Interpretable Deepfake Voice Detection: A Hybrid Deep-Learning Model and Explanation Evaluation

Jacob LaRock[1], Md Shajalal[1] and Gunnar Stevens[1,2]

[1]University of Siegen, Siegen, Germany
[2]Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany

Abstract

With the unprecedented advancement of Generative Artificial Intelligence (GenAI), the threat of voice scams using synthetic voices has become a serious concern across various sectors. Recent efforts have focused on identifying fake voices through handcrafted features, deep learning models, and hybrid approaches. However, most existing methods lack explainability, rendering their predictions non-transparent to users. This paper proposes a novel, interpretable, and transparent method for fake voice identification by introducing a hybrid deep learning model that leverages multiple extracted features. The hybrid model consists of two main components: the first component addresses heterogeneous feature spaces by employing deep convolutional sub-models tailored to individual features, while the second component, the terminus model, utilizes the concatenated representations from the final layers of each sub-model as input. The terminus model follows a typical multi-layer perceptron architecture, enabling effective integration and classification of the diverse feature representations. To enhance interpretability, we decompose the model's decisions using Local Interpretable Model-agnostic Explanations (LIME), taking advantage of the identical feature representation before the concatenation layers to address challenges related to multidimensional feature representations. To evaluate the features and assess the quality of the generated explanations, we propose two metrics: importance and trust. Extensive experiments are conducted on the In-the-Wild dataset, which is designed to test the generalization capability of synthetic audio detection methods. The experimental results demonstrate that our approach achieves performance comparable to benchmark methods. Furthermore, the results based on our proposed metrics conclude that certain perceptible features demonstrate promise for generating explanations that are meaningful to general users. For reproducibility, the source code for these experiments is available in the following repository: https://github.com/jacoblarock/fake_voices_xai

Keywords

Explainable AI (XAI), Fake Voice Detection, Hybrid Model, DeepFake Detection, Explanation Evaluation, Metrics

1. Introduction

Since the use of generative methods for creating synthetic voices has become more widespread, the need for reliable and usable detection methods to protect the security of individuals and businesses has grown. In particular, the rise of deepfake technology has raised concerns about its potential misuse in areas such as politics, entertainment, and national security. For instance, malicious actors could exploit this technology to create fake audio recordings that appear to be genuine statements made by public figures or to fabricate recordings of events that never occurred. Such misuse could have significant consequences, including the spread of misinformation, defamation of individuals, and the escalation of political tensions [1].

There has been considerable attention on identifying audio deepfakes using classical and sophisticated deep learning models [2, 3, 4, 5]. Methods for fake voice identification can be broadly categorized into two classes: methods with handcrafted feature extraction and methods with end-to-end deepfake detectors [6]. In the former, detection approaches first extract various features from the voices. These high-dimensional feature values are then passed through complex deep learning models to determine whether the voice is real or synthesized (i.e., fake). End-to-end fake voice identification methods generally optimize the feature extraction process and classification task jointly. Both methods, however, have shown promise in the field to produce reliable results.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

jacoblarock@gmail.com (J. LaRock); md.shajalal@uni-siegen.de (M. Shajalal)

The field of Explainable AI (XAI) shows promise for producing useful and interpretable results from such models. Explainable AI refers to the ability of artificial intelligence systems, such as machine learning models and neural networks, to provide understandable and interpretable explanations for their decisions or predictions. This means that XAI systems can articulate why they made a particular decision, what factors influenced it, and how confident they are in their conclusion.

In this work, we introduce a hybrid deep learning model leveraging perceptible features extracted from voice signals in order to use those features in interpretable explanations. The dimensions of each perceptible feature are not the same. Therefore, our hybrid model consists of a component sub-model for each input feature to address the problem of varying dimensionality of the extracted features without the drawbacks of costly transformations in the pre-processing phase. The concatenation of the sub-models' outputs is followed by a terminus model. The terminus model, which uses the concatenated outputs of the individual sub-models as inputs, distills its inputs into a singular output using a classic multi-layer perceptron architecture, allowing for an effective combination of the features into a final classification result.

To introduce explainability into our method, we generate explanations using the output of each sub-model before the concatenation layers and then apply Local Interpretable Model-agnostic Explanations (LIME) [7] on the terminus model. This allows us to assess the impact of each input feature on the final result through local explanations. Due to the independence of the sub-models, the importance of their outputs directly correlates with the importance of their inputs for the end classification made by the terminus model. We then introduce two metrics: *trust* and *importance*. These metrics allow for an aggregate evaluation of a large number of generated LIME explanations on a per-feature basis, enabling us to assess the usefulness of the features within a given feature set for producing useful and understandable explanations. We conducted a wide range of experiments to evaluate the performance of our hybrid audio deepfake detection approach on the *In-The-Wild* dataset. The experimental results demonstrate that our model performs effectively compared to state-of-the-art methods. The evaluation of the generated explanations using our proposed metrics also identifies which features have the highest impact on the model's overall predictions. The contributions of this paper are twofold: i) we propose an interpretable hybrid deep learning model to identify synthetic voices using perceptible features, and ii) we propose two metrics in order to perform aggregate evaluations of many explanations from our LIME-based method, in order to assess how useful the average explanations are.

2. Related Work

The most common approach in this domain involves using either learned or hand-crafted imperceptible features. Examples include spectrographic features such as mel-spectrograms and their hand-crafted derivatives, such as mel-frequency cepstral coefficients (MFCCs), both of which are widely used due to their effectiveness. Anagha et al. [8] utilized mel-spectrograms in combination with a convolutional neural network (CNN)-based architecture, achieving strong performance on the ASVSpoof2019 dataset [5]. Müller et al. [3] introduced the In-the-Wild dataset to assess the generalizability of deepfake voice detection methods. Their evaluation of various features and architectures revealed that existing methods struggle with generalization. Yang et al. [4] conducted a comparative analysis of multiple features and a feature selection method aimed at improving model efficiency. Their experiments, conducted on ASVSpoof2019[5], ASV2021 [9], and In-the-Wild [3] datasets, demonstrated the impact of different features on model performance and emphasized the benefits of feature selection and classification fusion techniques. Ranjan et al. [10] proposed a deep convolutional network designed for both spoof detection and source identification. Their evaluation across ASVSpoof2019[5], the FOR-Norm dataset [11], and the In-the-Wild dataset [3] showed high accuracy in individual dataset evaluations. However, cross-dataset evaluations revealed limitations in generalizability. Yi et al. [12] provided a comprehensive comparison of features and model architectures for synthetic voice detection. Their study, conducted on ASVSpoof2019 [5] and In-the-Wild [3] datasets, reinforced the observation that existing methods struggle with generalization across datasets.

Although less explored, several studies have investigated the use of perceptible features for detecting synthetic voices. Some of these features were incorporated into our experiments. Barrington et al. [13] examined the potential of perceptible features in deepfake audio classification, emphasizing their role in improving explainability. However, while they implemented a classifier, they did not develop an explainer. Their findings indicated a performance drop when using perceptible features compared to imperceptible hand-crafted and deep-learning-based features. Chaiwongyen et al. [14, 15] explored perceptible feature-based classification. Their initial perceptron model, trained and tested on the ADD2022 Challenge dataset [16], showed limited performance in 2022. However, with an expanded feature set in 2023, they achieved improved results. Li et al. [17] investigated a hybrid approach combining perceptible and imperceptible (referred to as "physical" in their work) features. Their experiments, conducted using various neural networks on the ASV2022 Challenge dataset [16], demonstrated that integrating both feature types yielded the best performance, outperforming models trained on only one type of feature.

Efforts have also been made to integrate explainability into synthetic voice detection models. Ge et al. [18] applied the SHAP (SHapley Additive exPlanations) method to analyze feature influence in deepfake audio detection. Using log-scaled power spectrograms as input features, they trained and tested their model on the ASV2019 dataset [5]. Their approach enabled graphical representation of feature importance on spectrograms and a global summary of SHAP values. Haq et al. [19] proposed an explainable approach by leveraging emotional state changes as input features. Their method visually represented "unlikely" emotional shifts to enhance interpretability for end users. By combining fake video and fake audio classifiers, they produced a final classification for video samples with audio. Their model, tested on the Presidential Deepfake dataset, achieved superior results compared to existing benchmarks. Also relevant are other hybrid deep-learning approaches. Concatenated sub-networks have been explored in some works such as with concept-based [20] models and neural additive models [21].

3. Method

Our proposed fake voice identification method consists of three major components including feature extractors, hybrid detection model, and the generation of explanations by modified LIME. After extracting features, we pass them individually through sub-models to overcome the multidimensionality problem. Then the concatenation of the outputs from the output layers of the sub-models are then passed through the *terminus* models. For generating explanations, we make use of the feature values returned from the output layers of the sub-models, which has same representation of every feature.

3.1. Feature Extraction

The extracted features from audio samples can be categorized into two different classes: perceptible and imperceptible features. The perceptible features are features that can be perceived by the human ear, often vocal qualities such as jitter, shimmer or pitch fluctuation that have a wide range of uses even outside of audio classification such as diagnosis of disease [15], while imperceptible features are typically out of the range of human hearing, and may not directly reflect a vocal quality, otherwise referred to as “speaker-independent” features [22]. In order to increase the likelihood that the explanations are useful and understandable to the end user, we focused on using multiple perceptible features as input to the classifier. However, we used two imperceptible features with the hypothesis, based on previous research [13, 14, 15, 17], that they would positively increase model performance.

The exact perceptible features we extracted are the following: Harmonic to noise ratios (HNRs), fundamental frequency lengths (f0 lengths), onset strengths, intensity, pitch-fluctuations, jitter and shimmer. The imperceptible features that we extracted for our analysis are the mel-spectrogram and their derivative mel-frequency cepstral coefficients (MFCCs). The features have varying dimensionalities from vectors to matrices of varying size.

3.2. Hybrid Fake Voice Detection Model

We propose an architecture for our hybrid fake voice detection model that allows for a combined classification based on features of different dimensionality, or in other words, regardless of whether the individual features have the same shape. To achieve this, we hypothesize that we can make use of individual separate models for each feature, which we will further refer to as *sub-models*, in a way resembling concept-based models . The main objective for having the individual sub-models is to have dimensionally similar representation of the features so that we can use them for generating understandable explanations. The structure of sub-model can be different based the dimension of the extracted features. For example, the dimension of convolution layers and max pooling layer would be different based on the dimension of the features. Sub-models may also omit the pooling layers all-together, in order to have a structure based purely on convolution. We hypothesize that the use of convolution within these sub-models will increase the localized pattern detection capability of our method. After the processing within the sub-models, we then needed to further process and distill the results into a singular output. To achieve this, followed by the concatenations of the output of all sub-models, we propose a terminus model. With the goal of maximizing performance, we tried various structures for the terminus models including single layer perceptron, convolutional neural networks and multi-layer perceptron.

3.3. Explaining the prediction

3.3.1. Generation of the Explanations

The choice to use the LIME method came from its implementability as well as its proven performance on tabular data, which is the most relevant approach to creating explanations for our method. Being that the sub-models are separate from one another, having no influence on each other before being concatenated at the terminus, only the terminus part of the model is relevant for assessing the weights of the features in a single evaluation. Therefore, we make use of the processed features' values after the output layers of individual sub-models. This does, however, pose two challenges: i) there are multiple input rows per sample because a sliding-window is used across each audio sample and ii) the input features cannot be used directly for assessment at the terminus input layer.

We solved the first problem by taking the mean of the weights returned by the LIME explainer of each feature produced by each row of the sample features. This results in an average contribution of each feature to the classification of the model. This is due to the fact that the end-classification of the surrogate is also a mean. Therefore, the averages of the weights from the LIME explainer represent aggregated influences of the features. In order to address the second problem, we needed to, for each input feature, figure out what the output of the respective sub-model would be. This would allow us to transform the input into the vector-shaped input for the terminus model from the given sample in the input layers in every sub-model. After the processing of the sub-model, we catch the representation before the processing of the terminus model. This allows us to use the explainer to find a weight of each input of the terminus model in relation to the end prediction. We generate the sub-model results for a random sample of the training set to use as a reference for the localized model approximations from the LIME explainer. We then summarized the results of the LIME explainer together on a per-feature basis and normalized them to produce a decimal number between -1 and 1 for every feature. In this case, a negative value implies that the given feature pushed the result of the model in the negative direction (i.e. not fake), while a positive value indicates the opposite.

3.3.2. Evaluation of the Explanations

In order to make an assessment the potential of this method of generating explanations, we needed a way to measure the influence of each feature on the average explanation generated using this method. For context, we make the following definitions: S as the set of samples with $s \in S$ as a sample; F as the set of features with $f \in F$ as a feature; w_{fs} is the weight of feature f in sample s as produced

by the explainer; L_s is the correct label for sample s. The first metric we propose is the mean of the absolute values of the weights of each feature within a set of explanations. We will further refer to this metric as *importance*, and we define it mathematically as follows for every feature $f \in F$ extracted from the samples in S: $Imp(f, S) = \frac{\sum_{s \in S} |w_{fs}|}{|S|}$. The second metric we will define is the average aggregate correctness on a per-feature basis, which we will further refer to as trust. The trust per feature $f \in F$ extracted from samples in S can be defined as $Trust(f, S) = \frac{\sum_{s \in S} w_{fs}(2l_s - 1)}{|S|}$.

4. Experimental Results and Evaluation

Dataset. We trained and tested our model employing the *In-the-Wild* dataset. This dataset focuses on the generalization of audio deepfake detection models by collecting real-world data, in comparison to other research that used more controlled laboratory conditions [3]. We have chosen it for these experiments because of the aforementioned focus on generalization, its relative recency compared to some others and the fact that it has also been used previously by some other experiments, which provides a useful perspective against which we can compare our method.

Experimental Results. We present experimental results for all variations of our proposed model in Table 1. The table details the following information: The type of terminus model used, where "perceptron" does not include hidden layers, MPL(3) is a multi-perceptron layer model with three hidden layers and CNN (3) is a convolutional neural network with three hidden layers; we used different set of features, where "standard" is the feature set as described above including only one pitch-fluctuation feature. However, with expanded pitch fluctuations, we combine the same feature set with the addition of more pitch fluctuation features with different comparison distances; the training batch size and the number of epochs per batch if there were more than one; followed by several performance metrics.

Table 1
Summary of the evaluation results of the experiments

Terminus	Features	Training	Accuracy	EER	AUC
Perceptron	standard	3474	95.02%	0.03702	0.90297
MPL (3)	standard	10000	**96.27%**	**0.03214**	0.81763
Perceptron	standard	23833 2 epochs	90.07%	0.90069	0.78889
Perceptron	standard	10000 2 epochs	94.43%	0.04408	0.89445
Perceptron	standard	10000	94.28%	0.04840	0.90182
CNN (3)	standard	10000	62.80%	0.37198	0.50051
Perceptron	expand pitchflucs	10000	93.31%	0.05661	**0.92727**
MPL (3)	expand pitchflucs	10000	93.87%	0.05317	0.83263

We can see that the hybrid model with an MLP terminus model trained with standard features set on first 100000 samples achieved higher accuracy (96.27%) and ERR (0.03214). However, the model with perceptron as a terminus applying features with expanded pitch fluctuations-based features performed better than other experimental settings in terms of AUC (0.92727). The other experimental settings also performed comparatively except with terminus models based on convolutional neural networks. Percepton-based terminus models with all variation also performed effectively with (2-3)% deviation compared to best performing model. The performance comparison among state-of-the-art methods is summarized in table 2.

Evaluating explanations generated by LIME. In order to obtain a comprehensive understanding of the quality of the explanations generated by LIME, we selected the three best-performing models and generated explanations for the first 500 samples of the testing data. These three models achieved the best performance in terms of Accuracy and Equal Error Rate (EER) or AUC. They include: a model using the standard feature set with a terminus model containing three hidden layers, model with expanded pitch-fluctuation features and a perceptron terminus, which achieved the best AUC result, and model with expanded pitch-fluctuation features and a terminus containing three hidden layers. We

Table 2
Performance comparison of our method with existing methods applied on the same dataset.

Author	Methods	Equal Error Rate (ERR)
Ranjan et al.[10]	STATNet	**0.00199**
Our Method	**Hybrid DL**	0.03214
Yang et al. [4]	Fusion	0.2427
	Selection	0.2598
	ResNet18	0.2748
Yi et al. [12]	ASSERT	0.2473
	LCNN	0.3514
	GMM	0.3749
Müller et al. [3]	RawGAT-ST	0.37154
	MesoInception	0.37414
	RawNet2	0.37819

selected these models to evaluate the best-performing configurations and to observe how the inclusion of expanded pitch-fluctuation features affects the evaluation of the explanations. We then aggregated and summarized the explanations using the two previously introduced metrics: *trust* and *importance*. A summary of the evaluation in terms of these metrics, grouped by feature, is presented in Table 3.

Table 3
Evaluation of generated explanations in terms of *trust* and *importance* for best performing model in terms of EER

Feature	Best-performing		Expanded pitch-flucs		Exp. pitch-Flucs MPL terminus	
	Imp	Trust	Imp	Trust	Imp	Trust
HNRs	0.1140	-0.0543	0.1072	-0.0148	0.2667	-0.0788
mel spectrogram	0.1084	-0.0378	0.0428	-0.0187	0.2667	-0.0788
MFCC	**0.7029**	**0.3753**	**0.4383**	**0.2345**	**0.2879**	**0.2408**
f0 lengths	0.0	0.0	0.0340	-0.0057	0.1690	-0.0521
onset strengths	0.0	0.0	0.0101	-0.0027	0.1384	-0.0613
intensities	0.0	0.0	0.0021	0.0021	0.1576	-0.0504
pitch fluctuations	0.0	0.0	0.0059	0.0020	0.0	0.0
jitter features	0.1277	0.0250	0.2439	0.0019	0.1444	0.0118
shimmer features	0.1227	0.0017	0.1664	0.0070	0.1747	0.0116

According to both metrics, we can see that in Table 3 , the MFCCs had the most positive and accurate influence on the classification results, with the jitter features being a distant second in terms of trust and importance. This indicates that the MFCCs not only had the greatest influence but also the most reliable impact on the classification outcomes. In contrast, other features had little, no, or even negative influence on the classification, depending on the experiment. However, the jitter and shimmer features are notable as perceptible features with a positive trust value. It is possible that the consistent correctness of the MFCC feature led to the other features having less influence over the course of training, thus decreasing their importance metric even when the average contribution is positive.

Viability and Potential of the Explanations. We hypothesized that explanations leveraging perceptible features have high potential to enhance the interpretability of complex predictions. However, as shown in the previous section in the Table 3, perceptible features contributed less to classification performance than their imperceptible counterparts, despite their increased presence in the overall set of input features. Two features—HNRs and mel-spectrograms—even consistently had a net-negative impact on classification accuracy in our experiments, as measured by our *trust* metric. This leaves an open question whether future experiments might perform better without these features. In the model without expanded pitch fluctuations, four perceptible features—f0 lengths, onset strengths, intensities, and pitch fluctuations—had an overall average influence of zero on classification across both metrics. This suggests that the model did not learn a meaningful correlation between these features and the authenticity of audio samples in the dataset. Although these features did have some impact on the

performance of the model with expanded pitch fluctuations, the effect remained minimal. For perceptible features with a positive *trust* score, there remains potential for their use in generating understandable explanations. Even though these features did not significantly influence the model's predictions, they still demonstrate a certain level of reliability. Nevertheless, we believe it is worth investigating whether other types of perceptible features might exert a stronger influence on classification or whether alternative imperceptible features—potentially replacing MFCCs—could achieve a better balance between classification accuracy and proportional influence on the final result. Furthermore, we believe that this model architecture, when paired with LIME and evaluated using the proposed metrics, has the potential to serve as a foundation for future research.

5. Conclusion

To conclude, we have presented a hybrid interpretable deep learning model that leverages a combination of heterogeneous features, both perceptible and imperceptible. We hypothesized that such a model could be used to generate explanations using LIME, which might be more useful for potential end users. The experimental results demonstrated top-tier performance in accurately identifying fake voices while mitigating the dimensionality problem in the input features. We observed that explanations leveraging the representations just before the model's final layer can generate technical insights. These explanations might be useful for improving model performance by modifying important parameters or as the basis for presenting an analysis of a sample to an end user, in the case where perceptible features have an influence on the classification, so they can understand if audible flaws in the sample are present. However, based on our proposed metrics, *trust* and *importance*, we did not observe a significant influence or usefulness of perceptible features, although they did contribute to explanations to some extent. A promising direction for future work would be integrating this method into different domains where feature-based local explanations have the potential to be effective. In such application domains, we could further validate both the accuracy and the usefulness of the generated explanations as well as the selected features using our proposed metrics. Additionally, it would be valuable to compare findings from a user study on the generated explanations with our metric-based evaluation.

Acknowledgments

This research has been funded by the AntiScam Project (Defense against communication fraud), funded by BMBF Germany, Grant reference 16KIS2214

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] N. Veerasamy, H. Pieterse, Rising above misinformation and deepfakes, in: International Conference on Cyber Warfare and Security, volume 17, Academic Conferences International Limited, 2022, pp. 340–348.

[2] R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma, Y. Liu, Deepsonar: Towards effective and robust detection of ai-synthesized fake voices, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1207–1216.

[3] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, K. Böttinger, Does audio deepfake detection generalize?, arXiv preprint arXiv:2203.16263 (2022).

[4] Y. Yang, H. Qin, H. Zhou, C. Wang, T. Guo, K. Han, Y. Wang, A robust audio deepfake detection system via multi-view feature, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 13131–13135.

[5] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, et al., Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech, Computer Speech & Language 64 (2020) 101114.

[6] A. Dixit, N. Kaur, S. Kingra, Review of audio deepfake detection techniques: Issues and prospects, Expert Systems 40 (2023) e13322.

[7] M. T. Ribeiro, S. Singh, 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.

[8] R. Anagha, A. Arya, V. H. Narayan, S. Abhishek, T. Anjali, Audio deepfake detection using deep learning, in: 2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART), IEEE, 2023, pp. 176–181.

[9] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, et al., Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild, IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2023) 2507–2522.

[10] R. Ranjan, M. Vatsa, R. Singh, Statnet: Spectral and temporal features based multi-task network for audio spoofing detection, in: 2022 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2022, pp. 1–9.

[11] R. Reimao, V. Tzerpos, For: A dataset for synthetic speech detection, in: 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), IEEE, 2019, pp. 1–10.

[12] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, Y. Zhao, Audio deepfake detection: A survey, arXiv preprint arXiv:2308.14970 (2023).

[13] S. Barrington, R. Barua, G. Koorma, H. Farid, Single and multi-speaker cloned voice detection: From perceptual to learned features, in: 2023 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2023, pp. 1–6.

[14] A. Chaiwongyen, N. Songsriboonsit, S. Duangpummet, J. Karnjana, W. Kongprawechnon, M. Unoki, Contribution of timbre and shimmer features to deepfake speech detection, in: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2022, pp. 97–103.

[15] A. Chaiwongyen, S. Duangpummet, J. Karnjana, W. Kongprawechnon, M. Unoki, Deepfake-speech detection with pathological features and multilayer perceptron neural network, in: 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2023, pp. 2182–2188.

[16] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, et al., Add 2022: the first audio deep synthesis detection challenge, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 9216–9220.

[17] M. Li, Y. Ahmadiadli, X.-P. Zhang, A comparative study on physical and perceptual features for deepfake audio detection, in: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, 2022, pp. 35–41.

[18] W. Ge, J. Patino, M. Todisco, N. Evans, Explaining deep learning models for spoofing and deepfake detection with shapley additive explanations, in: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2022, pp. 6387–6391.

[19] I. U. Haq, K. M. Malik, K. Muhammad, Multimodal neurosymbolic approach for explainable deepfake detection, ACM Transactions on Multimedia Computing, Communications and Applications 20 (2024) 1–16.

[20] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, P. Liang, Concept bottleneck models, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 5338–5348.

[21] R. Agarwal, N. Frosst, X. Zhang, R. Caruana, G. E. Hinton, Neural additive models: Interpretable machine learning with neural nets, CoRR abs/2004.13912 (2020). arXiv:2004.13912.

[22] X. Liu, Y. Tan, X. Hai, Q. Yu, Q. Zhou, Hidden-in-wave: A novel idea to camouflage ai-synthesized voices based on speaker-irrelative features, in: 2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE), IEEE, 2023, pp. 786–794.