Signal, Image and Video Processing (2023) 17:2211–2219
https://doi.org/10.1007/s11760-022-02436-4

ORIGINAL PAPER

Mel spectrogram-based audio forgery detection using CNN

Arda Ustubioglu[2] · Beste Ustubioglu[1] · Guzin Ulutas[1]

Received: 5 July 2022 / Revised: 23 October 2022 / Accepted: 4 December 2022 / Published online: 19 December 2022

Abstract
In this time of technology, digital speech can be created and falsified by a very diverse of hardware and software technologies. Audio copy-move forgery is an audio forgery technique that goals to create forged audio by hiding undesirable words or repeating wanted words in identical speech. Therefore, audio authentication has been a necessary requisition. In this study, an effective approach to spectral images based on audio copy-move forgery detection using convolutional neural networks (CNN) with data augmentation is proposed. There are only a few handcrafted methods conducted for the detection of audio copy-move forgery. None of the existing works on audio copy-move forgery detection has proposed deep feature learning from speech recording with Mel spectrogram. This is the first method to employ deep learning with Mel spectrogram of audio for the detection of audio copy-move forgery. The proposed CNN architecture classifies the suspicious Mel spectrogram images into two classes: original and forged. The proposed CNN system is successfully trained on these Mel spectrogram image feature extraction. The proposed algorithm has been tested on our datasets generated from Arabic Speech Corpus and TIMIT speech database. The results show the effectiveness, robustness of post-processing operations, and high accuracy of the proposed approach compared to other studies.

Keywords Copy-move forgery detection · Audio forgery · Audio forensic · Spectrogram-CNN

1 Introduction

The use of speech recordings in daily life has increased with the improvement of recent free-developed audio editing software like Adobe, and Audition CC. Thanks to the easy use of such software, even an ordinary user can easily create forged audio. Copy-move forgery process is one of the frequently used forged audio generation methods. The attacker copies one or more segments of the speech and pastes this segment or segments into different positions within the identical audio to create forged speech with copy-move forgery. As a result of this operation, the semantic meaning of the speech changes. Moreover, these speech recordings can be submitted to the court as digital proof. For example, the sentence "I will not say any more, I'm an offender." can be modified to "I will not say any more, I'm not the offender." by copying and pasting the word "not" said by the identical talker in the identical audio recording. Since speech records are also used as digital evidence, the determination of the authentication of these records has made this area active research. Because speech records are digital evidence, it is very important to authenticate these records.

In the literature, there are a lot of methods to detect audio forgery types such as splicing, resampling, deletion, re-quantizing, insertion, copy-move, and re-compression. However, among these methods, the number of methods detecting audio copy-move forgery is quite limited. This is mainly because it is usually unnoticeable because the forged segments are reproduced from the identical audio file. A malicious user can easily create a forged speech recording with copy-move forgery by copying some segments in the speech recording and pasting these segments into different parts of the same speech to change the original content of the speech recording. At the same time, the attacker applies post-processing operations such as adding noise, filtering, and compression to the forged speech recording to hide the

✉ Beste Ustubioglu
bustubioglu@ktu.edu.tr

Arda Ustubioglu
ardaustubioglu@trabzon.edu.tr

Guzin Ulutas
gulutas@ktu.edu.tr

[1] Department of Computer Engineering, Karadeniz Technical University, 61080 Trabzon, Turkey

[2] Department of Management Information Systems, Trabzon University, Trabzon, Turkey

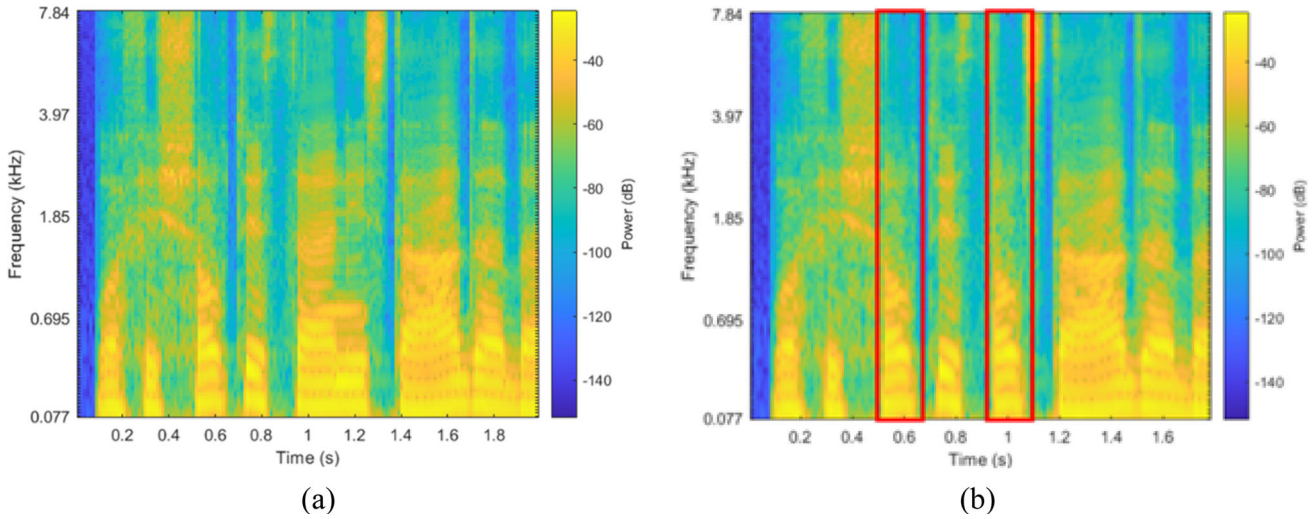

Fig. 1 **a** Original, **b** forged spectrogram image

traces of copy-move forgery. Figure 1 shows an example of a copy-move forgery operation. The spectrogram image of the forged speech (obtained from audio file ARANORM 0006.wav taken from the Arabic Speech Corpus database) is given in Fig. 1a, b shows the original speech spectrogram. The red arrows on the spectrogram indicate the start and end of the copy-pasted segments.

In this paper, we present a novel ACMFD methodology based on deep neural learning using CNN with data augmentation to detect an audio copy-move forgery in a fast and robust manner. The proposed method generates a Mel spectrogram color image from an input audio signal as a feature. Afterward, these spectrogram images are classified into two classes: forged and original with the developed CNN classification system. The development CNN system extracts spectrogram features and generates feature maps. The system finds the feature correspondences and dependencies using the average of the generated feature maps. After the developed CNN system is trained, suspicious audio files given as input are classified and labeled as fake or original. In addition, since there is no general database in the literature, in this study, an audio copy-move database was created using TIMIT and Arabic Speech Corpus databases to test the performance of the proposed methodology.

The major contributions of this article include:

1. This method uses Mel spectrogram features. The above methodologies all use speech directly to detect audio copy-move forgery. Our approach extracts feature from the speech in the form of spectrogram images.
2. In the literature, there are only a few handcrafted methods conducted for the detection of audio copy-move forgery. To the best of our knowledge, none of the existing works on audio copy-move forgery detection has proposed deep feature learning from speech recording. This is the first study to employ deep learning with Mel spectrogram of audio for the detection of audio copy-move forgery.
3. A copy-moved forged audio database is also created from two different speech databases in this study to evaluate the performance of the proposed method.
4. It was seen from the experimental results that the proposed method achieved superior accuracy rates for original and post-processed forged audios than other detection methods [1–5].

The rest of the paper is arranged as follows: Sect. 2 presents related works. The proposed approach is given in Sect. 3. Experiments and results are reported in Sect. 4. Finally, in Sect. 5, the study is concluded.

2 Related work

In the literature, many studies have been carried out in the field of detecting various audio forgeries such as resampling, splicing, deletion, insertion, re-quantizing, copy-move, and re-compression and the field of audio authentication has attracted great attention. Pan et al. [6] detected audio splicing forgery utilizing extraordinary differences in the local noise levels. Chen et al. [7] used DWT and examined singularity points of speech recording to detect audio forged operations in the time domain like insertion, deletion, substitution, and splicing. Gupta et al. [8] suggested a content-based method to detect audio copy-move forgery. This method compares the query and the test fingerprints and calculates a number of matching fingerprints. In the literature, studies in the field of splicing audio forgery detection are more than the detection of other audio forgery types, because it is relatively easier to

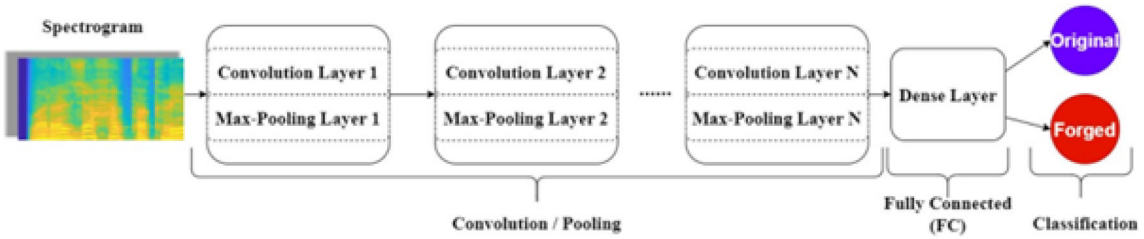

Fig. 2 The structure of the proposed algorithm

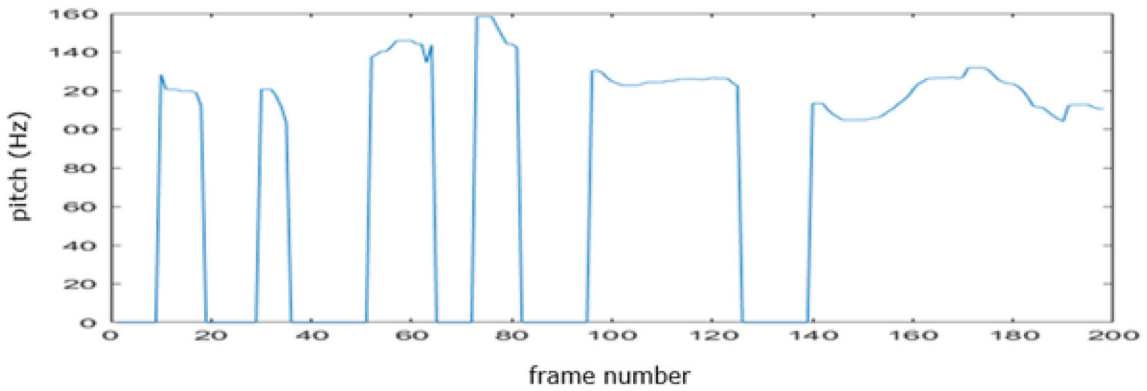

Fig. 3 Pitch sequence of original voiced parts speech

detect recordings of different environments [9, 10], microphones classification [11, 12] and speakers' identification [13–15].

Other studies are also available in the literature [6, 10, 16, 17]. Studies in the field of audio copy-move forgery are relatively few compared to other studies about audio forgeries. Xiao [18] split the speech into segments according to the constant time. This method calculated the similarity between all segments in a speech recording to detect duplicate segments. Huang's method [1] extracted voiced parts of audio utilizing a voice activity detection method. They obtained features from voiced parts with discrete Fourier transform (DFT) and Mel-frequency cepstral coefficients (MFCC) methods. After their method sorted all voiced parts according to the features, they compared one voice part with some adjacent parts instead of all parts in the sorted list to reduce the time complexity of the algorithm. Wang et al. [2] suggested a method using singular value decomposition (SVD) transform to detect audio copy-move forgery. They obtained voiced parts from speech recording using a voice activity detection (VAD) method. Their method extracted features from these voiced parts with discrete cosine transform (DCT). Xie et al. [19] proposed a method to detect audio copy-move forgery with a multi-feature decision. They obtain four features: gammatone, Mel-frequency cepstral coefficients (MFCCs), pitch, and discrete Fourier transform (DFT) coefficients. Their method fused the detection result of these four features with the C4.5 decision tree. Imran et al. [3] extracted features from the voiced parts after extracting the voiced parts with their proposed VAD algorithm. They utilized 1-D local binary patterns (LBP) to obtain features. Afterward, their method compared all LBP histograms to detect duplicated parts that have similar histograms. Yan [4] utilized a pitch tracking method YAAPT [20] to obtain voiced parts and computed the similarities of pitch sequences to detect audio copy-move forgery. In another study by the same authors, Yan et al. [5] utilized pitch and formant sequences. They split the speech into voiced and unvoiced parts, then obtained the pitch and formant sequences of voiced parts as features. Finally, it computed the similarities of each feature set with dynamic time warping (DTW) to detect duplicated parts.

In the literature, besides audio authentication, there are many studies in the field of audio speaker verification (ASV). The main objective of an ASV system is to detect whether the input audio is original or not. The methods in the literature consist of two stages to detect forged audio: Feature extraction and classification. In the feature extraction phase, existing approaches have employed either handcrafted features [21–23] or deep learning-generated features [24–27].

3 Proposed approach

The aim of this study is to build up an effective deep ACMFD (Audio Copy-Move Forgery Detection) method that can obtain high performance with noticeably low cost to

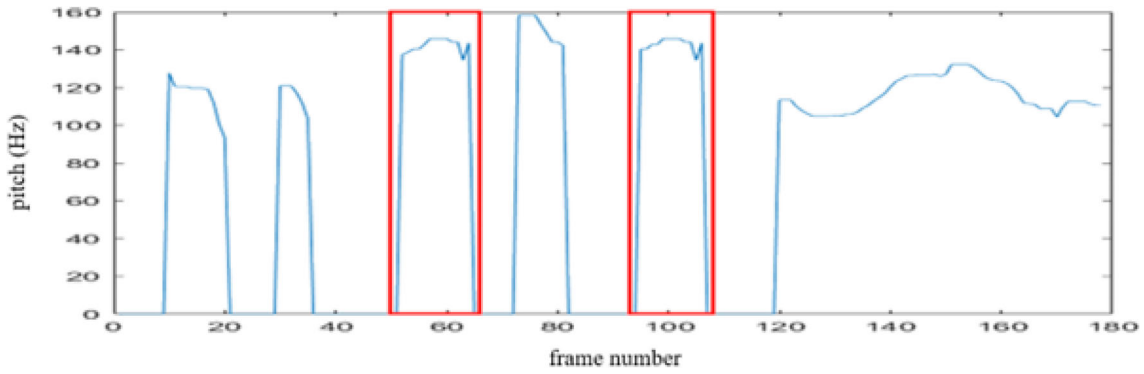

Fig. 4 The pitch sequence of forged speech

detect audio copy-move forgery. The proposed algorithm which uses the CNN model given in Fig. 2 consists of two stages: creation of the forged database and detection of audio copy-move forgery. CNN, which was developed based on a multi-layer perceptron network structure, has been used effectively in many fields in the literature and has become quite popular. A CNN model contains several layers. These layers are convolution, pooling, normalization, and fully connected layers.

The proposed method consists of two stages as stated before. While one of them is used to create a forged database, the other one detects copied and pasted parts on the audio file. In the first stage, speech taken from the Arabic speech corpus database is segmented into voiced and unvoiced parts with a pitch extraction method. Afterward, a forged speech is created by choosing one of the speech segments at random and pasting it on another randomly chosen segment. In the copy-move forgery detection stage of the proposed method, whether the speech is forged or not is determined by using the proposed CNN architecture. The spectrogram is used by the method to transform the audio file into an image with its spatial frequency representation. Specially designed CNN architecture gets the spectrogram image as input and decides the originality of the audio file. Details of these stages are given below.

3.1 Creation of the forged database

In this study, an audio forgery database is created to measure the performance of our study and to compare the study with another method about audio forgeries in the literature.

Speeches that are taken from the Arabic Speech Corpus [28] and TIMIT [29] databases are segmented into voiced and unvoiced parts at first to create audio forgery database. Details of original speech recordings in the Arabic Speech Corpus and TIMIT database are presented in Sect. 4. The proposed method extracts a pitch sequence from the audio for the segmentation of voiced and unvoiced parts. Pitch is a measure that refers to the fundamental frequency and represents the vibration frequency of the vocal utterance. Even if the same word is said twice in a speech, the pitch sequences of these repeated words will not be the same [5]. The proposed approach utilized from the YAAPT algorithm for segmentation of the audio file and the details of it are given below.

After extracting the pitch sequence from the speech with the YAAPT method, places with frequency values greater than zero in the pitch sequence are marked as the voiced part and the places with zero frequency are marked as the unvoiced part in the speech. Figure 3 shows the pitch sequence of audio shown in Fig. 1a. The dashed lines show the boundaries of the audio.

One of the voiced segments in the speech is copied and pasted in place of another voiced segment to create a forged speech. Two numbers (i, j) are randomly selected between 1 and N. N represents a total number of vocal segments in the speech. The first of these numbers, S_i, represents the voiced segment to be copied, and the second one denoted by S_j corresponds to the voiced segment to which the copied segment will be pasted. The S_j is deleted before the S_i is copied and pasted. Applying this process to all original speech recordings creates forged speeches and finally forged databases are created by using forged speeches. The pitch sequence of forged speech given in Fig. 1b is given in Fig. 4. The dashed lines show the duplicated regions in the forged audio.

3.2 Detection of audio copy-move forgery

The proposed algorithm uses the CNN model given in Fig. 2 to detect audio copy-move forgery. The deep learning-based method is proposed to detect whether the suspicious audio file is forged instead of handcrafted feature extraction methods. If the handcrafted feature extraction method were used, a method would have to be found that would be robust to all post-processing operations to hide any traces of forgery in forged audio. At the same time, since the selected feature extraction method will be applied to each frame after the

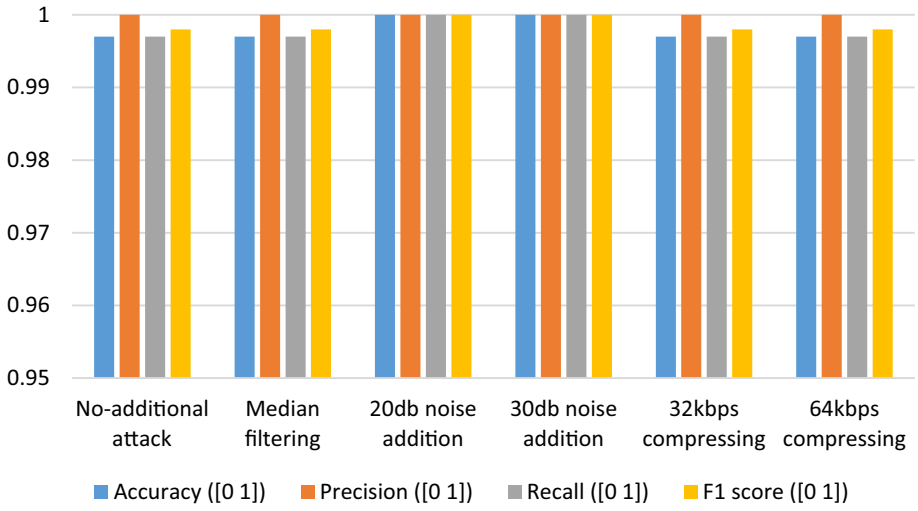

Fig. 5 The testing result of the proposed method on unseen data (generated from the TIMIT database)

Table 1 The result of proposed CNN system terms of accuracy ([0 1]]), precision ([0 1]]), recall([0 1]]), F1 score([0 1]]) and AUC([0 1]]) with/without data augmentation

tenfold	Accuracy	Precision	Recall	F1 score	AUC
Without data augmentation	0.95	0.91	0.94	0.98	0.96
With data augmentation	0.99	0.96	0.97	0.99	0.98

audio is divided into overlapping frames, the execution time and processing overhead of the proposed algorithm would increase. For this reason, a deep learning-based method has been developed to ensure that the proposed method is both robust to post-processing operations and fast and has low processing overhead.

It is a non-block-based algorithm, which works on the spectrogram image. Extracting features and classifying various speech records through short audio clips are not easy. Many speech recordings have background noises, very short intervals, and fast changes in the recordings. Those noises and attributes make it very difficult for the CNN model to classify them. The proposed method works on the spectrogram image instead of the audio. For these reasons, it is a very rare approach to convert audio files into spectrogram images for the detection of audio copy-move forgery. The spectrogram, which is the visual representation of audio, has two dimensions representing frequency and time on the vertical and horizontal axes, respectively. The amplitude of the signal at a given frequency at a specific time is represented by the color intensity in the spectrogram. For example, while the light blue in the spectrogram shows the lowest amplitude, the highest amplitude is displayed by dark red.

In the study, the short-time fast Fourier transform (STFT) was utilized to create the spectrogram of the speech signal. Firstly, the audio signal is first splitted into 30 ms frames and each frame is multiplied by the Hamming window. Afterward, FFT is applied to all subframes according to (1).

$$S(f,\, t) = \sum_{n=0}^{N-1} w_n x_t(n) \exp(-j2\pi(f/f_s)n) \tag{1}$$

$$f = kf_s/N$$

where w_n is the Hamming window function, x_n is the original speech signal, f is the frequency range for $k = 1, 2, \ldots N/2 + 1$ and$N$ refers to the number of samples in the frame.

Mel spectogram is STFT filtered with a Mel-frequency filter bank. Mel spectogram coefficients $S_{mel}(k,\, t)$ are calculated according to (2).

$$S_{mel}(k,\, t) = \sum_{l=0}^{L-1} m_k(l) |S(l,\, t)^2| \tag{2}$$

where L is the frequency component number and $m_k(l)$ refers to the k filter of the mel filter bank.

After creating the forged and original spectograms obtained from the speech signals received from the input, a training and test set was created from these spectograms for the proposed CNN architecture. However, due to the large number of parameter adjustments in CNN, a large amount of data is needed during the training stage. If an insufficient amount of data is used in training, there will be inconsistency between training accuracy and test accuracy which results in overfitting problem. For this reason, data augmentation technique was also used during creation of training set in the proposed method. With the data augmentation technique, which has become a very common approach in deep learning

Table 2 The result of proposed method with post-processing terms of accuracy ([0 1]), precision ([0 1]), recall ([0 1]), F1 score([0 1]) and AUC ([0 1]) on forged database generated from Arabic Speech Corpus

	Accuracy	Precision	Recall	F1-score
Median filtering	0.91	0.87	0.93	0.9
64kbps compressing	0.9	0.88	0.93	0.91
20db noise adding	0.94	0.88	1	0.94
30db noise adding	0.93	0.87	1	0.93

Table 3 The result of proposed method with post-processing terms of accuracy ([0 1]), precision ([0 1]), recall ([0 1]), F1 score([0 1]) and AUC ([0 1]) on all forged database generated from Arabic Speech Corpus and TIMIT

	Accuracy	Precision	Recall	F1-score
No additional attack	0.98	0.97	0.98	0.99
Median filtering	0.98	0.96	0.99	0.98
64 kbps compressing	0.98	0.96	0.99	0.98
32 kbps compressing	0.98	0.95	0.99	0.97
20 db noise adding	0.99	0.96	1	0.98
30 db noise adding	0.99	0.96	0.99	0.97

Table 4 Comparison of the accuracy of the proposed approach with other existing published studies on database generated from TIMIT

	Lbp [3]	Dft [1]	Formant [5]	DCT-SVD [2]	Pitch-sim [4]	Proposed
No additional attack	0.18	0.19	0.23	0,22	0.37	0.99
Median filtering	0.1	0.21	0.28	0,22	0.36	1
20 db noise adding	0.17	0.1	0.37	0,25	0.33	1
30 db noise adding	0.19	0.16	0.31	0,29	0.35	1
32 kbps compressing	0.2	0.2	0.31	0,29	0.31	0.99
64 kbps compressing	0.2	0.19	0.3	0,3	0.34	0.99

systems, more training data is generated and the overfitting problem in the training phase can be prevented.

In this study, frequently used image data augmentation techniques were used on spectrogram images using Kera's package. Augmentation operations applied to the spectrogram images are Width shift: 0.5, Height shift: 0.5, Fill mode: nearest, and Shear range: 0.30. By applying these augmentation operations to spectogram images, desired number of images can be produced.

After the number of spectrogram images is increased by data augmentation operations, spectrogram images are given to the proposed CNN architecture. For this, firstly, spectrogram images are divided into two classes as fake and original, and a training set is created. As a result of the CNN training, the label of the suspicious test speech is predicted. Training process stops after 140 epochs, the Adam optimizer and the categorical cross-entropy loss function were preferred, and the batch size for the dataset was selected to 64.

Table 5 Comparison of the accuracy of the proposed approach with other existing published studies on database generated from Arabic Speech Corpus

	Lbp [3]	Dft [1]	Formant [5]	DCT-SVD [2]	Pitch-sim [4]	Proposed
20 db noise adding	0.5	0.2	0.63	0.6	0.5	0.94
30 db noise adding	0.4	0.16	0.6	0.4	0.37	0.93
64 kbps compressing	0.3	0.2	0.57	0.3	0.3	0.9
Median filtering	0.1	0.2	0.57	0.4	0.3	0.91

4 Experimental results

This part presents an involved analysis of the results obtained with our method. Moreover, the proposed method was compared with other studies in this field in the literature to show the efficiency of the proposed method. The experiments have been performed using Python 3.5 and Keras with TensorFlow backend toolkits on a machine with Intel Core i5, 64 bits processor, 8 GB RAM, operating by Windows 10.

4.1 Dataset

In this study, the Arabic Speech Corpus and TIMIT database is utilized to produce the audio copy-move forgery databases. While creating the forged audio file, a random segment was taken in the speech and this segment was pasted on the segment at a random position in the same speech. The duration of each repeated segment thus formed is approximately between 0.2 and 0.6 s. To show the efficiency of our method, 368 forged audio files from the TIMIT database and 1329 forged audio files from the Arabic Speech Corpus are generated. As a result, forged audio files are produced. (created database is available at https://drive.google.com/file/d/1i9TmYy_tZw6PUBrVN8oO7KVAJqe14r93/view?usp=sharing).

4.2 Results with data augmentation

In this part is given obtained evaluation results for the proposed deep ACMFD algorithm. The proposed algorithm has been applied to the Arabic Speech Corpus database. Several experiments have been carried out using data augmentation to increase testing accuracy.

In the proposed method, firstly, the k-fold cross-validation method was applied to the training set without data augmentation. This technique was used to fully test the dataset and make an effective evaluation. When applying k-fold cross-validation to the dataset, the dataset is randomly divided into k groups of approximately equal size. Afterward, the proposed model is trained each time with the $(k-1)$ group and tested with the remaining group as well. This process is repeated k times. We firstly investigated the effect of data augmentation. For this purpose, the proposed algorithm was trained for tenfold cross-validation before applying data augmentation.

The obtained results are shown in Table 1. As can be seen in the table, the proposed method with a data augmentation approach for Mel spectrogram-based image classification gives the best performance according to accuracy, prediction, recall, $F1$-score, and AUC.

4.3 Results with post-processing operations

After the forged speech is created, various post-processing operations such as noise adding, compression and filtering are applied to the forged audio to remove traces of forgery. For this purpose, post-processing operations were applied to the test dataset to present the robustness of the proposed method to post-processing operations. We applied to the forged audios from the Arabic Speech Corpus and TIMIT databases some commonly used post-processing operations to create the audio copy-move forgery databases with post-processing. These post-processing operations are adding noise (30 dB and 20 dB white Gaussian noise), filtering (median filter), and Mp3 compression (32 kbps and 64 kbps). As a result, our copy-move forgery dataset contains a total of 1960 forged speeches with post-processing operations. Table 2 presents the result of the proposed method in terms of accuracy, precision, recall and F1-score on a forged database generated from Arabic Speech Corpus. (Prepared database is available at https://drive.google.com/drive/u/1/folders/1IilbrbxJG3SDqceo4K0DlLL95zZ4k9oc).

As can be seen from Table 2, the proposed method is highly robust to noise addition, compression, and median filtering post-processing operations. The $F1$-score and accuracy values of the proposed method are obtained 0.94 because of the noise post-processing operation. This shows that the proposed method is more resistant to noise attacks than other attacks. We also experiment by training and testing on different datasets. Our cross-dataset experiments are conducted with our copy-move forgery datasets generated from Arabic Speech Corpus and TIMIT datasets.

We train our proposed approach on Arabic Speech Corpus and test on TIMIT. The results in terms of accuracy, precision, recall and $F1$-score of the proposed method on forged database generated from TIMIT are provided in Fig. 5.

The experimental results presented in Fig. 5 show the robustness of our method for unseen forged audios in the training phase. At the same time, looking at the figure, the metric results obtained for all post-processing operations are 0.99 and above. This shows that the method is also very robust to post-processing operations.

The results of the proposed method on all databases (forged databases created from Arabic Speech Corpus and TIMIT) are given in Table 3.

When the results in Table 3 are examined, the accuracy and recall values of the proposed method are above 0.98, the precision value is above 0.95, and the $F1$-score value is above 0.97.

4.4 Comparison with traditional results

In this part, we compared the proposed method with other studies in this field in the literature to present the efficiency of our method. The results obtained as a result of the proposed method were compared with the results of Lbp method [3], pitch similarity method [4], formant method [5], DFT method [1] and DCT-SVD [2]. The proposed method and

other studies were coded on the same machine and tested in the audio database created by us. Table 4 summarizes the results of the comparison with post-processing operations on the database generated from TIMIT. When the results were examined, the accuracy values of the proposed algorithm in all post-processing operations were 0.99 and above.

Table 5 shows the results of the comparison with post-processing operations on database generated from the Arabic Speech Corpus. As can be seen from the table, the accuracy values of the studies in the literature cannot exceed 0.63, while the accuracy values obtained from the proposed method are 0.90 and above.

The results show that the proposed ACMFD algorithm with deep learning on database generated from TIMIT and Arabic Speech Corpus shows the highest performance in comparison with those of the other ACMFD algorithms.

5 Conclusion

In this paper, we present a new ACMFD method with deep neural networks. The motivation of this method is to build up a CNN architecture to categorize the suspicious Mel spectrogram images into two classes: original and forged. The proposed CNN model is successfully trained on these Mel spectrogram images feature extraction for audio copy-move forgery detection. The method also utilized data augmentation for the proposed architecture. We tested the proposed model with our datasets generated from Arabic Speech Corpus and TIMIT speech database. Several experiments have been conducted using data augmentation, cross-dataset, and post-processing operations. The obtained results showed that the best performance was obtained using data augmentation. For datasets with post-processing operations generated from Arabic Speech Corpus and TIMIT database, the best accuracies are 0.99 and 0.98, respectively, with no additional attack. When the results of other studies in the literature are examined, it will be seen that the proposed method gave the best and highest results in our datasets.

Acknowledgements Not applicable.

Author contributions AU contributed to conceptualization, methodology, and software. BU contributed to data curation, writing—original draft preparation, and software. GU contributed to visualization, investigation, and methodology.

Funding This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) with Project No: 122E013.

Availability of data and materials Not applicable.

Declarations

Competing interests The authors declare the following financial interests/personal relationships which may be considered as potential competing interests.

Consent for publication Not applicable.

Ethics approval and consent to participate The authors declare that this article is original, has not been published before, and is not currently being considered for publication elsewhere. The authors confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. The authors further confirm that the order of authors listed in the manuscript has been approved by all of them.

References

1. Huang, X., Liu, Z., Lu, W., Liu, H., Xiang, S.: Fast and effective copy-move detection of digital audio based on auto segment. In: Digital Forensics and Forensic Investigations: Breakthroughs in Research and Practice, pp. 127–142 (2020)
2. Wang, F., Li, C., Tian, L.: An algorithm of detecting audio copy-move forgery based on DCT and SVD. In: 2017 IEEE 17th International Conference on Communication Technology (ICCT), pp. 1652–1657 (2017).
3. Imran, M., Ali, Z., Bakhsh, S.T., Akram, S.: Blind detection of copy-move forgery in digital audio forensics. IEEE Access **5**, 12843–12855 (2017)
4. Yan, Q., Yang, R., Huang, J.: Copy-move detection of audio recording with pitch similarity. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1782–1786 (2015)
5. Yan, Q., Yang, R., Huang, J.: Robust copy—move detection of speech recording using similarities of pitch and formant. IEEE Trans. Inf. Forensics Secur. **14**(9), 2331–2341 (2019)
6. Pan, X., Zhang, X., Lyu, S.: Detecting splicing in digital audios using local noise level estimation. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1841–1844 (2012)
7. Chen, J., Xiang, S., Huang, H., Liu, W.: Detecting and locating digital audio forgeries based on singularity analysis with wavelet packet. Multimed. Tools Appl. **75**(4), 2303–2325 (2016)
8. Gupta, V., Boulianne, G., Cardinal, P.: Content-based audio copy detection using nearest-neighbor mapping. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 261–264 (2010)
9. Muhammad, G., Alotaibi, Y.A., Alsulaiman, M., Huda, M.N.: Environment recognition using selected MPEG-7 audio features and mel-frequency cepstral coefficients. In: 2010 Fifth International Conference on Digital Telecommunications, pp. 11–16 (2010)
10. Zhao, H., Chen, Y., Wang, R., Malik, H.: Audio splicing detection and localization using environmental signature. Multimed. Tools Appl. **76**(12), 13897–13927 (2017)
11. Cuccovillo, L., Mann, S., Tagliasacchi, M., Aichroth, P.: Audio tampering detection via microphone classification. In: 2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP), pp. 177–182 (2013)
12. Buchholz, R., Kraetzer, C., Dittmann, J.: Microphone classification using Fourier coefficients. In: International Workshop on Information Hiding, pp. 235–246 (2009)

13. Luo, D., Yang, R., Li, B., Huang, J.: Detection of double compressed AMR audio using stacked autoencoder. IEEE Trans. Inf. Forensics Secur. **12**(2), 432–444 (2016)
14. Lin, X., Kang, X.: Supervised audio tampering detection using an autoregressive model. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2142–2146 (2017)
15. Zhang, Y., Dai, S., Song, W., Zhang, L., Li, D.: Exposing speech resampling manipulation by local texture analysis on spectrogram images. Electronics **9**(1), 23 (2019)
16. Capoferri, D., Borrelli, C., Bestagini, P., Antonacci, F., Sarti, A., Tubaro, S.: Speech audio splicing detection and localization exploiting reverberation cues. In: 2020 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6 (2020)
17. Jadhav, S., Patole, R., & Rege, P.: Audio splicing detection using convolutional neural network. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–5 (2020)
18. Xiao, J.N., Jia, Y.Z., Fu, E.D., Huang, Z., Li, Y., Shi, S.P.: Audio authenticity: duplicated audio segment detection in waveform audio file. J. Shanghai Jiaotong Univ. (Science) **19**(4), 392–397 (2014)
19. Xie, Z., Lu, W., Liu, X., Xue, Y., Yeung, Y.: Copy-move detection of digital audio based on multi-feature decision. J. Inf. Secur. Appl. **43**, 37–46 (2018)
20. Zahorian, S.A., Hu, H.: A spectral/temporal method for robust fundamental frequency tracking. J. Acoust. Soc. Am. **123**(6), 4559–4571 (2008)
21. Todisco, M., Delgado, H., Evans, N.W.: A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients. In: Odyssey, vol. 2016, pp. 283–290 (2016)
22. Alluri, K.R., Achanta, S., Kadiri, S.R., Gangashetty, S.V., Vuppala, A.K.: Detection of replay attacks using single frequency filtering cepstral coefficients. In: Interspeech, pp. 2596–2600 (2017)
23. Das, R.K., Yang, J., Li, H.: Long range acoustic features for spoofed speech detection. In: Interspeech, pp 1058–1062 (2019)
24. Kumar, M.G., Kumar, S.R., Saranya, M.S., Bharathi, B., Murthy, H.A.: Spoof detection using time-delay shallow neural network and feature switching. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 1011–1017 (2019)
25. Chettri, B., Stoller, D., Morfi, V., Ramírez, M.A.M., Benetos, E., Sturm, B.L.: Ensemble models for spoofing detection in automatic speaker verification (2019). arXiv preprint arXiv:1904.04589
26. Chen, T., Kumar, A., Nagarsheth, P., Sivaraman, G., Khoury, E.: Generalization of audio deepfake detection. In: Odyssey, pp. 132–137 (2020)
27. Zhang, Y., Jiang, F., Duan, Z.: One-class learning towards synthetic voice spoofing detection. IEEE Signal Process. Lett. **28**, 937–941 (2021)
28. http://en.arabicspeechcorpus.com/
29. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S.: DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon technical report n, 93, 27403 (1993)