applied sciences

MDPI

Article

Source Microphone Identification Using Swin Transformer

Mustafa Qamhan [1,*], Yousef A. Alotaibi [1] and Sid-Ahmed Selouani [2]

[1] Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia; yaalotaibi@ksu.edu.sa
[2] LARIHS Laboratory, Campus Shippagan, Université de Moncton, Shippagan, NB E8S 1P6, Canada; sid-ahmed.selouani@umoncton.ca
* Correspondence: mqamhan@ksu.edu.sa

Abstract: Microphone identification is a crucial challenge in the field of digital audio forensics. The ability to accurately identify the type of microphone used to record a piece of audio can provide important information for forensic analysis and crime investigations. In recent years, transformer-based deep-learning models have been shown to be effective in many different tasks. This paper proposes a system based on a transformer for microphone identification based on recorded audio. Two types of experiments were conducted: one to identify the model of the microphones and another in which identical microphones were identified within the same model. Furthermore, extensive experiments were performed to study the effects of different input types and sub-band frequencies on system accuracy. The proposed system is evaluated on the Audio Forensic Dataset for Digital Multimedia Forensics (AF-DB). The experimental results demonstrate that our model achieves state-of-the-art accuracy for inter-model and intra-model microphone classification with 5-fold cross-validation.

Keywords: source identification; digital forensics; deep learning; transformer; audio analysis; spectral analysis; pattern recognition

Citation: Qamhan, M.; Alotaibi, Y.A.; Selouani, S.-A. Source Microphone Identification Using Swin Transformer. *Appl. Sci.* **2023**, *13*, 7112. https://doi.org/10.3390/app13127112

Academic Editors: Artur Janicki and Katarzyna Wasielewska

Received: 15 May 2023
Revised: 10 June 2023
Accepted: 12 June 2023
Published: 14 June 2023

1. Introduction

Source identification in digital forensics tries to connect multimedia content to a specific type of device that was used to acquire it. It is based on the idea that every device leaves its own unique trace in the content it captures (fingerprint). To identify the source device, specific characteristics are extracted from the unknown content and compared with a database of known characteristics from various devices. Therefore, forensic source identification is essentially a process of classifying content based on its acquisition characteristics.

There are several studies conducted on source identification for different fields, such as source camera identification [1,2], scanner and printer forensics [3], pen-based digitizer devices [4], and microphone and environment identification [5]. Microphone identification is a technique used in audio forensics to determine the specific type of microphone that was used to record a specific piece of audio evidence. It can be used in investigations to determine the origin of a recording or to identify the source of a recording in cases of fraud or forgery.

In order to classify a microphone, audio forensics experts may use a combination of techniques, including spectral analysis, impulse-response analysis, and pattern recognition. The spectral analysis involves analyzing the frequency content of the recording, while impulse-response analysis examines how the microphone responds to different types of sounds. Pattern recognition, on the other hand, involves identifying specific patterns or features in the recording that are unique to certain types of microphones.

Deep learning, which is a subfield of machine learning [6], has become one of the key development-drivers in the field of artificial intelligence in recent years because it

has achieved good results in many fields and applications, such as natural language processing [7], speech, language, and emotion recognition [8,9], map generalization [10], landform recognition [11], IoT security [12], etc. This has encouraged many researchers to use it in digital forensics, where high accuracy is required. Transformer deep learning models, such as the Transformer-XL [13] and the BERT [7], have been developed to efficiently process sequential data. These models use self-attention mechanisms to weigh the importance of different parts of the input sequence, which allows them to effectively handle long-term dependencies. This makes them well-suited for processing audio signals, which are inherently sequential in nature.

The Swin Transformer is a deep-learning model for computer vision introduced by Liu et al. [14]. It addresses some of the shortcomings of conventional transformer models, such as their difficulty scaling and computational inefficiency. The Swin Transformer uses a hierarchical architecture and shifts windows to achieve better scalability and efficiency. With its hierarchical structure, the Swin Transformer can capture multi-scale contextual information with remarkable effectiveness, much like a CNN. Moreover, the shifted-window-based self-attention mechanism drastically reduces the computational complexity of the self-attention operation, allowing the model to handle larger input sizes without incurring prohibitively high computational costs. By fusing features across different stages, the model can leverage low-level and high-level features, enhancing its capacity to tackle complex visual tasks.

In this paper, we propose a system for identifying microphones based on recorded audio using a Swin Transformer deep neural network [14], in which audio recordings are processed to extract features such as the Mel-scale frequency spectrum. These features are then used as inputs to the transformer model, which would learn to distinguish between different types of microphones based on their characteristic patterns.

Our Contribution: In this paper, we build a system using the Swin Transformer model for microphone classification that achieves state-of-the-art results on the Audio Forensic Dataset for Digital Multimedia Forensics (AF-DB). In addition, comprehensive experiments are conducted to find out what is the best input preprocessing and frequency band that improves the accuracy of the system.

Paper Structure: The rest of the paper is organized as follows: Section 2 provides an overview of the literature on audio source identification. Section 3 describes the general methodology, the database used, the input preprocessing, and the proposed Swin Transformer algorithms. Section 4 presents the experimental and results discussions. Finally, Section 5 presents the conclusions of this paper.

2. Related Work

Microphone identification for audio forensics is a rapidly growing field with a wide range of research and applications. The goal of microphone identification is to define the type of microphone used to record a specific audio clip, which can be used in forensic investigations to determine the source of a recording or to authenticate a recording. The researchers used different feature extraction methods and systems to verify the microphone's authenticity. The electric network frequency (ENF) signal results from electromagnetic interference from power lines, which manifests as an acoustic hum that is susceptible to being picked up by microphones close to power mains [15]. The ENF has been utilized in digital recording authentication [16], as evidenced by numerous early and current studies focusing on ENF-based techniques [9]. Nonetheless, there are situations where such techniques cannot be applied, such as audio devices powered by batteries in areas away from the power grid [17].

Buchholz R. et al. [18] propose a method for classifying microphones using Fourier coefficients, and the results suggest that Fourier coefficients can be useful features for microphone classification. Zhang et al. [19] present a method for identifying audio sources using a residual network. The network is trained on a dataset of mixed audio sources and is able to separate and identify the individual sources. A new approach to improve

the performance of identifying the source cell phone is demonstrated in [20]. The method employs a spatiotemporal representation learning approach. Baldini G. et al. [21] evaluate various entropy measures for identifying microphones based on their audio recordings. The study found that the Rényi entropy measure performs the best in terms of accuracy, followed by the Shannon entropy measure. The authors also find that using a combination of entropy measures improves the overall performance of microphone identification.

Luo D. et al. [22] present a method for determining the source of an audio recording using the "band energy difference" (BED) technique. The BED method compares the energy levels of different frequency bands in the recording to a database of known sources and uses the differences in energy levels to identify the most-likely source of the recording. The method in [23] uses deep-representation learning to extract features from the speech recordings and then applies spectral clustering to group the users into clusters based on those features. O. Eskidere [24] uses wavelet-based features to extract characteristics of the speech signal and then uses a machine-learning algorithm to classify the signal based on these features. In [25], a model for verifying a person's identity using speech recordings taken from their cell phone is presented. The technique involves sparse representation to compare the recording with a pre-existing reference sample to determine whether the speakers are who they claim to be.

Li Y et al.'s [26] method involves extracting features from speech recordings using a deep-Gaussian-supervector model and then using spectral clustering to group similar speakers together. O. Eskidere et al.'s [27] method uses a single-Gaussian-mixture model to represent the audio from a single microphone and then compares the model with a database of known microphone models to identify the source microphone. Jiang Y et al. [28] introduce a method for recognizing the source microphone using a kernel-based projection technique to map the audio data into a high-dimensional feature space, where a classifier can be trained to distinguish between different microphones. Zou L. et al. [29] propose a method for identifying the source cell phone used to make a speech recording by using sparse representation and the KISS metric. The method involves first extracting features from the speech recording, then representing these features using sparse coding, and finally comparing the sparse representations using the KISS metric to identify the source cell phone.

Cuccovillo L. et al. [30] present a method for identifying microphone devices in noisy environments using speaker-independent features. The proposed approach involves extracting features from the audio signal, such as the cepstral mean and variance, and then using machine-learning algorithms to classify the microphones. Qamhan et al. [5] present a method for classifying digital audio recordings based on the microphone and recording environment used to capture them. The authors use the CRNN model to train a model that can accurately classify audio recordings based on features such as spectrogram and noise level.

M. Pavlovic et al. [31] present a classification model for recognizing the type of microphone being used. The model uses a multi-layer perceptron network to determine the microphone type based on characteristics such as frequency response and noise level. The paper of Qin T et al. [32] presents a method for identifying the source cell phone of a recorded audio signal in the presence of additive noise using the constant-Q transform (CQT) domain. The proposed method first applies the CQT to the recorded signal and then extracts a set of features, including the mean and standard deviation of the CQT magnitude coefficients. These features are then used to train a machine-learning classifier, such as a support vector machine (SVM), to recognize the source cell phone. Simeng Qi et al. [33] present a method for identifying the device used to record audio based on the unique characteristics of the background noise present in the recording. Kurniawan F et al. [34] propose a method that involves analyzing statistical properties of the audio, such as frequency content and noise levels, and comparing them to known characteristics of the microphone.

Baldini G et al. [35] identify microphones using spectral entropy and a convolutional neural network (CNN). The authors extract spectral entropy features from audio recordings made with different microphones and use a CNN to classify the recordings based on their spectral entropy features. Zeng C et al. [36] propose a method for identifying

recording devices using deep learning. The method uses CNN to extract features from audio recordings and a support vector machine (SVM) to classify the recording devices based on these features. The authors [37] present a method for identifying smartphones based on their microphone characteristics when recording audio in different environments. The authors collected data from various smartphones in different environments and used a linear discriminant classifier (LDA) and a CNN to analyze the data. Lin X et al. [38] propose a sub-band-aware CNN model for cell phone recognition. The model uses sub-band information to improve the accuracy of cell phone recognition by capturing frequency-specific features of the audio signal. Baldini G et al. [39] collect a dataset of microphone recordings and use it to train a CNN to classify microphone brands. The results indicate that CNNs can be effectively used for microphone identification. The paper [40] proposes a method for identifying smartphones using the built-in microphone and a CNN. The authors collected a dataset of audio recordings from different smartphone models and used it to train the neural network.

Previous research has shown varying levels of accuracy for proposed systems in digital audio forensics, with many of these studies relying on local databases that cannot be generalized for several reasons. For instance, some of these methods utilize databases with significant classes variations, making the classification task relatively simple. Conversely, other databases are recorded using pre-set parameters that do not reflect the actual conditions of audio recordings used in digital forensics. Furthermore, some databases incorporate multiple microphones connected to different recording devices, making it unclear whether the result was to classify the microphones or the recording devices. This study proposes a deep learning system (Swin Transformer) for microphone classification to achieve state-of-the-art results on the Audio Forensic Dataset for Digital Multimedia Forensics (AF-DB). Such a system is expected to help forensic experts validate evidence before presenting it in court.

3. Method

3.1. Audio Forensic Dataset for Digital Multimedia Forensics (AF-DB) [41]

The Audio Forensic Dataset for Digital Multimedia Forensics is a comprehensive collection of audio files that are specifically designed for the purpose of developing, training, and validating algorithms and tools in the field of digital multimedia forensics. This dataset is intended to provide researchers and practitioners with a diverse set of audio samples that can be used to investigate various aspects of audio forensics, such as audio tampering detection, speaker identification, and environment and recording device identification. The dataset includes audio samples recorded in six distinct settings (a soundproof room, classroom, lab, stairs, car parking lot, and garden). There are microphones from seven distinct brands totaling 22, which can be used to develop and evaluate algorithms for identifying the source microphone. The dataset consists of files that have a duration of about three minutes each. The initial minute of each recording is characterized by silence, while the remaining duration features a reading of a predefined set of sentences by the volunteers. The dataset includes 660 audio recordings that were obtained from a group of five volunteers. The volunteers were asked to read a specific set of sentences in several languages, such as English, Arabic, Indonesian, and Chinese, and their voices were recorded for the dataset [41]. Microphone models and quantities for the AF-DB are presented in Table 1.

Table 1. Microphone information that was utilized for the AF-DB dataset.

	Brand	Model	Transduction Method	# Mics	Frequency Response Hz
M1	Shure	SM 58	Dynamic	3	50–15,000
M2	Electro Voice	RE 20	Dynamic	2	45–18,000
M3	Sennheiser	MD-421	Dynamic	3	30–17,000
M4	AKG	C 451	Condenser	2	20–20,000
M5	AKG	C 3000 B	Condenser	2	20–20,000
M6	Neumann	KM184	Condenser	2	20–20,000
M7	Coles	4038	Ribbon	2	30–15,000
M8	t.bone	MB88U	Dynamic	6	20–16,000
		Total		22	

3.2. Input Preprocessing

As previously stated in the database description, each file comprises recordings of three minutes. These files were sliced into equal segments of five seconds, ensuring that each segment inherits the same label as the original file. After that, the audio files are transformed into a Mel spectrogram that serves as the input to the system.

A Mel spectrogram is a type of spectrogram that uses the Mel scale to represent frequency in the vertical axis instead of the traditional linear scale. The Mel scale is a perceptual scale of pitches that is based on the way humans hear sound. In a Mel spectrogram, the horizontal axis represents time, and the color or intensity of each point in the spectrogram represents the amplitude of the sound at a particular frequency and time. The Mel spectrograms have many advantages, such as the ability to mimic human auditory perception, reduce dimensionality, preserve resilience in noisy conditions, allow simple visualization, and be widely adopted in diverse fields. Mel spectrograms are often used in voice recognition, music classification, and other audio-signal-processing applications, and they may be easily incorporated into vision-deep neural networks for classification.

To create a Mel spectrogram, the audio signal is first divided into short overlapping frames, and the Fourier transform is applied to each frame to obtain its frequency spectrum. The resulting spectrum is then transformed into the Mel scale using a bank of overlapping triangular filters. The logarithm of the power spectrum is taken, and the resulting values are plotted as a function of time and frequency as shown by the following equations:

$$X_k = \sum_{n=0}^{N-1} w_n x_n e^{-i2\pi kn/N}, \;\; k = 0 \ldots . N-1 \tag{1}$$

where w_n represents the window function (Hamming window), and x_n represents the original speech signal.

The process involves applying 128 triangular filters on a Mel frequency scale to the power spectrum to extract frequency bands. The Mel frequency scale is a perceptual scale that better matches how humans perceive frequency. The following formula converts between frequencies in Hertz (f) and the Mel scale (m):

$$m = 2595 \, log\left(1 + \frac{f}{700}\right) \tag{2}$$

3.3. Swin Transformer Model

The overall architecture of the proposed Swin Transformer is shown in Figure 1. It first splits an audio Mel spectrogram into different non-overlapping patch tokens with a CNN projection layer, as shown in Figure 1; each patch has a 96-dimensional vector feature, and the characteristics of each token are then projected to an arbitrary dimension represented by

C using a linear embedding layer, which will pass via a number of Swin Transformer blocks and patch merging layers to produce hierarchical representations. The 'Patch Merging' block and the 'Swin Transformer Block' are the two fundamental construction blocks of the Swin Transformer, as we can see in Figure 1. We will go through these two blocks in depth in the next subsections.

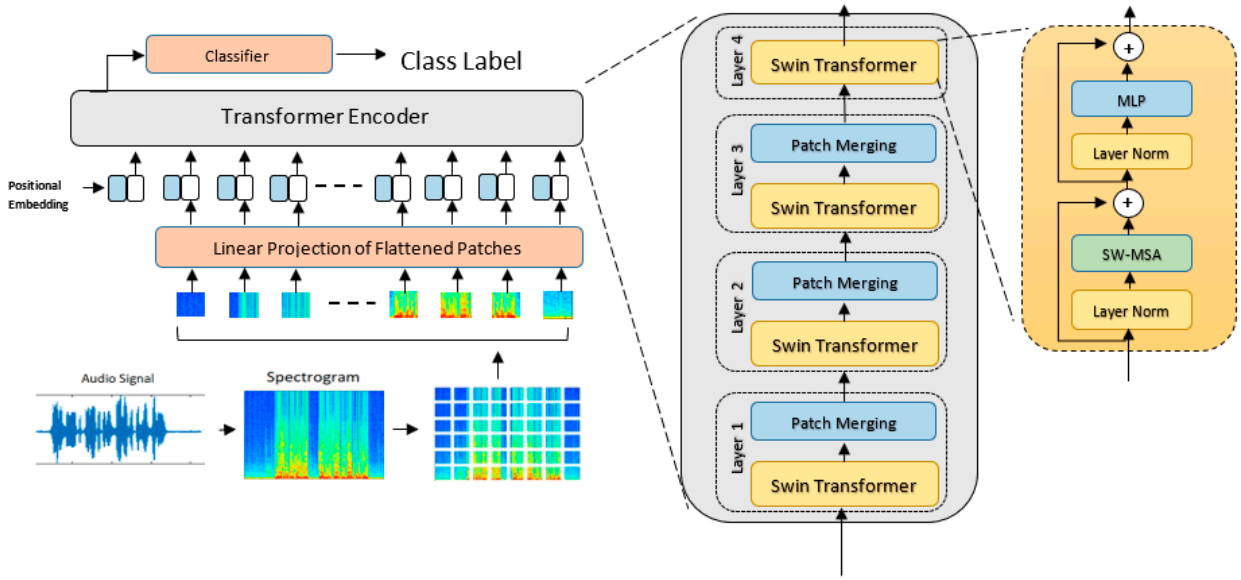

Figure 1. Audio spectrogram Swin Transformer architecture for microphone identification.

3.3.1. Swin Transformer Block

The Swin Transformer block consists of two sub-units, as illustrated in Figure 1, each of which has a normalization layer followed by an attention module, followed by a Multi-Layer Perceptron (MLP) with the GeLU activation function, and a residual connection is applied after each unit. The attention module is based on a modified multi-head self-attention mechanism that uses a shifted window (SW-MSA) approach to improve computational efficiency. The window-based multi-head self-attention (W-MSA) module is used in the first Transformer block, and the shifted window-based multi-head self-attention (SW-MSA) module is used in the second. Using the window-partitioning approach, we can express the continuous Swin Transformer blocks as follows:

$$\begin{aligned} \hat{z}^l &= W_{MSA(LN(z^{l-1}))} + z^{l-1} \\ z^l &= MLP\left(LN\left(\hat{z}^l\right)\right) + \hat{z}^l \\ \hat{z}^{l+1} &= SW\_MSA\left(LN\left(z^l\right)\right) + z^l \\ z^{l+1} &= MLP\left(LN\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \end{aligned} \tag{3}$$

where $\hat{z}$ and z^l represents the output features of the (S)W-MSA module and module and the MLP module for block l, respectively.

3.3.2. Patch Merging Block

Patch merging is a technique used in vision transformers to reduce the number of patches passed onto each individual transformer encoder block. As the depth of the network increases, the number of tokens is decreased via patch-merging layers to form a hierarchical representation. In the initial patch-merging layer, features from pairs of adjacent patches are concatenated together, and then a linear layer is applied to the resulting 4C-dimensional features. This reduces the input's resolution by a factor of n, changing its original dimensions (in this case, height, width, and channel depth) from $H \times W \times C$ to $(H/n) \times (W/n) \times (n^2 \times C)$.

4. Experiments and Discussion

This section will cover all the experiments conducted with the proposed Swin Transformer model (Swin_T), and we will compare the results with two previous studies that used the AF-DB database, discuss the same methodology, and have the same common goal. The first study employed a Gabor filter for feature extraction and K-nearest neighbors (KNN) as the classifier (Gabor-KNN) [41], while the second study utilized an audio spectrogram transformer (AST) for classification [14]. Furthermore, some experiments were performed to investigate how the accuracy of the system is affected by varying inputs.

4.1. Experimental Setup

Experiments were conducted on a DELL Precision workstation that was equipped with the following components: an Intel Xeon processor operating at 2.9 GHz, 32 gigabytes of random-access memory (RAM), an NVIDIA GeForce GTX 2080 Ti graphics processing unit with 11 gigabytes of GDDR6 memory and 4352 CUDA cores, and the Linux Ubuntu 18 operating system. PyTorch was used as the front end for the implementation of all of the programs. PyTorch is a free and open-source machine-learning library that was built in Python. For the mel spectrogram, extraction processes were generated with Python for each audio file in the database using a Hamming window with a length of 32 ms and a window step of 20 ms.

4.2. Inter-Model Classification

Inter-model source microphone classification is the task of identifying which microphone model was used to record a given audio signal when multiple microphones were used to capture the same source sound. The proposed system was tested on all microphones, and the results demonstrate its superior performance in all environments compared with the previous system, which relied on a conventional transformer. As shown in Figure 2, Swin_T outperforms AST in all environments, achieving high accuracy ranging from 97.6% to 1.0%, while AST ranges from 85.77% to 99.75%. The 'Lab' and 'Soundproof' environments are where both models perform the best, with Swin_T achieving nearly perfect scores. Conversely, the 'Stairs' and 'Class' tasks are the most challenging for the Swin_T model, resulting in the lowest scores. The most significant performance difference between the models is observed in the 'Class' and 'Garden' environments. On the 'Class' task, Swin_T outperforms AST by more than 14.21%, and on 'Garden', the difference is over 8.64%, suggesting that AST may struggle more in these settings. The accuracy of the system was negatively impacted in environments that had some form of noise, including the stairs, parking lots, and garden. Overall, the proposed Swin_T model achieves a substantial increase in classification accuracy when compared with the AST model.

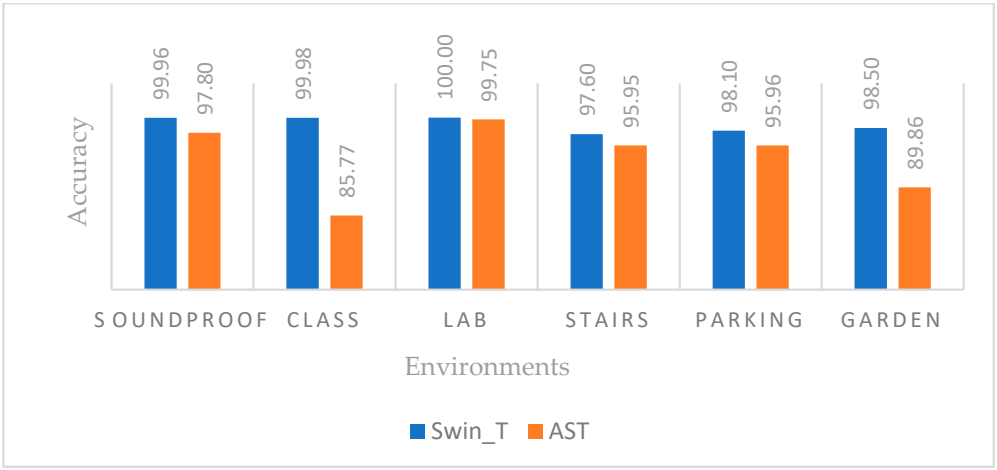

Figure 2. Classification of the performance of microphone models in different environments. Comparison among Swin_T and AST.

4.3. Intra-Model Classification

Intra-model source microphone classification is the task of identifying which specific microphone was used to record a given audio signal when multiple recordings were made using the same type of microphone. This problem is more challenging than inter-model source microphone classification because the differences between microphones of the same model are often subtle and may not be easily discernible. We performed experiments to classify the microphone model for all models in the database, with each model undergoing a separate experiment. The accuracy comparison between our Swin-T method and two other methods (AST and KNN-Gabor) is shown in Figure 3. The data reveals that the Swin_T model outperforms the other two methods in improving the accuracy of the classification of seven microphone models. In contrast, the Swin_T model accuracy is slightly lower for AKG_0451 model than the other methods. Figure 3 also illustrates the varying degrees of improvement in accuracy achieved by the Swin_T model compared with the other AST and KNN-Gabor across the seven microphones, where the microphone SHU-0058 exhibits the highest accuracy improvement, followed by the microphone ELE-0020, while the microphone AKG-3000 shows the least difference in accuracy improvement. These improvements in accuracy can be attributed to the fact that the Swin Transformer can capture multi-scale contextual information with remarkable effectiveness.

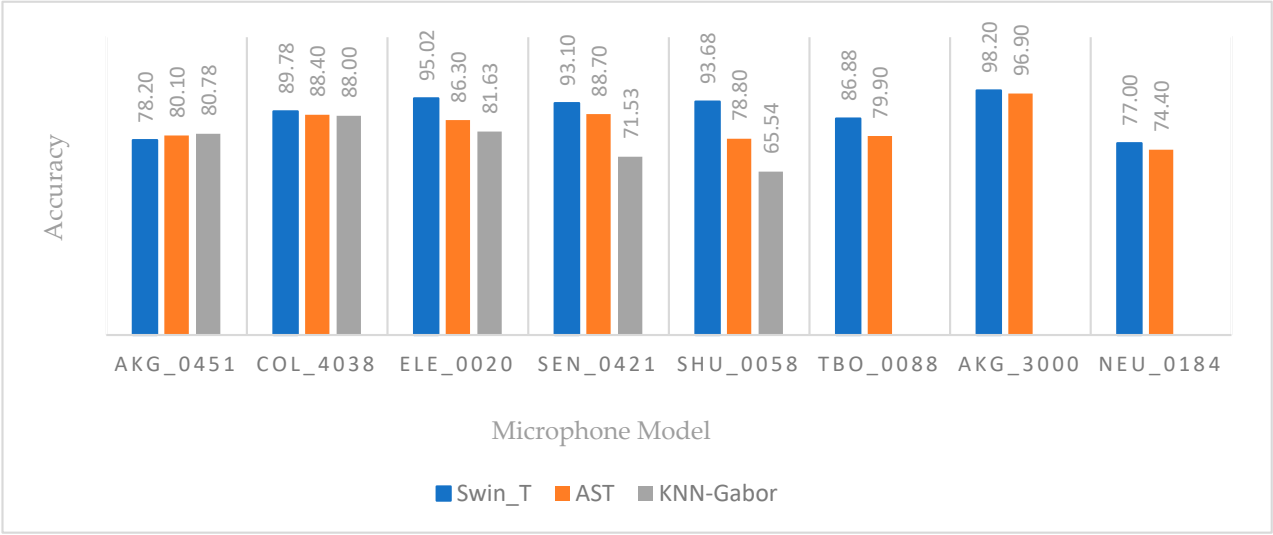

Figure 3. Classification of the intra-model performance of microphones. Comparison among Swin_T, AST, and KNN.

4.4. Investigating the Preprocessing of Inputs to Improve System Accuracy

4.4.1. Frequency Band Effect on Accuracy

This experiment aims to investigate the impact of bandwidth on classification accuracy for microphone intra-model classification. Various experiments were conducted with different frequency bands on all microphones in the database. The frequency range from 0 to 16 kHz was segmented, and the study was conducted by using a portion of the frequency spectrum while neglecting the remaining frequencies to determine the optimal frequencies that could potentially yield a microphone identification.

As shown in Table 2, the results indicate that the frequency range from 0 to 4 kHz produces the best outcomes for microphones M1, M2, M3, and M5, while the frequency range from 4 to 8 kHz yields the best results for microphones M4, M6, M7, and M8. High frequencies ranging from 8 to 12 kHz, as well as frequencies from 12 to 16 kHz, do not yield satisfactory results when compared with other frequencies. Additionally, another experiment was conducted using the entire frequency spectrum, and the system achieves better accuracy for microphones M7 and M8 compared with only using parts of the frequency spectrum.

Table 2. Impact of the frequency band on the classification accuracy with respect to the microphone model.

	Mic Model	# Mic	Frequency Band				
			(0–4) k	(4–8) k	(8–12) k	(12–16) k	All Bands (0–16) k
M1	AKG0451	2	**0.7208**	0.646	0.558	0.626	0.6752
M2	COL4038	2	**0.992**	0.697	0.637	0.603	0.976
M3	ELE0020	2	**0.8308**	0.823	0.599	0.722	0.7876
M4	SEN0421	3	0.8	**0.859**	0.747	0.711	0.8122
M5	SHU0058	3	**0.7444**	0.606	0.593	0.584	0.6874
M6	TBO0088	6	0.6002	**0.864**	0.457	0.453	0.8518
M7	AKG3000	2	0.5608	**0.783**	0.52	0.514	0.8388
M8	NEU0184	2	0.6242	**0.676**	0.346	0.304	0.7984

Bold for the best, underline for the worst.

Based on the previous experiments, it can be concluded that each microphone model may have a unique pattern that is more pronounced in a specific frequency range compared with others. In other words, it is impractical to specify a single frequency range that can be used for all microphones.

4.4.2. Input Projection Kernel Shape Effect on Accuracy

As was previously mentioned in the methodology, the input sound is converted into a spectrogram and then divided into 4×4 batches, and as it is known that the spectrogram differs from the images as there is a temporal relationship in the second axis, and accordingly, experiments were conducted to study the effect of the different shape. The batch is based on the accuracy of the classification. Experiments were conducted using different forms of batches, as shown in Figure 4. The results of experiments that were conducted on all microphones are shown in Figure 5. We note that the best results are at projection 4×4, and the results are uneven whenever the batch shape is changed from the shape on which the system was built, where the worst results are 1×16 in addition to 8×2. This implies that using frequency resolutions and more bands gives a better performance than time resolution and intervals.

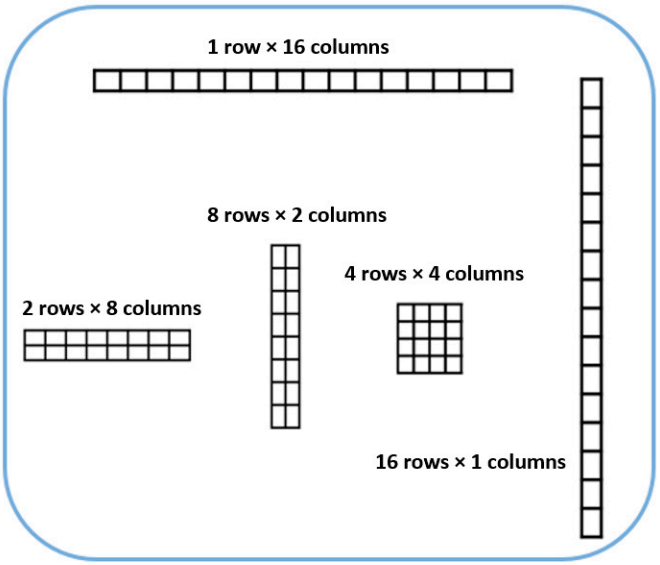

Figure 4. The shapes of the projection kernel for the input spectrogram are presented in five different shapes, denoted as (1×16), (2×8), (4×4), (8×2), and (16×1).

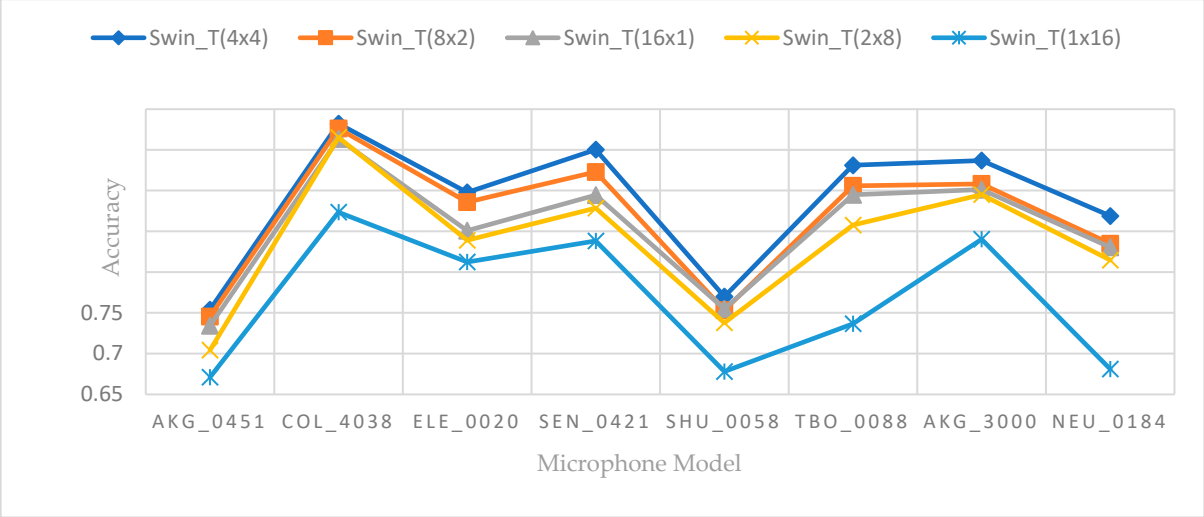

Figure 5. The effect of input projection kernel shapes on classification accuracy is investigated using five distinct shapes, which are illustrated in Figure 4.

4.4.3. Input Type Effect on Accuracy

In this experiment, the effect of changing the type of spectrogram on the system accuracy is investigated. Three different types of spectrograms were examined: spectrogram, Mel spectrogram, and linear spectrogram. As explained earlier in the methodology section on the method of calculating the Mel spectrogram, the linear spectrogram can be calculated in the same way with the difference in the filter bank where the linear scale is used instead of using the Mel scale. As for the spectrogram, it is obtained without applying any filter bank.

Figure 6, depicting the experimental results, reveals that the Mel spectrogram outperforms the other two methods in classifying all microphone models. This superiority could be attributed to the Mel spectrogram's emphasis on low frequencies over high frequencies, which is consistent with the findings of the previous section.

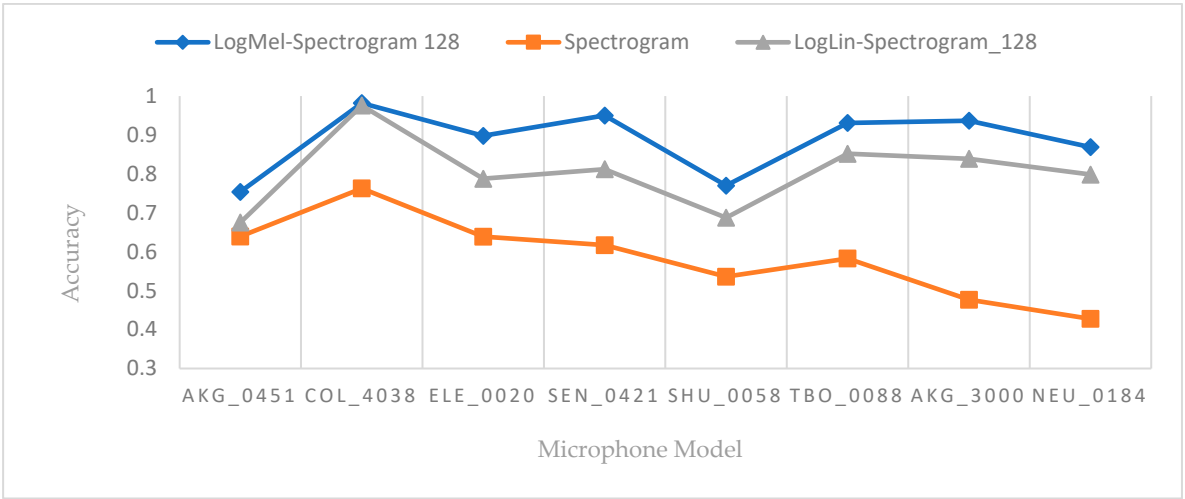

Figure 6. Impact of spectrogram type on classification accuracy.

5. Conclusions

In this paper, we introduce a Swin Transformer, deep-neural-network approach for classifying source microphones, specifically tailored for audio forensics applications. The proposed model was assessed using an audio forensic dataset. The classification process consisted of two experiments: inter-model classification, which focused on identifying the microphone manufacturer's model, and intra-model classification, which aimed to

differentiate similar microphones within the same model. In addition, experiments were conducted to study the effect of the system inputs on the accuracy of the system. The outcomes demonstrate that the proposed Swin Transformer framework effectively classifies source microphones, achieving state-of-the-art results. We recommend that future research investigate additional approaches with a comparative analysis to improve the system's accuracy. In particular, additional developments are required to further improve the accuracy of the intra-model classification performed.

Author Contributions: Conceptualization, M.Q., Y.A.A. and S.-A.S.; methodology, M.Q., Y.A.A. and S.-A.S.; software, M.Q.; validation, M.Q., Y.A.A. and S.-A.S.; resources, M.Q.; data curation, M.Q.; writing—original draft preparation, M.Q.; writing—review and editing, Y.A.A. and S.-A.S.; supervision, Y.A.A. and S.-A.S.; project administration, Y.A.A.; funding acquisition, Y.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Researchers Supporting Project number (RSP2023R322), King Saud University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Deep Blue Data (DBD) at [doi], reference number: doi.org/10.7302/Z2RJ4GCC.

Acknowledgments: This work was partially supported by the Researchers Supporting Project number (RSP2023R322), King Saud University, Riyadh, Saudi Arabia.

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

References

1. Jarosław, B. A survey on digital camera identification methods. *Forensic Sci. Int. Digit. Investig.* **2020**, *34*, 300983.
2. Manisha; Li, C.-T.; Lin, X.; Kotegar, K.A. Beyond PRNU: Learning Robust Device-Specific Fingerprint for Source Camera Identification. *Sensors* **2022**, *22*, 7871. [CrossRef] [PubMed]
3. Chiang, P.-J.; Khanna, N.; Mikkilineni, A.K.; Segovia, M.V.O.; Suh, S.; Allebach, J.P.; Chiu, G.T.-C.; Delp, E.J. Printer and scanner forensics. *IEEE Signal Process Mag.* **2009**, *26*, 72–83. [CrossRef]
4. Khanna, A.N.; Mikkilineni, A.K.; Chiu, G.T.-C.; Jan, P.; Delp, E. Survey of Scanner and Printer Forensics at Purdue University. In *International Workshop on Computational Forensics*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 22–34.
5. Qamhan, M.A.; Altaheri, H.; Meftah, A.H.; Muhammad, G.; Alotaibi, Y.A. Digital Audio Forensics: Microphone and Environment Classification Using Deep Learning. *IEEE Access* **2021**, *9*, 62719–62733. [CrossRef]
6. Dong, S.; Wang, P.; Abbas, K. A survey on deep learning and its applications. *Comput. Sci. Rev.* **2021**, *40*, 100379. [CrossRef]
7. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
8. Alashban, A.A.; Qamhan, M.A.; Meftah, A.H.; Alotaibi, Y.A. Spoken Language Identification System Using Convolutional Recurrent Neural Network. *Appl. Sci.* **2022**, *12*, 9181. [CrossRef]
9. Ali, M.; Mustafa, Q.; Yasser, S.; Yousef, A.; Sid Ahmed, S. King Saud University Emotions Corpus: Construction, Analysis, Evaluation, and Comparison. *IEEE Access* **2021**, *9*, 54201–54219.
10. Courtial, A.; Touya, G.; Zhang, X. Constraint-Based Evaluation of Map Images Generalized by Deep Learning. *J. Geovisualization Spat. Anal.* **2022**, *6*, 13. [CrossRef]
11. Shirani, K.; Solhi, S.; Pasandi, M. Automatic Landform Recognition, Extraction, and Classification using Kernel Pattern Modeling. *J. Geovisualization Spat. Anal.* **2023**, *7*, 2. [CrossRef]
12. Xia, Q.; Dong, S.; Peng, T. An Abnormal Traffic Detection Method for IoT Devices Based on Federated Learning and Depthwise Separable Convolutional Neural Networks. In Proceedings of the 2022 IEEE International Performance, Computing, and Communications Conference (IPCCC), Austin, TX, USA, 11–13 November 2022; pp. 352–359. [CrossRef]
13. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. In Proceedings of the ACL 2019—57th Annual Meeting of the Association for Computational Linguistics 2020, Florence, Italy, 28 July–2 August 2020; pp. 2978–2988.
14. Qamhan, M.; Alotaibi, Y.A.; Selouani, S.A. Transformer for authenticating the source microphone in digital audio forensics. *Forensic Sci. Int. Digit. Investig.* **2023**, *45*, 301539. [CrossRef]
15. Hua, G.; Zhang, H. ENF Signal Enhancement in Audio Recordings. *IEEE Trans. Inf. Forensics Secur.* **2019**, *15*, 1868–1878. [CrossRef]

16. Gerazov, B.; Kokolanski, Z.; Arsov, G.; Dimcev, V. Tracking of electrical network frequency for the purpose of forensic audio authentication. In Proceedings of the 2012 13th International Conference on Optimization of Electrical and Electronic Equipment (OPTIM), Brasov, Romania, 24–26 May 2012; pp. 1164–1169.
17. Zakariah, M.; Khan, M.K.; Malik, H. Digital multimedia audio forensics: Past, present and future. *Multimed. Tools Appl.* **2018**, *77*, 1009–1040. [CrossRef]
18. Buchholz, R.; Kraetzer, C.; Dittmann, J. Microphone Classification Using Fourier Coefficients. In *Information Hiding*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 235–246. [CrossRef]
19. Zhang, Y.; Luo, D. Audio source identification based on residual network. In Proceedings of the Second International Symposium on Computer Technology and Information Science (ISCTIS 2022), Guilin, China, 10–12 June 2022.
20. Zeng, C.; Feng, S.; Wang, Z.; Wan, X.; Chen, Y.; Zhao, N. Spatio-Temporal Representation Learning Enhanced Source Cell-phone Recognition from Speech Recordings. *arXiv* **2022**, arXiv:2208.12753.
21. Baldini, G.; Amerini, I. An Evaluation of Entropy Measures for Microphone Identification. *Entropy* **2020**, *22*, 1235. [CrossRef] [PubMed]
22. Luo, D.; Korus, P.; Huang, J. Band Energy Difference for Source Attribution in Audio Forensics. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2179–2189. [CrossRef]
23. Li, Y.; Zhang, X.; Li, X.; Zhang, Y.; Yang, J.; He, Q. Mobile Phone Clustering From Speech Recordings Using Deep Representation and Spectral Clustering. *IEEE Trans. Inf. Forensics Secur.* **2017**, *13*, 965–977. [CrossRef]
24. Eskidere, O. Identifying acquisition devices from recorded speech signals using wavelet-based features. *Turkish J. Electr. Eng. Comput. Sci.* **2016**, *24*, 1942–1954. [CrossRef]
25. Zou, L.; He, Q.; Wu, J. Source cell phone verification from speech recordings using sparse representation. *Digit. Signal Process* **2017**, *62*, 125–136. [CrossRef]
26. Li, Y.; Zhang, X.; Li, X.; Feng, X.; Yang, J.; Chen, A.; He, Q. Mobile phone clustering from acquired speech recordings using deep Gaussian supervector and spectral clustering. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2137–2141. [CrossRef]
27. Eskidere, O.; Karatutlu, A. Source microphone identification from speech recordings based on a Gaussian mixture model. *Turk. J. Electr. Eng. Comput. Sci.* **2014**, *22*, 754–767. [CrossRef]
28. Jiang, Y.; Leung, F.H.F. Source Microphone Recognition Aided by a Kernel-Based Projection Method. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 2875–2886. [CrossRef]
29. Zou, L.; He, Q.; Yang, J.; Li, Y. Source cell phone matching from speech recordings by sparse representation and KISS metric. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2079–2083. [CrossRef]
30. Cuccovillo, L.; Giganti, A.; Bestagini, P.; Aichroth, P.; Tubaro, S. Spectral Denoising for Microphone Classification. In Proceedings of the 1st International Workshop on Multimedia AI against Disinformation, Newark, NJ, USA, 27–30 June 2022.
31. Pavlovic, M.; Kupusinac, A.; Popovic, M. Classification model for microphone type recognition. *arXiv* **2019**, arXiv:1906.09958.
32. Qin, T.; Wang, R.; Yan, D.; Lin, L. Source Cell-Phone Identification in the Presence of Additive Noise from CQT Domain. *Information* **2018**, *9*, 205. [CrossRef]
33. Qi, S.; Huang, Z.; Li, Y.; Shi, S. Audio recording device identification based on deep learning. In Proceedings of the IEEE International Conference on Signal and Image Processing (ICSIP), Beijing, China, 13–15 August 2016; pp. 426–431. [CrossRef]
34. Kurniawan, F.; Rahim, M.; Khalil, M.S.; Khan, M.K. Statistical based audio forensic on identical microphones. *Int. J. Electr. Comput. Eng.* **2016**, *6*, 2211.
35. Baldini, G.; Amerini, I. Microphone Identification based on Spectral Entropy with Convolutional Neural Network. In Proceedings of the 2022 IEEE International Workshop on Information Forensics and Security (WIFS), Shanghai, China, 12–16 December 2022; pp. 1–6. [CrossRef]
36. Zeng, C.; Zhu, D.; Wang, Z.; Wu, M.; Xiong, W.; Zhao, N. End-to-end Recording Device Identification Based on Deep Representation Learning. *arXiv* **2022**, arXiv:2212.02084.
37. Berdich, A.; Groza, B.; Levy, E.; Shabtai, A.; Elovici, Y.; Mayrhofer, R. Fingerprinting Smartphones Based on Microphone Characteristics from Environment Affected Recordings. *IEEE Access* **2022**, *10*, 122399–122413. [CrossRef]
38. Lin, X.; Zhu, J.; Chen, D. Subband Aware CNN for Cell-Phone Recognition. *IEEE Signal Process Lett.* **2020**, *27*, 605–609. [CrossRef]
39. Baldini, G.; Amerini, I.; Gentile, C. Microphone Identification Using Convolutional Neural Networks. *IEEE Sensors Lett.* **2019**, *3*, 6001504. [CrossRef]
40. Baldini, G.; Amerini, I. Smartphones Identification Through the Built-In Microphones with Convolutional Neural Network. *IEEE Access* **2019**, *7*, 158685–158696. [CrossRef]
41. Khan, M.K.; Zakariah, M.; Malik, H.; Choo, K.-K.R. A novel audio forensic data-set for digital multimedia forensics. *Aust. J. Forensic Sci.* **2017**, *50*, 525–542. [CrossRef]