CoughLIME: Sonified Explanations for the Predictions of COVID-19 Cough Classifiers

Anne Wullenweber[1,3,*], Alican Akman[1,*], Björn W. Schuller[1,2]

***Abstract*— Since the emergence of the COVID-19 pandemic, various methods to detect the illness from cough and speech audio data have been proposed. While many of them deliver promising results, they lack transparency in the form of explanations which is crucial for establishing trust in the classifiers. We propose CoughLIME which extends LIME to explanations for audio data, specifically tailored towards cough data. We show that CoughLIME is capable of generating faithful sonified explanations for COVID-19 detection. To quantify the performance of the explanations generated for the CIdeR model, we adopt pixel flipping to audio and introduce a novel metric to assess the performance of the XAI classifier. CoughLIME achieves a $\triangle$AUC of 19.48 % generating explanations for CIdeR's predictions.**

I. INTRODUCTION

The worldwide coronavirus disease 2019 (COVID-19) pandemic has immensely impacted many sectors, among them the global economy and the mental health of the population [1]. To this date, 497,960,492 cases of COVID-19 and 6,181,850 deaths have been reported[1]. Arguably, frequent testing of major parts or, ideally, the entire population is an efficient way to tackle the pandemic [2]. However, clinical tests come with various challenges and open questions, among them implementing a regular testing strategy for large parts of the population [3]. Pathological changes in the lungs and vocal systems attributed to COVID-19 suggest that the disease could be detected solely from the voice of a patient [4]. This raises the question whether COVID-19 could be detected from speech or cough data employing Machine Learning (ML). An application that takes an audio sample as input and outputs the probability of a current COVID-19 infection could run on any smartphone and be combined with traditional tests to form an efficient testing strategy.

Some proposed applications detecting COVID-19 from speech or cough data deliver good results [5]; however, in order for them to be widely accepted, trust in the systems is crucial. To establish trust, users must understand how the model determines its predictions. Explainable artificial intelligence (XAI) aims at providing explanations for the output of an AI application that are understandable to humans. While many promising applications for detecting COVID-19 from audio data have been proposed, none of them focus on providing explanations for their models' decisions. We argue that in order for a COVID-19 detection application to be deployed, trust, hence explaining its predictions, is crucial. Additionally, methods to explain COVID-19 prediction models can help medical experts and users understand and accept the models' decisions. Intuitively, explanations for predictions on audio data should be provided in the form of audio. However, existing XAI techniques almost exclusively focus on visual and textual explanations. In this work, we introduce CoughLIME to obtain sonified explanations for the predictions of ML applications. It extends Ribeiro et al.'s approach providing Local Interpretable Model-agnostic Explanations (LIME) [6] for use with audio data, specifically tailored towards cough data. To demonstrate its functionality, we apply CoughLIME to explain the predictions of the CIdeR model [7].

II. RELATED WORK

Various approaches to detect active COVID-19 infections from breath, speech, or cough samples have been proposed. They rely on end-to-end learning and automatic feature engineering [8], [7], [9] or hand-crafted features such as Mel-Frequency Cepstral Coefficients (MFCCs) and their deltas or Mel-spectrograms with dedicated classifiers [10].

While many of these studies lead to high accuracies, they have various deficiencies. Given the novelty of COVID-19, only little training data with inconsistent ground truth annotations are available [5]. High performing COVID-19 detection models from audio data often perform poorly across data sets [11]. Furthermore, the vast majority of the architectures do not provide explanations for their predictions. However, given the detection being a healthcare application, explanations are crucial to establish trust in the model [12].

XAI methods proposed up to this date can be classified into four categories: interpretable local surrogates, occlusion analysis, integrated gradients/smoothGrad, and layer-wise relevance propagation [13]. An example for an interpretable local surrogate is LIME [6] which generates local explanations for a specific sample and is compatible with any model. Explanations are generated by perturbing the input and training a surrogate model which assigns weights to each input component. LIME's authors originally implemented the technique for image and text data [6]. Extensions of LIME for audio data that have been proposed in the literature focus on music analysis [14], [15]. While SoundLIME is generally applicable to audio data, it does not focus on providing listenable explanations [14]. AudioLIME does provide listenable explanations, but separates the audio data into different sources [15]. While this is applicable

*Equal contribution.
[1]GLAM – Group on Language, Audio, & Music, Imperial College London, London, UK
[2]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]Technical University of Munich, Germany
[1]https://covid19.who.int, accessed 12 April 2022

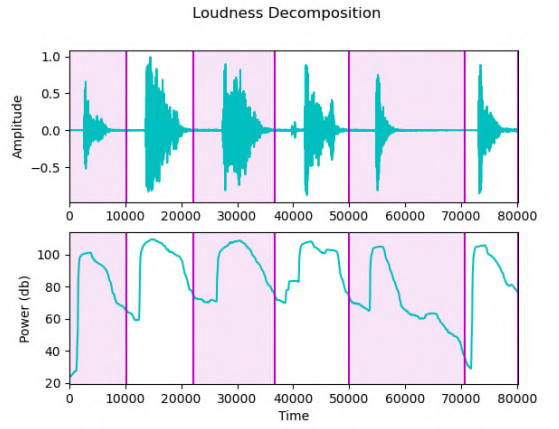
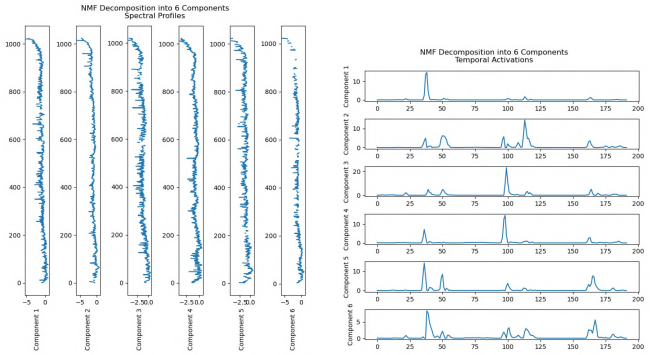


Fig. 1. Example loudness decomposition of a cough audio file with threshold 75 dB.



(a) (b)

Fig. 2. Example NMF decomposition of a cough audio file into 6 components consisting of spectral profiles (a) and temporal activations (b).

to instrumental music, it does not generalise to all audio data as, e. g., cough data only has one source. This work focuses on introducing CoughLIME to prove the feasibility of explaining the predictions of COVID-19 detection models in the form of listenable, hence true to audio, explanations.

III. CoughLIME

CoughLIME extends the basic LIME functionality developed by Ribeiro et al. [6]. It generates explanations by determining and highlighting the parts of the input data responsible for a model's decision. Calculating the weights $w_i \in R$, it determines both the components that tend the model towards a positive prediction as well as the ones that account for a negative prediction.

A. Audio decomposition

A crucial part of generating explanations with CoughLIME is decomposing the input audio into components which can be interpreted by humans. We additionally emphasise that, in order to produce listenable explanations, it must be possible to reconstruct an audio file from these components. To create the interpretable components, we experiment with different approaches for spectral and temporal masking. Employing these rather simple decompositions is motivated by the fact that they have led to excellent results for audio data augmentation [16]. To compare the obtained results with a more complex approach, we further implement a decomposition based on Non-negative Matrix Factorization (NMF). For space reasons, we only further illustrate the loudness decomposition and the NMF decomposition. Results for CoughLIME with the temporal, spectral and loudness-spectral decomposition as well as the code for all experiments is publicly available[2]. The proposed novel loudness decomposition aims at extracting individual cough sounds. For this purpose, we compute the power $\mathbf{p}$ in dB of the cough sounds. We round $\mathbf{p}$ to the nearest 10 to avoid an excess of components and calculate local minima.

$$\mathbf{z} =_x \mathbf{p}_{rounded}(x) \tag{1}$$

[2]https://github.com/glam-imperial/CoughLIME

To avoid creating components that do not entail an entire cough, we introduce a threshold γ as a hyperparameter and only generate a new component if $\mathbf{p}_{z(i)} < \gamma$. Component i is then created by taking the subset of the audio array between indices z_{i-1} and z_i along the temporal axis. Figure 1 illustrates this approach. We point out that the number of generated components n varies depending on the audio file and on γ.

NMF decomposes a matrix $\mathbf{V} \in R_{\leq 0}^{nxm}$ into two matrices $\mathbf{W} \in R_{\leq 0}^{nxk}$ and $\mathbf{H} \in R_{\leq 0}^{kxm}$ such that $\mathbf{V} = \mathbf{WH}$. As audio signals are not typically non-negative, a cough signal is first transformed to its spectrogram using the discrete Fourier transform. Taking the magnitude of the spectrogram then results in a non-negative matrix which can be decomposed into k components using NMF. The hyperparameter k is set by the user. We ensure the decomposed signal can be transformed back to audio to generate listenable explanations by retaining the phase information when calculating the spectrogram's magnitude. The described decomposition is implemented using the Python packages *librosa* and *sklearn.decomposition*. Figure 2 shows an example for the NMF decomposition of a cough audio file into six components.

B. Explanation generation

Once the input cough audio is decomposed into interpretable components, explanations are generated according to the procedure from [6]. The decomposed input is given by a binary vector $\mathbf{x}' \in \{0,1\}^{d'}$ indicating the presence or absence of the individual components. Hence, to obtain the representation of the original audio data, all entries in x' would be set to 1. To generate the training data for the sparse linear surrogate model, the input audio is perturbed by randomly setting entries in $\mathbf{x}'$ to 0 resulting in n training samples $\mathbf{z}' \in \{0,1\}^{d'}$. The loss function of the linear surrogate model is given by

$$\mathcal{L}(f, g, \pi_x) = \sum_{\mathbf{z},\mathbf{z}' \in \mathcal{Z}} \pi_x(\mathbf{z})(f(\mathbf{z}) - g(\mathbf{z}'))^2 \tag{2}$$

with f the black-box model and g the surrogate model [6]. $\pi_x(\mathbf{z})$ is a proximity measure which takes the distance of

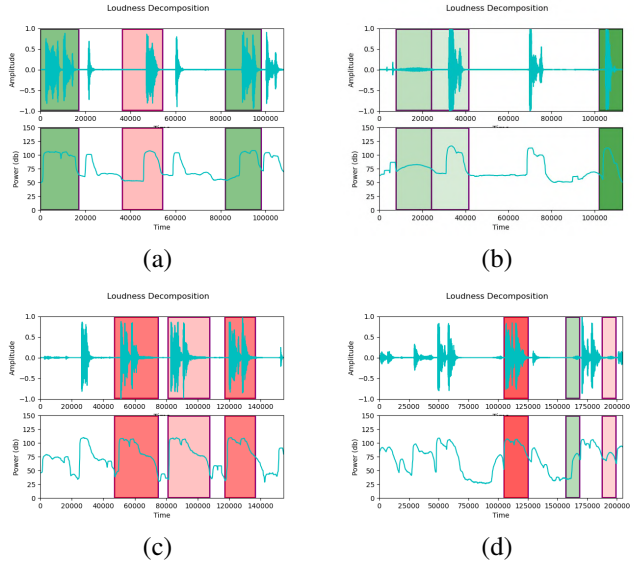


Fig. 3. Example explanations for COVID-19 positive (a), (b) and COVID-19 negative (c), (d) coughs generated with the loudness decomposition. Components highlighted in red (green) account for a COVID-19 negative (positive) prediction.

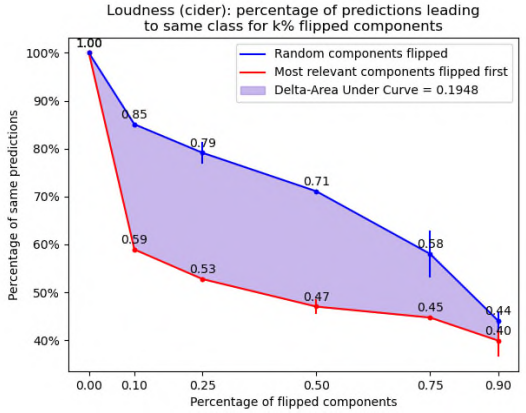


Fig. 4. Pixel flipping for audio for the loudness decomposition with $\gamma = 75dB$ shows CoughLIME is able to correctly identify the most important components. Results are averaged over 2 runs.

the perturbed training samples $\mathbf{z}$ to the original data instance $\mathbf{x}$ into account. As proposed in [6], we use an exponential kernel with cosine distance measure. The overall generated explanation is given by a dictionary associating the weights $w_i \in R$ of the different components with the component indices, sorted in descending order according to the absolute value of the weights. An explanation is always generated for a certain class label. Negative weights correspond to components that tend the model towards predicting that the component does not belong to the class whereas positive weights are associated with components that tend the model towards predicting that the sample is part of the class. Sonified explanations are obtained by factoring the k most important components together in the original representation. Listenable explanations were the main focus of this work, however, complementing them with visualised explanations can arguably be beneficial. It is therefore also possible to obtain a visualised explanation highlighting the most important components of the input representation and their respective weights $w_i \in R$. Hereby, focus was set on a modular implementation: CoughLIME which extends the LIME base class can be used with different decompositions to explain the predictions of any model taking audio as input data. We emphasise that the compatibility of the CoughLIME with any COVID-19 detection algorithm from audio data enables the comparison of different classifiers.

IV. Experiments, Results and Discussion

To assess CoughLIME's ability to faithfully explain COVID-19 prediction models, we generate explanations for the CIdeR model [7] trained on the publicly available data set from the DiCOVA challenge [17]. The data set consists of 1040 samples which can be split for a 5-fold cross-validation with instance-per-fold lists provided by the challenge. CIdeR is a deep ResNet model specifically developed for COVID-19 detection from audio data. It achieved an Area under the Curve (AUC) of 79.9% on the blind test set. We point out that CoughLIME is a model-agnostic approach – it is hence of minor importance which COVID-19 classification model we explain for the purpose of this demonstration. As explained above, we focus on sonified explanations as they are the most intuitive way to explain the behaviour of a model acting on audio data. Listenable examples can be found in our repository[3].

To quantify the faithfulness of the generated explanations, we adopt pixel flipping to audio. This evaluation method was originally proposed for image data and argues that if an XAI method correctly highlights the important parts of the input representation, 'flipping' these components to 0 should lead to a rapid decrease in performance [18]. To this end, we generate explanations using CoughLIME with the loudness decomposition for all 218 files contained in fold 1 of the DiCOVA data set. We then calculate the percentage of the predictions having flipped the most significant components to 0 that lead to the same class prediction as the entire file. We compare this percentage to the percentage of same class predictions having flipped the same number of randomly chosen components. As the loudness decomposition generates a variable number of components, we flip a percentage of all generated components n to establish a basis of comparison. The results and the standard deviation over two runs are reported in Figure 4. All results were generated with a threshold $\gamma = 75dB$. For all k and all decompositions tested, the predictions on the components found important by CoughLIME lead to superior results than the predictions on random components.

To establish a base of comparison for the pixel flipping results, we introduce a new metric: the Delta-Area Under

[3]https://github.com/glam-imperial/CoughLIME

TABLE I
COMPARISON OF THE ΔAUC ACHIEVED WITH DIFFERENT DECOMPOSITIONS.

Decomp-osition	Temporal	Spectral	Loudness	Loudness-Spectral	NMF
ΔAUC	13.26%	9.38%	19.48%	5.68%	4.13%

the Curve (ΔAUC)

$$\Delta AUC = \frac{\int_{c_1}^{c_n} y_2(x)\,\mathrm{d}x - \int_{c_1}^{c_n} y_1(x)\,\mathrm{d}x}{\int_{c_1}^{c_n} 1\,\mathrm{d}x} \tag{3}$$

with **c** being the percentages of components flipped, y_1 the curve generated by flipping the most relevant components first and y_2 the curve generated by flipping random components. The ΔAUC is given by the percentage of the entire graph area which is covered by the area between the curves generated for the most relevant and the random components. The higher the ΔAUC, the better the explanations. Table I compares the ΔAUC obtained for the proposed decompositions.

We argue that it can be beneficial to complement listenable explanations with visual or textual ones to obtain a better overall understanding of the system. Therefore, we implement methods to highlight the components found most important by the classification model together with their relative weights (Figure 3).

V. CONCLUSION

In this work, we proposed CoughLIME – a model-agnostic approach to generate sonified local explanations for COVID-19 detection models from audio data. CoughLIME decomposes the input into different interpretable components and fits a surrogate model to them to learn the importance of each component towards the model's prediction. We focus on obtaining sonified – hence true to audio – explanations and combine them with visual highlights of the most important components. Our results show that CoughLIME is capable of generating faithful explanations for the predictions of the CIdeR model by identifying the most important components of the input cough sample. This was also quantitatively verified by the newly suggested ΔAUC measure and pixel flipping for audio.

VI. FUTURE WORK

In the future, one should extend CoughLIME in various directions. Conducting an extensive qualitative usability evaluation both with medical and non-expert human users will provide valuable insights into the usefulness of the generated explanations in practice. Furthermore, one needs to experiment with different decompositions, possibly based on deep architectures. One further needs to investigate whether these lead to better quantitative and qualitative evaluation results than the decompositions proposed in this work. Additionally, extending CoughLIME to global explanations could provide more general insights into why models predict a cough as COVID-19 positive. Lastly, one has to compare the performance and results of CoughLIME with different COVID-19 detection models, among others to assess whether CoughLIME could be a remedy to the issue that some models do not generalise well to other data sets.

REFERENCES

[1] G. Serafini, B. Parmigiani, A. Amerio, A. Aguglia, L. Sher, and M. Amore, "The psychological impact of COVID-19 on the mental health in the general population," 2020.

[2] J. Peto, "Covid-19 mass testing facilities could end the epidemic rapidly," 2020.

[3] M. J. Binnicker, "Challenges and controversies to testing for COVID-19," 2020.

[4] K. D. Bartl-Pokorny, F. B. Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler, and B. W. Schuller, "The voice of COVID-19: Acoustic correlates of infection in sustained vowels," *The Journal of the Acoustical Society of America*, vol. 149, no. 6, 2021.

[5] K. Qian, B. W. Schuller, and Y. Yamamoto, "Recent advances in computer audition for diagnosing Covid-19: An overview," in *LifeTech 2021 - 2021 IEEE 3rd Global Conference on Life Sciences and Technologies*, 2021.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, vol. 13-17-August-2016.

[7] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. Schuller, "End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: A pilot study," *BMJ Innovations*, vol. 7, no. 2, 2021.

[8] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, 2020.

[9] J. Laguarta, F. Hueto, and B. Subirana, "COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, 2020.

[10] A. Hassan, I. Shahin, and M. B. Alsabek, "COVID-19 Detection System using Recurrent Neural Networks," in *Proceedings of the 2020 IEEE International Conference on Communications, Computing, Cybersecurity, and Informatics, CCCI 2020*, 2020.

[11] A. Akman, H. Coppock, A. Gaskell, P. Tzirakis, L. Jones, and B. W. Schuller, "Evaluating the COVID-19 Identification ResNet (CIdeR) on the INTERSPEECH COVID-19 from Audio Challenges," *arXiv: 2107.14549*, 2021.

[12] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," 2021.

[13] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K. R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proceedings of the IEEE*, vol. 109, no. 3, 2021.

[14] S. Mishra, B. L. Sturm, and S. Dixon, "Local interpretable model-agnostic explanations for music content analysis," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017*, 2017.

[15] V. Haunschmid, E. Manilow, and G. Widmer, "audiolime: Listenable explanations using source separation," *arXiv preprint arXiv:2008.00582*, 2020.

[16] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, vol. 2019-September.

[17] A. a. Muguli, L. Pinto, N. Sharma, P. Krishnan, P. K. Ghosh, R. Kumar, S. Bhat, S. R. Chetupalli, S. Ganapathy, S. Ramoji, and others, "DiCOVA Challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics," *arXiv preprint arXiv:2103.09148*, 2021.

[18] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, 2017.