*Proceedings of the 11th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'25)*
*Paris, France – August, 2025*
*Paper No. ICBES 206*
*DOI: 10.11159/icbes25.206*

# Explainable Heart Failure Voice Prediction using Machine Learning Ensembles

**Muniba Ashfaq[1], Petar Vračar[1], Borut Flis[1], Matej Pičulin[1], Amy Fuller[2], Nduka Okwose[2], Nenad Filipović[3], Djordje Jakovljević[2], Zoran Bosnić[1]**
[1]University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana, Slovenia
muniba.ashfaq@fri.uni-lj.si; petar.vracar@fri.uni-lj.si; borut.flis@fri.uni-lj.si; matej.piculin@fri.uni-lj.si; zoran.bosnic@fri.uni-lj.si
[2]Clinical Sciences and Translational Medicine, RC for Health and Life Sciences, Coventry University, UK
ae2713@coventry.ac.uk; ad6707@coventry.ac.uk; ad5287@coventry.ac.uk
[3]Bioengineering Research and Development Center, BioIRC, Kragujevac, Serbia
fica@kg.ac.rs

***Abstract*** **-** Heart failure (HF) detection is one of the challenging health concerns as early diagnosis may reduce mortality rate and improve the quality of life. In this aspect, we propose a novel biomedical voice signal processing framework including multiple voice tasks. The extracted features from multiple voice tasks are used to classify the HF detection using an ensemble of explainable AI (XAI) integrated machine learning classifiers. A key innovation of our methodology is the implementation of a two-stage majority voting strategy to consolidate the predictions of the diverse classifiers across heterogeneous voice tasks. In the first stage, each voice task is independently processed, and predictions from all classifiers are aggregated using standard majority voting; in the second stage, these task-level decisions are integrated via an additional majority voting layer to produce the final HF prediction. This hierarchical voting mechanism is motivated by the need to mitigate bias from any single classifier or voice task, thus enhancing predictive robustness and ensuring that the final decision reflects a consensus derived from multi-task auditory inputs.

***Keywords*****:** Voice signal processing, heart failure prediction using machine learning, heart failure recognition using voice signals, machine learning ensembles, classification, majority voting, SHAP, voice tasks.

## 1. Introduction

In recent years, advances in biomedical signal processing, along with machine learning (ML) techniques, have been used to extract acoustic features from voice recordings for heart disease detection [1]. The extracted features from the voice signals are highly correlated with heart failure detection and prognosis [2, 3]. Heart failure is a complex clinical syndrome that affects millions worldwide, leading to considerable morbidity and mortality. The conventional diagnostic methods are often invasive, expensive, and challenging for remote monitoring. The diagnostics may include chest imaging, biomarker assays such as B-type natriuretic peptide (BNP) [4] and N-terminal pro b-type natriuretic peptide (NT-proBNP) [5], and physical examinations. Hence, there is a substantial clinical need for non-invasive, cost-effective, and easily accessible methods for heart failure detection. Recent research has focused on employing voice and speech analysis as potential digital biomarkers in HF owing to their non-invasive nature, ease of collection via smartphones, and the possibility of continuous monitoring [6]. Heart failure leads to physiological changes that affect not only the cardiovascular system but also the respiratory and laryngeal systems. Fluid overload, pulmonary congestion, and altered autonomic regulation in HF patients can cause changes in vocal fold function, phonation stability, and speech patterns. These voice related changes are quantifiable using extraction of acoustic and prosodic features from the recordings. The changes in fundamental frequency (F0), cepstral peak prominence (CPP), jitter, shimmer, and maximum phonation time (MPT) have been used to recognize HF status [7].

To convert the high-dimensional information in the form of extracted features of voice signals into actionable clinical insights, a robust machine learning framework is essential [8]. Ensemble learning, which combines the predictions of multiple classifiers on the set of the voice signal task to yield a more accurate and stable outcome than any individual model. It is useful in diverse medical diagnostic applications to reduce the potential model-based biases and overfitting problems [9]. An

ensemble approach for heart failure prediction that utilizes the strengths of various classifiers such as Support Vector Machine (SVM), Decision Tree, Random Forest, Logistic Regression, and Gradient Boosting has been proposed [10]. Moreover, the integration of explainable AI (XAI) methods like SHAP (SHapley Additive exPlanations) identifies the importance scores to features, which are based on their contribution in predicting model output [11–13]. SHAP-based feature selection and training models with reduced features, thus effectively used to improve interpretability and dimensionality reduction.

In this work, we propose a hybrid two-stage ensemble framework for heart failure prediction using voice signal data. The key contributions of our work are:

- We design a SHAP-guided feature selection mechanism integrated into the hybrid ML ensembles for model transparency and identify the most impactful acoustic features.
- We used novel voice tasks combination set to analyse and predict heart failure on limited dataset
- We implement a nested leave-one-out cross-validation (LOOCV) strategy to evaluate the generalization.
- We introduce a two-stage majority voting mechanism, where individual classifiers' predictions for each task are used for task-based heart failure prediction via majority voting (Stage-1), and the final prediction of heart failure is concluded by voting across all the task-level predictions via majority voting (Stage-2).

This hierarchy majority voting scheme along with SHAP based machine learning classification is more consistent and reduces false predictions from weak classifiers for each task separately.

## 2. Methodology

In this research, real clinical data is collected from the heart failure patients in the form of biomedical voice signals. The voice signals of heart failure patients are associated with 5 different tasks. The patients' voices are collected from multiple clinical centers across Europe. The dataset contains 4 confirmed heart failure patients and 4 suspected. Each patient with 5 tasks makes a dataset with a total of 20 voice records for confirmed and 20 for suspected heart failure. Each feature set contains 94 features, hence the cumulative feature set for both individual classes is of dimension 20x94. The dimension of the feature set, including both classes, is 40x94. The voice recorder, along with the application for feature extraction, is the same across the medical centers. The research is based on the novel voice signals tasks combination for the analysis of heart failure patients' voices and the prediction of disease. The details of the tasks are as follows.

Task 1 is reading a specific paragraph, task 2 is at least 30 second free speech, task 3 is counting number 1 to 30 as fast and accurate as possible, task 4 is counting numbers from 30 to 1 as fast and accurate as possible, and the task 5 is the phonation "aaa" as long as possible three times with breath in and out between the sounds. All 5 voice tasks are recorded in the local disk drive for further biomedical voice signal processing. Each medical center is given an exe file (application) that takes all 5 voice tasks one by one and extracts the features from each task separately. The extracted features are again stored in the local disk drive in the form of CSV files. The extracted features are further passed through different machine learning classifiers for prediction.

The features considered for the heart failure prediction are 94 in total. The diverse range of features considered for extraction includes total time of phonation, first and last phonation, number of pauses greater than 100ms and 500ms, longest pause, longest continuous phonation, standard deviation of the pause and phonation length, total number of phonation segments, mean and standard deviation of the pitch, loudness, F0, [1-13] MFCCs (Mel-frequency cepstral coefficients), jitter, shimmer, CPP (Cepstral Peak Prominence). Moreover, mean and standard versions of [1-13] (5-95) MFCCs, 5-95 pitch, 5-95 jitter, and shimmer.

Let $m$ be the number of voice signal tasks $T_i = [T_1, T_2, \dots, T_m]$, with associated feature vector for each task as $F(T_i) \in \mathbb{R}^d$, where $i = 1, 2, \dots, m$. Let $N$ be the number of candidate machine learning classifiers $C_j$, where $j = 1, 2, \dots, N$, e.g., Random Forest, Gradient Boosting, SVM, Logistic Regression, and Decision Trees.

To deeply understand the importance of the features using explainable AI (XAI), the SHAP (SHapley Additive exPlanations) explainer is used. The SHAP values are used to explain and quantify the relevant important features in all tasks using each classifier for heart failure prediction. The threshold value is used to retain the most important features above it. The increasing

threshold value reduces the selected feature vector length with increasing relevant importance. The appropriate threshold value is important in terms of reducing feature vector length and increased performance.

Let $S_{ij} \in \mathbb{R}^d$ represent SHAP values for task $T_i$ and classifier $C_j$. We define a binary mask $M_{ij} \in \mathbb{R}\{0,1\}^d$, where $M_{ij}(k) = \begin{cases} 1, & if\ S_{ij}(k) \geq \theta \\ 0, & otherwise \end{cases}$ where $\theta$ is the threshold for SHAP importance. The reduced feature vector hence becomes $\tilde{F}_{ij}(T_i) = M_{ij} \odot F(T_i)$. The performance of the classifier for each task is computed with reduced feature set to form an integrated XAI based Hybrid ML ensemble for heart failure prediction using biomedical voice signals. Furthermore, the mathematical representation at each step of the process is explained as follows:

### 2.1. Feature Representation

Each classifier has an input of stacked features vectors as shown in Eq. (1)

$$\mathbf{F} = \begin{bmatrix} F(T_1) \\ F(T_2) \\ \vdots \\ F(T_m) \end{bmatrix} \in \mathbb{R}^{m \times d} \tag{1}$$

where **F** represents complete feature matrix containing feature vector $F(T_i) \in \mathbb{R}^d$ associated with task $T_i$, for $i = 1, 2, \dots, m$.

### 2.2. Classifier Prediction

Each classifier produces the predictions for all tasks as shown in Eq. (2)

$$P_j(T_i) = C_j(F(T_i)), \qquad \forall j = 1, 2, \dots, N, \forall i = 1, 2, \dots, m \tag{2}$$

where $C_j$ is the $j$-th classifier in the ensemble, for $j = 1, 2, \dots, N$. $P_j(T_i)$ is the predicted label for $T_i$ using classifier $C_j$. The predictions for each task from all classifiers is represented as a vector in Eq. (3)

$$\mathbf{P}(T_i) = [P_1(T_i), P_2(T_i), \dots, P_N(T_i)]^T \tag{3}$$

where $\mathbf{P}(T_i)$ is the vector of predictions from all the classifiers for task $T_i$, as defined in Eq. (2)

### 2.3. Stage 1: Task-Wise Majority Voting

The predictions from all the classifiers for each task undergo majority voting as shown in Eq. (4)

$$P'(T_i) = mode(\{P_1(T_i), P_2(T_i), \dots, P_N(T_i)\}) \tag{4}$$

where $P'(T_i)$ is the final label for task $T_i$ after majority voting across all classifiers. The cumulative task-wise prediction vector as a result of majority voting applied on each task is shown in Eq. (5)

$$\mathbf{P}' = \begin{bmatrix} P(T_1) \\ P(T_2) \\ \vdots \\ P(T_m) \end{bmatrix} \in \mathbb{R}^{m \times 1} \tag{5}$$

where $\mathbf{P}'$ is the vector of ensemble predictions of all the individual tasks after stage-1 majority voting.

### 2.4. Stage 2: Final Majority Voting Across Tasks

The predictions of all the tasks acquired after stage 1 majority voting undergoes another round. This is the final stage of prediction as stage 2 majority voting across all tasks for final prediction as shown in Eq. (6)

$$P_{final} = mode(\mathbf{P}') \tag{6}$$

where $P_{final}$ is the final prediction after stage-2 majority voting across all the tasks.

Fig. 1 shows the workflow of the non-invasive heart failure prediction using ensemble of machine learning classifiers. The patient undergoes $m$ different voice recordings denoted as $T_1, T_2, \dots, T_m$ for task 1, task 2, … task $m$ respectively. The

feature set represented as $F(T_1), F(T_2), \ldots, F(T_m)$ are extracted from $T_1, T_2, \ldots, T_m$ respectively. The feature set of all tasks are provided as input to each machine learning classifiers. In our experiments we used 5 different machine learning classifiers including Random Forest, Gradient Boosting, SVM, Decision Trees, and Logistic Regression. The predictions $P_j(T_{i=1,2,\ldots,m})$ for all the tasks are performed using each classifier. Stage 1 majority voting is applied on the predictions the classifiers for each task. The final prediction is computed based on the stage 2 majority voting applied on the of all tasks.

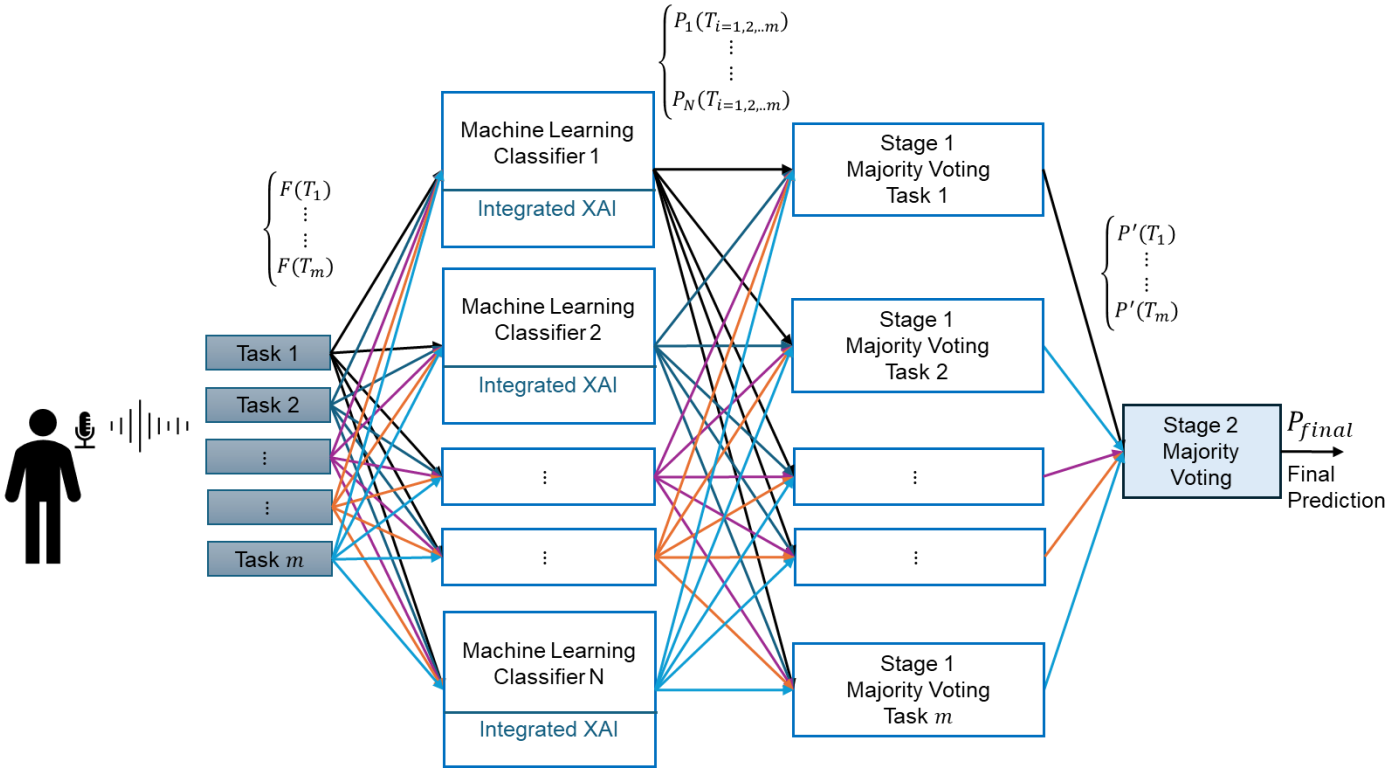


Fig. 1: Voice signal processing for multiple tasks and heart failure prediction using hybrid machine learning ensembles

The best parameters for each classifier for all tasks are found separately using Leave One Out (LOO) cross-validation. The hyperparameters include the model parameters and also reduced feature sets based on different thresholding of important features. The feature importance is explained using SHAP (SHapley Additive exPlanations) values, where thresholding is applied to select the top important features for prediction using each task-based classifier. The increasing thresholding reduces the feature vector set for all tasks, along with the selection of the most important corresponding features. The algorithm for multi-task explainable AI (XAI) integrated hybrid ML ensembles is described in Algorithm 1. The algorithm is based on the nested leave-one-out approach for heart failure prediction using multiple tasks. The outer leave-one-out loop splits the dataset into train and test data, while the inner leave-one-out loop splits the data further into train and validation to perform hyperparameter tuning using leave-one-out cross validation. In the first step of hyperparameter tuning, the model with the best parameters is found. Afterward, SHAP values are computed and a threshold is applied to filter out the most important features. Hence, it reduces the length of the feature vector and produces good performance. Based on the best average threshold accuracies, the best threshold is selected along with the already selected parameters using Grid Search CV for better performance. If the performance of the model, along with a threshold applied on SHAP values, is not better than without feature reduction, then the final test predictions are made without feature reduction. This process repeats for each task and all classifiers are applied one by one, computing

1st stage majority voting on all classifiers. After 1st stage majority voting, all the task-based predictions undergo another final 2nd stage majority voting. Final predictions are made after this final stage of majority voting is applied across all tasks.

**ALGORITHM 1: MULTI-TASKS XAI INTEGRATED HYBRID ML ENSEMBLES**

***Input:*** *Dataset* $D = \{(F(T_i), y_i)\}_{i=1}^{m}$, *where* $F(T_i)$ *is feature set for task* $T_i$ *and label* $y_i$, *classifier pool* $C = \{C_1, C_2, \cdots, C_N\}$, *SHAP thresholds* $\Theta = \{\theta_1, \theta_2, \dots, \theta_k\}$, *parameter grid for each classifier*

***Output:*** *Final prediction class label* $P_{final}$

***for*** *each task* $T_i$ *in* $[T_1, T_2, \dots, T_m]$
  ***for*** *each classifier* $C_j$ *in* $[C_1, C_2, \dots, C_N]$
    ***for*** *split data* $D$ *as* $D_{train+val}$ *and* $D_{test}$ *using Leave-One-Out data split approach*
      *Apply Leave-One-Out Grid Search CV with parameter grid, then train the model with best parameters*
      ***for*** *each threshold* $\theta$ *in* $\Theta$
        ***for*** *split data* $D_{train+val}$ *as* $D_{train}$ *and* $D_{val}$ *using Leave-One-Out data split approach*
          *Compute SHAP values and apply threshold* $\theta$ *to extract important feature set*
          *Record each accuracy* $Acc_{thresh}$ *after feature reduction*
        ***end for***
        ***if*** $mean(Acc_{thresh}) > Acc_{best}$, *then* $Acc_{best} \leftarrow mean(Acc_{thresh})$ *and* $\theta_{best} \leftarrow \theta$ ***end if***
      ***end for***
      *Train the model with best parameters, SHAP values with* $\theta_{best}$ *and record each test accuracy* $Acc_{test}$
    ***end for***
    *Record* $P_j(T_i)$ *and the final accuracy* $Acc_{final} \leftarrow mean(Acc_{test})$
  ***end for***
  *Compute the final prediction* $P'(T_i)$ *for task* $T_i$ *using Stage 1 Majority Voting as mode*$(P_j(T_i))$
***end for***
*Compute the final prediction* $P_{final}$ *as Stage 2 Majority Voting across all tasks as mode*$(P'(T_i))$

## 3. Results

The performance of the two-stage majority voting is evaluated as the final prediction of the biomedical voice signal as to whether it indicates heart failure or not. The leave-one-out cross-validation is used along with SHAP values computation. The final performance is evaluated using leave-one-out approach using the best model (parameters and threshold for SHAP feature reduction) resulting in nested leave-one-out as a complete strategy for heart failure prediction. In the first step, each task undergoes a performance evaluation applying all classifiers. In the second step, 1st stage of majority voting is applied on the predictions of all classifiers at the task level. This results in significantly improving the performance as compared to individual classifiers. The final step, the 2nd stage of majority voting, is performed across the predictions at the task level in heart failure prediction. The complete two-stage majority voting strategy with multiple tasks shows the higher performance of the prediction system as compared to individual classifiers in majority of the tasks. Table 1 shows the performance analysis of all the prediction steps of multi-task XAI integrated Hybrid ML Ensembles. The performance of the average of all the classifiers at the task level shows the diversity of individual classification responses to each task. The greater average performance of task 1 shows the capability of most of the classifiers to well predict at this level. The higher average performance of task 1 leads to higher stage 1 performance of task 1 using majority voting. The performance is less than 50% for task 5, hence the worst performance at stage 1 majority voting. Each classifier's performance is different for each task. The classifiers performing better in one task might not perform better in other tasks. Hence, creating ambiguity in recognizing one classifier as best for all tasks. Our approach, as shown in Table 1, the final prediction at the end of stage 2, majority voting across all tasks, is beneficial to cope with a variety of tasks for heart failure prediction using voice signals.

Table 1: Performance analysis of all the prediction steps of Multi-Tasks XAI integrated Hybrid ML Ensembles

| Prediction Steps | Accuracy | Precision | Sensitivity | Specificity | F1-Score |
|---|---|---|---|---|---|
| Average All Classifiers' Performance (Task 1) | 0.70 | 0.65 | 0.70 | 0.70 | 0.67 |
| Average All Classifiers' Performance (Task 2) | 0.63 | 0.64 | 0.65 | 0.60 | 0.64 |
| Average All Classifiers' Performance (Task 3) | 0.63 | 0.67 | 0.55 | 0.70 | 0.59 |
| Average All Classifiers' Performance (Task 4) | 0.58 | 0.53 | 0.60 | 0.55 | 0.56 |
| Average All Classifiers' Performance (Task 5) | 0.28 | 0.27 | 0.35 | 0.20 | 0.30 |
| Stage 1: Majority Voting (Task 1) | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Stage 1: Majority Voting (Task 2) | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 |
| Stage 1: Majority Voting (Task 3) | 0.63 | 0.66 | 0.50 | 0.75 | 0.57 |
| Stage 1: Majority Voting (Task 4) | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| Stage 1: Majority Voting (Task 5) | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 |
| Stage 2: Majority Voting (All Tasks) | 0.88 | 1.00 | 0.75 | 1.00 | 0.86 |

Fig 2. shows the mean SHAP values across all the training samples for task 1. It shows the feature importance in both classes with task 1 as a sample example using a SVM classifier. The SHAP values are further averaged across all the leave-one-out iterations. The top 15 features contribution for heart prediction is shown in Fig 2. The top feature contributing positively towards the prediction of heart failure in task 1 is the standard deviation of the length of all the phonation. The other features with decreasing positive contribution are mean of (5-95) percentile of 13 MFCC, (5-95) percentile of jitter, (5-95) percentile of 9 MFCC, standard deviation of (5-95) percentile of 13 MFCC, standard deviation of 13 MFCC, and mean of loudness.

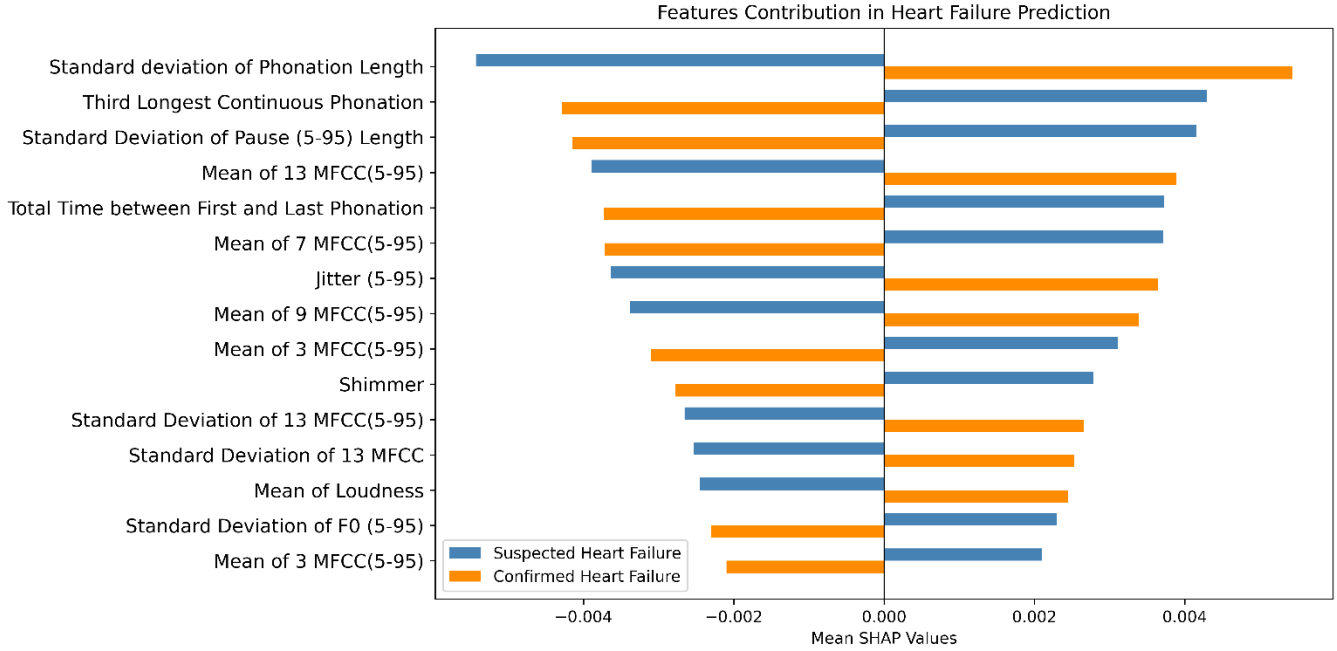


Fig. 2: Mean SHAP values for voice task 1 and SVM classifier in heart failure prediction

## 4. Conclusion

The biomedical voice signal processing for heart failure prediction represents a significant step in non-invasive cardiac diagnostics. We combined multi-task voice acquisition, feature extraction, ensemble machine learning, and a two-stage majority voting strategy to achieve high predictive accuracy and clinical reliability. The two-stage majority voting aggregates diverse classifier outputs to achieve consensus, consequently serving to dilute the impact of potential outliers or misclassifications that may occur when analysing any single voice task in isolation. Future work will focus on exploring additional features and investigating further enhancements of the ensemble strategy for heart failure detection.


## Acknowledgements

This publication is supported by the European Union's Horizon Europe research and innovation programme under the Marie Sklodowska-Curie Postdoctoral Fellowship Programme, SMASH co-funded under the grant agreement No. 101081355. The operation (SMASH project) is co-funded by the Republic of Slovenia and the European Union from the European Regional Development Fund. The study received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101080905.


## References


[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[2] J.V. Firmino, M. Melo, V. Salemi, K. Bringel, D. Leone, R. Pereira, and M. Rodrigues, "Heart failure recognition using human voice analysis and artificial intelligence," *Evolutionary Intelligence*, *16*(6), pp.2015-2027, 2023.

[3] N. Ahmadli, M.A. Sarsil, B. Mizrak, K. Karauzum, A. Shaker, E. Tulumen, D. Mirzamidinov, D. Ural, and O. Ergen, "Voice-driven mortality prediction in hospitalized heart failure patients: A machine learning approach enhanced with diagnostic biomarkers," *arXiv preprint arXiv:2402.13812*, 2024.

[4] Y. Yumita, Z. Xu, G.P. Diller, A. Kempny, I. Rafiq, C. Montanaro, W. Li, H. Gu, K. Dimopoulos, K. Niwa, and M.A. Gatzoulis, "B-type natriuretic peptide levels predict long-term mortality in a large cohort of adults with congenital heart disease," *European Heart Journal*, *45*(23), pp.2066-2075, 2024.

[5] L. Willinger, L. Brudy, A.L. Häcker, M. Meyer, A. Hager, R. Oberhoffer-Fritz, P. Ewert, and J. Müller, "High-sensitive troponin T and N-terminal pro-B-type natriuretic peptide independently predict survival and cardiac-related events in adults with congenital heart disease". *European Journal of Cardiovascular Nursing*, *23*(1), pp.55-61, 2024.

[6] Artificial Intelligence-Based Disease Management in the Vulnerable Period of Heart Failure (AIDMy-HF), clinicaltrials.gov, 2022.

[7] E. Maor, D. Perry, D. Mevorach, N. Taiblum, Y. Luz, I. Mazin, A. Lerman, G. Koren, and V. Shalev, "Vocal biomarker is associated with hospitalization and mortality among heart failure patients", *Journal of the American Heart Association*, 9(7), p.e013359, 2020.

[8] S. Ambesange, A. Vijayalaxmi, S. Sridevi, and B.S. Yashoda, "Multiple heart diseases prediction using logistic regression with ensemble and hyper parameter tuning techniques," in *2020 fourth world conference on smart trends in systems, security and sustainability (WorldS4)* (pp. 827-832), IEEE, 2020.

[9] N.D. Ariyanta, A.N. Handayani, J.T. Ardiansah, and K. Arai, "Ensemble learning approaches for predicting heart failure outcomes: A comparative analysis of feedforward neural networks, random forest, and XGBoost," *Applied Engineering and Technology*, *3*(3), pp.173-184, 2024.

[10] D. Asif, M. Bibi, M.S. Arif, and A. Mukheimer, "Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization," *Algorithms*, *16*(6), p.308, 2023.

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

[12] C. Molnar, *Interpretable machine learning*. Lulu. Com, 2020.

[13] 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* (pp. 1135-1144), 2016.