples into real or synthesized audio. The structure includes three convolutional layers, each followed by a max-pooling layer; after the convolutional layers, the output is flattened and passed through two fully connected (dense) layers and a 10% dropout rate to prevent overfitting. The final dense layer has a single output unit with a sigmoid activation function, providing a probability score for classifying the input as deepfake. Figure 3 illustrates the architecture of our custom CNN model.

For training the deepfake audio detection models, the dataset was split into train, validation, and test subsets with a ratio of 8.5:1:0.5.

The models were trained for 100 epochs using the following configurations:

1. Early stopping: An early stopping technique was implemented with a patience of 10 epochs to prevent overfitting. This means the training process would be stopped if the validation loss did not improve for 10 consecutive epochs.

2. Optimizer: The Adam optimizer was used for updating the model's weights during training.
3. Loss function: Binary cross-entropy loss was used as the loss function, which is suitable for binary classification problems.
4. Class weights: Class weights were calculated and applied during the training process to handle the class imbalance in the dataset. Class weights adjust the importance of each class during the loss calculation, helping the model learn from imbalanced datasets more effectively while avoiding bias.

By incorporating these techniques and configurations, the model was trained to learn discriminative features from the audio spectrograms and classify them as real or deepfake with improved performance and generalization to new, unseen data.

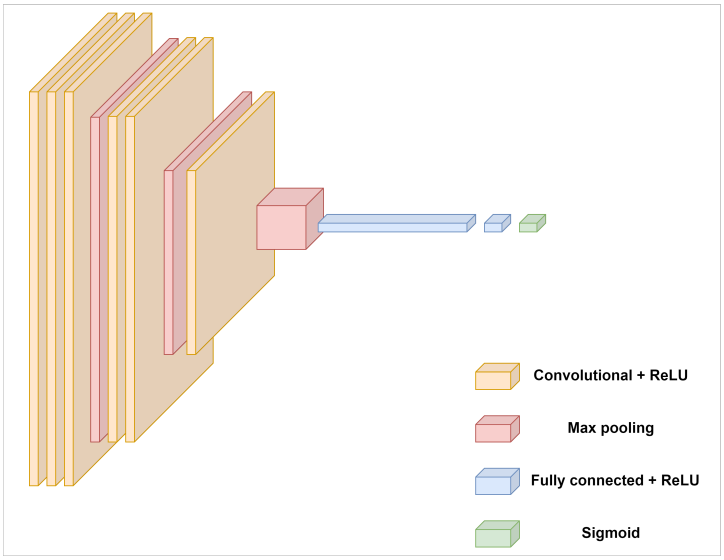

Figure 3. Sonic Sleuth architecture.

3.5. Evaluation Metrics

Equal error rate (EER) and accuracy are two commonly used metrics for evaluating the performance of classification models, especially in binary classification tasks. EER represents the point on the receiver operating characteristic (ROC) curve where the false positive rate (FPR) equals the false negative rate (FNR). It is a helpful metric for assessing the balance between the two types of errors, making it the standard metric for evaluating verification systems. A lower EER value indicates better performance, with 0 indicating perfect classification.

Accuracy is a more straightforward metric that measures the proportion of correctly classified instances out of the total number of instances. While accuracy is easy to interpret and widely used, it can be misleading, especially in imbalanced datasets where one class dominates the other. We avoid this mistake in our implementation by ensuring a balanced test dataset and including class weights during training.

In addition to EER and accuracy, the F1 score is another important metric that combines both precision and recall into a single measure. The F1 score is the harmonic mean of precision (the proportion of true positives among the instances classified as positive) and recall (the proportion of true positives among the actual positives). This metric is particularly useful in scenarios where both false positives and false negatives are important, providing a balance between them. A higher F1 score indicates better model performance, especially in cases of class imbalance, where precision and recall are critical.

4. Results and Discussion

To accurately represent the results of our experiments, we used several evaluation metrics, including balanced accuracy, F1 score, and equal error rate (EER).

Table 3 presents the results of testing our models on the test subset of our dataset. LFCC demonstrated the best performance across all metrics, achieving an EER of 0.0160, an accuracy of 98.27%, and an F1 score of 98.65%. MFCC also performed well, with an EER of 0.0185, an accuracy of 98.04%, and an F1 score of 98.45%. In contrast, CQT exhibited the lowest performance, with an EER of 0.0757, an accuracy of 94.15%, and an F1 score of 95.76%. These results indicate that LFCC and MFCC are superior feature extraction methods for this specific classification task, providing higher accuracy and lower error rates compared to CQT.

Table 3. Model performance metrics.

Feature	EER	Accuracy	F1
CQT	0.0757	94.15%	95.76%
LFCC	0.0160	98.27%	98.65%
MFCC	0.0185	98.04%	98.45%

4.1. Test on External Dataset

To assess the generalizability of our models, we tested them on an external dataset, Fake-or-Real (FoR). The testing was conducted on a balanced subset of 2116 samples of 4 s of audio. The performance metrics are reported in Table 4.

Table 4. Model performance metrics on external dataset.

Feature	EER	Accuracy	F1
CQT	0.0942	82.51%	83.19%
LFCC	0.1718	75.28%	72.63%
MFCC	0.3165	61.24%	54.94%

While the performance of the models is decreased compared to the initial evaluation, it is noteworthy that CQT, which previously had the lowest accuracy among the three methods, now shows improved performance. With an EER of 0.0942, an accuracy of 82.51%, and an F1 score of 83.19%, CQT has made significant strides, indicating its viability as a feature extraction method even on diverse datasets.

Similarly, LFCC, despite experiencing a decline in performance, continues to exhibit its utility in real-world scenarios. LFCC, with an EER of 0.1718, an accuracy of 75.28%, and an F1 score of 72.63%, maintains its effectiveness in capturing relevant audio features. Meanwhile, MFCC, with an EER of 0.3165, an accuracy of 61.24%, and an F1 score of 54.94%, appears not to be the best for real-life applications due to weak generalizability.

4.2. Ensemble Approach

In machine learning, the ensemble model approach combines the output of multiple models by combining the output probabilities (percentages) from the three models, each contributing up to 100%. The final decision is based on the combined score, normalized by dividing the total by 300 to yield a final classification for models trained on the same task, which can reduce error.

Table 5 shows the results of the ensemble approach that combines the CQT and LFCC models.

Table 5. Model performance metrics on external dataset (ensemble approach).

Feature	EER	Accuracy	F1
CQT and LFCC	0.0851	84.92%	84.73%

The combination of CQT and LFCC on the external dataset presents promising results, with an equal error rate (EER) of 0.0851, an accuracy of 84.92%, and an F1 score of 84.73%. This joint approach showcases slight improvement compared to the individual methods, indicating the complementary nature of CQT and LFCC in capturing relevant audio features. These results provide valuable insights for further exploration and optimization, highlighting the potential for synergy between different feature extraction methods to achieve superior performance across diverse datasets and application contexts.

4.3. Model Performance Comparing to Recent Deepfake Detection Methods

The table below, Table 6, summarizes the performance of various models evaluated on different datasets for audio deepfake detection. Each model's equal error rate (EER) and accuracy metrics are highlighted, along with key observations.

Table 6. Summary of model performance for audio deepfake detection.

Dataset	Model	Features	EER	Accuracy
FakeAVCeleb	Xception [20]	MFCC	0.2472	73.06%
In-the-Wild	MesoInception [10]	Log-spectrogram	0.37414	-
ASVspoof2019	CQCC-ResNet [21]	CQCC	0.0769	-
Combined	Sonic Sleuth	CQT	0.0942	82.51%
		LFCC	0.1718	75.28%
		MFCC	0.3165	61.24%

- **Xception model performance:** The Xception model demonstrated the best performance on the FakeAVCeleb dataset, achieving an EER of 0.2472 and an accuracy of 73.06%. This model's strong performance indicates its robustness in detecting audio deepfakes. However, further refinement is needed for broader applicability across datasets.
- **MesoInception insights:** For the In-the-Wild dataset, the MesoInception model showed an EER of 0.37414 when evaluated with log-spectrogram features. The results indicate that longer audio inputs enhance detection capabilities. However, the model faced generalization challenges, suggesting that performance may vary in real-world scenarios.
- **CQCC-ResNet analysis:** The CQCC-ResNet model, tested on the ASVspoof2019 dataset, achieved an EER of 0.0769 on unknown attacks. This highlights the model's effectiveness in recognizing familiar patterns but also its struggle with novel spoofing techniques, suggesting the need for better generalization capabilities.

Overall, the table encapsulates the strengths and weaknesses of each model in detecting audio deepfakes. The findings underscore the need for continued research and development in this field, particularly regarding model robustness and generalization capabilities.

5. Conclusions

In this research, we developed the AI model Sonic Sleuth for detecting audio deepfakes. Using a custom convolutional neural network (CNN), we trained the model on three diverse datasets—ASVspoof2019, In-the-Wild, and FakeAVCeleb—comprising a total of 78,725 audio samples. Feature extraction techniques such as linear frequency cepstral coefficients (LFCCs), mel-frequency cepstral coefficients (MFCCs), and constant-Q transform (CQT) were employed to convert the audio signals into spectrograms for detailed analysis.

Our results, shown in Table 7, show that the LFCC-based model performed optimally on the training dataset, achieving an equal error rate (EER) of 0.0160 and an accuracy of 98.27%. However, CQT exhibited superior performance on the external dataset, indicating stronger generalization, with an EER of 0.0942 and an accuracy of 82.51%. Additionally, combining CQT and LFCC in an ensemble approach further improved generalization performance, achieving an EER of 0.0851 and an accuracy of 84.92%.

Table 7. Summary of key results.

Feature Extraction	Dataset	EER	Accuracy
LFCC	Training	0.0160	98.27%
CQT	External	0.0942	82.51%
LFCC + CQT (Ensemble)	External	0.0851	84.92%

While these results are promising, several limitations remain. Notably, the performance varied across datasets, and the model's ability to generalize to real-world scenarios with noisy or low-quality audio needs further enhancement. For future work, we propose a deeper exploration of advanced signal processing techniques and ensemble methods to address noise and other environmental factors. Expanding the feature set and optimizing combinations of features could further enhance the model's generalization and performance in diverse applications.

Author Contributions: Conceptualization, A.A. and D.A.; methodology, A.A. and D.A.; software, A.A. and D.A.; validation, E.A. and S.A.; formal analysis, A.A. and D.A.; investigation, A.A. and D.A.; resources, A.A. and D.A.; data curation, A.A. and D.A.; writing—original draft preparation, A.A. and D.A.; writing—review and editing, E.A. and S.A.; visualization, A.A. and D.A.; supervision, E.A. and S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was conducted without external financial support. The project relied on the advanced computational tools, software, and datasets that were integral to the development and testing of the Sonic Sleuth model. The research team's dedication, combined with access to high-quality resources such as deep learning frameworks and substantial datasets, was pivotal in achieving the study's goals. The institution supported us by providing the necessary infrastructure and opportunities to demonstrate our capabilities, which played a crucial role in the successful completion of this research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Acknowledgments: This work was written by Anfal Alshehri and Danah Almalki, under the supervision of Eaman Alharbiand and Somayah Albaradei. We are grateful for the guidance and support provided throughout this project.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
ASV	Automatic speaker verification
CQT	Constant-Q transform
CNN	Convolutional neural network
DL	Deep learning
DNN	Deconvolutional neural network
EER	Equal error rate

FNR	False negative rate
FPR	False positive rate
GAN	Generative adversarial network
LFCCs	Linear frequency cepstral coefficients
MFCCs	Mel-frequency cepstral coefficients
ROC	Receiver operating characteristic
TTS	Text-to-speech
VC	Voice conversion

References

1. Oh, S.; Kang, M.; Moon, H.; Choi, K.; Chon, B.S. A demand-driven perspective on generative audio AI. *arXiv* **2023**, arXiv:2307.04292.
2. Deepfakes (a Portmanteau of "Deep Learning" and "Fake"). Images, Videos, or Audio Edited or Generated Using Artificial Intelligence Tools. *Synthetic Media*, 2023. Available online: https://en.wikipedia.org/wiki/Deepfake (accessed on 4 May 2020).
3. Gu, Y.; Chen, Q.; Liu, K.; Xie, L.; Kang, C. GAN-based Model for Residential Load Generation Considering Typical Consumption Patterns. In Proceedings of ISGT 2019, Washington, DC, USA, 18–21 February 2019; IEEE: Piscataway, NJ, USA, November 2018. [CrossRef]
4. Camastra, F.; Vinciarelli, A. *Machine Learning for Audio, Image and Video Analysis: Theory and Applications*; Springer: London, UK, 2015.
5. Tenoudji, F.C. *Analog and Digital Signal Analysis: From Basics to Applications*; Springer International Publishing: Cham, Switzerland, 2018.
6. Natsiou, A.; O'Leary, S. Audio Representations for Deep Learning in Sound Synthesis: A Review. *arXiv* **2022**, arXiv:2201.02490.
7. Marcus, G. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. *arXiv* **2020**, arXiv:2002.06177.
8. Frank, J.; Schönherr, L. WaveFake: A Data Set to Facilitate Audio Deepfake Detection. *arXiv* **2021**, arXiv: 2111.02813
9. Kawa, P.; Plata, M.; Syga, P. Attack Agnostic Dataset: Towards Generalization and Stabilization of Audio DeepFake Detection. In Proceedings of Interspeech 2022, ISCA, Incheon, Republic of Korea, 18–22 September 2022. [CrossRef]
10. Müller, N.M.; Czempin, P.; Dieckmann, F.; Froghyar, A.; Böttinger, K. Does audio deepfake detection generalize? *arXiv* **2024**, arXiv:2203.16263.
11. Almutairi, Z.; Elgibreen, H. A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. *Algorithms* **2022**, *15*, 155. [CrossRef]
12. Sun, C.; Jia, S.; Hou, S.; AlBadawy, E.; Lyu, S. Exposing AI-Synthesized Human Voices Using Neural Vocoder Artifacts. *arXiv* **2023**, arXiv:2302.09198.
13. Zhang, C.; Zhang, C.; Zheng, S.; Zhang, M.; Qamar, M.; Bae, S.-H.; Kweon, I.S. A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI. *arXiv* **2023**, arXiv:2303.13336.
14. 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. *arXiv* **2020**, arXiv:1911.01601. [CrossRef]
15. Khalid, H.; Tariq, S.; Kim, M.; Woo, S.S. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. In Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021. Available online: https://openreview.net/forum?id=TAXFsg6ZaOl (accessed on 29 September 2023).
16. Abdeldayem, M. The Fake-or-Real Dataset. *Kaggle Dataset*, 2022. Available online: https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-dataset (accessed on 28 May 2024).
17. Sahidullah, M.; Kinnunen, T.; Hanilçi, C. A Comparison of Features for Synthetic Speech Detection. *Interspeech* **2015**, *2015*, 2087–2091. [CrossRef]
18. Zheng, F.; Zhang, G. Integrating the energy information into MFCC. In Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China, 16–20 October 2000; Volume 1, pp. 389–392. [CrossRef]
19. Todisco, M.; Delgado, H.; Evans, N. Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification. *Comput. Speech Lang.* **2017**, *45*, 516–535. [CrossRef]
20. Khalid, H.; Kim, M.; Tariq, S.; Woo, S.S. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In Proceedings of the 1st Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection, Virtual Event, 24 October 2021; pp. 7–15.
21. Alzantot, M.; Wang, Z.; Srivastava, M.B. Deep residual neural networks for audio spoofing detection. *arXiv* **2019**, arXiv:1907.00501.

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.