

A lightweight predictive model for the detection of heart sound abnormalities from heart sound recordings.

By Dagmawi W. Tesfaye, MD

Overview

The variability of clinical auscultation and its dependence on examiner experience were documented in prior literature and were taken as the motivating clinical problem (Davidsen *et al.*, 2023; Gardezi *et al.*, 2018). To address this variability, a reproducible data-driven pipeline for the automated analysis of clear heart sound recordings was implemented. The implemented pipeline aimed to provide a computationally lightweight, interpretable classifier that could be executed on modest hardware while preserving clinically useful sensitivity for detecting abnormal heart sounds (Kotsiantis *et al.*, 2006).

Background and significance

Detecting cardiac lesions accurately was recognized as a critical task because early diagnosis enables timely intervention and improved outcomes. Traditional auscultation was inexpensive and bedside-ready but suffered from inter-examiner variability: auscultatory findings depended strongly on clinician experience, and subtle murmurs could be missed or misinterpreted (Davidsen *et al.*, 2023; Gardezi *et al.*, 2018). A data-driven approach was therefore implemented to standardize screening and scale access to specialist-level triage in low-resource settings by analyzing high-quality phonocardiogram (PCG) recordings (Brodersen *et al.*, 2010).

Although cardiac auscultation remains a foundational component of bedside cardiovascular assessment, its diagnostic performance varies widely across clinicians. Multiple studies show that auscultatory accuracy for valve disease is inconsistent and often inadequate for reliable triage, especially in primary-care or resource-limited settings where specialist access is limited (Davidsen *et al.*, 2023; Gardezi *et al.*, 2018). Existing automated approaches either rely on computationally expensive deep-learning architectures or require large, heterogeneous datasets that are often unavailable in low-resource environments. Consequently, a gap exists between the need for reproducible, accessible decision-support tools and the practical constraints of implementing such systems in real-world, resource-constrained clinical workflows.

This project addresses that gap by developing a lightweight classification pipeline grounded in well-established spectral features of heart sounds (Davis and Mermelstein, 1980). The system was intentionally designed for modest hardware, enabling deployment in settings where computational resources, network access, and high-volume datasets are unavailable. By combining four classical ML models through a probabilistic soft-voting ensemble (Kuncheva, 2004; Kotsiantis *et al.*, 2006). The pipeline provides a reproducible, interpretable framework that mitigates the limitations of purely auscultation-based assessment while avoiding the inflexibility and resource burden of large deep-learning systems. The study thus fills a practical and methodological gap by demonstrating that clinically useful abnormal heart sound detection can be achieved using structured features and transparent algorithms rather than black-box methods (Breiman, 2001; Cortes and Vapnik, 1995; Chen and Guestrin, 2016).

Methods

Dataset generation and description

The primary dataset for this study was derived directly from the raw WAV heart sound recording of the Wadahupy/Heartbeat-Sounds-Dataset (Pereira, 2013). To enable machine learning, each audio file was converted into a fixed-length numerical representation capturing both spectral and temporal characteristics. The resulting feature table was saved as a .csv file, which served as the input for all subsequent modeling and analysis.

This file was chosen for training the model because it was readily available and contains high-quality WAV recordings with clearly defined and clinically interpretable labels. It covers distinct categories that correspond closely to physiological phenomena clinicians must differentiate during auscultation (rather than just discerning between normal and abnormal), making the dataset directly aligned with the intended use case of the algorithm.

The primary dataset contained 878 observations, each corresponding to a single heart sound recording, with one label column indicating the heart sound class [artifact, extrastole, murmur, or normal]. Per-class counts were [40, 19, 46, 129, 351] respectively, totaling 585. Recordings with missing or ambiguous labels were assigned unknown [293 samples] and excluded from supervised training and primary evaluation to focus on the five clinically interpretable classes.

The feature creation pipeline followed a programmatic workflow:

1. **Audio ingestion and channel standardization:** Each WAV file was loaded and converted to a single mono channel to ensure consistent processing across recordings.
2. **Sampling rate handling:** the original sampling rate of each recording was captured and used in subsequent spectral analyses.
3. **MFCC extraction:** For each recording, 13 Mel-Frequency Cepstral Coefficients (MFCCs) were computed for every frame. The resulting matrix was standardized to have frames as rows and coefficients as columns to ensure consistency across recordings (Davis and Mermelstein, 1980).
4. **Feature aggregation:** The frame-level MFCCs were summarized for each recording by calculating the mean and standard deviation of each coefficient, producing 26 features (13 means + 13 SDs).
5. **Temporal derivatives:** First-order (delta) and second-order (delta-delta) changes of each MFCC were computed across frames to capture temporal patterns. The means of each derivative were added to the feature set, contributing another 26 features. Each recording was therefore represented by a total of 52 numeric features.
6. **Label Inference:** Labels for each recording were automatically determined from the file name or path using pattern matching. For instance, files containing “normal” were labeled as “normal”, “murmur” as “murmur”, etc.
7. **Final output:** The final dataset, including the 52 numeric features and the label column, was saved as features_labeled.csv.

A secondary dataset was used for external testing, which is appropriate to evaluate generalization and guard against overfitting. However, it is important to note that this external validation was performed for only the normal heart sound classes due to data scarcity, with a consistent labeling approach as the primary dataset. (Liu *et al.*, 2016).

Dataset preprocessing

Before model training, numeric features were standardized using the caret package's function `preprocess` (method = c("center", "scale")). The mean and standard deviation for each feature were calculated from the training data and saved as a .rds file to ensure consistent scaling during predictions. The standardized features were then combined with the class labels to create the final dataset for model training.

Handling class imbalance

Because class imbalance was present (notably small extrahls, artifact, and extrastole counts relative to normal), the training procedure applied within-fold upsampling during cross-validation by specifying `sampling = "up"` inside caret's `trainControl()` function. The upsampling operation was performed for each training fold independently so that validation (hold-out) folds remained unbiased by resampled duplicates. The repeated stratified cross-validation strategy thus combined resampling and fold stratification to produce robust, unbiased performance estimates on an imbalanced class distribution (He and Garcia, 2009).

Model Training

Four supervised classification models were trained using the caret framework with identical resampling and control settings. The explicit training configuration was as follows:

- `trainControl` parameters: `method = "repeatedcv"`, `number = 10`, `repeats = 3`, `sampling = "up"`, `classProbs = TRUE`, and `savePredictions = "final"`.

These settings yielded stratified 10-fold cross-validation repeated three times, with class balancing performed inside each sampling iteration and out-of-fold predictions retained for each model.

The following classifiers were trained via `caret::train()`:

- k-Nearest Neighbors (method = "knn")
- Support Vector Machine (Radial kernel) (method = "svmRadial") (Cortes and Vapnik, 1995).
- Random Forest (method = "rf") (Breiman, 2001).
- XGBoost (method = "xgbTree") (Chen and Guestrin, 2016).

Model artifacts and the preprocessing object were then saved as rds files. Storing both preprocessing and model objects ensured that the same transformations were applied during later inference and enabled reproducibility of predictions (Kuncheva, 2004).

Combining model predictions for the ensemble

During model training, each model made predictions on the data that was not used for its own training (out-of-fold predictions). These predictions, along with the predicted probabilities for each class, were saved for later analysis. This approach allowed us to evaluate each model on data it had not seen during training, providing a fair estimate of performance.

The class probabilities from all models were then combined to form the ensemble. Each model's contribution was proportional to its macro F1 score on the out-of-fold predictions, so models that performed better had more influence on the final prediction. The combined probability for each class was calculated as a weighted average of the model's probabilities (soft-voting ensemble) (Kuncheva, 2004; Kotsiantis *et al.*, 2006), and the class with the highest combined probability was chosen as the ensemble prediction. The weights were saved as an rds file for reproducibility.

Finally, the ensemble predictions provided a single, consolidated label for each recording. Per-model probabilities were retained and printed to maintain transparency and allow inspection of how each individual model contributed to the final decision.

Performance metrics and evaluation

A custom function was applied to the out-of-fold prediction for each model and for the ensemble. The function computed the following metrics:

1. **Confusion matrix:** counts of correct and incorrect predictions for each class.
2. **Per-class precision and recall:** precision measures the proportion of correct predictions among all predictions for a class, while recall measures the proportion of actual instances of a class that were correctly identified. This was the main metric used for the external validation dataset.
3. **Per-class F1 score:** the harmonic mean of precision and recall for each class, with any undefined or missing values treated as zero.
4. **Macro F1 score:** the average of the F1 scores across all classes, giving equal weight to each class regardless of its size.
5. **Balanced accuracy:** the average of per-class recall values, ensuring that performance on minority classes contributes equally to the overall score (Brodersen *et al.*, 2010).

Prediction function and reproducibility for new audio

A new `predict_new_audio()` function was used to perform end-to-end prediction on a new WAV recording. It performed the following steps:

1. Loaded the saved models, preprocessing scaler, and ensemble weights.
2. Extract features from the recording using the same MFCC + delta/delta-delta pipeline, producing a fixed-length feature vector (Davis and Mermelstein, 1980).
3. Apply the saved scaler to standardize the features.
4. Predict class probabilities using each of the four trained models.
5. Combine the model probabilities using the stored ensemble weights and assign the class with the highest combined probability as the final prediction.
6. Return and print both the per-model probabilities and the ensemble prediction for transparency and interpretability.

Visualization

Performance was summarized in a combined bar plot for macro F1 and balanced accuracy across the four base models and the ensemble. Confusion-matrix heatmaps were plotted for each model and for the ensemble using tile plots with overlaid counts to facilitate inspection of which classes were commonly confused. These visualizations were printed to the R graphics device and were saved as figures for reporting (Kotsiantis *et al.*, 2006).

For the external validation dataset, a bar plot of the normal class metrics was produced for each model. The metrics used were the F1 score, precision, and recall because they assessed the accuracy for the normal class only.

Results

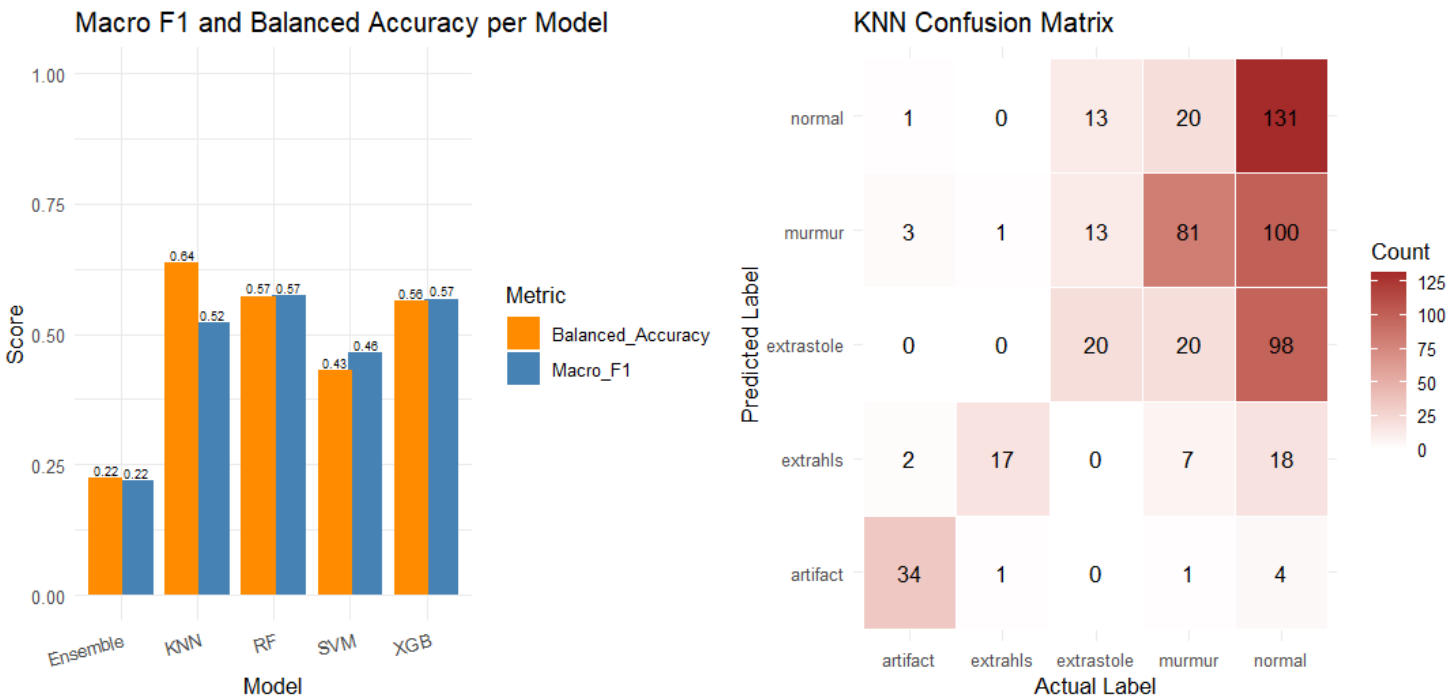


Figure 1 (left): a bar plot depicting per-model performance based on Macro F1 scores and balanced accuracy. **Figure 2 (right):** a confusion matrix showing per-class prediction against the actual classes for the KNN model.

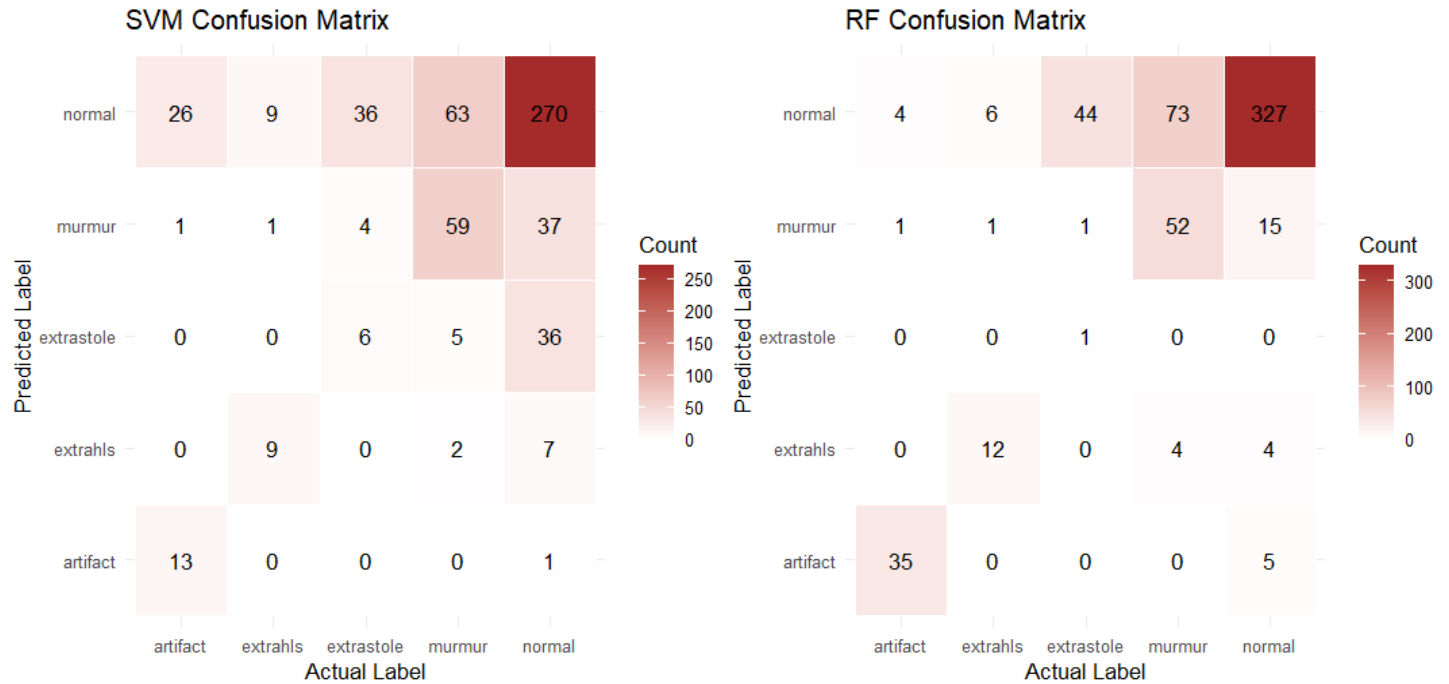


Figure 3 (left): a confusion matrix showing per-class prediction against the actual classes for the SVM model. **Figure 4 (right):** a confusion matrix showing per-class prediction against the actual classes for the Random Forest model.

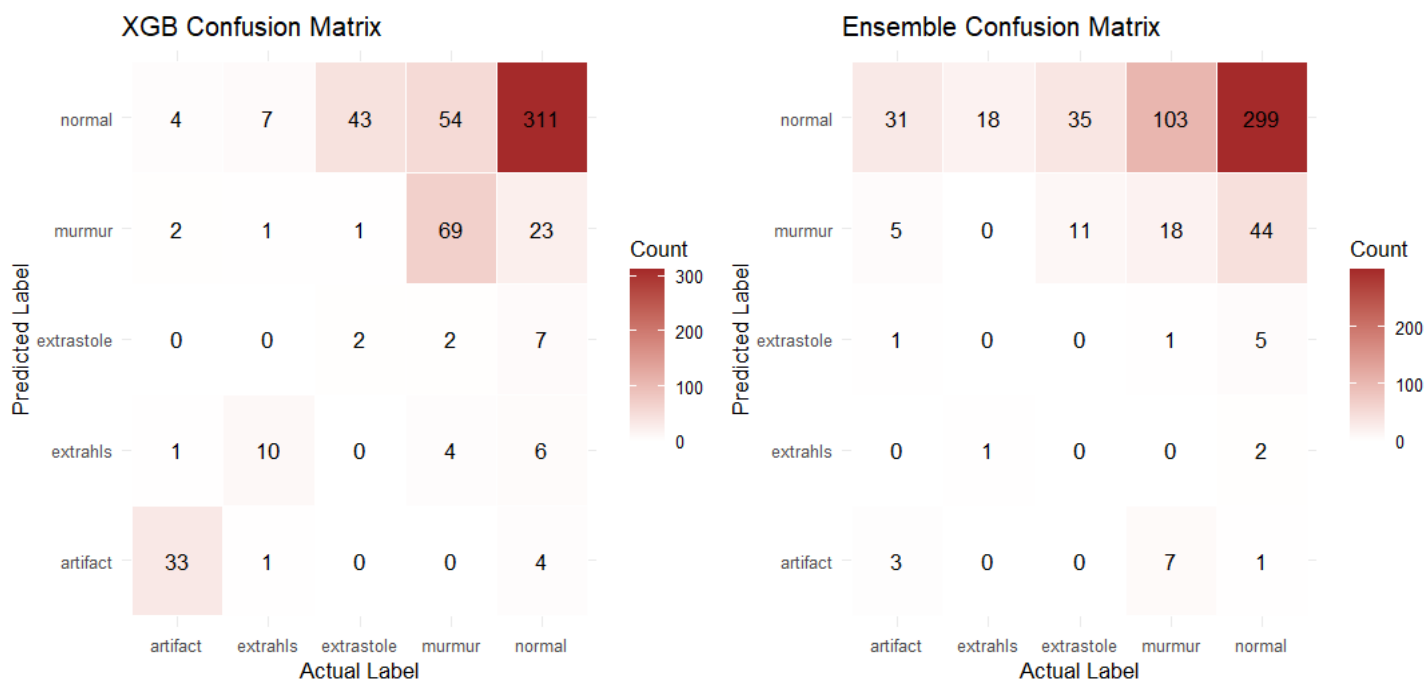


Figure 5 (left): a confusion matrix showing per-class prediction against the actual classes for the XGBoost model. **Figure 6 (right):** a confusion matrix showing per-class prediction against the actual classes for the Ensemble model.

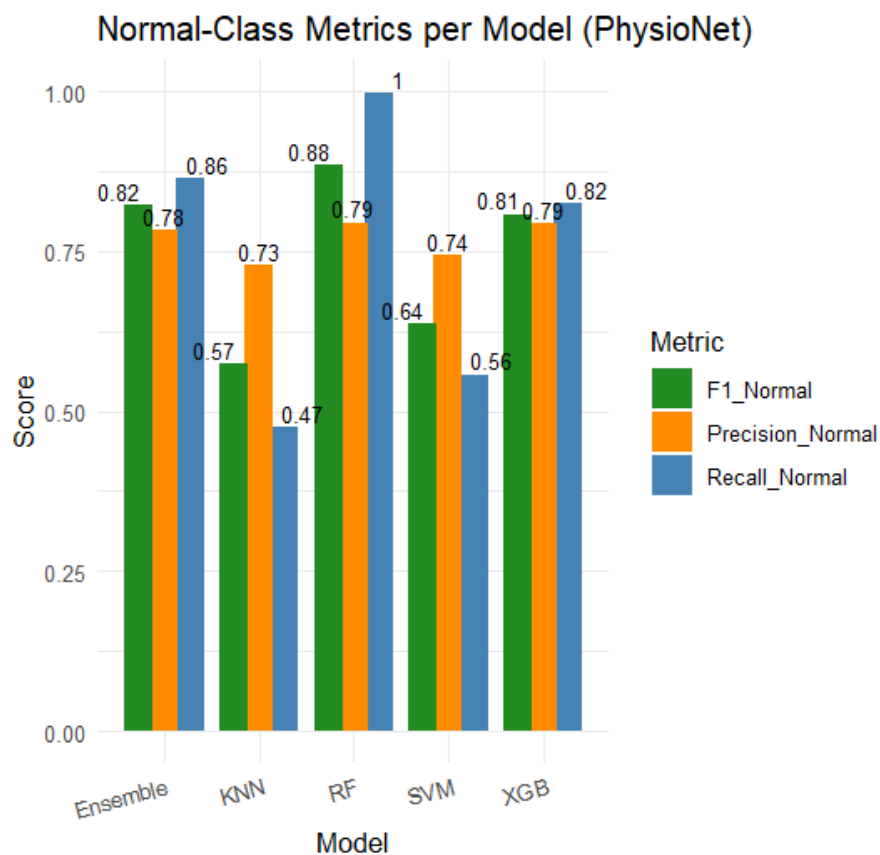


Figure 7: A bar plot depicting per-model performance of the F1, the precision, and the recall scores for the normal classes on the PhysioNet (testing) dataset.

Discussion

In cross-validation on the primary dataset, classifier performance varied widely. We report both macro-F1 and balanced accuracy (the average recall across all classes) (Brodersen *et al.*, 2010), to account for class imbalance. Individual classifiers (KNN, RF, SVM, XGB) all achieved substantially higher scores than the ensemble: Macro F1 and Balanced accuracy, respectively. Ensemble (0.22, 0.22), KNN (0.52, 0.64), Random Forest (0.57, 0.57), SVM (0.43, 0.48), and XGBoost (0.57, 0.58).

On the external PhysioNet test set, all models did better. The ensemble now reached an F1 of 0.82 with a precision of 0.78 and a recall of 0.86 for the normal class. RF had the strongest performance (F1 of 0.88, precision of 0.79, and recall of 1.00), followed by the ensemble, while KNN lagged with (F1 of 0.57, precision of 0.73, and recall of 0.47). These results show that although our ensemble was weak on the original dataset, it generalized reasonably well on the new data, where it was the second-best performer for identifying normal heart sounds.

The ensemble's poor performance on the primary dataset was most likely related to the limited diversity of its component models. Because all classifiers were trained on the same feature representation and showed similar error patterns, the soft-voting step effectively averaged highly correlated outputs rather than combining complementary information. Prior work has shown that averaging the predictions of classifiers with strongly correlated errors offers little benefit and can even reduce overall accuracy. The error reduction expected from combining classifiers depends strongly on the independence of their errors. When errors are correlated, an ensemble does not necessarily outperform its individual members (Tumer and Ghosh, 1996). Similarly, another study showed that an ensemble's performance improves only when its members make different, partially uncorrelated mistakes, highlighting diversity as a key ingredient for successful model combination (Krogh and Vedelsby, 1994).

In our case, the probability averaging procedure diluted the high confidence prediction from models that were correct on a certain class, while reinforcing the shared weakness of the more dominant models. Weighting models by macro-F1 further amplified these effects by assigning greater influence on classifiers that were already making similar decisions. As a result, the ensemble behaved much like a single classifier with its errors preserved and amplified rather than corrected.

Importantly, this behavior does not mean the ensemble will always perform worse than the individual models. On the external PhysioNet dataset, the ensemble recovered substantially stronger performance and achieved a high, stable recall for the normal class. This level of reliability in detecting normal recordings is clinically meaningful in screening settings, where correctly identifying healthy patients helps avoid unnecessary follow-up while focusing attention on potentially abnormal cases. Thus, even though the ensemble underperformed on the primary dataset, its improved generalization on new, unseen data suggests that it still has practical value as part of a lightweight screening tool.

Limitations and observed caveats

Several limitations were encountered during the development and evaluation of the heartbeat classification system. The dataset exhibited pronounced class imbalance, with "extrahls" and "extrastole" categories represented by only 19 and 46 recordings, respectively. Although upsampling was applied to compensate for these disparities (He and Garcia, 2009). The limited number of unique examples restricted the model's ability to learn broadly generalizable patterns. This scarcity also introduced a greater risk of overfitting to duplicated minority examples during training.

In addition, the decision to summarize each recording into mean and standard deviation statistics of MFCCs and their temporal derivatives provided a lightweight and interpretable representation but came at the expense of temporal resolution. This approach contrasts with sequence-based or beat-level

models such as convolutional or recurrent neural networks, which can exploit finer temporal dynamics (Chen and Guestrin, 2016). The trade-off was intentionally accepted to maintain computational efficiency and interpretability, both of which were key design priorities of this project.

The emphasis on transparency and reproducibility also guided the choice of using classical machine-learning models rather than deep learning architectures (Breiman, 2001; Cortes and Vapnik, 1995). While these models are faster to train and explain, they may achieve slightly lower predictive performance compared to more complex, data-intensive approaches. Finally, although the soft-voting ensemble generally improved stability and robustness, it occasionally underperformed individual base models due to imperfect weighting (Kuncheva, 2004) or variability in per-class probability distributions, particularly for minority classes.

Future directions

Future work should explore multimodal data integration to improve heart sound classification. For example, recent studies show that combining PCG (phonocardiogram) features with patient demographic information can significantly boost performance (Despotovic *et al.*, 2026). One could integrate age, gender, medical history, or laboratory values alongside the sound recordings. Likewise, fusing heart sound signals with other cardiac data, such as ECG waveforms or echocardiogram findings, could provide complementary insights. In fact, adding ECG signals to PCG analysis has been shown to capture more aspects of heart function and improve diagnostic accuracy (Mains and Kshirsagar, 2024). The same can be said about incorporating key lab measurements, such as cardiac biomarkers or imaging features that could make the model more sensitive to disease detection. There is also a huge gap in the literature when it comes to combining such datasets with omics data. In summary, a multimodal approach that combines phonocardiogram features with demographic data, lab values, ECG, or echo is a promising direction for more accurate and robust heart disease screening.

Acknowledgements

This thesis would not be possible without the guidance and contributions of a few individuals who played key roles in the completion of my work. It is a pleasure for me to thank those who made it possible for me.

I am indebted to my advisor, Marcos Perez-Losada (Ph.D.), and research supervisor Ali Rahnavard (Ph.D.), whose support, suggestions, and flexibility allowed me to complete my thesis on time. I am grateful for your guidance, insights, and encouragement that shaped the direction and quality of this thesis.

Additionally, I would like to extend my sincere thanks to Adam Ciarleglio (Ph.D.), whose expertise and feedback were instrumental to my project. It was genuinely reassuring to be able to discuss the thesis results with you, and the clarity gained from those conversations was invaluable.

Thank you, Alireza Taheriyoun (Ph.D.), for taking your valuable time to meet with me every week to discuss my progress and for providing continuous guidance and assistance. I would also like to thank Keith A. Crandell (Ph.D.) as program director for a smooth introduction into the world of bioinformatics/health data science research. Without him, I would not be in this position in the first place.

Lastly, but most importantly, none of this could have been possible without the support of my family. Thank you to my parents and brothers for providing me with moral support whenever I felt disheartened and stressed. This thesis is a testament to their dedication and support.

Resources

The source code and trained models are publicly available on GitHub. The bot link is also available.

GitHub: <https://github.com/Dagm-Workalemahu/lightweight-heart-sound-classification>

Telegram bot: <https://t.me/LubDub2bot>

References

- Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5–32.
- Brodersen, K.H. et al. (2010) The Balanced Accuracy and Its Posterior Distribution. In, *2010 20th International Conference on Pattern Recognition.*, pp. 3121–3124.
- Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. In, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*, pp. 785–794.
- Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Mach Learn*, **20**, 273–297.
- Davidson, A.H. et al. (2023) Diagnostic accuracy of heart auscultation for detecting valve disease: a systematic review. *BMJ Open*, **13**, e068121.
- Davis, S. and Mermelstein, P. (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **28**, 357–366.
- Despotovic, V. et al. (2026) CardioPHON: Quality assessment and self-supervised pretraining for screening of cardiac function based on phonocardiogram recordings. *Biomedical Signal Processing and Control*, **113**, 109047.
- Gardezi, S.K.M. et al. (2018) Cardiac auscultation poorly predicts the presence of valvular heart disease in asymptomatic primary care patients. *Heart*, **104**, 1832–1835.
- He, H. and Garcia, E.A. (2009) Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1263–1284.
- Kotsiantis, S.B. et al. (2006) Machine learning: a review of classification and combining techniques. *Artif Intell Rev*, **26**, 159–190.
- Krogh, A. and Vedelsby, J. (1994) Neural network ensembles, cross validation and active learning. In, *Proceedings of the 8th International Conference on Neural Information Processing Systems*, NIPS'94. MIT Press, Cambridge, MA, USA, pp. 231–238.
- Kuncheva, L.I. (2004) Combining Pattern Classifiers: Methods and Algorithms 1st ed. Wiley.
- Liu, C. et al. (2016) An open access database for the evaluation of heart sound algorithms. *Physiol. Meas.*, **37**, 2181.
- Mains, T. and Kshirsagar, S. (2024) A Machine Learning Approach for Integrating Phonocardiogram and Electrocardiogram Data for Heart Sound Detection.
- Pereira, E. (2013) Classifying Heart Sounds - Approaches to the PASCAL Challenge. *Proceedings of the International Conference on Health Informatics*.
- Tumer, K. and Ghosh, J. (1996) Error Correlation And Error Reduction In Ensemble Classifiers. *Connection Science*, **8**.