High-Performance Fake Voice Detection on Automatic Speaker Verification Systems for the Prevention of Cyber Fraud with Convolutional Neural Networks

Ricardo Buettner
University of Bayreuth
ricardo.buettner@uni-bayreuth.de

Jan Gross
Aalen University
jan.gross@hotmail.de

Philipp Roessler
Aalen University
philipp.roessler@studmail.htw-aalen.de

Julia Winter
Aalen University
julia.winter@studmail.htw-aalen.de

Daniel Sauter
Aalen University
mail@danielsauter.de

Hermann Baumgartl
Aalen University
hermann.baumgartl@hs-aalen.de

Patrick Ulrich
Aalen University
patrick.ulrich@hs-aalen.de

Abstract

This study proposes a highly effective data analytics approach to prevent cyber fraud on automatic speaker verification systems by classifying histograms of genuine and spoofed voice recordings. Our deep learning-based lightweight architecture advances the application of fake voice detection on embedded systems. It sets a new benchmark with a balanced accuracy of 95.64% and an equal error rate of 4.43%, contributing to adopting artificial intelligence technologies in organizational systems and technologies. As fake voice-related fraud causes monetary damage and serious privacy concerns for various applications, our approach improves the security of such services, being of high practical relevance. Furthermore, the post-hoc analysis of our results reveals that our model confirms image texture analysis-related findings of prior studies and discovers further voice signal features (i.e., textural and contextual) that can advance future work in this field.

1. Introduction

Between 2013 and 2017, the rate of voice fraud increased by over 350%, while studies assume that the rate is even higher today due to the increasing use of voice-assisted technologies in various application areas. While it is estimated that voice-related fraud costs call center organizations in the United States alone $14 billion each year, the overall monetary damage is assumed to be significantly higher. As an example, criminals used artificial intelligence techniques to impersonate a CEO's voice on the phone and grifted more than $243,000 [1, 2]. Although automatic speaker verification systems enable biometric authentication for a variety of voice-related services and applications (e.g., online banking and shopping, social networking sites, telecommunication) [3], contributing to the popularity of cloud-based services such as voice-activated personal assistants in private households, the possibility of unauthorized access and control inhibits their use [4, 5], making fraud resulting from fake voices a severe security and privacy concern to today's society. Accordingly, criminal activities have grown substantially with the development and easy access to spoofing methods [6].

Various spoofing methods can be used to generate spoofed voices. They can be categorized into physical access (PA), including impersonation and replay methods, and into logical access (LA), including speech synthesis and voice conversion methods [7]. As in real-world applications, it is indeterminable which spoofing method is applied during an attack [8], the main challenge in the development of suitable countermeasures is to detect spoofed voices independently of the underlying method [9].

Furthermore, as automatic speaker verification systems are mainly used on small devices, characterized by limited performance capabilities, another challenge is that approaches for the detection of voice-related fraud must follow the design requirements of mobile and embedded applications such as smartphones or smart speakers [10].

While related studies either proposed individual models for each spoofing method (i.e., PA or LA)

URI: https://hdl.handle.net/10125/80104
978-0-9981331-5-7

[11, 12, 13], are limited to a specific language (e.g., English), require additional hardware components due to the model's size [14], or are limited to specific applications (impersonation [15, 16], replay [17, 18, 19], speech synthesis [20], or voice conversion [21]), further research is required to advance the accuracy, and universal practicability of such fake voice detection approaches.

However, prior studies have revealed the potential of using visually represented voice recordings (e.g., histograms, spectrograms) to extract meaningful features (e.g., textural and contextual) for differentiating genuine and spoofed voices. Furthermore, convolutional neural networks (CNNs) have established themselves as a widely used method in the area of voice command recognition [22, 23] and image processing tasks in general due to their superior performance and advantage of automated feature extraction (i.e., descriptive and discriminative) [24]. Inspired by the findings in the field of image texture analysis [25] and following the current research trend, we contribute to cyber security and reduce the risk and damage of fake voice-related fraud [26].

Finally, theories in cyber fraud and deception concordantly indicate that while the world is becoming increasingly reliant on computers for critical infrastructures, poor architectural decisions result in weak points that allow malicious actors to gain access to sensitive information. Furthermore, cyber-attacks can affect human decision-making by exploiting social and cognitive biases [27, 28, 29]. To counteract these problems and contribute to preventing cyber criminology in the cyber-physical space [27], we present a technical-oriented design science approach in this work [30] by investigating the following research question:

Can a CNN-based deep learning approach accurately differentiate genuine from spoofed voices by automatically analyzing histograms of audio recordings, independently of the underlying spoofing method? The most important contributions of our study are:

1. We present a highly effective data analytics approach for detecting and preventing cyber fraud with spoofed voices [31], setting a benchmark with a balanced accuracy of 95.64% and an equal error rate (EER) of 4.43%.
2. By proposing a lightweight CNN architecture that advanced the application of spoofing attacks detection approaches on embedded systems [32], we contribute to the adoption of artificial intelligence technologies in small and medium-sized enterprises [33].
3. Our CNN-based detection model is accurate, cost-efficient, and allows to objectively distinguish between genuine and spoofed voices by only using a histogram visualization of the underlying voice recording.
4. As our approach allows to lower the risk of unauthorized access to voice-activated personal assistants [4, 5], improving the security and reducing the fraud-related monetary damage [1] for a variety of services applications (e.g., banking, telecommunication) [3], our approach is of high practical relevance.
5. Using GRAD-CAM heatmaps for the post-hoc analysis of our CNN-based classification results [34], our findings confirm the outcome of prior studies. Following the findings of image texture analysis, we identified both textural and contextual [25] histogram-based features (i.e., long-term temporal dependencies [35]). Therefore, our study contributes to future research in the field of histogram-based detection of fake voices.

Our work is organized as follows: First, the research background is presented, followed by the study's methodology, including the evaluation data and preprocessing labeling procedure of the genuine and spoofed voice recordings and the applied machine learning (ML) method. Subsequently, the model's results, performance evaluation, and post-hoc analysis using GRAD-CAM heatmaps are presented. Next, the results and implications are discussed. Finally, we conclude our work, followed by limitations and suggestions for future research.

2. Research Background

According to Wu et al. [7] spoofing methods can be categorized into four groups (i.e., impersonation, replay, speech synthesis, and voice conversion). Additionally, impersonation and replay are assigned to the subordinate category of PA, whereas speech synthesis and voice conversion are assigned to the subordinate category of LA [12]. While in PA-based scenarios, a microphone captures the speech data in a physical, reverberate space, LA scenarios refer to attacks that are injected directly into the automatic speaker verification system [36].

2.1. Related Work on Method-Independent Detection of Spoofed Voices

Due to the prevalence of cyber fraud in fake voices and automatic speaker verification systems, previous research already investigated and proposed countermeasures for the detection and prevention of fake voices. Table 1 gives an overview of related work, where the corresponding performance of each approach is presented as either accuracy or EER, based on the underlying spoofing method (i.e., PA or LA).

Rahmeni et al. [37] used histograms as images for the classification based on support vector machine (SVM). By applying a local binary pattern as a feature extraction technique on a spectrogram, these images show only the texture of genuine and spoofed voices, respectively. They achieved a spoofed voice detection accuracy of 71.67% with the applied feature extraction method following their approach. Hanilci [38] used the gaussian mixture model (GMM) with maximum likelihood criterion for classification on linear prediction-based feature extraction. More precisely, he used the linear prediction residual phase cepstral coefficients and the linear prediction residual Hilbert envelope cepstral coefficients on a previously applied pre-emphasis filter for higher-frequencies. Similar to Sahidullah et al. [39], he experiences that higher frequencies are more significant for the classification of spoofed voices. His approach achieved an EER of 6.52% on the evaluation data.

Jelil et al. [40] focused on the question of how silence in audio files influences the results of spoofed voice detection. Therefore, they built three individual CNN-based systems, which handle silence in three different ways. First, they kept all silent features for the input. Second, they removed all silent features. Third, they used the voice activity detector for the extraction of features. Further, they cut each audio file into several pieces to train and test their model with the corresponding spectrogram images. In order to classify the whole audio file, the most important image from the cut-off pieces was identified and consulted for the final decision. Regarding the inspection of the silent features, the best results were obtained with the first method (i.e., keeping all silent features for the input) with an EER of 18.15%. Additionally, they proposed a baseline-designed binary classifier using GMM and constant Q cepstral coefficients as a feature extraction method. Their model achieved an EER of 8.09% on unknown LA attacks without considering their silence approach.

Monteiro and Alam [9] built five different systems to classify genuine from spoofed voices and compared their results to find the best approach. First, they built a fully connected neural network that uses low-level descriptors like linear-frequency cepstral coefficients, product spectrum-based cepstral coefficients, and constant Q transform-based cepstral coefficients as input [41]. Second, they used a linear discriminant analysis for the classification, which was based on constant Q transform-based cepstral coefficients and the universal background model [42]. Third, they used the fisher vector [43] which is also based on the low-level descriptors (cepstral coefficients). Fourth, they designed a CNN based on Light CNN-29 [39] and ResNet-18 architecture. Fifth, they used a stand-alone system based on GMM. The first method achieved the best performance on unknown attacks with an EER of 4.50% (LA) and 0.96% (PA). Finally, Monteiro et al. [44] introduced an end-to-end ensemble-based approach to detect various spoofing methods, achieving an EER of 1.75% for PA attacks and 9.87% for LA attacks.

Table 1. Related work performance (i.e., accuracy (ACC), EER) using spoofing (SM) and ML methods (MLM) for voice detection.

Source	SM	MLM	Performance
[37]	LA	SVM	71.67% (ACC)
[38]	LA	GMM	6.52% (EER)
[40]	LA	CNN	9.57% (EER)
[9]	LA, PA	CNN	4.50%, 0.96% (EER)
[44]	LA, PA	CNN	9.87%, 1.75% (ERR)

Analyzing related work shows that the considered studies consistently show better results in detecting PA-based than in detecting LA-based attacks. Although previous studies show promising results in detecting spoofed voices, they are either limited to a specific language or application, designed for a specific spoofing method (i.e., PA or LA) or require additional hardware. Further research is required, focusing on developing a fake voice detection approach that allows differentiating between genuine and spoofed voice recordings, independently of the underlying spoofing method, with high accuracy and lightweight architecture, making the approach applicable for organizations with limited hardware capabilities. With our work, we propose a deep CNN-based data analytics model which uses an image (i.e., histogram) classification to identify spoofed voices generated from various spoofing methods, fully automated and in real-time. The presented approach is accurate, practically relevant for organizational systems, and outperforms the current benchmark in this domain. As the implementation of artificial intelligence technologies in enterprises of international economies is rather sluggish [33], access to such technologies has to be designed as efficiently as possible (e.g., cost-efficiency). Addressing this requirement, our lightweight model approach allows us to reduce the total voice-related fraud costs and the costs required to implement and execute corresponding countermeasures.

3. Methodology

3.1. Model Architecture

To follow our study's aim, we used a CNN-based approach [24] for the automated detection of fake voices. CNN's have established themselves as a widely used method in the area of spoofing and voice command recognition [22, 23] The lightweight architecture of our network is illustrated in figure **??**. It consists of three convolutional layers, where the input images (here: histograms based on voice recordings) are represented by four-dimensional arrays, characterized by the number of images (n), size (x, y. here: 224, 224), and channels (here: 3 (RGB)). For every convolutional layer, we applied filters of size 64 for the first and second layer and 128 for the third layer so that the spatial information remains unchanged. The kernel size was set to 2x2 for the first and 3x3 for the following two convolutional layers to define the intensity of vector components [45]. We used the rectified linear unit activation function to pass the results of every convolutional layer to the inherent pooling layer [46]. This step allowed us to reduce the image size and makes the CNN model more location-independent.

A max-pooling layer follows every convolutional block. Through this combination of convolutional and pooling layers, the images (i.e., histograms) can be transformed into a more abstract feature presentation [24]. As soon as convolutional blocks have transformed the input into a proper downsize image, the flattening layer can be used to convert the input of three dimensions into a one-dimensional array [32]. In our example, the input of the flatten layer is 26x26x128, resulting in a corresponding output of 86,528 parameters. The flattening layer passes its output to the first dense layer. We used a dropout layer with a rate of 30% to control overfitting and achieve better performance. The output of the dropout layer with 64 was committed directly to the second and last dense layer [47]. For the last classification, we used a filter size of one and the Sigmoid activation function, which takes the range into a value between 0 and 1 [48]. For stochastic gradient descent, we used Adam optimizer with a default learning rate of 1e-3 [49].

3.2. Histogram-Based Visualization of Voice Signals

In the area of spoofing and voice command recognition, spectrograms are commonly used to represent speech signals and are usually the basis for the detection of spoofed voices [1]. Spectrograms reflect the time, frequency, and magnitude of the signal from audio or video recordings. However, by using spectrogram images with CNN architectures, it is challenging to capture long-term temporal dependencies using a smaller filter size [13]. Another limitation of spectrograms is the need for feature extraction methods such as Fourier transformation in order to determine and extract the most relevant sub-bands [35]. On the other side, histograms allow identifying the probability distribution of different such as frequencies in the case of voice signals [50]. Furthermore, histograms address the limitations of spectrograms as no feature extraction is necessary, and long-term temporal dependencies can be depicted. Therefore, all information can be considered for the classification of spoofed voice recordings. Furthermore, studies have already shown that histograms based on voice recordings manifest characteristic differences between genuine and spoofed voices and have also highlighted that the classification of histograms represents a promising approach for the detection of spoofed voices due to their advantages over spectrograms in the respective domain [20].

3.3. Evaluation Data and Preprocessing

We used the latest (i.e., fourth) version of the H-Voice dataset for the evaluation of our detection approach [51]. The dataset consists of 3,268 histograms based on genuine voice recordings and 3,404 histograms based on spoofed voice recordings, resulting in a total of 6,672 histograms which we used for training and evaluating our fake voice classification approach. The spoofed voices were generated with different spoofing methods, following the research need of identifying fake voices independently from the underlying spoofing method. First, the imitation method, which belongs to the category of voice conversion. The imitation method follows the transformation procedure proposed in the study by Ballesteros and Moreno [52].

While the efficient wavelet masking scheme assumes that any speech signal may seem similar to a speech host signal if its wavelet coefficients are sorted, they delimitate the conditions under which the above hypothesis is true. To create the imitation voices, an algorithm according to [52] has been used, providing validity for this study's training and evaluation data. Second, the deep voice method belongs to the category of speech synthesis [53]. After the voice recordings were created, all genuine and spoofed voice recordings were converted into histograms. Therefore, the voice recordings were re-quantized to 16 bits. The histograms were organized with 65,536 bins and stored as files in the size of 875x656 pixels each. Figure 1 shows two examples of the voice signal-based histograms.

Apparent differences between the genuine and spoofed histograms, resulting from a visual inspection, are encircled in red. As shown in figure 1, the middle area of the peak in (B) is falling steeper than in (A). This difference stands out even stronger in the bottom area of the right slope of the histograms (A) and (B).

For our approach, we tested the classification for two different cases:

1. Genuine voices (i.e., positive class): Histograms based on real voice recordings.
2. Spoofed voices (i.e., negative class): Histograms based on fake voice recordings.

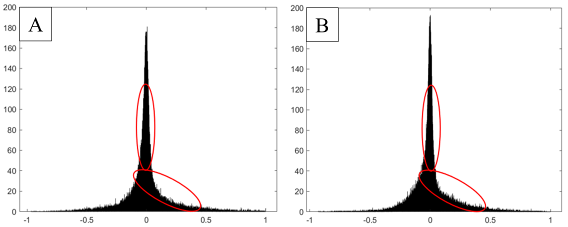

Figure 1. Examples of histograms based on (A) a genuine and (B) spoofed voice recordings [51].

For the preprocessing, we used a train-validation-test split to decrease overfitting and increase both the robustness and validity of our classification results [54]. While we used 60% (train: 4002 histograms) of the dataset for the training of our approach, 20% (validation: 1335 histograms) were used for providing an unbiased evaluation of our model's fit on the training dataset, and the remaining unseen 20% (test: 1335 histograms) were used to perform the final evaluation. Accordingly, all reported performance indicators and corresponding values are based on the evaluation of the test split. We highlight that we ensured the validity of our study's evaluation by using unseen testing data (i.e., test-set). While the validation-set has been used to assess the algorithm's training performance, the test-set (i.e., 1335 unseen histograms) contains randomly selected voice recordings shown to the model only for evaluation. Subsequently, we resized the images (i.e., histograms) to fixed input size of 224x224 pixels to improve the computational efficiency and rescaled the input RGB channels (i.e., normalization) to [0,1].

4. Results

To build and train our CNN-based detection model, we used Python 3.6.9 with the Keras 2.2.4 package [55], and TensorFlow 1.15.2 as backend [56]. We trained our model for 100 epochs with 128 samples per batch, running on an NVIDIA Tesla K80 12GB graphics processing unit (GPU).

Following prior research, particularly studies related to ASVspoof (i.e., series of challenges, promoting the development of countermeasures to protect automatic speaker verification from spoofing threats), the EER has been established as a widely used evaluation metric in the detection of spoofed voices [19, 57] The EER denotes the threshold at which the false positive rate and the false-negative rate are approximately equal [44]. The false-positive rate (i.e., miss rate) represents the ratio between misclassified spoofed test samples and the total number of spoofed test samples. Meanwhile, the false-negative rate (i.e., false alarm rate) refers to the ratio between misclassified genuine test samples and the total number of genuine test samples. Therefore, the lower the EER value, the better the developed countermeasure performs [58].

Thus, to make our results comparable with previous studies, we assessed the model's final classification results in terms of accuracy, balanced accuracy, Cohen's Kappa score, precision, recall, area underneath the receiver operating characteristic curve (AUC-ROC), and EER. The model results show that our proposed spoofing attack detection approach sets a new benchmark in distinguishing between genuine and spoofed voice recordings, characterized by a very good balanced accuracy of 95.64%, EER of 4.43%, and AUC-ROC of 0.98. Table 2 provides all the values of the considered performance indicators.

Table 2. Performance indicators.

Performance Indicator	Value
Accuracy	95.66%
Balanced Accuracy	95.64%
Kappa	91.31%
Precision	96.13%
Recall	94.95%
AUC-ROC	0.9837
EER	4.43%

Table 3 (i.e., confusion matrix) reports the number of false positives, false negatives, true positives, and true negatives, whereas we defined genuine recordings as positive and spoofed recordings as negative class. Out of the 654 samples labeled as genuine voice recordings (i.e., positive), our model correctly classified 621 histograms, resulting in a true positive rate of 94.95%. Of the 681 spoofed voice recordings, 656 got classified correctly, leading to a true negative rate of 96.33%.

Table 3. Confusion matrix (n_{test}=1,335).

		Predicted	
		Spoofed	Genuine
Actual	Spoofed	**49.14%**	1.87%
	Genuine	2.47%	**46.52%**

To foster our model's architecture, we compared the performance of classifying histogram representations of genuine and spoofed voice recordings of the H-Voice dataset [51], with other pre-trained networks. Following the purpose of differentiating between genuine and spoofed recordings, we excluded the model's top layer of each considered pre-trained network and added our classifier [59]. Depending on the pre-trained network, we adjusted the number of trained layers and the input shapes of the images. We then first froze the layers of each base model for training, followed by unfreezing and reduction of the learning rate for fine-tuning of each model [60].

We highlight that we define the overall efficiency by the ratio between the performance (i.e., classification accuracy) and resources required (i.e., network parameters). The results in Table 4 show that in terms of balanced classification accuracy and EER, our lightweight model architecture outperforms all of the investigated architectures. With an amount of 5.6M trainable layers, slightly larger than the mobile architecture MobileNetV2 [61], our architecture manifests a significantly smaller amount of parameters than the remaining architectures ResNet50 [62], Xception [63], VGG16, and VGG19 [64], contributing to the overall efficiency of our model, and the application of spoofing attack detection approaches on embedded systems [32]

Table 4. Performance of other pre-trained networks.

Architecture	EER	Parameters	Bal. Acc.
Our Model	**4.43%**	**5.6M**	**95.64%**
MobileNetV2	5.34%	4.5M	94.24%
ResNet50	5.35%	25.8M	92.36%
Xception	6.11%	23.2M	95.14%
VGG16	7.65%	14.7M	95.14%
VGG19	8.26%	20.1M	95.14%

4.1. Post-Hoc Analysis of Predictive Areas for Detecting Fake Voices

The gradient-weighted class activation mapping Grad-CAM heatmap algorithm for visualization was used to analyze our model in detail [34]. The algorithm visualizes the predictive areas of our classification model, allowing us to reveals characteristic features for the differentiation between genuine and spoofed voices. To distinguish predictive from non-predictive areas, we applied a pseudo/false-color to the heatmap, where the areas highlighted in red (encircled in red for both figures 3 and 4) refer to important areas, whereas less relevant regions are marked in blue. We highlight that due to the up-and downscaling of the images for the training of the model and the Grad-CAM algorithm, the heatmaps are not pixel-accurate.

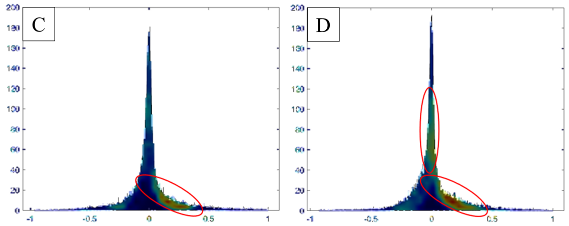

Figure 2. Post-hoc analysis of correctly classified (C) genuine and (D) spoofed recordings using heatmaps.

On the one hand, figure 2 shows two heatmaps of correctly classified histograms of (C) a genuine and (D) a spoofed voice recording. They correspond to the example histograms which were illustrated in figure 1. Both heatmaps indicate that the model primarily looks at the bottom area of the right slope of the histograms. Furthermore, the model considers some parts of the peak in the middle as relevant for its decision for the spoofed voice recording (D), which is highlighted in yellow and green (i.e., moderately relevant) by the Grad-CAM algorithm.

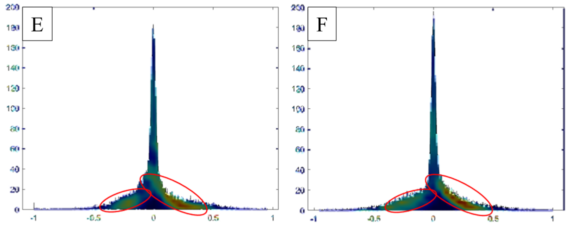

Figure 3. Post-hoc analysis of misclassified (E) genuine and (F) spoofed recordings using heatmaps.

On the other hand, figure 3 shows two heatmaps of misclassified histograms of (E) a genuine and (F) a spoofed voice recording. Similar to the correct predicted histograms, the bottom area of the right slope is mainly important for the decision. In contrast, the model looks at parts of the left slope of the histograms (E) and (F).

5. Discussion

Several studies have already demonstrated the effectiveness of ML techniques in the detection of spoofed voices in general. In contrast, our work concentrates on detecting spoofing attacks generated by various spoofing methods. While the ASVspoof dataset was used for the evaluation by most of the previously developed countermeasures, we addressed the lack of external validation with our approach. Our proposed model achieves excellent performance in detecting spoofed voices under consideration of two different LA spoofing attacks. With an EER of 4.43% based on only unseen testing data, we outperformed the model by Monteiro and Alam [9] and set a new benchmark.

As shown in table 3, our model did only misclassify 2.47% genuine and 1.87% spoofed voice recordings from a total of 1,335 tested histograms. Therefore, we revealed differences in the histogram representations between the different spoofing methods, advancing future research in fake voice detection for organizational systems and technology. The four heatmaps (i.e., (A), (B), (C), (D)) in figures 2 and 3 disclose that our model mainly focuses on the bottom area of the right slope of the histograms for the classification. On the one hand, the two correctly classified histograms reveal that some parts of the peak in the middle were also decisive for classifying the spoofed voice (D) but not for the genuine voice (C).

On the other hand, the two wrongly classified histograms (E) and (F) reveal that the model focuses on little parts of the left bottom area of the left slope of the histograms. Furthermore, the peak is not considered here. These two findings might be the reason for the wrong classification. The heatmaps indicate that our model predominantly predicts genuine and spoofed voices based on the correct parts of the histograms. Automatic speaker verification systems are often built in small devices with a relatively low available computational power and memory (e.g., smartphones, smart speakers). Therefore, it is essential to consider this aspect in the development and following the design requirements by providing a lightweight CNN architecture [4, 14]. A prominent example in such applications is the pre-trained architecture MobileNetV2 [61]. As shown in Table 4, while the parameters are similar in size, our model (i.e., 5.6M parameters) manifests a better classification accuracy by achieving a lower EER and higher accuracy. Therefore, our model follows the design requirements of embedded systems, being applicable on smaller devices, which is of high practical relevance for organizational systems with limited hardware capabilities [32].

Further, the preprocessing of the images is comparatively small to other approaches since the voice recordings must only be transformed into histograms without additional feature extraction (i.e., no information loss) [50], allowing the approach to be highly objective and reproducible. While the predominantly used spectrograms in the voice signal domain do not offer these advantages [13, 35], studies already indicated that it is possible to detect spoofed voices using histograms [37]. Therefore, we confirm previous scientific findings with our results and show that this approach (i.e., using image classification on histogram representations of voice recordings) can outperform previously proposed methods.

6. Conclusion

Automatic speaker verification systems grant access to sensitive organizational systems and technologies such as online banking or telecommunication services, making them a popular target for spoofing attacks (i.e., cyber fraud). While most of the previous studies which deal with the detection of spoofing attacks proposed countermeasures for only one specific spoofing method (i.e., LA or PA), real-world scenarios show that automatic speaker verification systems must be protected against all spoofing methods, as the type of the spoofing attack is unknown. Following our initial research question, we show with our CNN-based fake voice detection approach that it is possible to classify voice recording-based histograms to differentiate between genuine and spoofed voice signals. With a balanced accuracy of 95.64% and an EER of 4.43%, our results outperform the current benchmark in this domain and reveal additional findings that can advance future research to develop additional countermeasures for the fight against cyber fraud. We have shown that histograms can be used for the detection of spoofed voices, which confirms prior research [37] and opens new research directions for future investigations. Our approach is well suited for real-time classification of spoofed and genuine voice recordings and is promising to adapt to other spoofing methods from LA or PA attacks. This opportunity is crucial because of the varying spoofing methods which are used for criminal activities.

Characterized by minor preprocessing measures and lightweight network architecture (see: embedded systems [32]), our approach manifests strong practical applicability. It allows organizations and various application scenarios with limited hardware capabilities to adapt artificial intelligence technologies (here: deep learning-based fraud detection approach) into their operational routines [33]. Furthermore, we contribute to the reduction of unauthorized access and monetary damage caused by cyber fraud resulting from fake voices, which is also of high practical relevance [4, 1].

By proposing a model that is fast, cost-efficient, objective, and reproducible, our CNN-based data analytics model manifests high practical relevance as it allows firms and organizations to effectively control their environments by analyzing voice-related data in real-time, serving both economic and security goals [65]. While Sahidullah et al. [39] highlight three main categories (i.e., Short-Term Power Spectrum Features, Short-Term Phase Features, and Spectral Features with Long-term Processing) of features commonly used for synthetic speech detection (i.e., spoofed voices), the current state of research only reveals very few findings of histogram-based features. Inspired by image texture analysis-related features (here: textural and contextual) [25], we contribute to the field and allow future research to investigate the usage of histogram representation further to identify fake voices.

As fake voice-related fraud causes monetary damage and serious privacy concerns for various applications, it is of great importance to improve corresponding prevention measures, significantly reducing the resources for subsequent limitation of possible loss and damage [1, 2].

Furthermore, as cyber fraud- and deception-related theories reveal concordantly, technical artifacts are required to prevent malicious actors from gaining access to sensitive information as the world is becoming increasingly reliant on computers for infrastructures in the cyberspace [27, 28, 29]. With our design science approach [30], we follow theories in cyber fraud and deception by presenting a CNN-based prevention model. As proposed by Ganesan and MSK [66], a two-tier model can be used, which distinguished between genuine and spoofed voice recordings in a first step, followed by data structure (e.g., bloom filter) which prevents the attack before reaching the victim (e.g., customer). Therefore, our CNN-based classification approach can be used as a first instance of a holistic prevention approach, allowing to improve the perceived quality of service and tackle cyber criminology and deception issues.

6.1. Limitations

The main limitation of our work is based on the H-Voice dataset [51], which was used to train and evaluate our spoofed voice detection algorithm. The dataset only provides two spoofing methods from the LA category, including speech synthesis and voice conversion, used to generate spoofed voices. There are further spoofing methods that can be used for attacks on automatic speaker verification systems. Furthermore, our model only contains knowledge derived from the H-Voice dataset, so our approach lacks external validity. Although the H-Voice dataset includes a relatively large number of samples (i.e., histograms), it still manifests a CNN-related limitation as deep learning algorithms require large quantities of data for training.

Another limitation of this study is the hardware as calculations (i.e., training and evaluation) were performed on a hosted GPU. Primarily as we aim to contribute to the development of fraud detection approaches suitable for embedded systems [32], tests need to be performed that adapt the implementation of our fake voice classification approach to computers used in the organizational environment.

Due to the implemented method of the H-Voice dataset and the general idea of imitating target voices, the two classes (i.e., genuine and spoofed) already manifest systematic differences in their visual appearance and, therefore, features. Although the implemented scheme (i.e., efficient wavelet masking) represents a common procedure in the field of voice imitation, the identified features using our post-hoc analysis lack external validity and need to be compared with features deriving from other datasets (i.e., different imitation methods).

6.2. Future Work

Based on our work, some further steps need to follow to tackle the limitations and improve our approach. The main goal of our future work is to re-evaluate the algorithm using larger datasets that contain a greater variety of spoofing methods. As the evaluation data we used in this study uses the same type of imitated voices for all spoofed recordings, datasets for future research need to feature different imitation methods. This investigation will allow us to assess and improve the robustness and external validity of our approach. Furthermore, as our model only misclassified spoofed voices generated with the imitation method, a multi-class classification could be another way of investigating our result in more detail.

While both the private and industrial usage of voice-activated personal assistants is continuously growing, our approach is of high practical relevance as it contributes to lower the risk of unauthorized access and control [4, 5]. As studies have shown that the implementation of artificial intelligence technologies (here: deep learning-based spoofing attack detection approach) in enterprises of international economies (e.g., Germany) is rather sluggish, we will also investigate the adoption of our approach in German small and medium-sized enterprises [33].

Acknowledgements

We want to thank the reviewers and mini-track chairs, who provided very helpful comments on the refinement of this paper. This research is funded by the German Federal Ministry of Education and Research (13FH4E03IA, 13FH566KX9, 13FH176PX8, 13FH4E05IA).

References

[1] N. Subramani and D. Rao, "Learning efficient representations for fake speech detection," in *AAAI '20 Proc.*, pp. 5859–5866, 2020.

[2] R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma, and Y. Liu, "Deepsonar: towards effective and robust detection of ai-synthesized fake voices," in *ACM-MM '20 Proc.*, pp. 1207–1216, 2020.

[3] J. Li, M. Sun, X. Zhang, and Y. Wang, "Joint decision of anti-spoofing and automatic speaker verification by multi-task learning with contrastive loss," *IEEE Access*, vol. 8, pp. 7907–7915, 2020.

[4] M. Haug, P. Roessler, and H. Gewald, "Identification and influence of perceived risks on smart speaker use behavior," in *WI '20 Proc.*, pp. 1325–1331, 2020.

[5] S. Han and H. Yang, "Understanding adoption of intelligent personal assistants: A parasocial relationship perspective," *Ind. Manag. Data Syst.*, vol. 118, no. 3, pp. 618–636, 2018.

[6] K. M. Malik, H. Malik, and R. Baumann, "Towards vulnerability analysis of voice-driven interfaces and countermeasures for replay attacks," in *MIPR '19 Proc.*, pp. 523–528, 2019.

[7] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Commun.*, vol. 66, pp. 130–153, 2015.

[8] Y. Wang, M. Zhang, and Z. Zhu, "Detection of voice transformation disguise based on deep residual net," in *ICCSP '20 Proc.*, pp. 126–130, 2020.

[9] J. Monteiro and J. Alam, "Development of voice spoofing detection systems for 2019 edition of automatic speaker verification and countermeasures challenge," in *ASRU '19 Workshop*, pp. 1003–1010, 2019.

[10] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

[11] C. Zhang, C. Yu, and J. H. L. Hansen, "An investigation of deep-learning frameworks for speaker verification antispoofing," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 4, pp. 684–694, 2017.

[12] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "Presentation attack detection using long-term spectral statistics for trustworthy speaker verification," in *BIOSIG '15 Proc.*, pp. 1–6, 2015.

[13] I. Himawan, F. Villavicencio, S. Sridharan, and C. Fookes, "Deep domain adaptation for anti-spoofing in speaker verification systems," *Comput. Speech Lang.*, vol. 58, no. 11, pp. 377–402, 2019.

[14] Y. Wang, W. Cai, T. Gu, W. Shao, Y. Li, and Y. Yu, "Secure your voice: an oral airflow-based continuous liveness detection for voice assistants," *ACM IMWUT*, vol. 3, no. 4, pp. 1–28, 2019.

[15] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *ISIMP '04 Proc.*, pp. 145–148, 2004.

[16] V. M. Unnikrishnan and R. Rajeev, "Mimicking voice recognition using mfcc-gmm framework," in *ICEI '17 Proc.*, pp. 301–304, 2017.

[17] J. Shang, S. Chen, and J. Wu, "Defending against voice spoofing: a robust software-based liveness detection system," in *MASS '18 Proc.*, pp. 28–36, 2018.

[18] Z. Chen, W. Zhang, Z. Xie, X. Xu, and D. Chen, "Recurrent neural networks for automatic replay spoofing attack detection," in *ICASSP '18 Proc.*, pp. 2052–2056, 2018.

[19] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: assessing the limits of replay spoofing attack detection," in *ISCA '17 Proc.*, pp. 2–6, 2017.

[20] R. Rahmeni, A. B. Aicha, and Y. B. Ayed, "Speech spoofing countermeasures based on source voice analysis and machine learning techniques," *Procedia Comput. Sci.*, vol. 159, pp. 668–675, 2019.

[21] H. Liang, L. Xiaodan, Q. Zhang, and X. Kang, "Recognition of spoofed voice using convolutional neural networks," in *GlobalSIP '17 Proc.*, pp. 293–297, 2017.

[22] A. Azarang, J. Hansen, and N. Kehtarnavaz, "Combining data augmentations for cnn-based voice command recognition," in *HSI '19 Proc.*, pp. 17–21, 2019.

[23] H. Dinkel, Y. Qian, and K. Yu, "Small-footprint convolutional neural network for spoofing detection," in *IJCNN '17 Proc.*, pp. 3086–3091, 2017.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[25] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.*, no. 6, pp. 610–621, 1973.

[26] J. Canelón, E. Huerta, N. Leal, and T. Ryan, "Unstructured data for cybersecurity and internal control," in *HICSS-53 Proc.*, 2020.

[27] K. Jaishankar, "Establishing a theory of cyber crimes," *Int. J. Cyber Criminol.*, vol. 1, no. 2, pp. 7–9, 2007.

[28] K. E. Heckman, F. J. Stech, R. K. Thomas, B. Schmoker, and A. W. Tsow, "Cyber denial, deception and counter deception," *Advances in Information Security*, 2015.

[29] E. A. Cranford, C. Gonzalez, P. Aggarwal, M. Tambe, S. Cooney, and C. Lebiere, "Towards a cognitive theory of cyber deception," *Cogn. Sci.*, vol. 45, no. 7, p. e13013, 2021.

[30] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Q*, pp. 75–105, 2004.

[31] K. Rainer, M. Frolick, and J. Ryan, "Introduction to the minitrack on topics in organizational systems and technology," in *HICSS-53 Proc.*, 2020.

[32] A. Canziani, E. Culurciello, and A. Paszke, "Evaluation of neural network architectures for embedded systems," in *ISCAS '17 Proc.*, pp. 1–4, 2017.

[33] P. Ulrich, V. Frank, and M. Kratt, "Adoption of artificial intelligence technologies in German SMEs — Results from an empirical study," in *Corporate Governance: A Search for Emerging Trends in the Pandemic Times*, 2021.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," in *ICCV '17 Proc.*, pp. 618–626, 2017.

[35] B. Chettri, T. Kinnunen, and E. Benetos, "Subband modeling for spoofing detection in automatic speaker verification," in *SPLC '20 Proc.*, pp. 1–8, 2020.

[36] C. Veaux, J. Yamagishi, and K. MacDonald, "Asvspoof 2019: future horizons in spoofed and fake audio detection," in *ISCA '19 Proc.*, pp. 1–5, 2019.

[37] R. Rahmeni, A. Ben Aicha, and Y. Ben Ayed, "On the contribution of the voice texture for speech spoofing detection," in *STA '19 Proc.*, pp. 501–505, 2019.

[38] C. Hanilçi, "Linear prediction residual features for automatic speaker verification anti-spoofing," *Multimed. Tools. Appl.*, vol. 77, no. 13, pp. 16099–16111, 2018.

[39] M. Sahidullah, T. Kinnunen, and C. Hanilc¸i, "A comparison of features for synthetic speech detection," in *ISCA '15 Proc.*, pp. 1–6, 2015.

[40] T. Nosek, S. Suzić, B. Papić, and N. Jakovljević, "Synthesized speech detection based on spectrogram and convolutional neural networks," in *TELFOR '19 Proc.*, pp. 22–26, 2019.

[41] M. J. Alam and P. Kenny, "Spoofing detection employing infinite impulse response - constant q transform-based feature representations," in *EUSIPCO '17 Proc.*, pp. 101–105, 2017.

[42] T. J. Watson and J. Hopkins, "Universal background model based speech recognition," in *ICASSP '08 Proc.*, pp. 4561–4564, 2008.

[43] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR '07 Proc.*, pp. 1–8, 2007.

[44] J. Monteiro, J. Alam, and T. H. Falk, "An ensemble based approach for generalized detection of spoofing attacks to automatic speaker recognizers," in *ICASSP '20 Proc.*, pp. 6599–6603, 2020.

[45] F. Carrara, F. Falchi, R. Caldelli, G. Amato, R. Fumarola, and R. Becarelli, "Detecting adversarial example attacks to deep neural networks," in *CBMI '17 Proc.*, pp. 1–7, 2017.

[46] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *ICASSP '13 Proc.*, pp. 8609–8613, 2013.

[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[48] C.-H. Tsai, Y.-T. Chih, W. H. Wong, and C.-Y. Lee, "A hardware-efficient sigmoid function with adjustable precision for a neural network system," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 62, no. 11, pp. 1073–1077, 2015.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR '15 Proc.*, pp. 1–15, 2015.

[50] D. Y. Loni and S. Subbaraman, "Singing voice identification using harmonic spectral envelope," in *ICIP '15 Proc.*, pp. 119–123, 2015.

[51] D. M. Ballesteros L, Y. Rodriguez, and D. Renza, "A dataset of histograms of original and fake voice recordings (h-voice)," *Data Brief*, vol. 29, no. 4, pp. 1–6, 2020.

[52] D. M. Ballesteros L and J. M. Moreno A, "Highly transparent steganography model of speech signals using efficient wavelet masking," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9141–9149, 2012.

[53] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep voice: real-time neural text-to-speech," in *ICML '17 Proc.*, pp. 1–17, 2017.

[54] H.-C. Shin, H. R. Roth, M. Gao, Le Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[55] F. Chollet, "Keras," *https://keras.io*, 2015.

[56] M. Abadi *et al.*, "Tensorflow: a system for large-scale machine learning," in *USENIX-OSDI '16 Proc.*, pp. 265–283, 2016.

[57] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *ISCA '15 Proc.*, pp. 2037–2041, 2015.

[58] B. T. Balamurali, K. E. Lin, S. Lui, J.-M. Chen, and D. Herremans, "Toward robust audio spoofing detection: a detailed comparison of traditional and learned features," *IEEE Access*, vol. 7, pp. 84229–84241, 2019.

[59] H. Liang, W. Fu, and F. Yi, "A survey of recent advances in transfer learning," in *ICCT '19 Proc.*, pp. 1516–1523, 2019.

[60] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, "Spottune: transfer learning through adaptive fine-tuning," in *CVPR '19 Proc.*, pp. 4805–4814, 2019.

[61] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: inverted residuals and linear bottlenecks," in *CVPR '18 Proc.*, pp. 4510–4520, 2018.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR '16 Proc.*, pp. 770–778, 2016.

[63] F. Chollet, "Xception: deep learning with depthwise seperable convolutions," in *CVPR '17 Proc.*, pp. 1251–1258, 2017.

[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR '15 Proc.*, pp. 1–14, 2015.

[65] C. Y. Jeong, S.-Y. T. Lee, and J.-H. Lim, "Information security breaches and it security investments: impacts on competitors," *Inf. Manag.*, vol. 56, no. 5, pp. 681–695, 2019.

[66] V. Ganesan and M. Msk, "A scalable detection and prevention scheme for voice over internet protocol (voip) signaling attacks using handler with bloom filter," *Int. J. Netw. Manag.*, vol. 28, no. 2, pp. 1–18, 2018.