EXPLAINING DEEP LEARNING MODELS FOR SPOOFING AND DEEPFAKE DETECTION WITH SHAPLEY ADDITIVE EXPLANATIONS

Wanying Ge, Jose Patino, Massimiliano Todisco and Nicholas Evans

EURECOM, Sophia Antipolis, France

ABSTRACT

Substantial progress in spoofing and deepfake detection has been made in recent years. Nonetheless, the community has yet to make notable inroads in providing an explanation for how a classifier produces its output. The dominance of black box spoofing detection solutions is at further odds with the drive toward trustworthy, explainable artificial intelligence. This paper describes our use of SHapley Additive exPlanations (SHAP) to gain new insights in spoofing detection. We demonstrate use of the tool in revealing unexpected classifier behaviour, the artefacts that contribute most to classifier outputs and differences in the behaviour of competing spoofing detection models. The tool is both efficient and flexible, being readily applicable to a host of different architecture models in addition to related, different applications. All results reported in the paper are reproducible using open-source software.

Index Terms— spoofing, presentation attack detection, explainability, Shapley additive explanations, ASVspoof

1. INTRODUCTION

Having begun over two decades ago [1, 2], research in spoofing or presentation attack detection for automatic speaker verification and speech deepfake detection has gained traction only in the last six years [3]. Judging by the estimates of performance reported in the literature, typically measured through spoofing detection metrics such as the minimum tandem detection cost function [4] or the equal error rate, a great deal of progress has been achieved in recent years.

While the community has been largely successful in developing effective detection solutions, it has yet to make substantial inroads to properly understanding and shedding light upon the artefacts used by spoofing countermeasures to distinguish between bona fide and spoofed speech. This is especially the case for so-called *black-box* solutions which typify probably all current, state-of-the-art solutions to spoofing and deepfake detection [5–9].

The growing trend toward explainable artificial intelligence (xAI) [10] has not only the potential to produce machine learning models that provide some justification for scores or decisions, i.e. in the general vein of improving trustworthiness, but also to help researchers learn more about the problem in hand as well as the characteristics, behaviour, strengths and weaknesses of different solutions. xAI approaches can also help reveal paths for future research and, ultimately, to expedite the development of better performing, more reliable and efficient machine learning solutions.

We have hence set out to explore explainability in anti-spoofing. Our goals are to better understand what artefacts are being captured or discarded by different solutions based upon different input features, different machine learning solutions or different components of ensemble systems [11, 12]. We are also hopeful of understanding why some spoofing attacks are more difficult to detect than others, e.g. the infamous A17 attack contained within the ASVspoof 2019 logical access (LA) database [13]. We hope too to understand why solutions that operate directly upon raw waveform inputs [8, 14, 15] can perform better than systems that operate upon hand-crafted spectro-temporal representations [12, 16], but worse for others, e.g. the A08 attack [15]. Additional motivation comes from the numerous reports in the literature which show that some solutions apply greater attention to non-speech intervals than to speech intervals [17–19]. Here we hope to better understand whether or not these issues are evidence of database design shortcomings and/or whether they point towards issues relating to the behaviour of certain, specific countermeasure solutions. In case of the former, we hope that studies of explainability might contribute to the development of tools to assist with database quality control.

This paper describes our exploration of SHapley Additive exPlanations (SHAP) [20] to different deep neural network solutions for spoofing and speech deepfake detection. As our first work in this direction, the emphasis is on classifier behaviour rather than attack analysis. While there are some related works [21–23], to the best of our knowledge, it is the first application of SHAP to the problem of spoofing detection. We describe SHAP and demonstrate what can be learned from such tools and advocate for their broader adoption. All results presented in this paper are fully reproducible using open-source codes.

2. RELATED WORK

Several related explainability studies have been reported previously. Using an approach based upon the attenuation of distinct spectral components, [24] shows that artefacts indicative of different spoofing attacks are located within different sub-band intervals, and hence that they can be detected more reliably with front-ends that emphasise the same frequency range. The use of Gradient-weighted Class Activation Mapping (Grad-CAM) [25] to explain spoofing classifier behaviour is reported in [26]. It is applied to generate a binary saliency map for the network input layer. Input audio is then reconstructed using spectrograms masked with the binary saliency map. Listening experiments show that the model uses the buzziness and rhythmic quality of speech sounds to distinguish between bona fide and spoofed speech. A study of replay detection [17] shows the impact of different replay attack configurations upon detection performance. The use of Local Interpretable Model-agnostic Explanations (LIME) [27] to generate both temporal and spectral explanations of model prediction behaviour for voice replay detection is reported

in [28]. The input speech spectrogram is first segmented into a number of temporal or spectral segments, before LIME is applied to learn their relative importance through experiments with and without their use for modelling. These works show that non-speech intervals can provide discriminative information for spoofing detection. Related work in [19] shows that the duration of non-speech intervals in a synthetic speech and converted voice detection task can also be indicative of whether an utterance is bona fide or spoofed.

Drawing from concepts in cooperative game theory [29] whereby payoffs are distributed fairly to each player based on their individual contributions, SHapley Additive exPlanations (SHAP), introduced in [20], offers a more elegant and powerful approach to explainability. SHAP values reflect the influence of particular features to a classifier output. The work in [23] reports the use of DeepSHAP [20] to help explain the behaviour of speech enhancement models. SHAP values, estimates of Shapley values [29] derived using DeepSHAP, are used to determine regions of the input feature, expressed in the form of spectrograms, that impact most upon on the model output. Results show that better performing models tend to rely more on information within speech-dominated spectro-temporal intervals. SHAP provides a unified and theoretically grounded approach to explain the relative importance of particular features to classifier outputs. We have sought to apply SHAP analysis to help explain the behaviour of different spoofing detection models. We provide a brief overview of SHAP before describing its application to speech data and then spoofing detection.

3. SHAPLEY ADDITIVE EXPLANATIONS (SHAP)

SHAP values can be both positive and negative and reflect the relative (un)importance of a particular feature to a classifier output. Given a prediction function (i.e., a model) $f(x)$ and a feature subset $S \subseteq F$, where F is the full set of features, the SHAP value ϕ_i is derived using a pair of models, one learned with the inclusion of feature i, the other without. For an arbitrary input x_S, the predictions derived using both models are compared according to:

$$\delta_i(S) = f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \quad (1)$$

where $f_{S \cup \{i\}}$ is the model trained with the feature subset supplemented with i and f_S is the model trained on the same subset without the inclusion of i. Eq. (1) is computed for all possible subsets $S \subseteq F \setminus \{i\}$ (subsets S not containing i) in order to gauge the impact of withholding i from the pool of features used for model training. The SHAP value is then given by:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! \, (|F| - |S| - 1)!}{|F|!} \delta_i(S) \quad (2)$$

which is a normalised average over the different permutations of $S \subseteq F \setminus \{i\}$, where $|S|$ and $|F|$ are the number of features in S and F.

By way of a simple, intuitive example, we consider an image (later a speech spectrogram) with a large number of pixels (the features, later short-time spectro-temporal magnitude estimates). Consider the selection of one pixel i in the image. If the model predictions obtained by the model when it is trained with or without the inclusion of i do not differ, then pixel i bears little, specific relevance to the output. On the other hand, differences between the two predictions would suggest that pixel i is comparatively informative and bears a stronger influence upon the model output.

In the case that $f(\mathbf{x})$ is a complex model, such as a deep neural network, the calculation of ϕ_i according to Eq. (2) is computationally prohibitive; the model must be trained twice for each feature

subset S. An alternative, more efficient approach is then needed. The input $\mathbf{x}$ is first simplified to take the form $\mathbf{x}' = \{x'_1, ..., x'_D\}$, where $x'_i \in \{0, 1\}$ implies either the absence (0) or presence (1) of the corresponding feature in $\mathbf{x}$, and where D is the feature dimension. An explanation model $g(\mathbf{x}')$ is then used as an approximation of $f(\mathbf{x})$:

$$f(\mathbf{x}) \approx g(\mathbf{x}') = \phi_0 + \sum_{i=1}^{D} \phi_i x'_i \quad (3)$$

where $\phi_0 = f(h_x(\mathbf{0}))$ (all-zero input) and where h_x is a mapping function that converts $\mathbf{x}'$ to $\mathbf{x}$, i.e., $\mathbf{x} = h(\mathbf{x}')$. The model output is hence approximated by the sum of SHAP values corresponding to the features for which $x'_i = 1$. $g(\mathbf{x}')$ is then trained to approximate the original network output $f(\mathbf{x})$ and the coefficients ϕ_i of the model $g(\mathbf{x}')$ are used in place of true SHAP values [20, 30].

Even with these approximations, the calculation of SHAP values for deep neural networks remains computationally challenging. DeepSHAP [20] provides an efficient method to estimate SHAP values for deep models. With assumptions of feature independence and model linearity, DeepSHAP approximates absent features with expected values. This approach mitigates the need for repetitive model retraining in Eqs. (1) and (2). SHAP values for simple network components (linear, max pooling, activation) are then estimated. The definition in Eq. (3) connects SHAP with DeepLIFT [31], an additive feature attribution method. DeepLIFT multipliers can then be passed via backpropagation to estimate SHAP values at the model input level. Full details can be found in [20, 30, 32].

4. APPLICATION TO SPEECH DATA

We describe how DeepSHAP can be applied to speech data and an arbitrary binary classifier. A time-series speech waveform $\mathbf{x}(t)$ sampled at 16 kHz is first transformed into a short-term spectro-temporal decomposition using the short-time Fourier transform (STFT) giving a spectrogram $\mathbf{X}(m, n)$, where m is the spectral bin and where n is the frame index. An example time waveform and spectrogram are shown in Fig. 1a and 1b. DeepSHAP is then applied treating a given (pre-trained) classifier (here a spoofing detection system) or network model as $f(\cdot)$ in Eq. (2) and each spectro-termporal bin in $\mathbf{X}(m, n)$ individually for every pair (m,n) as a feature, akin to i in Eq. (1), and the whole spectrogram as the full feature set F. For a given classifier, SHAP values ϕ_i in Eq. (2) are calculated to determine the relative contribution of each point (m, n) in $\mathbf{X}$ to the classifier output. For a binary classifier (later spoofing detection), a pair of ϕ_i are obtained, each one representing support for one of the two classes (later bona fide and spoof classes).

SHAP values $\phi_{(m,n)}$ can then be visualised in a similar manner to the spectrogram $\mathbf{X}$. This process can be applied individually to any input $\mathbf{x}(t)$. An example for an arbitrary speech signal (here a bona fide speech utterance) is shown in Fig. 1c. It shows the degree to which each spectro-temporal bin contributes to the classifier output. Darker red points indicate the spectro-temporal bins which lend stronger support for the positive class (here bona fide). In contrast, darker blue points indicate greater support for the negative class (here, spoofed speech).

SHAP visualisations such as that in Fig. 1c can be sparse, indicating that only few spectro-temporal bins contribute to the classifier output. A comparison of the time waveform in Fig. 1a and the SHAP values in Fig. 1c shows that this particular classifier essentially ignores information contained in non-speech regions, focusing instead upon the speech interval between approximately 1 and 2 seconds and, furthermore between frequencies mostly below 1.5 kHz.
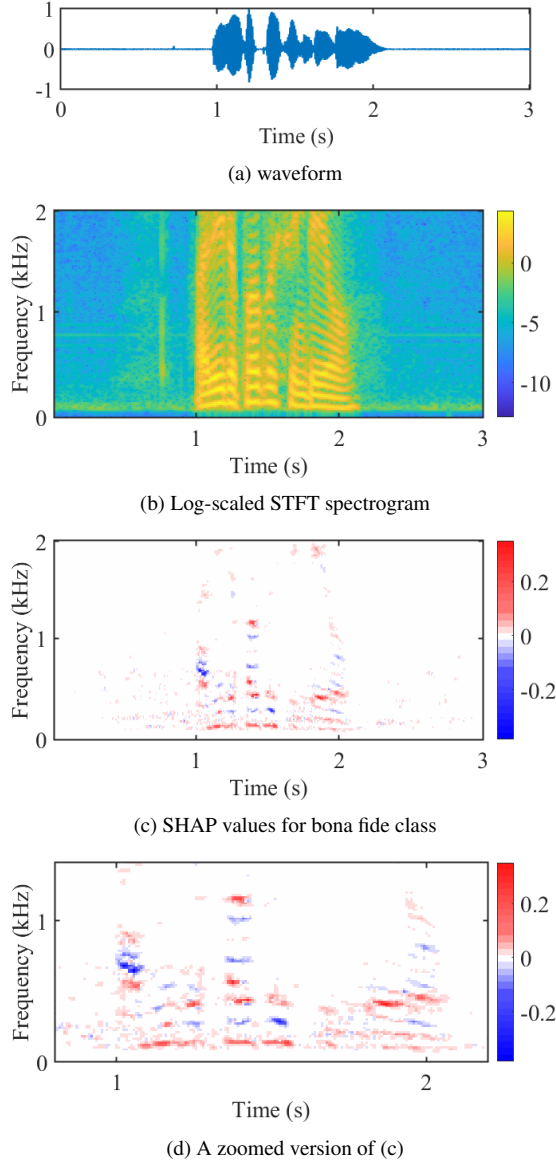
(a) waveform

(b) Log-scaled STFT spectrogram

(c) SHAP values for bona fide class

(d) A zoomed version of (c)

Fig. 1: Illustrations of SHAP analysis for the 'LA_E_3757378' bona fide utterance contained within the ASVspoof 2019 logical access evaluation dataset.

A second visualisation focusing on this specific region is displayed in Fig. 1d. Ignoring for now whether or not the SHAP values are positive or negative, it exhibits a high degree of correlation to the fundamental frequency and harmonics in the spectrogram, indicating the focus of the classifier on these same components. Last, while the presence of dark blue traces in Fig. 1d indicate components of the spectrogram which favour the negative class, the overall dominance of red colours (though not all *dark* red) indicate a greater support for the positive class (the classifier output correctly indicates bona fide speech). Plots of SHAP values such as those shown in Fig. 1c are not easily visualised without the use of dilation operations or some other such smoothing operations which distort the results. While they offer interesting insights, we need more easily visualised means with which to explore results.

5. EXPLAINABILITY FOR SPOOFING DETECTION

In the remainder of this paper we describe our use of DeepSHAP to help explain the behaviour of spoofing detection systems. We show a number of illustrative examples for which the inputs, all A17 utterances drawn from the ASVspoof 2019 LA database [13], are chosen specially to demonstrate the potential insights which can be gained. Given the difficulty in visualising true SHAP values, in the following we present average temporal or spectral results. Given our focus on spoofing detection, we present results for both bona fide and spoofed utterances and the temporal or spectral regions which favour either bona fide or spoofed classes. Results hence reflect where, either in time or frequency, the model has learned to focus attention and hence help to explain its behaviour in terms of how the model responds to a particular utterance.

5.1. Spoofing detection system

While the goal of this work is to demonstrate how DeepSHAP can be applied to the spoofing detection task, rather than an explainability study for a particular classifier, results are nonetheless specific to the latter. We used two different classifiers in this work. Source code is available for both and can be used with the open-source DeepSHAP toolkit[1] to reproduce our results. The first model is our own partial channel differentiable architecture search (PC-DARTS) model described in [33].[2] The second is the 2D-Res-TSSDNet model presented in [8].[3] Both are built upon the same base concept of ResNets [34], though the convolutional block architectures of the PC-DARTS model are optimised (searched) automatically rather than manually. The models operate upon log-scaled power spectrograms extracted from 4-second utterances (PC-DARTS) and magnitude spectrograms extracted from 6-second utterances (2D-Res-TSSDNet), both of which are formed in the usual way from the concatenation or truncation of source data, and using 64 ms Hamming windows with a 8 ms shift and a 1024-point FFT. Both systems are competitive with other state-of-the-art, single systems [8, 33]. The expected value of a given input feature is computed from 20 different utterances of the same class (bona fide or the particular spoofing attack).

5.2. Classification using non-speech intervals

Fig. 2 shows the results of SHAP analysis for the 'LA_E_1832578' utterance and the PC-DARTS classifier. The plot shows the time waveform (a) and the temporal variation in SHAP values averaged across the full spectrum (b). This first example shows that the classifier has learned to focus predominantly upon non-speech intervals. The support in speech intervals for either class is comparatively lower. These observations are unexpected; it is assumed a priori that spoofed speech detection systems should operate upon *speech*. This observation corroborates the findings in [17], and also [19] which shows that reliable bona fide/spoof decisions might even be inferred from the *length* of the non-speech interval.

5.3. Sub-band artefacts

Fig. 3 shows time-averaged SHAP values against frequency for the spoofed 'LA_E_2634822' utterance and the 2D-Res-TSSDNet model, for frequencies up to 4 kHz. Here we see which spectral

---
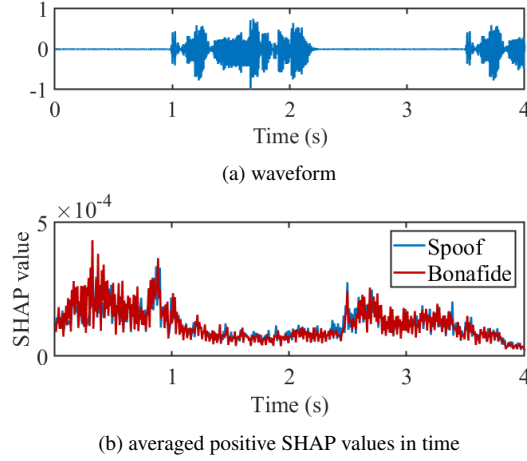
[1] https://github.com/slundberg/shap
[2] https://github.com/eurecom-asp/pc-darts-anti-spoofing
[3] https://github.com/ghuawhu/end-to-end-synthetic-speech-detection

(a) waveform

(b) averaged positive SHAP values in time

Fig. 2: SHAP values for the 'LA_E_1832578' utterance and the PC-DARTS model.
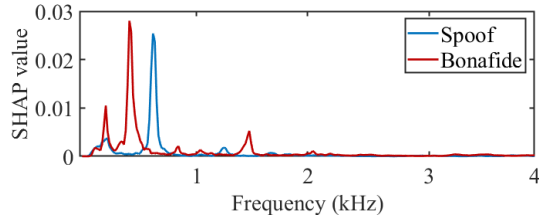
Fig. 3: SHAP values for the 'LA_E_2634822' utterance and the 2D-Res-TSSDNet model.

(a) waveform

(b) 2D-Res-TSSDNet

(c) PC-DARTS

Fig. 4: SHAP values for the 'LA_E_4428024' utterance and both the 2D-Res-TSSDNet and PC-DARTS models.

bins lend the most support on average to either bona fide or spoofed classes. In addition to other less substantial differences, there is predominantly greater support for the bona fide class at 0.5 kHz but substantially greater support for the spoofed class at 0.6 kHz. We observed many instances of such differences suggesting that some spoofing attacks leave artefacts in specific spectral intervals while they are largely effective in replicating the characteristics of bona fide speech in others. Similar observations have been reported previously [7, 24]. Such characteristics may help not just to distinguish between bona fide and spoofed speech, but also to identity a particular spoofing attack algorithm or its nature, e.g., whether it is a synthetic speech, converted voice or replay attack.

5.4. Classifier differences

Fig. 4 shows an example for which SHAP analysis reveals differences in classifier behaviour. The two plots show frequency-averaged SHAP values against time for the 2D-Res-TSSDNet classifier (middle) and the PC-DARTS classifier (bottom) and in both cases the support for the spoof class (blue) and bona fide class (red). The classifiers are shown to apply attention to different intervals. The PC-DARTS model derives its output mostly from non-speech segments whereas the 2D-Res-TSSDNet model applies greater attention to speech intervals. We observed other, more subtle differences in classifier behaviour. SHAP values for the PC-DARTS model are notably more noisy than those of 2D-Res-TSSDNet model. We believe that these differences stem from the relative simplicity of the PC-DARTS model which contains only 0.1 M network parameters, 0.87 M fewer than the 2D-Res-TSSDNet
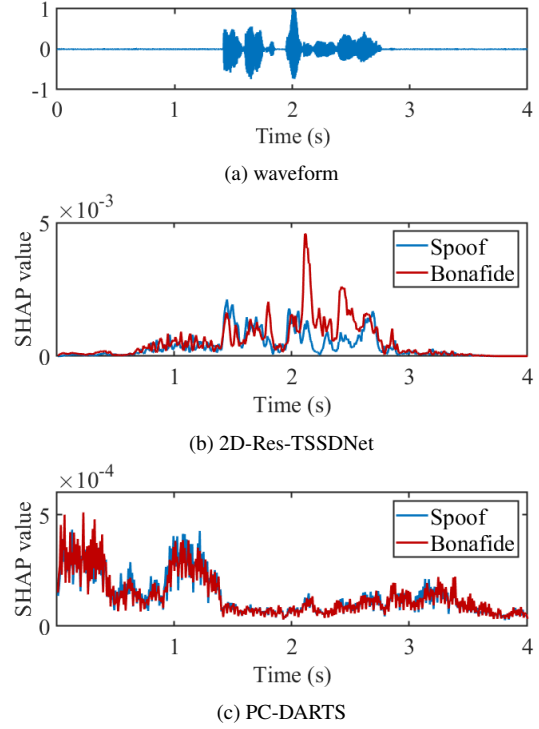
model, possibly implying that the former has insufficient learning capacity, leading to noisier outputs. These characteristics might also be caused by the high number of dilated convolutions in the learned architecture. We observed more stable SHAP values when dilated convolution operations are excluded from the architecture search space. SHAP analysis can hence also be used to explore lower-level classifier behaviours. Instances where different classifiers exhibit different behaviour might also help to improve performance in cases where single system solutions are preferred to system fusion.

6. CONCLUSIONS

This paper demonstrates how DeepSHAP can be applied to explain what influences the outputs produced by a spoofing detection model. The examples shown in the paper show how SHAP analysis can be used to highlight the attention applied by a given classifier at low-level spectro-temporal intervals. Nonetheless, the tool offers the basis for what is needed to explore higher-level explanations. It will be of interest to explore, e.g., whether we can make the link between SHAP results, classifier outputs and specific speech units or spoofing attack algorithms (e.g., synthetic speech, converted voice and replay) and their algorithmic components (e.g., waveform models, recording devices and microphones). Other future directions include the use of SHAP to explore classifier differences in an attempt to explain why some classifiers perform better than others, and under which conditions. In this context, it will be of interest to develop SHAP-based approaches that can be used to compare systems that operate upon hand-crafted spectro-temporal decompositions to those that operate directly upon raw waveforms. Ultimately, of course, the goal is to exploit what we can learned from SHAP analysis to design better performing, more reliable models.

7. REFERENCES

[1] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification - A study of technical impostor techniques," in *Proc. EUROSPEECH*, 1999, pp. 1211–1214.

[2] T. Masuko, T. Hitotsumatsu, K. Tokuda et al., "On the security of HMM-based speaker verification systems against imposture using synthetic speech," in *Proc. EUROSPEECH*, 1999, pp. 1223–1226.

[3] J. Yamagishi, X. Wang, M. Todisco et al., "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *Proc. ASVspoof 2021 Workshop*, 2021, pp. 47–54.

[4] T. Kinnunen, K. A. Lee, H. Delgado et al., "t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proc. Odyssey*, 2018.

[5] X. Li, X. Wu, H. Lu et al., "Channel-wise gated Res2Net: Towards robust detection of synthetic speech attacks," in *Proc. INTERSPEECH*, 2021, pp. 4314–4318.

[6] H. Tak, J.-w. Jung, J. Patino et al., "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. ASVspoof 2021 Workshop*, 2021, pp. 1–8.

[7] Y. Zhang, W. Wang and P. Zhang, "The effect of silence and dual-band fusion in anti-spoofing system," in *Proc. INTERSPEECH*, 2021, pp. 4279–4283.

[8] G. Hua, A. B.-j. Teoh and H. Zhang, "Towards end-to-end synthetic speech detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1265–1269, 2021.

[9] A. Luo, E. Li, Y. Liu et al., "A capsule network based approach for detection of audio spoofing attacks," in *Proc. ICASSP*, 2021.

[10] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[11] G. Lavrentyeva, S. Novoselov, A. Tseren et al., "STC anti-spoofing systems for the ASVspoof2019 challenge," in *Proc. INTERSPEECH*, 2019, pp. 1033–1037.

[12] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Proc. INTERSPEECH*, 2021, pp. 4259–4263.

[13] M. Todisco, X. Wang, V. Vestman et al., "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. INTERSPEECH*, 2019, pp. 1008–1012.

[14] H. Tak, J. Patino, M. Todisco et al., "End-to-end anti-spoofing with RawNet2," in *Proc. ICASSP*, 2021, pp. 6369–6373.

[15] W. Ge, J. Patino, M. Todisco et al., "Raw differentiable architecture search for speech deepfake and spoofing detection," in *Proc. ASVspoof 2021 Workshop*, 2021, pp. 22–28.

[16] T. Chen, A. Kumar, P. Nagarsheth et al., "Generalization of audio deepfake detection," in *Proc. Odyssey*, 2020, pp. 1–5.

[17] B. Chettri, E. Benetos and B. Sturm, "Dataset artefacts in anti-spoofing systems: A case study on the ASVspoof 2017 benchmark," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3018–3028, 2020.

[18] H. Dinkel, N. Chen, Y. Qian et al., "End-to-end spoofing detection with raw waveform CLDNNS," in *Proc. ICASSP*, 2017, pp. 4860–4864.

[19] N. Müller, F. Dieckmann, P. Czempin et al., "Speech is silver, silence is golden: What do ASVspoof-trained models really learn?," in *Proc. ASVspoof 2021 Workshop*, 2021, pp. 55–60.

[20] S. M. Lundberg, S.-i. Lee and D. Fohr, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[21] S. Becker, M. Ackermann, S. Lapuschkin et al., "Interpreting and explaining deep neural networks for classification of audio signals," *CoRR*, vol. abs/1807.03418, 2018.

[22] H. Muckenhirn, A. Abrol, M. Magimai-Doss et al., "Understanding and visualizing raw waveform-based CNNs," in *Proc. INTERSPEECH*, 2019, pp. 2345–2349.

[23] S. Sivasankaran, E. Vincent and D. Fohr, "Explaining deep learning models for speech enhancement," in *Proc. INTERSPEECH*, 2021, pp. 696–700.

[24] H. Tak, J. Patino, A. Nautsch et al., "An explainability study of the Constant Q Cepstral Coefficient spoofing countermeasure for automatic speaker verification," in *Proc. Odyssey*, 2020, pp. 333–340.

[25] R. R. Selvaraju, M. Cogswell, A. Das et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[26] B. Halpern, F. Kelly, B. van Son et al., "Residual networks for resisting noise: Analysis of an embeddings-based spoofing countermeasure," in *Proc. Odyssey*, 2020, pp. 326–332.

[27] M. T. Ribeiro, S. Singh and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[28] B. Chettri, S. Mishra, B. L. Sturm et al., "Analysing the predictions of a CNN-based replay spoofing detection system," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 92–97.

[29] L. S. Shapley, "A value of n-person games," *Contributions to the Theory of Games*, pp. 307–317, 1953.

[30] M. Christoph, "Interpretable machine learning: A guide for making black box models explainable," 2020.

[31] A. Shrikumar, P. Greenside and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, vol. 70, pp. 3145–3153.

[32] S. M. Lundberg, G. G. Erion and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[33] W. Ge, M. Panariello, J. Patino et al., "Partially-connected differentiable architecture search for deepfake and spoofing detection," in *Proc. INTERSPEECH*, 2021, pp. 4319–4323.

[34] K. He, X. Zhang, S. Ren et al., "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.