An Explainable AI-Integrated Diagnostic System for Voice Analysis in Heart Failure Patients

Mikolaj Najda[1, 2], Milosz Dudek[3], Olgierd Unold[2], Tomasz Jadczyk[4, 5], Krzysztof Swierz[4], Grzegorz Swiatek[4], Daria Hemmerling[3*]

[1]Institute of Data Science, Maastricht University, Maastricht, Limburg, The Netherlands
[2]Department of Computer Engineering, Wroclaw University of Science and Technology, Wroclaw, Poland
[3]AGH University of Krakow, Krakow, Poland
[4]Department of Cardiology and Structural Heart Diseases, Medical University of Silesia, Katowice, Poland
[5]International Cardiac Electrophysiology Group, International Clinical Research Center, St. Anne's University Hospital in Brno, Brno, Czech Republic
∗hemmer@agh.edu.pl

Abstract

Integrating Explainable Artificial Intelligence to analyse voice characteristics is an essential topic for future research. We explore the utility of tree-based machine learning models, including Random Forest, XGBoost, and LightGBM, in distinguishing between two groups: 100 participants with heart failure and 100 healthy controls. The acoustic features extracted from sustained vowel recordings are used to differentiate between the two groups. The evaluation shows that the Random Forest model performs better, especially with the vowel /i/, achieving Accuracy, Precision, Recall, and F1 score over 0.80. We investigate the interpretability of these models using SHapley Additive exPlanations values, which reveal the essential acoustic features that influence model predictions and provide insights into their clinical relevance. This research highlights the potential of interpretable vocal biomarkers in remote monitoring and diagnosing heart failure.

Introduction

Vocal biomarkers for remote monitoring of essential health parameters present a potentially revolutionary solution for advancing telemedicine. This approach involves analysing patients' voice patterns to detect subtle changes that may indicate deterioration of health status, enabling early intervention and continuous follow-up without physical consultations. The application of vocal biomarkers in clinical practice extends beyond monitoring, covering diagnosis and treatment plan management, seamlessly integrating into the healthcare delivery model (Fagherazzi et al. 2021). Several research studies have been carried out to validate the effectiveness and reliability of this innovative approach. These studies mainly use deep learning algorithms to analyse the effect of speech patterns on the diagnosis and progression of neurodegenerative diseases. The focus has primarily been on conditions such as Parkinson's and Alzheimer's diseases, which are known to affect speech and can, therefore, be potentially monitored through changes in vocal characteristics (Wodzinski et al. 2019; Bertini et al. 2022). Through speech analysis, researchers aim to identify biomarkers that can predict the onset of diseases, monitor their progression, and potentially respond to treatment outcomes.

There is a growing interest in vocal biomarkers for cardiovascular disease research combining different approaches for conversational tasks and analysis. Heart failure (HF) often results in systemic fluid accumulation, which can lead to edema in the vocal folds, impacting voice quality and production. Such edema increases the phonation threshold pressure, affecting vocal fold vibration and causing changes in vocal attributes like pitch and stability (Murton et al. 2023a; Watson et al. 2024). Schöbi et al. (Schöbi et al. 2022) conducted speech pattern analysis in patients with heart failure (HF). The study involved 68 patients with acute HF and 36 patients with stable HF who read five sentences using a web browser application. The results revealed that the pause ratio was almost 15% higher in the group of patients with decompensated HF, showing the feasibility of voice analysis to detect early signs of disease exacerbation. Offer et al. (Amir et al. 2022) analysed recorded audio at hospital admission and after treatment. They collected audio of 5 repeated sentences in three languages from 40 subjects. The results obtained via five distinct speech measures showed that 94% of the discharge recordings significantly differed from their respective baseline admission recordings. This suggests the potential for non-invasive monitoring of HF patients. In line, Murton et al. (Murton et al. 2023b) successfully demonstrated a non-invasive approach to evaluate the effectiveness of pharmacological treatment for acute HF through the analysis of 52 patients' voices recorded during the course of hospitalization, leveraging the applicability of speech technology in medical diagnostics. The study achieved a 69% accuracy rate using a Logistic Classifier to classify patients' voices before and after admission to the hospital. This indicates potential for further development and clinical application. Furthermore, a method that distinguishes between HF patients and healthy controls using speech samples could be a valuable tool for the emerging field of telemedicine. Reddy et al. (Kiran Reddy et al. 2021) tested their approach on 25 healthy individuals and 20 HF patients, achieving 95% and 81% accuracy, respectively,

in speaker-dependent and speaker-independent approaches with a Feed-Forward Neural Network. The study's results suggest that this method could be a promising approach for diagnosing HF patients remotely. Priyasad et al. proposed the Self-Supervised Mode-Based Memory Fusion techniquefo, which achieved 90% accuracy results with a balanced dataset (37 healthy individuals and 37 subjects diagnosed with congestive heart failure, in discriminating patients from healthy individuals using preprocessed audio from a reading task in a subject-independent evaluation method (Priyasad et al. 2022).

The studies referenced in the text demonstrate encouraging results in classifying HF and treatment monitoring. However, to achieve more reliable outcomes, it is essential to have a deeper understanding of the model. Our work contributes to this area by employing SHAP (SHapley Additive exPlanations) to identify the acoustic features with the greatest impact on the models, providing the necessary explainability crucial in the medical domain. We also offer visual interpretations of feature changes over time and explain the influence of selected features on voice production and pathologies. Additionally, our unique contribution lies in exploring HF possibilities within the Polish language, which was not undertaken in this way. Furthermore, we determine which vowel has the most significant impact on the models. The extensive dataset enables a detailed analysis of acoustic features that can differentiate between patients with heart failure and healthy individuals.

Methodology

This section outlines the methodology used in our study, including the systematic approach for data collection, preprocessing, and analysis. We provide technical specifications and procedural steps to ensure the integrity and reproducibility of our research findings. Figure 1 summarises these steps.

Dataset

The dataset includes 100 voice recordings from HF patients (mean age 68 ± 12, 18 women, 82 men) hospitalised for HF exacerbation (data collected in non-acute clinical settings) and 100 samples from healthy participants (mean age 43 ± 7, 46 women, 54 men), all of Polish nationality. This distribution reflects a representative sample typical of a cardiology department, with an uneven gender ratio that realistically mirrors the selection bias shaped by epidemiological factors in heart failure prevalence. We acknowledge that other comorbidities could influence the final analysis; therefore, we made efforts to carefully select participants with relevant health profiles to ensure that these variables minimally impact the study's outcomes.

An anechoic chamber was constructed for this study (Figure 2), designed to attenuate sounds from the corridor and adjacent rooms. The reverberation studies were conducted within the chamber to minimise additional acoustic reflections, and ventilation was prepared to maintain a constant temperature of 21 degrees Celsius. Data were collected using a SHURE SM7B microphone with an active microphone preamplifier +28dB sE Electronics DM1 Dynamite with a sampling frequency of 48,000 Hz and 32-bit resolution. The data collection process involved recording stable, sustained vowel articulations /a/, /e/, /i/, /o/ and /u/ from both groups while the subject was in a seated position. This choice of analysis is grounded in the fact that vowels serve as robust acoustic markers, as they maintain high individual consistency and demonstrate considerable reproducibility across populations. The duration of the recorded audio varies from 1 second to 6 seconds for the longest one. Data are protected by privacy and GDPR regulations. Therefore, the dataset remains non-public.

Audio preprocessing

To ensure equal power distribution across all voice samples in our analysis, we employed loudness normalisation (Figure 1A) based on the ITU-R BS.1770 standard, using the pyloudnorm (Steinmetz and Reiss 2021). Studies have shown that this approach yields positive outcomes in speech prediction (Liebig et al. 2022) and recognition (Shivaprasad and Sadanandam 2021) tasks. The process measures the loudness level of each audio sample and adjusts it to a consistent target loudness level of -23 Loudness Unit Full Scale (LUFS). By standardising the loudness level of each sample we ensured that our models could accurately analyse the nuances of speech patterns and linguistic features without being biased by amplitude variations.

Features Extraction

Audio feature extraction is pivotal for understanding and analysing the characteristics of audio signals in various domains, such as speech recognition, music analysis, environmental sound classification, and medical voice analysis. This work selected audio features based on ComParE 2016 set (Schuller et al. 2016) from openSMILE (Eyben, Wöllmer, and Schuller 2010). The configuration files were not modified, and the low-level descriptors were chosen to improve the model's explanatory power based on a smaller number of features (64), even with the possibility of extracting more than 6.000. When extracting features from audio that varied in time, we averaged them over the entire length. In addition, the original naming convention was mapped to improve clarity. Processes included in this step are as follows: extraction using openSMILE, averaging across the entire outcome and normalization for improving the model's performance (Figure 1B).

Model Evaluation

The presented research investigates the use of tree-based models, such as Random Forest (Breiman 2001), XGBoost (Chen and Guestrin 2016), and LightGBM (Ke et al. 2017), for binary classification. These models suit Explainable Artificial Intelligence methods due to their decision-making characteristics (Iban and Bilgilioglu 2023; Guldogan et al. 2023). The binary classification approach was intentionally chosen to address the complexity associated with comorbid conditions. By carefully selecting patients with heart failure and minimizing the influence of other coexisting diseases,

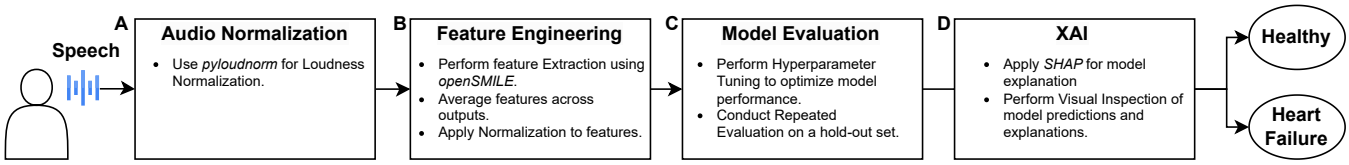

Figure 1: Methodology pipeline summary.

Vowel	Metric	Model			Average ± SD
		Random Forest	XGBoost	LightGBM	
/a/	Accuracy ± SD	0.77 ± 0.05	0.76 ± 0.06	0.76 ± 0.06	0.76 ± 0.00
	F1 score ± SD	0.77 ± 0.05	0.76 ± 0.06	0.75 ± 0.07	0.76 ± 0.00
	Precision ± SD	0.78 ± 0.05	0.77 ± 0.06	0.76 ± 0.07	0.77 ± 0.00
	Recall ± SD	0.77 ± 0.06	0.76 ± 0.06	0.76 ± 0.07	0.76 ± 0.00
/e/	Accuracy ± SD	0.81 ± 0.06	0.80 ± 0.06	0.76 ± 0.05	0.79 ± 0.02
	F1 score ± SD	0.81 ± 0.06	0.80 ± 0.07	0.76 ± 0.05	0.79 ± 0.02
	Precision ± SD	0.81 ± 0.06	0.80 ± 0.07	0.77 ± 0.05	0.79 ± 0.01
	Recall ± SD	0.81 ± 0.06	0.80 ± 0.06	0.76 ± 0.05	0.79 ± 0.02
/i/	Accuracy ± SD	**0.82 ± 0.04**	0.80 ± 0.04	0.78 ± 0.04	0.80 ± 0.01
	F1 score ± SD	**0.82 ± 0.04**	0.80 ± 0.04	0.78 ± 0.05	0.80 ± 0.01
	Precision ± SD	**0.82 ± 0.04**	0.80 ± 0.04	0.79 ± 0.04	0.80 ± 0.01
	Recall ± SD	**0.82 ± 0.04**	0.80 ± 0.04	0.78 ± 0.05	0.80 ± 0.01
/o/	Accuracy ± SD	0.78 ± 0.06	0.76 ± 0.07	0.73 ± 0.05	0.76 ± 0.02
	F1 score ± SD	0.78 ± 0.06	0.76 ± 0.07	0.73 ± 0.05	0.76 ± 0.02
	Precision ± SD	0.79 ± 0.06	0.77 ± 0.07	0.73 ± 0.05	0.76 ± 0.02
	Recall ± SD	0.77 ± 0.06	0.76 ± 0.07	0.73 ± 0.05	0.75 ± 0.01
/u/	Accuracy ± SD	0.77 ± 0.07	0.77 ± 0.06	0.78 ± 0.06	0.77 ± 0.00
	F1 score ± SD	0.77 ± 0.07	0.77 ± 0.06	0.78 ± 0.06	0.77 ± 0.00
	Precision ± SD	0.78 ± 0.07	0.78 ± 0.06	0.78 ± 0.06	0.78 ± 0.00
	Recall ± SD	0.77 ± 0.07	0.77 ± 0.06	0.78 ± 0.06	0.77 ± 0.00

Table 1: Averaged results within 20 repetitions of model evaluation with highlighted text indicating best performance.

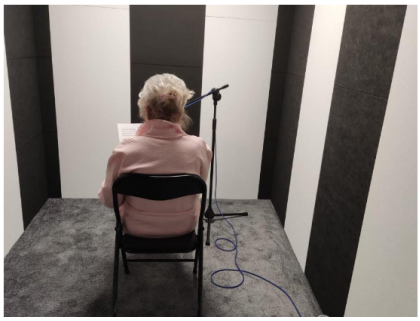

Figure 2: Patient during the recording session in the hospital.

we aimed to isolate and examine whether discernible vocal differences specifically associated with heart failure could be identified. We recognize that a multi-class approach could offer broader applications, especially in telematics. However, to achieve preliminary insights into the impact of heart failure on vocal biomarkers, this binary design allowed us to focus on a clear, controlled comparison.

Before model training, the dataset was split into training and test sets in an 80/20 ratio, maintaining a balance of 50% for each class in both sets (Figure 1C). To increase the credibility of our research, we conducted 20 repeated training and testing evaluations. We used the scikit-learn library for the entire model evaluation procedure (Pedregosa et al. 2011). We used Accuracy, Precision, Recall and F1 score metrics to assess the models' performance. Accuracy quantifies the overall effectiveness of the model by calculating the proportion of true results, both true positives and true negatives, in relation to all evaluated cases. Precision measures the proportion of true positive results in all positive predictions made by the model, indicating its ability to return relevant results. On the other hand, Recall quantifies how well the model identifies all actual positives, reflecting its capacity to capture all relevant instances within the dataset. The F1 score synthesised the model's Precision and Recall by computing the harmonic mean of these two metrics, thus integrating the proportion of accurately predicted positive observations with the rate at which the model correctly identified all relevant instances. The hyperparameter tuning procedure has been evaluated with a randomised search. The estimator parameters used to apply these methods are optimised by cross-validated search over parameter settings within validation data with a number of iterations set to 500, 5-fold cross-validation and scoring method as Accuracy due to balanced dataset across both classes.

XAI

In our research, we used SHAP, which is a model-agnostic approach to explaining the performance of machine learning models, introduced by Lundberg and Lee (Lundberg and Lee 2017). It is based on game theory, with SHAP scores quantifying the contribution of each feature to a given instance's prediction relative to the model's average prediction. In addition, we display the acoustic feature differences over time that had the greatest impact on the best-performing model, as determined by the SHAP method (Figure 1D). To account for variations in audio length and parameter values the features were not averaged as in a model evaluation process, but instead, they were interpolated to the same length using linear interpolation. Their values were then normalized to fit within the range of -1 to 1. We have included an example of the waveform for the vowel articulation to enhance the clarity of the proposed method.

Evaluation

Statistical Analysis of Features

The statistical analyses of the extracted acoustic features from the vowels /a/, /e/, /i/, /o/, and /u/ revealed non-normality in the data distribution, which was tested using the Shapiro-Wilk test with a α of 0.05. Therefore, a Mann-Whitney U test was conducted (α=0.05). Significant differences were observed in the following features: Log Harmonic-to-Noise Ratio (HNR), which is a vocal feature that measures the ratio of harmonics (voice) to noise (breathiness, raspiness) in the speech signal. A lower HNR value compared to a healthy control may indicate voice disorders or pathologies (Lee, Kim, and Kang 2014). Additionally, significant disparities were observed in various auditory spectral bands (from Band 0 to Band 14), suggesting distinct variations in vocal and auditory characteristics between the groups. Furthermore, there were notable variations in the acoustic properties, such as the Zero Crossing Rate (ZCR), which indicates a higher frequency of amplitude changes, and FFT Magnitude in the 1000-4000 Hz Band, suggesting differences in the sound's texture and harmonic structure. Significant differences were observed in the spectral features, including the spectral roll-off and spectral centroid, indicating variations in the energy distribution across the spectrum. In addition, psychoacoustic measures such as Psychoacoustic Sharpness and Mel-Frequency Cepstral Coefficients (MFCCs) showed significant differences, indicating unique perceptual characteristics and timbral textures between the groups. On the other hand, features like Fundamental Frequency, referencing the lowest frequency of a periodic waveform and Voicing Final Unclipped, which relates to the degree to which vocal folds are engaged at the end of speech sounds, frequently do not show significant differences between classes ($p \gg \alpha$), suggesting that they may not be dependable for distinguishing between vowel sounds.

Results

In conclusion, all three models showed their highest performance metrics with the vowel /i/, achieving Average metric scores (counted as the average of the three methods) of 0.80, as shown in Table 1. However, the Random Forest model outperformed XGBoost and LightGBM, achieving an Accuracy, Precision, Recall, and F1 score of 0.82 $\pm$ 0.04. The statistical significance tests performed between the results of the models for each vowel show that Random Forest performs significantly better than other approaches. The p-values for all t-tests are below the α set at 0.05, indicating that the means between the models for each metric and vowel are statistically significant.

Best Parameters

The selected Random Forest model achieved best results, with key parameters: bootstrap=True, criterion=entropy, max_depth=None, min_samples_leaf=8, min_samples_split=7, and n_estimators=100. Parameter variations from evaluations included: max_depth (5 to 25 or None, average 15), min_samples_leaf (1 to 10, average 4).

Interpreting Features Impact on a Model

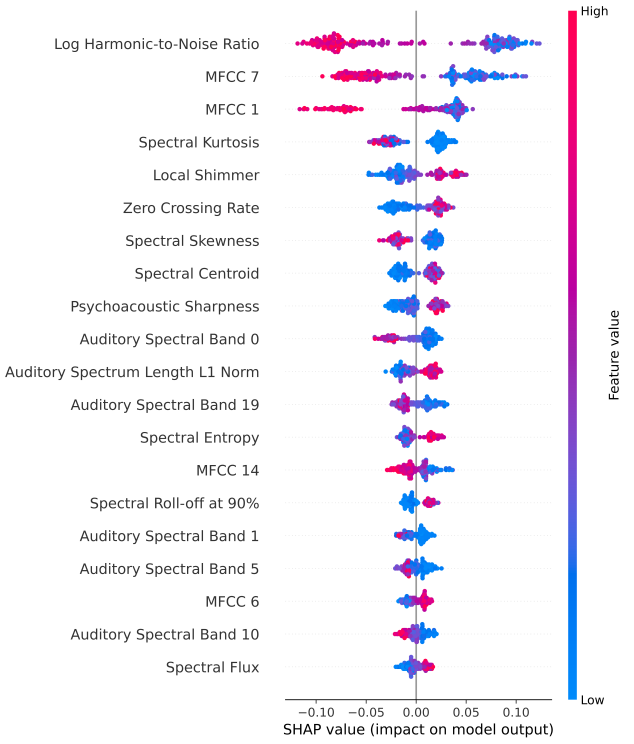

Figure 3: SHAP values illustrating feature impact on Random Forest model predictions.

This SHAP summary plot (Figure 3) illustrates the impact of various features on the model's output for a positive class prediction (HF). Each point represents a SHAP value for a specific feature and instance, indicating the degree to which that feature contributed to pushing the model's prediction towards heart failure classification. High feature values are represented in red, while low values are shown in blue. For example, features like low Harmonic-to-Noise Ratio (HNR), specific Mel-Frequency Cepstral Coefficients

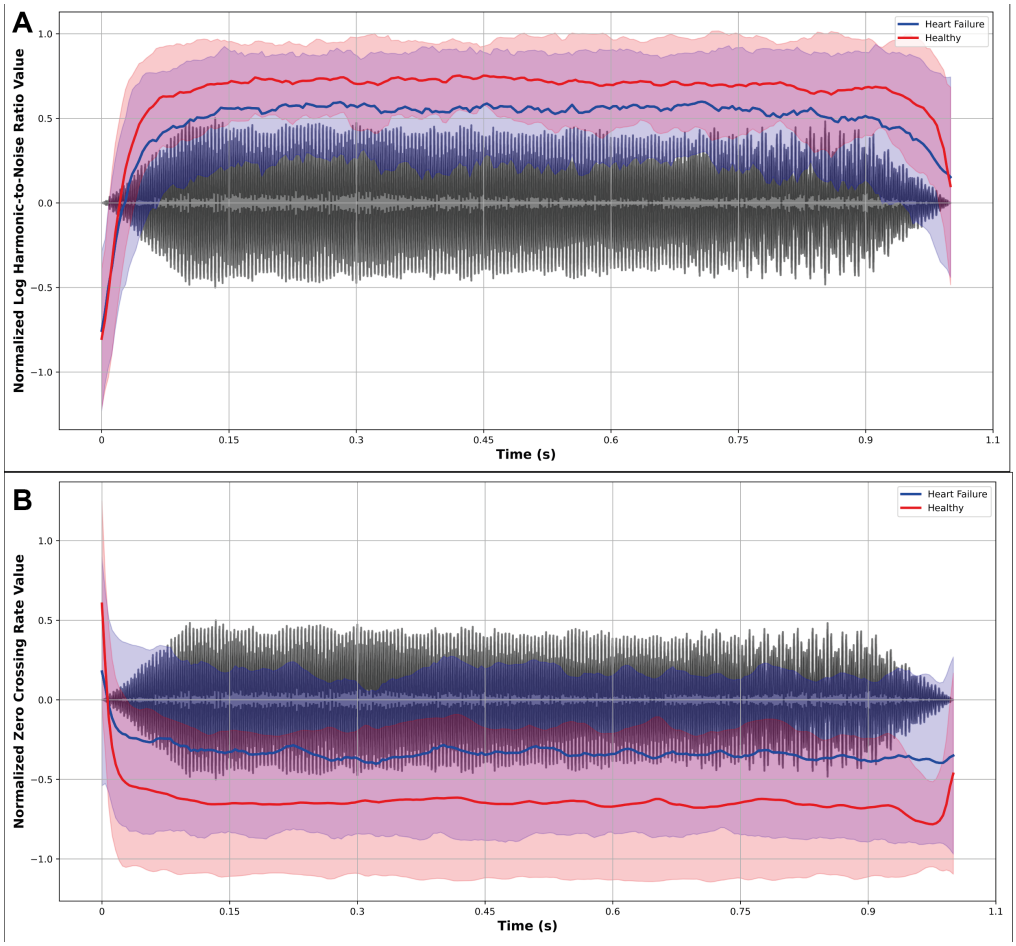

Figure 4: Temporal differences in normalized feature values between classes over time for vowel /a/. A - Log Harmonic-to-Noise ratio, B - Zero Crossing Rate.

(MFCC 7, MFCC 1), and Spectral Kurtosis, as well as high Local Shimmer and Zero Crossing Rate (ZCR), increase the likelihood of a heart failure prediction. In contrast, features such as Auditory Spectral Band 10 and Spectral Flux appear to have a minimal effect, as their SHAP values cluster near zero.

Local Shimmer, which often appears elevated in dysphonic patients, signals irregular vocal fold vibration, a characteristic linked to various voice pathologies (Teixeira and Gonçalves 2016). Notably, high ZCR values during vowel articulation are associated with model predictions for heart failure, suggesting that affected portions of the signal resemble unvoiced sounds (Jalil, Butt, and Malik 2013). Spectral Kurtosis, used to detect irregular or random signals, helps characterize spectral sparsity often found in dysarthric speech (Vrabie, Granjon, and Serviere 2003; Kodrasi and Bourlard 2020). MFCC features, extracted from the Mel frequency domain, capture the non-linear characteristics of human auditory perception and are widely employed in models for diagnosing voice disorders, including in our model (Chen and Chen 2022).

It is plausible that these acoustic changes reflect systemic alterations in heart failure patients. For instance, fluid retention common in heart failure can cause laryngeal tissue edema, which may disrupt vocal fold movement, leading to irregular vibration patterns. This effect could contribute to increased shimmer and ZCR values in the model, as these features often capture irregular vocal fold behavior. Additionally, reduced circulation and oxygenation levels associated with heart failure could impair neuromuscular control, potentially weakening the stability and closure strength of the vocal folds. Such impairments may correlate with lower HNR values, as vocal fold instability introduces more aperiodic noise into the voice signal.

To enhance the interpretability of SHAP results and visibly highlight the differences, we conducted a visual inspection (Figure 4) to illustrate how heart failure affects vocal characteristics over time, specifically distinguishing between patients and healthy individuals based on Harmonic-to-Noise Ratio (HNR) and Zero Crossing Rate (ZCR). Heart failure patients tend to exhibit a lower HNR, with this difference becoming more pronounced at the start of vocalization and remaining consistently distinct, suggesting potential vocal impairments linked to heart failure. The graph

(Figure 4B) also shows that heart failure patients have a higher ZCR than healthy individuals, indicating more frequent signal fluctuations.

These signal variations may reflect systemic physiological factors associated with heart failure, such as increased inflammation and vagus nerve involvement. Chronic heart failure is often accompanied by systemic inflammation, which can affect the structure and resilience of vocal fold tissues, potentially leading to swelling or weakening of the vocal folds. Combined with potential disruptions in autonomic nerve function, particularly vagus nerve involvement, this inflammation may impair the precision of vocal fold control, resulting in the elevated ZCR and signal instability observed in the Figure 4. Thus, these vocal features could serve as indirect markers of the broader systemic effects of heart failure, extending beyond localized changes in the vocal folds. The observed irregularities in vocal fold vibration patterns may provide a potential diagnostic marker for distinguishing between heart failure patients and healthy subjects based on speech signals.

Conclusion

The evaluation demonstrates the effectiveness of the Random Forest model, particularly for the vowel /i/. It outperformed gradient-boosted XGBoost and LightGBM models across Accuracy, Precision, Recall, and F1 score metrics. SHAP values provided deep insights into the model's decision-making process, underscoring the importance of specific acoustic features in distinguishing between heart failure patients and healthy individuals—a significant step in advancing explainability in the medical domain. Critical features such as Log Harmonic-to-Noise Ratio, Local Shimmer, and Zero Crossing Rate influenced the model's predictions towards heart failure, revealing health status changes likely reflected in vocal characteristics as a result of physiological impacts of heart failure, such as fluid retention and tissue changes in the vocal folds. Visual inspection highlights changes in acoustic features over time, further revealing differences between the groups. This approach provides an additional way to understand how heart failure impacts specific quantified features in speech.

Moreover, successfully applying this methodology to the Polish language broadens its potential applications, which have traditionally focused on English, thereby expanding the scope of acoustic analysis for clinical use. This study validates the potential of machine learning models in enhancing auditory health diagnostics and paves the way for future explorations into integrating explainable AI frameworks. By shedding light on the interpretability of model predictions, we contribute to the broader field of biomedical signal processing and analysis.

Acknowledgements

We would like to express our gratitude to all participants in our study. Your contributions have been invaluable in advancing medical knowledge and understanding of conditions such as heart failure.

The study adhered to the principles of the Declaration of Helsinki and was approved by the Ethics Committee of the Medical University of Silesia in Katowice (PCN/CBN/0022/KB1/41/II/21/22). All subjects gave written informed consent to participate in the study.

The study was funded by Miniatura 6 Grant, no 2022/06/X/ST6/01191, funded by the National Science Centre in Poland, and by Statutory funds of the Medical University of Silesia in Poland (no. (PCN-1-005/N/0/K and PCN-1-139/N/2/K).

Moreover this research project was supported by "The Excellence Initiative - Research University" program for the AGH University of Krakow.

References

Amir, O.; Abraham, W. T.; Azzam, Z. S.; Berger, G.; Anker, S. D.; Pinney, S. P.; Burkhoff, D.; Shallom, I. D.; Lotan, C.; and Edelman, E. R. 2022. Remote Speech Analysis in the Evaluation of Hospitalized Patients With Acute Decompensated Heart Failure. *JACC: Heart Failure*, 10(1): 41–49.

Bertini, F.; Allevi, D.; Lutero, G.; Calzà, L.; and Montesi, D. 2022. An automatic Alzheimer's disease classifier based on spontaneous spoken English. *Computer Speech & Language*, 72: 101298.

Breiman, L. 2001. Random Forests. *Mach. Learn.*, 45(1): 5–32.

Chen, L.; and Chen, J. 2022. Deep neural network for automatic classification of pathological voice signals. *J. Voice*, 36(2): 288.e15–288.e24.

Chen, T.; and Guestrin, C. 2016. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, 1459–1462. New York, NY, USA: Association for Computing Machinery. ISBN 9781605589336.

Fagherazzi, G.; Fischer, A.; Ismael, M.; and Despotovic, V. 2021. Voice for health: The use of vocal biomarkers from research to clinical practice. *Digit. Biomark.*, 5(1): 78–88.

Guldogan, E.; Yagin, F. H.; Pinar, A.; Colak, C.; Kadry, S.; and Kim, J. 2023. A proposed tree-based explainable artificial intelligence approach for the prediction of angina pectoris. *Scientific Reports*, 13(1): 22189.

Iban, M. C.; and Bilgilioglu, S. S. 2023. Snow avalanche susceptibility mapping using novel tree-based machine learning algorithms (XGBoost, NGBoost, and LightGBM) with eXplainable Artificial Intelligence (XAI) approach. *Stochastic Environmental Research and Risk Assessment*, 37(6): 2243–2270.

Jalil, M.; Butt, F. A.; and Malik, A. 2013. Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals. In *2013 The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE)*, 208–212.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kiran Reddy, M.; Helkkula, P.; Madhu Keerthana, Y.; Kaitue, K.; Minkkinen, M.; Tolppanen, H.; Nieminen, T.; and Alku, P. 2021. The automatic detection of heart failure using speech signals. *Computer Speech & Language*, 69: 101205.

Kodrasi, I.; and Bourlard, H. 2020. Spectro-Temporal Sparsity Characterization for Dysarthric Speech Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 1210–1222.

Lee, J.-W.; Kim, S.; and Kang, H.-G. 2014. Detecting pathological speech using contour modeling of harmonic-to-noise ratio. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Liebig, L.; Wagner, C.; Mainka, A.; and Birkholz, P. 2022. An investigation of regression-based prediction of the femininity or masculinity in speech of transgender people. In *Proc. Interspeech 2022*, 4676–4680.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 4765–4774. Curran Associates, Inc.

Murton, O. M.; Dec, G. W.; Hillman, R. E.; Majmudar, M. D.; Steiner, J.; Guttag, J. V.; and Mehta, D. D. 2023a. Acoustic Voice and Speech Biomarkers of Treatment Status during Hospitalization for Acute Decompensated Heart Failure. *Applied Sciences*, 13(3): –.

Murton, O. M.; Dec, G. W.; Hillman, R. E.; Majmudar, M. D.; Steiner, J.; Guttag, J. V.; and Mehta, D. D. 2023b. Acoustic Voice and Speech Biomarkers of Treatment Status during Hospitalization for Acute Decompensated Heart Failure. *Applied Sciences*, 13(3).

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Priyasad, D.; Partovi, A.; Sridharan, S.; Kashefpoor, M.; Fernando, T.; Denman, S.; Fookes, C.; Tang, J.; and Kaye, D. 2022. Detecting Heart Failure Through Voice Analysis using Self-Supervised Mode-Based Memory Fusion. In *Proc. Interspeech 2022*, 2848–2852.

Schuller, B.; Steidl, S.; Batliner, A.; Hirschberg, J.; Burgoon, J. K.; Baird, A.; Elkins, A.; Zhang, Y.; Coutinho, E.; and Evanini, K. 2016. The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language. In *Proc. Interspeech 2016*, 2001–2005.

Schöbi, D.; Zhang, Y.; Kehl, J.; Aissani, M.; Pfister, O.; Strahm, M.; van Haelst, P.; and Zhou, Q. 2022. Evaluation of Speech and Pause Alterations in Patients With Acute and Chronic Heart Failure. *Journal of the American Heart Association*, 11.

Shivaprasad, S.; and Sadanandam, M. 2021. Dialect recognition from Telugu speech utterances using spectral and prosodic features. *Int. J. Speech Technol.*

Steinmetz, C. J.; and Reiss, J. D. 2021. pyloudnorm: A simple yet flexible loudness meter in Python. In *150th AES Convention*.

Teixeira, J. P.; and Gonçalves, A. 2016. Algorithm for jitter and shimmer measurement in pathologic voices. *Procedia Comput. Sci.*, 100: 271–279.

Vrabie, V.; Granjon, P.; and Serviere, C. 2003. Spectral kurtosis: from definition to application. *6th IEEE International Workshop on Nonlinear Signal and Image Processing*.

Watson, K.; Oates, J.; Sinclair, C.; Smith, J. A.; and Phyland, D. 2024. Associations Between Immunological Biomarkers, Voice Use Patterns, and Phonotraumatic Vocal Fold Lesions: A Scoping Review. *Journal of Voice*.

Wodzinski, M.; Skalski, A.; Hemmerling, D.; Orozco-Arroyave, J. R.; and Nöth, E. 2019. Deep learning approach to Parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, 717–720. IEEE.