Towards Frequency Band Explainability in Synthetic Speech Detection

Davide Salvi, Paolo Bestagini, Stefano Tubaro
Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano
Piazza Leonardo Da Vinci 32, 20133 Milano, Italy
{davide.salvi, paolo.bestagini, stefano.tubaro}@polimi.it

***Abstract*—Recent advancements in deep learning techniques have brought remarkable developments in synthetic media generation, leading to the creation of forged contents that are almost indistinguishable from real data. This phenomenon poses a new challenge for the multimedia forensics community, as the misuse of synthetic media can potentially cause adverse consequences. Regarding the audio field, several methods have been proposed to detect synthetic speech, but due to their data-driven nature, their results are often little interpretable. To overcome this limitation, the scientific community is focusing on Explainable AI (XAI) aimed at understanding the critical elements in a speech track that drive the predictions of the detectors. In this work, we address the task of XAI in synthetic speech detection and explore the critical factors that allow us to detect forged tracks generated by unseen techniques. Our results suggest that the artifacts of synthetic speech are contained in specific frequency bands and show how we can make the detection process more accurate by focusing on single spectral bands. We also generalize our findings to other detectors, showing how these can benefit them and improve their final classification performances.**

***Index Terms*—Multimedia Forensics, Audio, Synthetic Speech, Explainability**

I. Introduction

Recent advances in deep learning and artificial intelligence have led to incredible developments in synthetic media generation. The ability to generate forged content that is indistinguishable from real one has spread across various domains (e.g., images, videos, audio) and has captured the attention of both researchers and specialists [1]. While this phenomenon opens the doors to exciting unexplored scenarios, it also poses a new challenge for the multimedia forensics community. There is a need to develop robust methods able to detect synthetic media and discriminate them from authentic content. This is a problem that cannot be overstated, as the misuse of synthetic data can lead to adverse consequences, as already seen in numerous cases of fraud and blackmail [2].

This material is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and AFRL or the U.S. Government. This work was supported by the PREMIER project, funded by the Italian Ministry of Education, University, and Research within the PRIN 2017 program. This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART").

Focusing on the audio field, several methods have been proposed in the last few years to tackle synthetic speech detection [3]. The goal of these systems is to distinguish between real and synthetic speech, and they do so based on several strategies, ranging from pure deep learning techniques to more semantic approaches. For instance, the authors of [4] employ a ResNet model with multi-head attention pooling to learn the distinctions between real and fake audio representations, while in [5] a model is trained solely on real speech data to exploit the speaker's biometric characteristics. This leads to improved generalization capabilities since no specific synthesis method has been considered during training. On the other hand, the methods proposed in [6] and [7] perform synthetic speech detection by analyzing the emotional and prosodic content of speech. Also, numerous datasets have been presented in this field to increase the interest of the scientific community on the topic and push the research toward the development of new detection methods [8], [9].

The performances of the proposed detectors are remarkable, especially in controlled scenarios. However, since most of these methods rely solely on data-driven approaches, the interpretability of their predictions is somehow limited. This makes these systems unsuitable for real-world applications where thoroughly understanding what is driving the detection process is essential. Indeed, what makes a prediction trustworthy is understanding the motivation that led to it, and the *black box* nature of data-driven systems compromises their reliability.

To overcome this limitation, the scientific community has increased its attention towards Explainable AI (XAI), with the intent of understanding the critical elements in an audio track that cause the detector's predictions. For example, the authors of [10] and [11] use the SHapley Additive exPlanations (SHAP) [12] method to analyze the artifacts generated by synthetic speech systems, while those of [13] and [14] utilize Gradient-weighted Class Activation Mapping (Grad-CAM) [15] and Local Interpretable Model-agnostic Explanations (LIME) [16] to gain insight into the decision-making process of synthetic speech detectors, respectively. Finally, other studies such as [17] and [18], have focused on explaining the results obtained using specific audio features, focusing on different frequency bands and speech formants.

Even though these studies provide valuable insights into the functioning of synthetic speech detectors and the elements that influence their predictions, they are often restricted in scope,

focusing on only a few audio samples or speech generation methods. This makes their findings limited and hard to extend to other detectors or generation techniques.

In this paper, we tackle the problem of XAI in synthetic speech detection and explore which are the critical factors that drive the detection process. In particular, we take into account speech data generated by synthesis techniques unseen during the training step of a detector, and we adopt two independent approaches to identify the frequency bands that prove most helpful for synthetic speech detection. Both techniques confirm the same finding: we can improve the final classification accuracy by focusing on a reduced frequency range of the speech signal under analysis. Furthermore, we show how the results obtained are not significant only for one considered detector, but can also be extended to other classifiers, providing valuable insights into the generalization of the synthetic speech detection process.

II. Proposed Method

In this paper we consider the problem of synthetic speech detection and investigate two different XAI approaches to understand which are the most relevant frequency bands to accomplish this task. The work is based on the hypothesis that not all frequency bands of the audio spectrum have equal relevance for synthetic speech detection. In other words, we suspect that speech generators leave more artifacts in some bands than others, so we can exploit this aspect to enhance our classification capabilities.

In the current state of the art, most synthetic speech detectors are data-driven and take as input standard acoustic features (e.g., LFCC, MFCC, CQCC, etc.) that are empirically chosen based on the outcome of the models. This is not an optimal approach since the classifiers are intended as black boxes and the decision of which audio representation to use as their input is not always well-motivated [19]. Contrariwise, suppose we manage to understand which are the most relevant frequency bands to perform the detection task. In that case, we could feed the developed detectors with an appropriate set of features based on our findings, i.e., focusing on the frequency bands that we know contain the most artifacts, thus improving the final classification accuracy of the developed systems.

A. Problem formulation

The problem we address can be formally defined as follows. Let us consider a discrete-time input speech signal $\mathbf{x}$ sampled with a sampling frequency f_s. This means that the informative part of the Discrete Fourier Transform (DFT) of the signal is defined over a set of frequencies $\mathcal{F} = \{f_1, f_2, ..., f_\mathrm{N}\}$ spanning the interval $[0, f_\mathrm{s}/2)$. The track $\mathbf{x}$ belongs to a class $y \in \{0, 1\}$, where 0 means that the signal is authentic while 1 indicates synthetic data. Let us consider a synthetic speech detector $\mathcal{D}$ trained to estimate the class of a signal $\mathbf{x}$ as $\hat{y} = \mathcal{D}(\mathbf{x})$, where $\hat{y} \in [0, 1]$ indicates the likelihood that the signal $\mathbf{x}$ is fake.

We define as $\mathbf{x}_\mathcal{S}$ a filtered version of the signal $\mathbf{x}$, defined only on a subset of frequencies $\mathcal{S}$, where $\mathcal{S} \subset \mathcal{F}$. Our goal is to find which is the most relevant set of frequencies $\hat{\mathcal{S}}$ to perform the synthetic speech detection task, thus minimizing the difference between the estimated class $\hat{y} = \mathcal{D}(\mathbf{x}_\mathcal{S})$ and the actual class y. Formally, we want to estimate

$$\hat{\mathcal{S}} = \arg\min_{\mathcal{S}} \|y - \mathcal{D}(\mathbf{x}_\mathcal{S})\|. \quad (1)$$

B. Considered detectors

During this study, we consider two different synthetic speech detectors proposed in the literature.

ResNet. The first one is proposed in [20] and is based on ResNet [21]. This is a residual Convolutional Neural Network (CNN) that creates shortcuts between layers by skipping connections that help stabilize training. The network is fed with a spectrogram representation of the input audio signal, and its architecture includes 6 residual blocks.

RawNet2. To broaden the scope of the results, we extend our findings to another detector, i.e., RawNet2 [22]. This is an end-to-end neural network that operates on raw waveform inputs. It was first proposed for the ASVspoof 2019 challenge [23] and included as a baseline in the ASVspoof 2021 challenge [8]. Its architecture includes Sinc filters taken from SincNet, followed by two Residual Blocks with skip connections on top of a Gated Recurrent Unit (GRU) layer to extract frame-level representations of the input signal.

C. Explainability methods

We investigate the influence of different frequency bands for the synthetic speech detection task considering two different XAI approaches. In doing so we can compare the outcomes of both methods and determine if they are coherent with each other in showing which are the predominant bands. The main difference between the two methods is that the former performs an *a posteriori* interpretation of a pre-trained detector, while the latter is an *active* approach that involves the training of several classifiers.

A posteriori interpretation. This method involves the use of Local Interpretable Model-agnostic Explanations (LIME) [16], which creates an interpretable, local approximation of a black box deep learning model to explain individual predictions. Given an input to the network of shape $M \times N$, this algorithm modifies it by altering its feature values and measuring the effect of these changes on the prediction result. The output is a binary mask of the same shape as the input, highlighting the most critical factors driving the predictions. In our case, we apply the method to iteratively mask parts of the spectrograms used as input to a trained ResNet model.

Active interpretation. The second approach involves considering various spectrogram masking configurations and retraining a ResNet version for each of them. We mask the spectrograms given as input to the model by showing only a chosen frequency band $\mathcal{S}$ and zeroing out all the frequency bins outside of it. In this way, each masking configuration presents the same dimension $M \times N$ as the original spectrogram but shows only a limited portion of its content. When dealing with this approach, we apply the same masking operation to both training and test data. After training and testing the

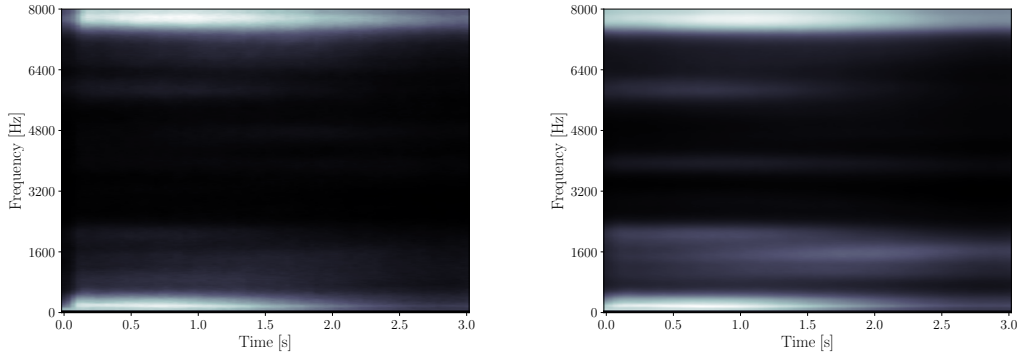

Fig. 1. Colormaps highlighting the most critical elements to perform the synthetic speech detection task, according to LIME algorithm. The two images depict the relevant components for detecting real (left) and synthetic (right) speech classes, respectively. The lightest areas are the most relevant for the detection task.

models on different masking configurations, we compare all the obtained results, analyzing if individual frequency bands are sufficient to perform the synthetic speech detection task, and if the results obtained on the masked spectrograms are comparable with those achieved with the original method.

III. Experimental Setup

In this section, we outline the evaluation setup employed in our experiments. We start by introducing the dataset that was used for training and testing the systems. Next, we detail the training parameters that we assumed. Finally, we explain the implementations of the two XAI approaches we consider.

A. Dataset

During all the experiments, we used the ASVspoof 2019 dataset [23]. This is a speech audio dataset that contains both real and synthetic tracks generated based on the VCTK corpus. The dataset has been released for the homonymous challenge, where participants had to compete to implement the best detector for Automatic Speaker Verification (ASV). It has been proposed to address two different tasks and here we consider the Logical Access (LA) one, which relates to the synthetic speech detection problem. The LA dataset is further divided into three sub-partitions, called *train*, *dev* and *eval*, which all include authentic signals along with synthetic speech samples generated with various methods. All the audio signals are released considering a sampling frequency of f_s = 16 kHz. The *train* and *dev* partitions have been created using the same set of 6 synthesis algorithms (named $A01$, $A02$, ..., $A06$), while the *eval* partition includes samples generated with 13 different techniques ($A07$, $A08$, ..., $A19$). We considered the *train* and *dev* partitions to train the considered detectors and the *eval* set to test them. The distinction between the speech generation techniques included in the training and test partitions allows us to perform analyses in an open set scenario and evaluate the considered detectors on data generated only by unseen synthesis algorithms.

B. Training Strategy

During the training phase, we trained the detectors to discriminate between real and synthetic speech data. We only consider the voiced segments of the analyzed tracks, as we want to focus on detecting the artifacts that appear in synthetic speech and not be biased by the presence of silence segments. To do so, we trim all the silences using a Voice Activity Detector (VAD). For both ResNet and RawNet2 models, we assumed a time window of 3.0 s. We used these values because, from preliminary experiments, they turned out to be the best compromise between the shortness of the windows and the performance, which is ideal in a real-world scenario.

All the hyperparameters of the networks have been fine-tuned to maximize their accuracy. These are the sets of parameters used, chosen after verifying the convergence of the models. For both models, we considered a maximum number of epochs equal to 100 and an early stopping patience of 20, weighted cross-entropy as loss function and Adam optimization, and a batch size of 128. The only difference between the two is the learning rate value, which is equal to 10^{-5} in the case of ResNet and 10^{-4} for RawNet2.

Regarding the computation of the ResNet input features, we calculate the Spectrograms by performing the log-magnitude representation of the Short-Time Fourier Transform (STFT) of the input audio, considering a Hamming window of 2048 samples and 25% overlap, which is a typical solution when computing this kind of audio representation [20].

C. XAI methods

Regarding the LIME approach, we applied perturbations to the input spectrograms by randomly obscuring parts of them at each iteration of the algorithm. We considered 1000 perturbations per test sample, which in previous experiments has been found to be an optimal balance between computational time and quality of results obtained.

As for the frequency masking of the spectrograms fed into the ResNet model, we consider 5 non-overlapping frequency bands $\mathcal{S}$. We do so for two reasons. First, given the impossibility of extensively investigating a large number of different masking configurations, we had to reduce the number of masking considered. Since this is an exploratory study, we decided to investigate a regular case. Second, based on the preliminary outcomes provided by the LIME approach, we found that this implementation was functional for our final

goal. We divided the frequency spectrum into 5 equally spaced bands, with the following ranges: $\mathcal{S}_1 = [0, 1600]$ Hz, $\mathcal{S}_2 = [1600, 3200]$ Hz, $\mathcal{S}_3 = [3200, 4800]$ Hz, $\mathcal{S}_4 = [4800, 6400]$ Hz, $\mathcal{S}_5 = [6400, 8000]$ Hz.

IV. Results

In this section we analyze and discuss the results of the two XAI approaches we applied to synthetic speech detection. In doing so, we evaluate the performances of the detectors in terms of Receiver Operating Characteristic (ROC) curves, Area Under the Curve (AUC) and Detection Rate (DR), defined as the percentage of samples of a single class that are correctly classified by the detector.

As a first experiment, we investigate the significance of the various frequency bands $\mathcal{S}$ of the spectrogram for synthetic speech detection using the LIME algorithm. Figure 1 shows the results of this analysis, divided by the two classes (i.e., real and synthetic speech). To create these plots, we generated a binary mask for each test track, highlighting the most relevant elements for the detection task. These masks were then averaged across all test signals to eliminate any dependency on specific signals or generation methods. We only considered tracks from content that were correctly classified by the detector, to avoid examining any misleading trace. In computing the masks, we emphasized elements that contributed both positively and negatively to the final prediction of a class. We did so since, in a binary classification task, positive and negative contributions are closely linked to each other.

The results show that two frequency bands $\mathcal{S}$ are significantly more influential than the others for the synthetic speech detection task. These involve the highest ($\approx$8000 Hz) and lowest ($<$1000 Hz) frequencies. Additionally, as regards the classification of fake tracks, even the medium-low frequencies ($\approx$2000 Hz) can help to drive the classification.

These results are further supported by the second XAI approach that we consider. Figure 2 shows the results of the detection task comparing the model fed with full-band signals with the models trained on masked spectrograms. In this case, the models trained on the lowest ($\mathcal{S}_1 = [0, 1600]$ Hz) and highest ($\mathcal{S}_5 = [6400, 8000]$ Hz) frequency bands not only have performances comparable to those of the original model, but they perform even better ($\text{AUC}_{\mathcal{F}}$=0.76 compared to $\text{AUC}_{\mathcal{S}_1}$=0.86 and $\text{AUC}_{\mathcal{S}_5}$=0.79). Conversely, the models trained on the central bands of the spectrum have modest performances, again demonstrating the poor information content included at these frequencies.

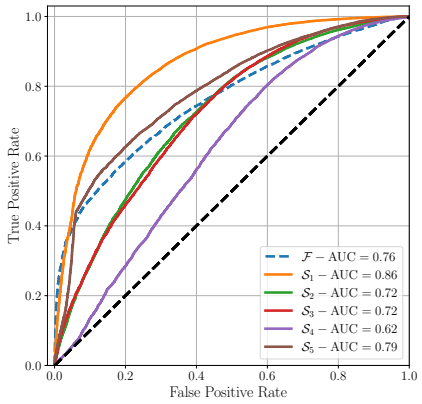

Fig. 2. ROC curve showing the synthetic speech detection performances of the ResNet model when working on audio data defined over the entire spectrum $\mathcal{F}$ compared with data filtered on frequency subsets $\mathcal{S}_i$.

We now perform an ablation study, to fully understand the contributions of different frequency bands to the detection task. As discussed in [17], different generation algorithms introduce artifacts to other spectral bands, so that by analyzing all the frequencies thoroughly, we can increase the discriminative power of the detectors. Table I shows the DR of each of the ResNet models we trained for the previous experiment on both real speech and synthetic speech generated by all the algorithms of the test dataset. We recall that none of these algorithms has been considered during training, making them all unseen for the detectors.

The results of the ablation study indicate that the performances of the detectors are quite different from each other, with significant variations in some generation algorithms. For instance, generator $A13$ is poorly identified by analyzing the whole spectrum $\mathcal{F}$ (DR = 0.30), but can be easily detected by investigating only the frequency band $\mathcal{S}_1$ (DR = 0.87). Similar behavior also appears for generators $A14$, $A17$ and $A18$. Conversely, other systems such as $A07$ and $A16$ can be better identified by looking at the $\mathcal{S}_5$ frequency band instead of $\mathcal{S}_1$. In general, our findings suggest that most artifacts can be detected by analyzing the low frequencies, as also proved by the previous experiments.

As a final experiment, we aim to investigate if our findings are specific to the ResNet model or can be generalized to other synthetic speech detectors. To do so, we consider RawNet2

TABLE I
Detection rate (DR) of the ResNet models trained on different spectral bands, measured on both real and synthetic speech, generated by all the algorithms of the test dataset. The DR values greater than 0.85 are shown in bold.

	REAL	**A07**	**A08**	**A09**	**A10**	**A11**	**A12**	**A13**	**A14**	**A15**	**A16**	**A17**	**A18**	**A19**
$\mathcal{F}$	0.69	**1.00**	**1.00**	**0.90**	0.54	0.60	0.31	0.30	0.67	0.50	**1.00**	0.56	0.53	**1.00**
$\mathcal{S}_1$	0.78	0.63	**0.97**	**0.94**	0.59	0.63	0.54	**0.87**	**0.92**	0.68	0.80	**0.87**	**0.89**	**0.86**
$\mathcal{S}_2$	0.66	0.73	0.80	0.80	0.70	0.69	0.73	0.76	0.82	0.75	0.71	0.40	0.43	0.31
$\mathcal{S}_3$	0.66	0.67	0.51	**0.88**	0.70	0.66	0.67	0.76	0.81	0.76	0.68	0.53	0.47	0.44
$\mathcal{S}_4$	0.58	0.56	0.71	**0.85**	0.54	0.57	0.50	0.51	0.78	0.71	0.58	0.51	0.43	0.36
$\mathcal{S}_5$	0.71	**1.00**	**1.00**	**0.93**	0.58	0.67	0.31	0.53	0.64	0.36	**1.00**	0.74	0.45	**1.00**

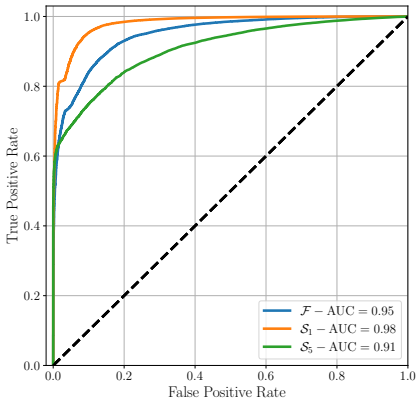

Fig. 3. ROC curve showing the synthetic speech detection performances of the RawNet2 model when working on audio data defined over the entire spectrum $\mathcal{F}$ compared with filtered data.

as classifier and we train and test it considering audio data filtered on specific frequency bands. We consider two types of filtering, low pass and high pass, to analyze the two frequency bands that proved most significant in the results shown in Figure 2 and Table I. We filter the audio data considering a 13th-order digital Butterworth filter, with cut-off frequencies equal to 1600 Hz (low-pass) and 6400 Hz (high-pass) to match the frequency bands $\mathcal{S}_1$ and $\mathcal{S}_5$, respectively.

Figure 3 shows the results of this analysis. These validate the outcomes of the previous studies, exhibiting that focusing solely on low frequencies leads to a more effective detection compared to using the entire spectrum $\mathcal{F}$. This is a significant finding for two reasons. First, it indicates that the results of this paper can be generalized to synthetic speech detectors other than ResNet. Second, it demonstrates that an effortless operation, such as filtering the input signals to focus on low frequencies, can result in an increase of AUC (in our case 0.03) compared to using unprocessed data.

V. Conclusions

In this paper, we considered the problem of synthetic speech detection and used two different XAI approaches to find which are the most relevant frequency bands to accomplish this task.

The main contributions of the work include a broader focus on speech synthesis methods unseen during training.We validate our findings by obtaining the same outcomes from two different XAI approaches. We also generalize the results to multiple synthetic speech detectors, accentuating the role of explainability studies in improving detection performance.

We believe these results have important implications for synthetic speech detection and we hope they can help in developing detectors that are not purely data-driven but also pay attention to data preprocessing to improve their results.

In future studies, we plan to expand the work by considering other acoustic features, other XAI techniques and detectors. A possible development involves implementing more punctual analyses over time and using different approaches between voiced and unvoiced segments of the audio track, which presumably contain different types of artifacts.

References

[1] S. J. Nightingale and H. Farid, "AI-synthesized faces are indistinguishable from real faces and more trustworthy," *Proceedings of the National Academy of Sciences*, vol. 119, pp. 1–3, 2022.

[2] Forbes, "Fraudsters Cloned Company Director's Voice In 35$ Million Bank Heist, Police Find," https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions.

[3] L. Cuccovillo, C. Papastergiopoulos *et al.*, "Open challenges in synthetic speech detection," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2022.

[4] R. Yan, C. Wen *et al.*, "Audio Deepfake Detection System with Neural Stitching for ADD 2022," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[5] A. Pianese, D. Cozzolino *et al.*, "Deepfake audio detection by speaker verification," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2022.

[6] E. Conti, D. Salvi *et al.*, "Deepfake Speech Detection Through Emotion Recognition: a Semantic Approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[7] L. Attorresi, D. Salvi *et al.*, "Combining Automatic Speaker Verification and Prosody Analysis for Synthetic Speech Detection," in *International Conference on Pattern Recognition (ICPR)*, 2022.

[8] J. Yamagishi, X. Wang *et al.*, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.

[9] D. Salvi, B. Hosler *et al.*, "TIMIT-TTS: a Text-to-Speech Dataset for Multimodal Synthetic Media Detection," *IEEE Access*, 2023.

[10] W. Ge, J. Patino *et al.*, "Explaining deep learning models for spoofing and deepfake detection with SHapley Additive exPlanations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[11] W. Ge, M. Todisco, and N. Evans, "Explainable deepfake and spoofing detection: an attack analysis using SHapley Additive exPlanations," *arXiv preprint arXiv:2202.13693*, 2022.

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[13] B. M. Halpern123, F. Kelly *et al.*, "Residual networks for resisting noise: analysis of an embeddings-based spoofing countermeasure," in *Speaker and Language Recognition Workshop (Odyssey)*, 2020.

[14] B. Chettri, S. Mishra *et al.*, "Analysing the predictions of a cnn-based replay spoofing detection system," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018.

[15] R. R. Selvaraju, M. Cogswell *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in *ACM International Conference on Knowledge Discovery and Data Mining*, 2016.

[17] H. Tak, J. Patino *et al.*, "An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification," in *Speaker and Language Recognition Workshop (Odyssey)*, 2020.

[18] S.-Y. Lim, D.-K. Chae, and S.-C. Lee, "Detecting Deepfake Voice Using Explainable Deep Learning Techniques," *Applied Sciences*, vol. 12, p. 3926, 2022.

[19] N. M. Müller, P. Czempin *et al.*, "Does audio deepfake detection generalize?" in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2022.

[20] M. Alzantot, Z. Wang, and M. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

[21] K. He, X. Zhang *et al.*, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[22] H. Tak, J. Patino *et al.*, "End-to-end anti-spoofing with RawNet2," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[23] M. Todisco, X. Wang *et al.*, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.