



Research article

Normal-abnormal heart sound classification using MFCC-based acoustic features and gradient boosting

Jingyu Fang and Lijun Pei*

School of Mathematics and Statistics, Zhengzhou University, Zhengzhou 450001, China

* **Correspondence:** Email: peilijun@zzu.edu.cn.

Abstract: Automated heart-sound classification can be affected by duplicate recordings, class imbalance, and information leakage during data preparation. We evaluated an MFCC-based acoustic-feature pipeline for recording-level normal-abnormal heart-sound classification. The assembled dataset initially contained 3906 WAV files drawn from a subset of the PhysioNet/Computing in Cardiology Challenge 2016 and the PASCAL Classifying Heart Sounds Challenge 2011. Byte-level deduplication removed 266 copies before partitioning, leaving 3640 unique recordings. Eighteen global acoustic features were extracted from each unique recording: 13 mean mel-frequency cepstral coefficients, three MFCC dispersion or temporal-difference summaries, mean spectral centroid, and mean zero-crossing rate. Logistic regression, a multilayer perceptron, LightGBM, and XGBoost were compared, with Bayesian hyperparameter optimization conducted by five-fold grouped cross-validation within the training partition for both gradient-boosting models. On the held-out set, Bayesian-optimized LightGBM achieved an area under the receiver operating characteristic curve of 0.965 (recording-level 95% bootstrap interval 0.951–0.976), accuracy of 0.920, and balanced accuracy of 0.886. The results support the usefulness of the implemented feature set and gradient-boosting models for internal classification within the assembled two-source dataset.

Keywords: heart sound; normal–abnormal classification; MFCC; gradient boosting; Bayesian optimization

1. Introduction

Automated heart-sound classification offers a practical complement to cardiac auscultation, whose interpretation varies with examiner training, recording quality, and clinical context. Public phonocardiogram (PCG) resources, including the PhysioNet/Computing in Cardiology Challenge 2016 [1, 2] and the PASCAL Classifying Heart Sounds Challenge [3], have enabled normal-abnormal classification from segmented and unsegmented recordings [4]. Because these resources combine

recordings acquired under different devices, settings, and labeling practices, transparent data curation and source-aware evaluation are important for reproducible model development. Cross-corpus and validation studies further emphasize grouped partitioning, training-fitted preprocessing, and explicit reporting in heterogeneous PCG datasets [5–8].

The analytical pipeline combines compact acoustic features with conventional machine learning. Class imbalance is addressed by applying the synthetic minority over-sampling technique (SMOTE) within the training data [9]. MFCCs summarize the spectral envelope on a perceptually motivated frequency scale [10], while LightGBM and XGBoost capture nonlinear interactions among tabular acoustic features [11, 12]. Because model performance depends on the selected hyperparameters, manually testing many candidate combinations can be time-consuming. Bayesian optimization therefore provides an efficient strategy for searching the parameter space [13].

Here, we first curated public heart-sound recordings and removed byte-identical duplicates, then extracted 18 acoustic features from 3640 unique recordings and established a fixed group-aware training-test split. Standardization, SMOTE, and Bayesian hyperparameter optimization were performed within the training data before comparing logistic regression, a multilayer perceptron, LightGBM, and XGBoost. Feature ablation and recording-level bootstrap intervals supplemented the model comparison. This workflow establishes a reproducible internal classification baseline for the assembled public recordings and supports subsequent evaluation on independent data sources. Transparent provenance, software environments, and preprocessing implementations further support reproducibility of the workflow [14–17].

2. Related work

Early PCG classification pipelines often segmented recordings into S1, systole, S2, and diastole before extracting cycle-level descriptors. Duration-dependent hidden Markov and hidden semi-Markov models established influential formulations for this step [18, 19]. Challenge systems then combined time–frequency, cepstral, and wavelet features with AdaBoost, convolutional neural networks, support-vector machines, or neural ensembles [20–24]. Alongside these segmented approaches, unsegmented spectral and wavelet analyses demonstrated the feasibility of whole-recording representations [4]. Collectively, this work shows that handcrafted acoustic features can support heart-sound classification while reducing reliance on exact cycle segmentation. Table 1 summarizes representative methodological choices and their relevance to the present study.

Studies have shifted representation learning toward deep models. Improved MFCC features, convolutional recurrent networks, wavelet scattering, learnable front ends, and unsegmented neural systems have been used to preserve local temporal structure beyond global summary statistics [25–28]; a survey further highlights the continuing importance of data quality and generalization [29]. The evidence base has also expanded through richly annotated pediatric recordings in CirCor DigiScope and multi-location recordings in PhysioNet Challenge 2022 [30, 31], while methods such as PCGmix illustrate ongoing interest in data augmentation [32]. Fuadah et al. present a closely related MFCC-based conventional-learning baseline with grid-search tuning on public challenge recordings [33]. Recent work has explored temporal convolution and asynchronous channel fusion across public PCG databases to address domain adaptation and external performance [34]. These developments broaden the available tasks and resources, although

differences in data sources, label definitions, and endpoints require method-focused rather than direct score-based comparisons.

Table 1. Representative heart-sound studies and their relevance to this study.

Study	Primary representation	Evaluation setting	Relevance or distinction
Schmidt et al. [18]	Duration-dependent HMM	Heart-state segmentation	Established cycle-state modeling used by later PCG classifiers.
Springer et al. [19]	Logistic regression–HSMM	Heterogeneous PCG segmentation	Widely adopted segmentation front end for comparison with the whole-recording feature extraction used here.
Potes et al. [20]	Handcrafted features + CNN ensemble	PhysioNet/CinC hidden challenge test	Segmented ensemble for comparison with the whole-recording MFCC-based representation evaluated here.
Rubin et al. [21]	MFCC heat maps + CNN	PhysioNet/CinC Challenge	Retains local time–frequency structure; our features temporally aggregate each recording.
Tschannen et al. [22]	Wavelet-CNN features + SVM	PhysioNet/CinC hidden challenge test	Hybrid learned and summary representation requiring cycle segmentation.
Zabihi et al. [23]	Feature-based neural ensembles	PhysioNet/CinC Challenge	Demonstrated a segmentation-free alternative on heterogeneous recordings.
Langley and Murray [4]	Spectral and wavelet summaries	Unsegmented public PCGs	Closest methodological precedent for whole-recording acoustic-feature analysis.
Yaseen et al. [24]	Multiple handcrafted feature families	Normal/abnormal classification	Supports complementary temporal, spectral, and cepstral information.
Deng et al. [25]	Improved MFCC + CRNN	PCG classification experiments	Learns temporal structure at greater model complexity than the present tabular pipeline.
Mei et al. [26]	Quality assessment + wavelet scattering	PCG classification experiments	Explicitly incorporates signal quality, which our source labels do not provide uniformly.
Liu et al. [27]	Learnable front end + attention	Deep heart-sound classification	Illustrates contemporary end-to-end feature learning.
Koike et al. [5]	Cross-corpus transfer + augmentation	Training and testing on distinct corpora	Shows why pooled-source internal validation should not be interpreted as external generalization.

Across these directions, evaluation depends on how recordings are curated, grouped, preprocessed, and used for model selection. Cross-corpus transfer and augmentation studies underscore the value of independent evaluation [5, 32], while methodological work on grouping and leakage motivates source-aware partitioning and training-fitted operations [6–8]. In this context, we removed byte-identical duplicates, extracted 18 recording-level acoustic features, and compared logistic regression, a multilayer perceptron, LightGBM, and XGBoost under a fixed group-aware held-out design. Standardization, SMOTE, and Bayesian hyperparameter optimization are performed within the training data, with feature ablation and recording-level bootstrap intervals used to complement the internal model comparison.

3. Materials and methods

3.1. Study design and target definition

The analysis was a retrospective computational study of publicly available, de-identified PCG recordings. The prediction unit was one WAV recording and the binary target was the directory label, normal (0) or abnormal (1). The positive class was not interpreted as congenital heart disease, coronary heart disease, or any single pathology because the source datasets contained heterogeneous abnormalities and did not support a uniform disease-specific label. All sample counts referred to original WAV recordings rather than participants or synthetic observations.

The analytical workflow comprised file inventory and hashing, removal of byte-identical duplicates, assignment of source and grouping identifiers, recording-level feature extraction, creation of a fixed grouped hold-out split, and training-only fitting of all learned preprocessing, resampling, and hyperparameter-selection procedures. Figure 1 makes the train–test boundary explicit, and Figure 2 reports the effect of deduplication on dataset composition.

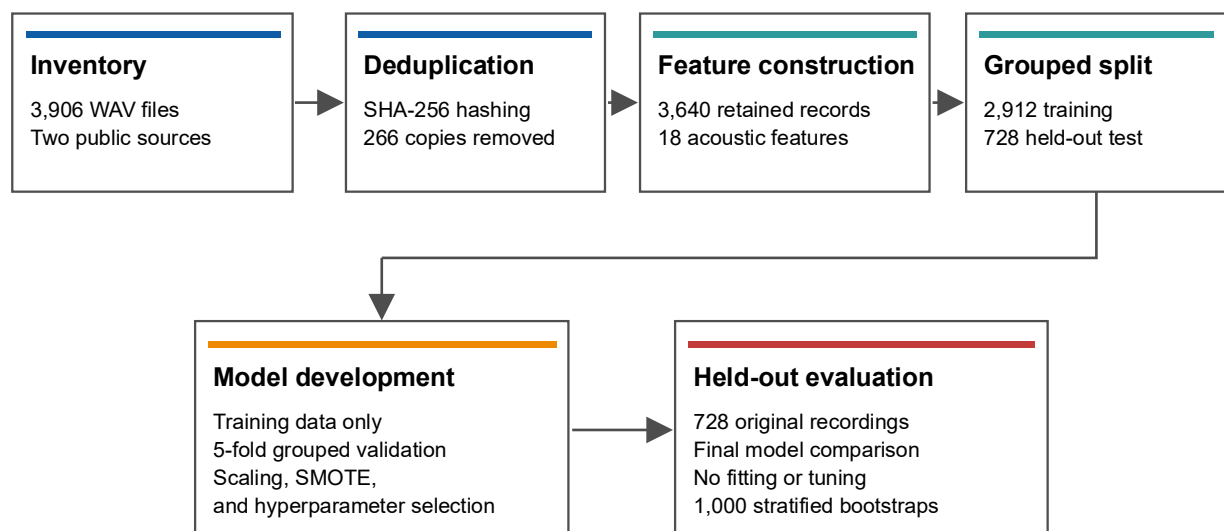


Figure 1. Leakage-controlled study workflow.

3.2. Data sources and deduplication

The assembled corpus contained 3906 files: 3160 files with PhysioNet/CinC-style identifiers and 746 files attributed to the PASCAL/Bentley collection. The PhysioNet portion comprised a subset of the Challenge 2016 training resource. PhysioNet provided open access to the recordings under the terms shown on the resource page [1, 2, 14]. The PASCAL portion comprised available normal and murmur challenge recordings [3]. This was a secondary analysis of de-identified public recordings with no new recruitment, intervention, participant contact, or re-identification; accordingly, no new ethics approval was sought for this analysis.

For every file, a SHA-256 digest was calculated from the raw bytes. Files with identical digests and the same label were treated as duplicates, and one deterministic representative was retained. No identical digest occurred across conflicting labels. This procedure removed 266 duplicate copies and left 3640 recordings. The removed copies all came from repeated PASCAL/Bentley files in the assembled corpus. Table 2 gives the resulting counts.

Table 2. Recording counts before and after deduplication.

Source	Before deduplication		After deduplication		Removed copies
	Normal	Abnormal	Normal	Abnormal	
PhysioNet/CinC 2016 subset	2495	665	2495	665	0
PASCAL/Bentley 2011	551	195	351	129	266
Total	3046	860	2846	794	266

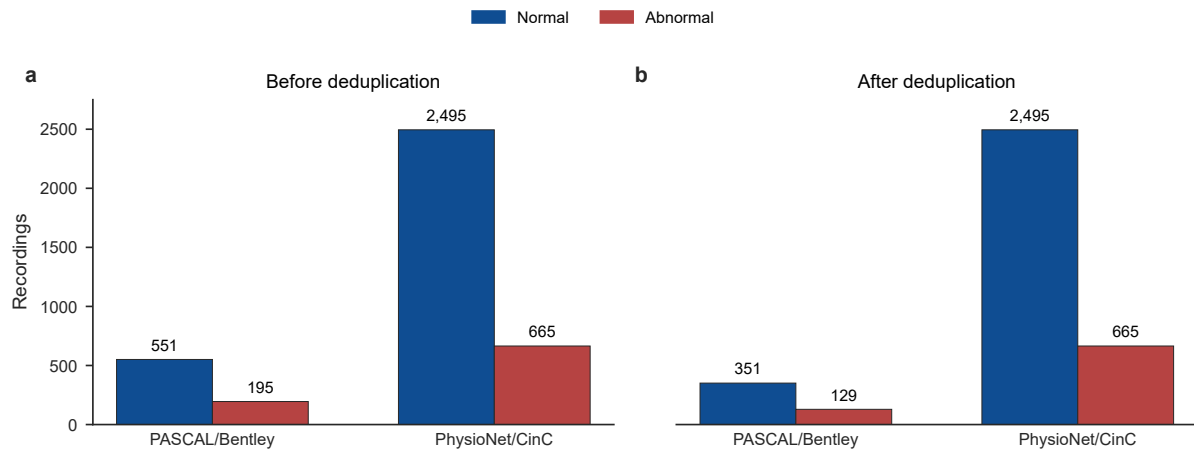


Figure 2. Dataset composition before and after deduplication.

3.3. Grouping and fixed held-out split

PASCAL/Bentley filenames that encoded a three-digit participant identifier were grouped by that identifier; multiple recordings from the same encoded participant were therefore kept together. When no participant identifier was recoverable, a recording-level group was used. The local PhysioNet subset did not include a reliable patient mapping in the assembled directory, so each PhysioNet file was treated as a recording-level group. This approach was stricter than row-level splitting but did not guarantee patient-level separation for every PhysioNet record; this limitation is noted in the interpretation.

A StratifiedGroupKFold splitter with five folds, shuffling, and random seed 42 was applied to strata defined jointly by corpus and class. The first fold was designated as the held-out test set, giving 2912 training recordings and 728 test recordings. The test set contained 569 normal and 159 abnormal recordings, with both sources represented (Table 3). The test set was excluded from scaling, SMOTE, early stopping, pipeline fitting, and hyperparameter selection. It was used only for the final comparison of the prespecified model configurations.

Table 3. Fixed grouped data split by source and class.

Partition	Source	Normal	Abnormal	Total
Training	PhysioNet/CinC 2016 subset	1996	532	2528
Training	PASCAL/Bentley 2011	281	103	384
Held-out test	PhysioNet/CinC 2016 subset	499	133	632
Held-out test	PASCAL/Bentley 2011	70	26	96
Total	Both sources	2846	794	3640

3.4. Acoustic feature construction

Each recording was converted to mono and resampled to 22.05 kHz using librosa 0.11.0 [15]. The implementation used its defaults of a centered 2048-sample Hann window, a 512-sample hop, a power spectrogram, and 128 area-normalized mel filters. The default Slaney mel scale was therefore

$$m(f) = \begin{cases} 3f/200, & 0 \leq f < 1000, \\ 15 + 27 \ln(f/1000)/\ln(6.4), & f \geq 1000, \end{cases} \quad (3.1)$$

where f denotes acoustic frequency in hertz and $m(f)$ is the corresponding frequency on the mel scale. Thirteen MFCC sequences were calculated from each complete recording. Let $E_j(t)$ be the log energy in mel filter j at frame t . With the orthonormal type-II discrete cosine transform used by the implementation, a cepstral coefficient can be written as

$$c_k(t) = \alpha_k \sum_{j=0}^{M-1} E_j(t) \cos \left[\frac{\pi k}{M} \left(j + \frac{1}{2} \right) \right], \quad (3.2)$$

where $\alpha_0 = M^{-1/2}$ and $\alpha_k = (2/M)^{1/2}$ for $k > 0$. Here, M is the number of mel filters, j indexes the filters, k indexes the cepstral coefficients, t indexes the analysis frames, and T is the total number of frames in a recording. The temporal mean of coefficient k is:

$$\bar{c}_k = \frac{1}{T} \sum_{t=1}^T c_k(t), \quad k = 0, \dots, 12. \quad (3.3)$$

Dynamic cepstral coefficients were calculated with librosa's local polynomial derivative operator. Here, $\Delta c_k(t)$ and $\Delta^2 c_k(t)$ denote the first- and second-order temporal differences of the k th MFCC sequence, respectively. Such delta features were introduced to capture spectral change that static coefficients omit [35]. The three global MFCC dispersion/dynamics summaries are defined as follows:

$$s_c = \left[\frac{1}{13T} \sum_{k=0}^{12} \sum_{t=1}^T (c_k(t) - \mu_c)^2 \right]^{1/2}, \quad (3.4)$$

$$\mu_{\Delta} = \frac{1}{13T} \sum_{k=0}^{12} \sum_{t=1}^T \Delta c_k(t), \quad (3.5)$$

$$v_{\Delta^2} = \frac{1}{13T} \sum_{k=0}^{12} \sum_{t=1}^T (\Delta^2 c_k(t) - \mu_{\Delta^2})^2, \quad (3.6)$$

where μ_c and μ_{Δ^2} denote the corresponding matrix means. Two complementary time–frequency descriptors were then appended. Given the short-time Fourier magnitude $A_t(b)$ and the bin-center frequency f_b , where b indexes the frequency bins, the frame spectral centroid was

$$SC(t) = \frac{\sum_b f_b A_t(b)}{\sum_b A_t(b)}, \quad \overline{SC} = \frac{1}{T} \sum_{t=1}^T SC(t). \quad (3.7)$$

For a waveform frame $x_t(n)$ of length N , where n is the within-frame sample index, the zero-crossing rate was

$$ZCR(t) = \frac{1}{2N} \sum_{n=1}^{N-1} |\text{sgn}(x_t(n)) - \text{sgn}(x_t(n-1))|, \quad (3.8)$$

followed by its frame mean. Spectral centroid and zero-crossing rate are low-cost audio descriptors; in PCG recordings, they should be regarded as signal summaries, not direct physiological measurements.

Table 4 maps the implemented columns to their dimensions and intended information. The final 18-dimensional vector comprised 13 MFCC means, the MFCC distribution/difference summaries, mean spectral centroid, and mean zero-crossing rate. These are global recording summaries rather than a frame sequence; accordingly, the comparison focused on classifiers designed for fixed-length tabular inputs.

Table 4. Acoustic feature vector.

Feature group	Dimension	Information summarized
MFCC temporal means ($\bar{c}_0$ – $\bar{c}_{12}$)	13	Average mel-cepstral spectral-envelope shape over the recording.
MFCC matrix standard deviation (s_c)	1	Overall dispersion across coefficient and frame indices.
Mean first difference (μ_{Δ})	1	Average first-order local cepstral change.
Second-difference variance (v_{Δ^2})	1	Variability of cepstral acceleration across the recording.
Mean spectral centroid ($\overline{SC}$)	1	Magnitude-weighted center of spectral energy.
Mean zero-crossing rate ($\overline{ZCR}$)	1	Average rate of waveform sign changes.
Total	18	Fixed-length vector supplied to every classifier.

3.5. Leakage-controlled preprocessing and classifiers

Every estimator was implemented as a single imbalanced-learn pipeline [16]. Within each training fold, StandardScaler first estimated the feature mean and standard deviation, after which SMOTE with five neighbors generated minority-class feature vectors. Outer validation-fold and held-out test

recordings were transformed using training-fold statistics and were never oversampled. For the MLP, scikit-learn drew its 15% early-stopping subset from the already standardized and SMOTE-resampled training matrix; this internal subset contained no outer-validation or held-out-test rows.

For feature j , standardization used only the n_{tr} recordings in the active training fold:

$$z_{ij} = \frac{x_{ij} - \mu_j^{(\text{tr})}}{\sigma_j^{(\text{tr})}}, \quad \mu_j^{(\text{tr})} = \frac{1}{n_{\text{tr}}} \sum_{i \in \text{tr}} x_{ij}. \quad (3.9)$$

Here, i indexes recordings, j indexes features, and $\mu_j^{(\text{tr})}$ and $\sigma_j^{(\text{tr})}$ are the mean and standard deviation of feature j , estimated exclusively from the n_{tr} recordings in the active training fold. SMOTE then selected a minority observation z_i , one of its minority-class neighbors $z_{i,\text{nn}}$, and $\lambda \sim U(0, 1)$ to form [9]

$$z_{\text{new}} = z_i + \lambda(z_{i,\text{nn}} - z_i). \quad (3.10)$$

This ordering is important: neighbor geometry was defined in the training-standardized space, and no validation/test vector could act as an interpolation endpoint.

Four model families were compared. Logistic regression provided a linear reference. The multilayer perceptron (MLP) used two hidden layers of 128 and 64 rectified-linear units, L_2 coefficient 10^{-3} , Adam optimization, and early stopping on the internal post-SMOTE subset described above. The LightGBM baseline used 400 trees, learning rate 0.05, 31 leaves, and L_1/L_2 regularization of 0.1. The XGBoost baseline used 400 trees, learning rate 0.05, maximum depth 5, row and column subsampling of 0.8, and fixed regularization. In generic gradient boosting, an additive predictor is updated as [36]

$$F_m(x) = F_{m-1}(x) + \eta h_m(x), \quad (3.11)$$

where $F_m(x)$ is the additive predictor after boosting iteration m , $h_m(x)$ is the tree fitted at that iteration, and η is the learning rate. LightGBM and XGBoost use different tree-growing and system-level strategies around this common additive principle [11, 12].

LightGBM and XGBoost were additionally tuned by Gaussian-process Bayesian search with 20 proposed configurations. The search objective was mean receiver operating characteristic area under the curve (AUROC) across an inner five-fold stratified group cross-validation of the training set. The searched parameters covered tree depth or leaf count, learning rate, child-size controls, row/feature subsampling, and L_1/L_2 penalties (Table 5). Log-uniform priors were used for learning rate and regularization magnitudes. The held-out test set was evaluated once after refitting the selected configuration to all training data. This separation guards against selection-induced optimism: reusing the same validation results to choose a configuration and report its error can bias performance estimates [37, 38].

3.6. Evaluation and feature ablation

The primary metric was AUROC, supplemented by accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, F1 score, average precision, and Matthews correlation coefficient. AUROC summarizes ranking across all thresholds [39], whereas average precision emphasizes positive-class retrieval and is informative under class imbalance [40]. For a confusion

Table 5. Model configurations and Bayesian search domains.

Model	Fixed baseline configuration	BO configuration or search domain
Logistic regression	Maximum 3000 iterations	Not tuned
MLP	Hidden layers 128/64; $\alpha = 10^{-3}$; learning rate 10^{-3} ; 15% post-SMOTE internal validation; maximum 500 iterations	Not tuned
LightGBM	400 trees; learning rate 0.05; 31 leaves; $\lambda_1 = \lambda_2 = 0.1$	500 trees; leaves [16,96]; depth [3,12]; learning rate [0.005,0.2]; minimum child samples [10,100]; feature fraction [0.6,1.0]; λ_1 [10^{-4} ,1]; λ_2 [10^{-4} ,2]
XGBoost	400 trees; learning rate 0.05; depth 5; row/column fractions 0.8; $\lambda_1 = 0.1$, $\lambda_2 = 1$	500 trees; depth [3,10]; learning rate [0.005,0.2]; child weight [1,10]; row/column fractions [0.6,1.0]; λ_1 [10^{-4} ,1]; λ_2 [10^{-3} ,5]

matrix with true positives TP , true negatives TN , false positives FP , and false negatives FN , the two principal threshold-dependent summaries were

$$BAcc = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad (3.12)$$

and

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \quad (3.13)$$

MCC is a correlation-like summary that remains informative when class sizes differ [41]. The decision threshold was 0.5 and was not optimized on the held-out set. Two-sided 95% percentile intervals were obtained from 1000 class-stratified bootstrap samples of the held-out recordings, following the nonparametric bootstrap principle [42]. Because this bootstrap resampled recordings rather than confirmed patients, the intervals quantified recording-level sampling uncertainty only.

Feature ablation was restricted to the training data and used five-fold stratified group cross-validation with the baseline LightGBM model. Three nested feature sets were compared: 13 MFCC means; the MFCC means plus three dispersion/difference summaries (16 features); and all 18 acoustic features after adding spectral centroid and zero-crossing rate.

3.7. Reproducibility and software

All analyses were conducted in Python 3.12 with NumPy, pandas, SciPy, librosa, scikit-learn [17], imbalanced-learn, LightGBM, XGBoost, and scikit-optimize. Package versions were fixed with uv; random seeds were set to 42 where supported. The supplementary package contained the analysis script, environment lock file, file-level deduplicated index, derived feature table, split manifest, model metrics, Bayesian-search parameters, and figure-generation code. Raw audio was not redistributed and remains available from the original providers.

4. Results

4.1. Dataset composition and held-out set

Byte-level deduplication identified 266 duplicate copies among the 3906 assembled files, corresponding to 6.81% of the corpus. After their removal, 3640 recordings remained across 3385 grouping identifiers. The class distribution was 2846 normal (78.19%) and 794 abnormal (21.81%). Deduplication affected only the PASCAL/Bentley portion, reducing it from 746 files to 480 recordings; no byte-identical conflict crossed the normal–abnormal boundary.

The fixed held-out set contained 728 original recordings and no synthetic samples. It preserved the assembled class proportions and included 632 PhysioNet/CinC-style recordings and 96 PASCAL/Bentley recordings. All subsequent held-out results refer to this same fixed set.

4.2. Held-out test performance

Table 6 reports the held-out metrics. The Bayesian-optimized LightGBM produced the highest held-out AUROC of 0.965 (95% recording-level bootstrap interval: 0.951–0.976) and accuracy of 0.920 (interval: 0.901–0.941). Its balanced accuracy was 0.886 (0.854–0.917), with sensitivity 0.824 and specificity 0.947. It correctly classified 539 of 569 normal recordings and 131 of 159 abnormal recordings, giving 30 false positives and 28 false negatives. The untuned LightGBM was nearly identical in ranking performance (AUROC 0.964) and balanced accuracy (0.886), so the observed benefit of Bayesian search was incremental rather than transformative.

Table 6. Held-out test-set performance.

Model	Accuracy	BAcc	Sensitivity	Specificity	Precision	F1	AUROC	AP	MCC
Logistic regression	0.680	0.673	0.660	0.685	0.370	0.474	0.755	0.420	0.293
MLP	0.880	0.862	0.830	0.895	0.688	0.752	0.944	0.857	0.680
LightGBM	0.913	0.886	0.836	0.935	0.782	0.809	0.964	0.896	0.753
LightGBM +									
Bayesian optimization	0.920	0.886	0.824	0.947	0.814	0.819	0.965	0.899	0.768
XGBoost	0.916	0.894	0.855	0.933	0.782	0.817	0.962	0.884	0.764
XGBoost +									
Bayesian optimization	0.913	0.883	0.830	0.937	0.786	0.807	0.963	0.890	0.752

Untuned XGBoost achieved the highest balanced accuracy (0.894) and sensitivity (0.855), while its AUROC was 0.962. Bayesian tuning increased XGBoost AUROC by less than 0.001 but reduced its threshold-dependent balanced accuracy to 0.883. The MLP remained competitive (AUROC 0.944), whereas logistic regression was substantially weaker (AUROC 0.755 and balanced accuracy 0.673). These comparisons were descriptive for one fixed held-out set and were not treated as pairwise significance tests. The complete ROC curves and the confusion matrices for the two highest-AUROC models are shown in Figures 4 and 5, respectively.

Figure 3 shows that the four boosted configurations had substantially overlapping recording-level intervals. For example, the Bayesian LightGBM AUROC interval (0.951–0.976) overlapped the untuned LightGBM (0.952–0.975), untuned XGBoost (0.949–0.974), and Bayesian XGBoost (0.949–0.975). The nonlinear models showed markedly higher point estimates than the linear

reference, but the overlapping intervals did not establish a clear performance advantage of one boosted configuration over another.

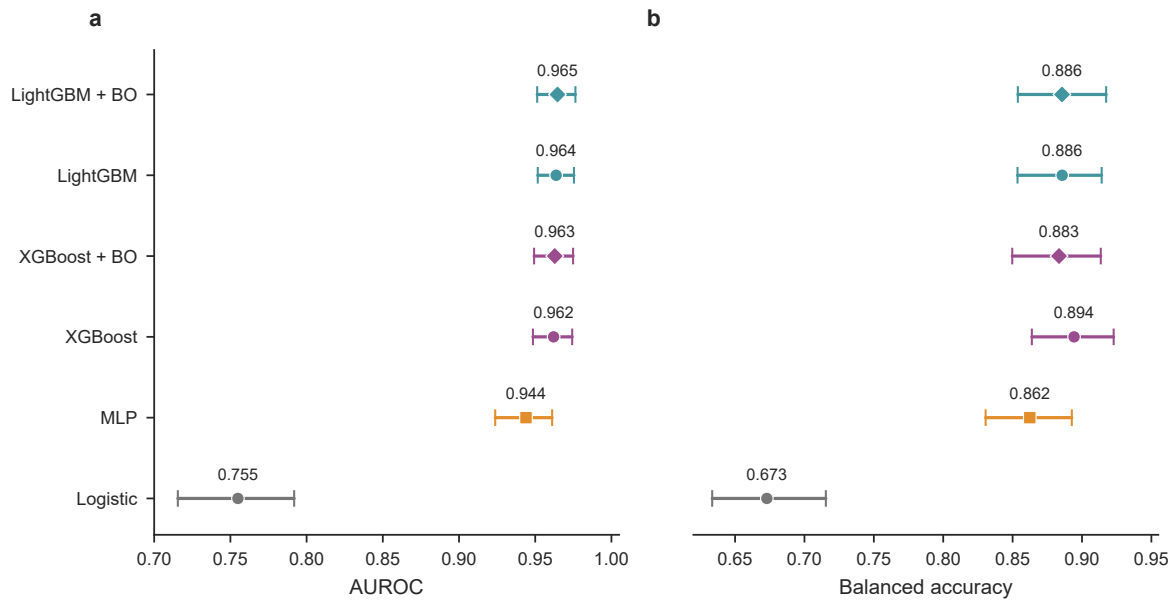


Figure 3. Held-out AUROC and balanced accuracy with bootstrap intervals.

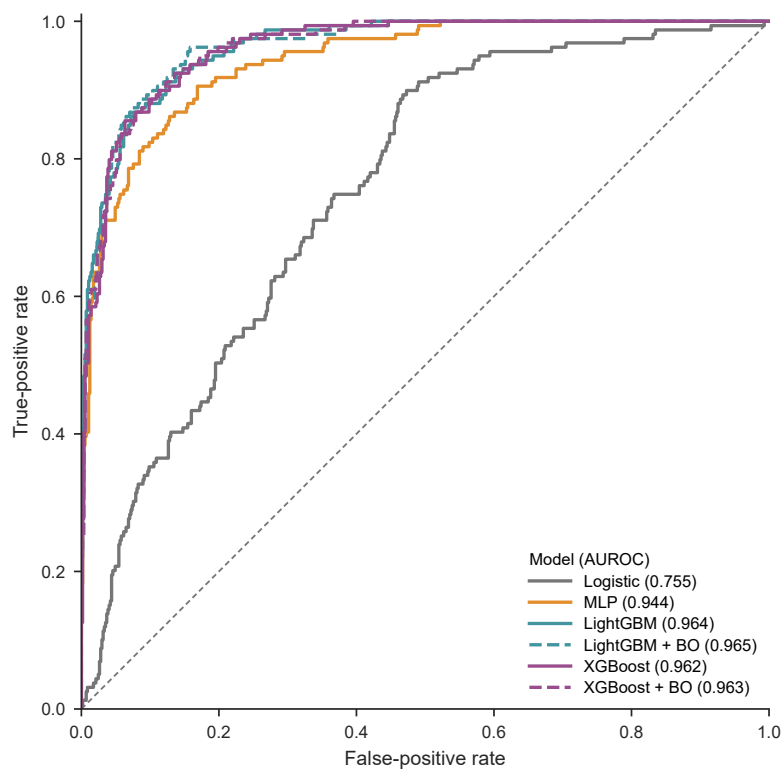


Figure 4. Held-out receiver operating characteristic curves.

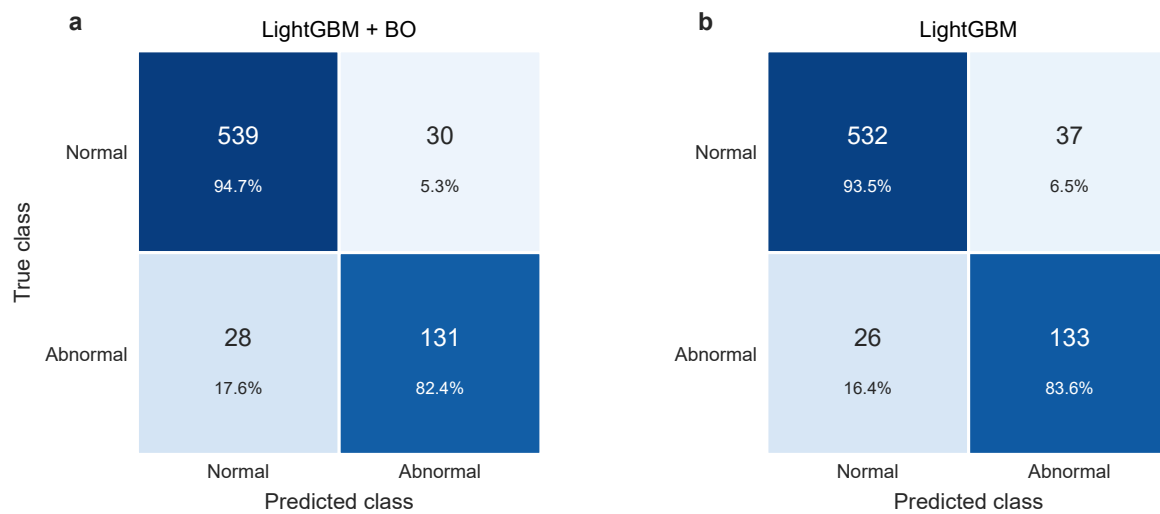


Figure 5. Confusion matrices of the two highest-AUROC models.

The selected Bayesian configurations are reported in Table 7 for reproducibility. Their inner grouped cross-validation AUROCs were also similar (0.945 for LightGBM and 0.946 for XGBoost). The selected regularization magnitudes for LightGBM were close to the lower search boundary, while the XGBoost search selected stronger L_2 regularization and full row sampling. These are empirical settings for the development corpus, not universal recommendations.

Table 7. Bayesian-optimized model configurations.

Model	Selected configuration	Inner CV AUROC
LightGBM + BO	leaves 46; depth 11; learning rate 0.0663; minimum child samples 47; feature fraction 0.982; $\lambda_1 = 0.000701$; $\lambda_2 = 0.000100$	0.945
XGBoost + BO	depth 8; learning rate 0.0205; minimum child weight 1; row fraction 1.0; column fraction 0.6; $\lambda_1 = 0.0001$; $\lambda_2 = 5.0$	0.946

4.3. Feature ablation and model-specific importance

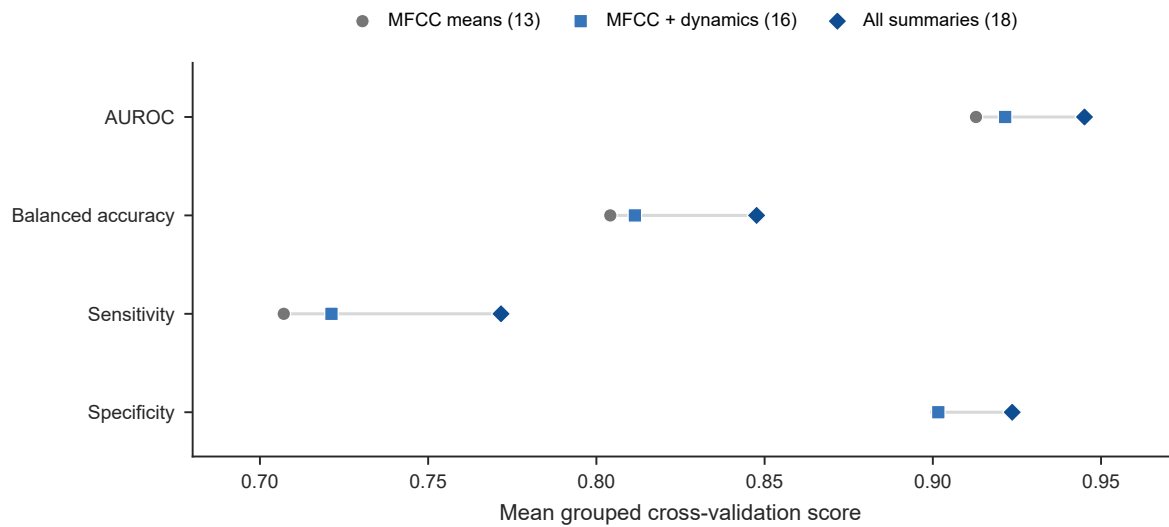
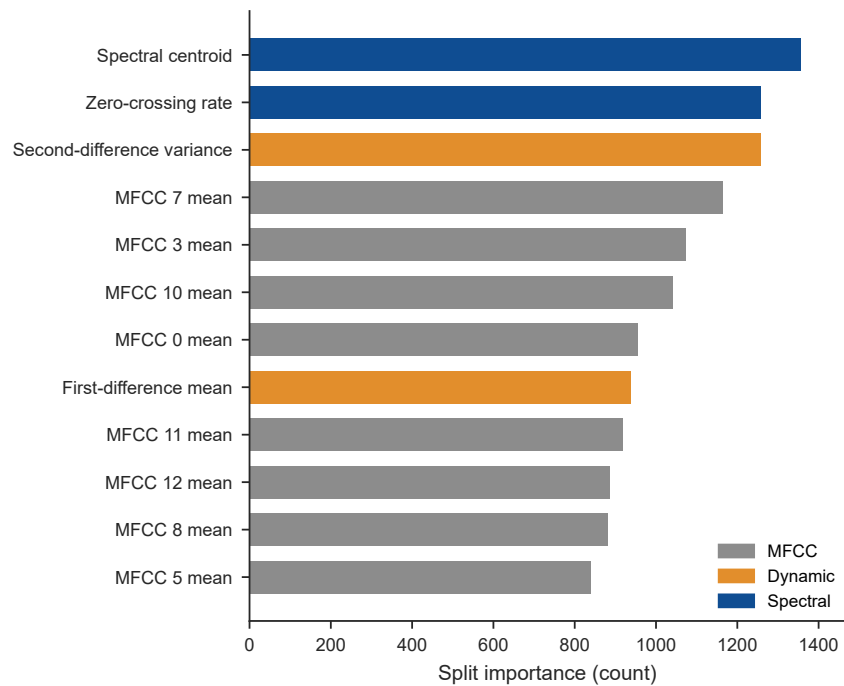
Table 8 summarizes the feature ablation. Adding the MFCC dispersion and difference summaries increased mean cross-validated AUROC from 0.913 to 0.921. Adding spectral centroid and zero-crossing rate produced a further increase to 0.945, together with balanced accuracy of 0.848 versus 0.804 for the 13 MFCC means alone. The 18-feature set improved sensitivity and specificity in these folds. Because the ablation was performed inside the model-development data, it supported retention of the full 18-feature vector but was not an independent estimate of its performance gain.

The staged gains were not confined to a single metric (Figure 6). Relative to the 13 MFCC means, the full vector increased mean sensitivity by 0.065 and specificity by 0.022. This pattern suggests that the additional summaries helped identify abnormal recordings and that the AUROC improvement was not solely due to shifting predictions toward the minority class. Fold-level variability and the absence of an outer ablation test nevertheless preclude a causal claim about any individual descriptor.

Figure 7 shows model-specific split-count importance for the highest-AUROC tree pipeline. It is provided to inspect the fitted model, not as evidence for any feature's causal physiological role.

Table 8. Feature ablation with grouped cross-validation.

Feature set	Accuracy	BACC	Sensitivity	Specificity	F1	AUROC	MCC
MFCC means (13)	0.859	0.804	0.707	0.901	0.686	0.913	0.596
MFCC means + dispersion /difference summaries (16)	0.862	0.811	0.721	0.902	0.695	0.921	0.608
All acoustic summaries (18)	0.890	0.848	0.772	0.924	0.755	0.945	0.685

**Figure 6.** Cross-validation scores for the nested feature sets.**Figure 7.** Feature importance of the highest-AUROC tree model.

5. Discussions

After duplicate removal and leakage-controlled evaluation, the MFCC-based feature set supported strong internal discrimination in the assembled dataset. Bayesian-optimized LightGBM achieved an AUROC of 0.965 and accuracy of 0.920, while untuned XGBoost achieved the highest balanced accuracy of 0.894. Feature ablation showed that adding MFCC dynamics, spectral centroid, and zero-crossing rate increased mean grouped cross-validated AUROC from 0.913 to 0.945. These estimates describe the fixed held-out partition of the same two-source corpus. Removing 266 repeated PASCAL/Bentley files before partitioning and applying SMOTE only within training folds kept duplicate recordings and synthetic observations outside the held-out evaluation.

The AUROC and threshold-dependent metrics provide complementary views of model behavior. AUROC evaluates ranking across thresholds, whereas balanced accuracy, sensitivity, and specificity describe the selected 0.5 operating point; this distinction explains the different leading models for AUROC and balanced accuracy. Average precision reached 0.899 against an abnormal-recording prevalence of 21.81%, while retaining the expected dependence on class prevalence. The 18-feature vector summarizes spectral-envelope shape, MFCC dynamics, brightness, and sign-change rate; MFCC filter-bank compression offers a practical PCG representation [10], and the nonlinear interactions in the table are well suited to LightGBM and XGBoost [11, 12]. Bayesian search improves configuration selection without exhaustive enumeration [13], although the bootstrap intervals quantify recording-level variation conditional on the fixed data construction and fitted pipelines.

The results are consistent with the broader development of PCG classification. PhysioNet/CinC challenge systems combined segmentation, engineered features, and learned representations, while other studies established whole-recording spectral, wavelet, and ensemble alternatives [1, 2, 4, 20–23]. Fuadah et al. present a close conventional-learning precedent by combining MFCCs with grid-search tuning, whereas Shin et al. addressed cross-database adaptation with temporal convolution and asynchronous channel fusion [33, 34]. Contemporary murmur datasets and augmentation methods extend the available tasks, but their data sources, endpoints, and evaluation protocols differ from the binary labels used here [30, 31]. The appropriate comparison is therefore methodological: this workflow documents source composition, duplicate removal, grouping, training-fitted preprocessing, separation of tuning from held-out scoring, and recording-level uncertainty. These choices apply established guidance on leakage and structured-data validation to the assembled PCG corpus [6–8].

Clinical translation requires alignment among intended use, target population, data-generating process, and prospective evaluation, together with transparent reporting of data flow, outcome definitions, model development, and uncertainty [43, 44]. This analysis is defined as normal–abnormal recording classification from public labels, with partial patient-level provenance and global acoustic summaries; clinical calibration, subgroup performance, workflow integration, and comparison with clinicians remain topics for a patient-level external study. A common benchmark could compare multidomain SVM features, denoising-autoencoder representations, compact one-dimensional CNNs, engineered-feature convolutions, and learnable filterbanks on the same deduplicated patient-grouped development set and untouched external corpora [45–49]. These estimates also reflect a finite Bayesian-search budget, a single random seed and a single held-out split, and recording-level bootstrap sampling; cross-validation uncertainty can remain substantial even

when point estimates appear precise [50]. Acoustic fingerprinting, calibration analysis, decision-curve evaluation, and prospective utility assessment would extend this baseline in future work.

6. Conclusions

In a deduplicated, group-held-out dataset of 3640 public heart-sound recordings, Bayesian-optimized LightGBM achieved an AUROC of 0.965 and an accuracy of 0.920, while untuned XGBoost achieved the highest balanced accuracy of 0.894. The complete 18-feature vector also outperformed the MFCC means alone in the training-only grouped ablation. These results demonstrate that the MFCC-based representation contains useful class-separating information for internal normal-abnormal classification within the audited pooled dataset.

The broader conclusion is methodological: a train–test boundary must be established before any operation that learns from the dataset or creates synthetic observations. Deduplication, patient/record grouping, training-only preprocessing, and transparent reporting are necessary to distinguish a useful acoustic model from an optimistic evaluation. The resulting analysis provides a reproducible internal baseline for future patient-level and external validation, but it does not support disease-specific or clinical diagnostic claims.

Ethics statement

This study performed secondary computational analysis of de-identified recordings released through public research challenges and involved no new participant recruitment, intervention, contact, or re-identification. The Methods state why no new ethics approval was sought. This statement concerns the present secondary analysis; the original datasets were obtained from the cited public providers under their posted access terms.

Data availability

The PhysioNet/CinC Challenge 2016 recordings are available from PhysioNet at <https://physionet.org/content/challenge-2016/1.0.0/>. The PASCAL Classifying Heart Sounds Challenge recordings are available from <https://www.peterjbentley.com/heartchallenge/>. The supplementary submission contains the deduplicated file index, derived feature table, fixed split manifest, and aggregate result files; raw audio is not redistributed.

Code availability

The supplementary submission contains the complete analysis script, dependency lock file, random seed, tuned parameters, and code used to generate all tables and figures.

Use of AI tools declaration

The authors used OpenAI Codex (GPT-5) to assist with material organization, code review and reproducibility auditing, language revision, reference checking, and LaTeX typesetting. The tool was

not used to assign clinical labels or modify heart-sound recordings. The authors reviewed the resulting material and remain responsible for the full content of this article.

Acknowledgments

This work was supported by the College Students' Innovative Entrepreneurial Training Plan Program under Grant No. 202410459130. The authors thank the PhysioNet/CinC and PASCAL challenge organizers and data contributors for making the heart-sound resources available.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Clifford GD, Liu C, Moody B, Springer D, Silva I, Li Q, et al. (2016) Classification of normal/abnormal heart sound recordings: The physioNet/computing in cardiology challenge 2016. *Comput Cardiol* 43: 609–612. <https://doi.org/10.22489/CinC.2016.179-154>
2. Liu C, Springer D, Li Q, Moody B, Abad Juan R, Chorro FJ, et al. (2016) An open access database for the evaluation of heart sound algorithms. *Physiol Meas* 37: 2181–2213. <https://doi.org/10.1088/0967-3334/37/12/2181>
3. Bentley P, Nordehn G, Coimbra M, Mannor S (2011) *The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results*. Available from: <https://www.peterjbentley.com/heartchallenge/>.
4. Langley P, Murray A (2017) Heart sound classification from unsegmented phonocardiograms. *Physiol Meas* 38: 1658–1670. <https://doi.org/10.1088/1361-6579/aa724c>
5. Koike T, Qian K, Schuller BW, Yamamoto Y (2021) Transferring cross-corpus knowledge: An investigation on data augmentation for heart sound classification. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*: 1976–1979. <https://doi.org/10.1109/EMBC46164.2021.9629714>
6. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Aroita G, et al. (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40: 913–929. <https://doi.org/10.1111/ecog.02881>
7. Kaufman S, Rosset S, Perlich C, Stitelman O (2012) Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans Knowl Discov Data* 6: 15. <https://doi.org/10.1145/2382577.2382579>
8. Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4: 100804. <https://doi.org/10.1016/j.patter.2023.100804>
9. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* 16: 321–357. <https://doi.org/10.1613/jair.953>

10. Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans Acoust Speech Signal Process* 28: 357–366. <https://doi.org/10.1109/TASSP.1980.1163420>
11. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. (2017) LightGBM: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 30: 3146–3154.
12. Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
13. Snoek J, Larochelle H, Adams RP (2012) Practical Bayesian optimization of machine learning algorithms. *Adv Neural Inf Process Syst* 25: 2951–2959.
14. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. (2000) PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* 101: e215–e220. <https://doi.org/10.1161/01.CIR.101.23.e215>
15. McFee B, Raffel C, Liang D, Ellis DPW, McVicar M, Battenberg E, et al. (2015) Librosa: Audio and music signal analysis in Python. *SciPy* 2015: 18–24. <https://doi.org/10.25080/Majora-7b98e3ed-003>
16. Lemaitre G, Nogueira F, Aridas CK (2017) Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res* 18: 1–5.
17. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12: 2825–2830.
18. Schmidt SE, Holst-Hansen C, Graff C, Toft E, Struijk JJ (2010) Segmentation of heart sound recordings by a duration-dependent hidden Markov model. *Physiol Meas* 31: 513–529. <https://doi.org/10.1088/0967-3334/31/4/004>
19. Springer DB, Tarassenko L, Clifford GD (2016) Logistic regression-HSMM-based heart sound segmentation. *IEEE Trans Biomed Eng* 63: 822–832. <https://doi.org/10.1109/TBME.2015.2475278>
20. Potes C, Parvaneh S, Rahman A, Conroy B (2016) Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. *Comput Cardiol* 43: 621–624. <https://doi.org/10.22489/CinC.2016.182-399>
21. Rubin J, Abreu R, Ganguli A, Nelaturi S, Matei I, Sricharan K (2016) Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients. *Comput Cardiol* 43: 813–816. <https://doi.org/10.22489/CinC.2016.236-175>
22. Tschannen M, Kramer T, Marti G, Heinzmann M, Wiatowski T (2016) Heart sound classification using deep structured features. *Comput Cardiol* 43: 565–568. <https://doi.org/10.22489/CinC.2016.162-186>
23. Zabihi M, Rad AB, Kiranyaz S, Gabbouj M, Katsaggelos AK (2016) Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. *Comput Cardiol* 43: 613–616. <https://doi.org/10.22489/CinC.2016.180-213>
24. Yaseen Y, Son GY, Kwon S (2018) Classification of heart sound signal using multiple features. *Appl Sci* 8: 2344. <https://doi.org/10.3390/app8122344>

25. Deng M, Meng T, Cao J, Wang S, Zhang J, Fan H (2020) Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. *Neural Networks* 130: 22–32. <https://doi.org/10.1016/j.neunet.2020.06.015>
26. Mei N, Wang H, Zhang Y, Liu F, Jiang X, Wei S (2021) Classification of heart sounds based on quality assessment and wavelet scattering transform. *Comput Biol Med* 137: 104814. <https://doi.org/10.1016/j.compbimed.2021.104814>
27. Liu A, Zhang S, Wang Z, Tang Y, Zhang X, Wang Y (2023) A learnable front-end based efficient channel attention network for heart sound classification. *Physiol Meas* 44: 095003. <https://doi.org/10.1088/1361-6579/acf3cf>
28. Krishnan PT, Balasubramanian P, Umapathy S (2020) Automated heart sound classification system from unsegmented phonocardiogram (PCG) using deep neural network. *Phys Eng Sci Med* 43: 505–515. <https://doi.org/10.1007/s13246-020-00851-w>
29. Ren Z, Chang Y, Nguyen TT, Tan Y, Qian K, Schuller BW (2024) A comprehensive survey on heart sound analysis in the deep learning era. *IEEE Comput Intell Mag* 19: 42–57. <https://doi.org/10.1109/MCI.2024.3401309>
30. Oliveira J, Renna F, Costa PD, Nogueira M, Oliveira C, Ferreira C, et al. (2022) The CirCor DigiScope dataset: From murmur detection to murmur classification. *IEEE J Biomed Health Inform* 26: 2524–2535. <https://doi.org/10.1109/JBHI.2021.3137048>
31. Reyna MA, Kiarashi Y, Elola A, Oliveira J, Renna F, Gu A, et al. (2023) Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022 *PLOS Digit Health* 2: e0000324. <https://doi.org/10.1371/journal.pdig.0000324>
32. Susic D, Gradisek A, Gams M (2024) PCGmix: A data-augmentation method for heart-sound classification. *IEEE J Biomed Health Inform* 28: 6874–6885. <https://doi.org/10.1109/JBHI.2024.3458430>
33. Fuadah YN, Pramudito MA, Lim KM (2023) An optimal approach for heart sound classification using grid search in hyperparameter optimization of machine learning. *Bioengineering* 10: 45. <https://doi.org/10.3390/bioengineering10010045>
34. Shin JM, Park S, Shin K, Seo WY, Kim HS, Kim DK, et al. (2025) Temporal convolutional neural network-based feature extraction and asynchronous channel information fusion method for heart abnormality detection in phonocardiograms. *Comput Methods Programs Biomed* 269: 108871. <https://doi.org/10.1016/j.cmpb.2025.108871>
35. Furui S (1986) Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans Acoust Speech Signal Process* 34: 52–59. <https://doi.org/10.1109/TASSP.1986.1164788>
36. Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Ann Stat* 29: 1189–1232. <https://doi.org/10.1214/aos/1013203451>
37. Cawley GC, Talbot NLC (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 11: 2079–2107.
38. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinf* 7: 91. <https://doi.org/10.1186/1471-2105-7-91>

39. Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognit Lett* 27: 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
40. Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 10: e0118432. <https://doi.org/10.1371/journal.pone.0118432>
41. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405: 442–451. [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9)
42. Efron B (1979) Bootstrap methods: Another look at the jackknife. *Ann Stat* 7: 1–26. <https://doi.org/10.1214/aos/1176344552>
43. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D (2019) Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 17: 195. <https://doi.org/10.1186/s12916-019-1426-2>
44. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. (2024) TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385: e078378. <https://doi.org/10.1136/bmj-2023-078378>
45. Tang H, Dai Z, Jiang Y, Li T, Liu C (2018) PCG classification using multidomain features and SVM classifier. *BioMed Res Int* 2018: 4205027. <https://doi.org/10.1155/2018/4205027>
46. Li F, Liu M, Zhao Y, Kong L, Dong L, Liu X, et al. (2019) Feature extraction and classification of heart sound using 1D convolutional neural networks. *EURASIP J Adv Signal Process* 2019: 59. <https://doi.org/10.1186/s13634-019-0651-3>
47. Xiao B, Xu Y, Bi X, Zhang J, Ma X (2020) Heart sounds classification using a novel 1-D convolutional neural network with extremely low parameter consumption. *Neurocomputing* 392: 153–159. <https://doi.org/10.1016/j.neucom.2018.09.101>
48. Li F, Tang H, Shang S, Mathiak K, Cong F (2020) Classification of heart sounds using convolutional neural network. *Appl Sci* 10: 3956. <https://doi.org/10.3390/app10113956>
49. Humayun AI, Ghaffarzadegan S, Ansari MI, Feng Z, Hasan T (2020) Towards domain invariant heart sound abnormality detection using learnable filterbanks. *IEEE J Biomed Health Inform* 24: 2189–2198. <https://doi.org/10.1109/JBHI.2020.2970252>
50. Varoquaux G (2018) Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* 180: 68–77. <https://doi.org/10.1016/j.neuroimage.2017.06.061>



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)