1. **ML problem:**
   Three class classification problem: the goal is to classify heartbeat audio recordings into one of three classes: healthy, artifact, murmur. A key focus is on **minimizing murmur misclassified as healthy**, as such errors could lead to delayed diagnosis or treatment of underlying cardiac conditions.
2. **ML pipeline and design decisions:**
   a. **Data preprocessing steps**
      - **Downsampling:** All .wav files were **downsampled** from 44100 Hz to 1000 Hz, which satisfies Nyquist criterion for common range of healthy and abnormal heart sounds (25-400 Hz) [3]. The original high sampling rate increases numerical instability during filtering and imposes unnecessary computational load.
      - **Filtering:** Downsampled signal was **bandpass filtered** in **25 – 400 Hz** with **4th order butterworth filter** to remove low-frequency drift and only extract relevant physiological signals. A **zero-phase (non-causal) filter** was applied via forward-backward filtering to avoid phase distortion and improve signal fidelity, which is acceptable here since true real-time operation is not required in this context.
      - **NaN handling:** Each signal is then checked for NaN values, if present, missing samples were filled using forward imputation. No NaNs were found in the provided data.
      - **Segmentation:** Each recording was then **segmented** to **3 seconds trials** with 50% overlap. This increases training size, which may help with generalization. 3 seconds was chosen to ensure that each trial contains at least a few cardiac cycles (average 0.8 s per cycle).
      - **Normalization:** Each segment was **z-score normalized** using its own mean and standard deviation. This corrects for inter-trial amplitude variations (e.g., from motion artefacts or variabilities in sensor placement).
      - **Class imbalance:** The dataset is imbalanced across classes (Table 1), which was considered in later stages of the pipeline design.

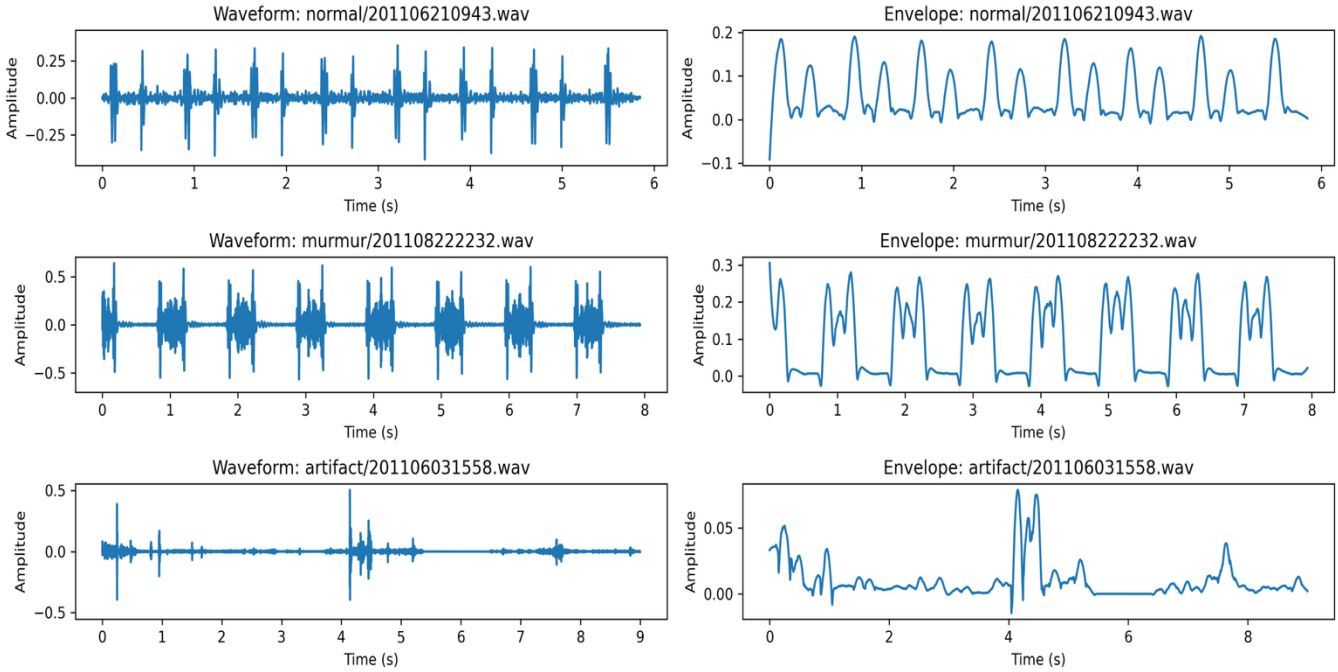
Table 1: Trial count and distribution across classes

Class	# of files	# of trials	Trial Distribution
Healthy	31	122	27.05%
Murmur	33	129	28.60%
Artifact	39	200	44.35%

   b. **Features extraction**
      - A combination of **time** and **frequency domain features** were extracted, informed from literature and physiological visualizations of recordings from each class [1-4].

- Since there is no time alignment across trials, extracted features were designed to be **time-invariant.** In addition, physiological visualizations cannot rely on grand averages.

Figure 1: Bandpass-filtered signals and smoothed envelopes for one recording per class.

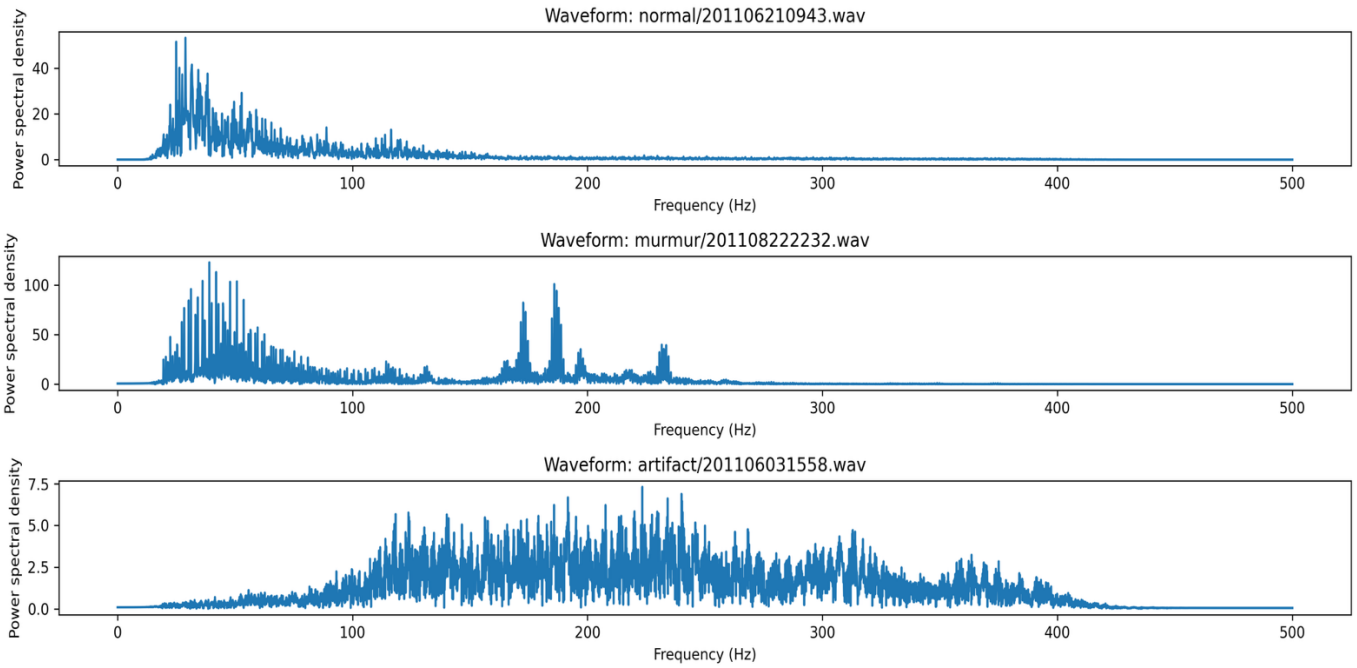


- **Time domain visualizations** of bandpass-filtered signals and their smoothed envelopes revealed class-specific patterns (Figure 1):
    1. Healthy/normal: clean, periodic impulse-like S1, S2 events, separated by low-amplitude regions.
    2. Murmur: Additional, turbulent low- to mid-amplitude noise between S1, S2 events.
    3. Artifact: missing S1, S2 events. Irregular, non-physiological spikes.
- These observations informed the extraction of these time-domain features (Table 2): **Root mean square (RMS) energy, peak-to-peak amplitudes, signal entropy, envelope mean, envelope variance, envelope dominant frequency, number of peaks in envelope** [1, 3, 4]**.**
- **Envelope** was used because it better captures cardiac events, whilst suppressing high-frequency oscillations and non-physiological events [3].
- Envelope was extracted using **Hilbert transform** and smoothed with the **Savitzky-Golay filter** to reduce noise and better emphasize cardiac events. Dominant frequencies were extracted with **fast fourier transform**, and signal entropy was computed based on the histogram distribution of signal amplitudes.

Table 2: Class-wise differences in time-domain features

	Normal	Murmur	Artifact
Envelope mean	Lower than murmur	Higher than normal	Highly variable
Envelope variance	Higher than murmur	Lower than normal	Highly variable
Envelope dominant frequency	Physiologically valid, approx.1-2Hz	Slightly higher and less stable than normal	Highly variable
Number of envelope peaks	Lower than murmur	Higher than normal	Highly variable
RMS energy	Less energy than murmur	More energy than normal	Highly variable
Signal entropy	Less than murmur	More than normal	Highly variable
Peak-to-peak amplitude	Higher than murmur	Less than normal	Highly variable, can have extreme swings

Figure 2: PSD for one recording per class



- **Frequency domain (PSD) visualizations** also showed class specific differences (Figure 2):
    1. Healthy/normal: Narrowband energy with a clear peak between 20-100 Hz

2. Murmur: There still exists a dominant frequency between 20-100 Hz, but elevated high frequency energy 150-200 Hz compared to normal.
3. Artifact: No dominant frequency < 100 Hz. Flat PSD from 100 to 300 Hz.

- These observations informed the extraction of these frequency-domain features (Figure 3)**: Bandpower ratio (25–150 Hz vs 150–400 Hz), spectral bandwidth, spectral entropy, spectral dominant frequency** [1, 3, 4]**.**
- Frequency domain features were computed from PSD estimated using the the **Welch** method.

Figure 3 Class-wise differences in frequency domain features

	Normal	Murmur	Artifact
Bandpower ratio (25–150 Hz vs 150–400 Hz)	Higher than murmur	Lower than normal	Low
Spectral bandwidth	Narrow	moderate	wide
Spectral entropy	Low	moderate	high
Spectral dominant frequency	Lower than murmur	Higher than normal	Highly variable

c. Features selection

- Feature selection not done here because feature set kept at low dimension (11 features), less than 10% of training set size to avoid **overfitting**.

d. Data splitting

- **Train-test set splitting at the file level** was done to prevent **data leakage** and robustly evaluate model's generalization to new patients or recordings. Hence, trials from the same file do not appear in both training and test sets. **This reflects real-world deployment.**
- **Stratification** splitting was done to ensure class representations in each set matches that of original data distribution.
- **Outlier trials removal (training data only):** Trials with raw signal amplitudes above the 99.5th percentile of the training set distribution were excluded as noisy outliers. This step eliminated 2 trials from the training set.
- **80-20 random split, 357 trials in the training set, 92 trial in test set.**

e. Model building

- **Two-stage hierarchical classification strategy:**
    - **Stage 1**: 3-class classifier (Healthy vs Murmur vs Artifact) to detect artifact class.
    - **Stage 2**: Binary classifier (Healthy vs Murmur) for fine-grained classification.
    - Hence**, if a trial is not identified as Artifact in stage 1 classifier, it will then be passed to the stage 2 classifier to be classified as either Healthy or Murmur.**

    - The stage 1, three-class model consistently identifies **artifact trials with high true positive rate and minimal misclassification**. However, performance in distinguishing **healthy vs murmur** is comparatively lower in this stage. Therefore, a second-stage **binary classifier**, trained only on artifact-free trials, allows for **more focused and fine-tuned discrimination** between healthy and murmur classes.
    - **The 2 classifiers followed the same design decisions.**
- Feature were **normalized** using **z-scoring,** with normalization parameters only extracted from the training folds.
- **Classifiers:** SVM or XGBoost
    - SVM was picked because it allows for nonlinear decision boundaries (helpful for healthy vs murmur) and effective in small datasets like the one provided. It has also been used in prior, relevant literature [2, 3].
    - XGBoost was picked because of its ability to capture interactions between features and robustness to overfitting.
- **Hyperparameter tuning** using Optuna (Bayesian optimization) with **stratified cross validation** at the **file level** on the train set.
    - **Stratified cross validation (CV)** was done to respect class distribution during training and validation.
    - **File level splitting** during CV prevents data leakage by ensuring that trials from the same file do not appear in both folds.
    - **Class weights** tuned to address class imbalance.
    - **Macro F1 score** was used as the optimization objective to ensure balanced performance across all classes, and to penalize **false positives and false negatives (**i.e. misclassifying healthy as murmur, false alarm**)**

f. Re-training and Evaluation

- Model was then **re-trained** on the **full train set** with the optimal hyperparameters derived from Optuna, and evaluated on the **test set.**
- Performance metrics:
    - Macro F1 score: Balanced metric across all classes.
    - Per-class accuracy, overall accuracy
    - Confusion matrix
    - Precision and Recall per class: evaluate number of false positives (i.e. detect murmur when normal) and false negatives (i.e. missing a murmur case)

Stage 1 classifier (Ternary, healthy vs murmur vs artifact) performance on test set:

	Accuracy (%)	Precision (%)	Recall (%)
Healthy	62.50	75.00	62.50
Murmur	85.71	72.72	85.71
Artifact	97.5	100.00	97.50

Macro F1 score: 81.87

Overall accuracy: 84.78%

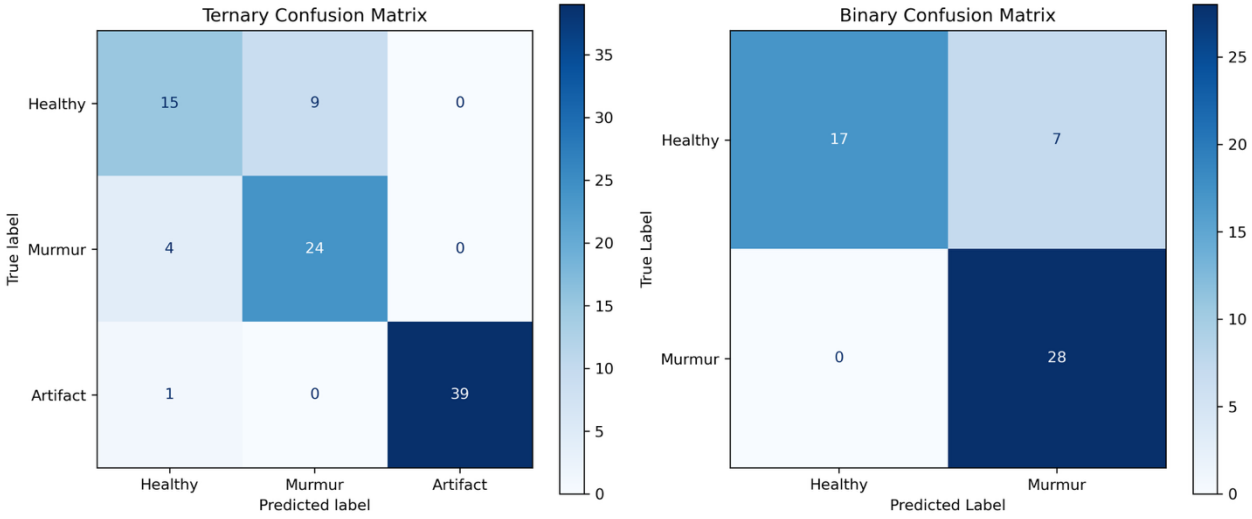
Stage 2 classifier (Binary, healthy vs murmur) performance on test set:

	Accuracy (%)	Precision (%)	Recall (%)
Healthy	70.83	100.00	70.83
Murmur	100.00	80.00	100.00

Macro F1 score: 85.91

Overall accuracy: 86.54%

Figure 4: Test Set Confusion Matrices for Stage 1 and 2 Classifiers



- **Observations:**
  - According to results above and Figure 4, the pipeline achieved **strong, balanced classification performance**.
    - Stage 1 (three-class classifier): Detected artifact trials with **very high precision and recall (>97%).** Healthy vs murmur classification was more challenging.
    - Stage 2 (binary classifier): **improved** on the **healthy vs murmur** classification relative to stage 1. Achieved **precision**

and recall above 80% for both classes. However, recall for healthy class needing improvement (70.83%) to avoid false alarms.

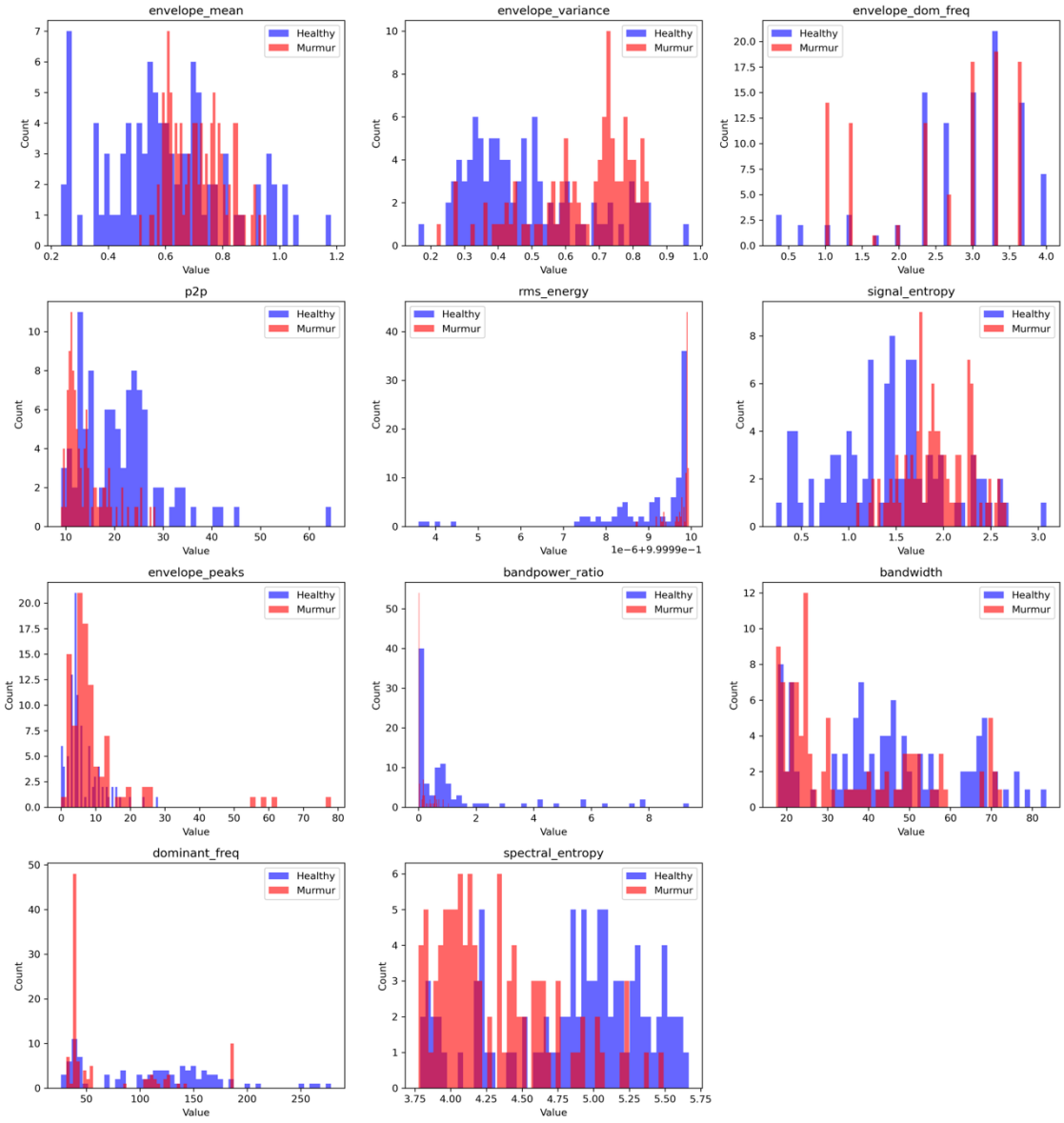
- Note that the pipeline performs very with segment level data splitting, surpassing 95% classification accuracy in all three classes.

3. **Challenges encountered and how you addressed them**
   - **Challenge 1:** Noisy input signals
     - **Solution:** Robust preprocessing pipeline included bandpass filtering + z-score normalization + removal of outlier trials
   - **Challenge 2:** class imbalance can lead to biased models
     - **Solution:** Stratified train-test splitting + stratified cross validation during hyperparamter tuning + tuned class weights + macro F1 score used as Optuna objective. **Balanced, high classification performances were achieved** as a result.
   - **Challenge 3:** Misclassifications between healthy and murmur
     - **Solution 1:** Used a cascaded classification approach, with a secondary binary, healthy vs murmur classifier trained and fine-tuned only on the two classes. **This improved precision and recall in both healthy and murmur classes.**
     - **Solution 2:** Used **per-feature histograms** and **Cohen's *d*** effect sizes to **assess class separability** on the training set **and guide feature refinement** (Table 3, Figure 5). RMS energy was found to be highly discriminative, which motivated the inclusion of peak-to-peak amplitude. In the frequency domain, the discriminative power of the bandpower ratio (low vs high frequency energy) led to the addition of spectral entropy. **These newly added features were equally discriminative and enhanced classification performance.**

Table 3: Cohen's d Effect Sizes for Feature Separability (Healthy vs Murmur) on Training Set

Feature	**Cohen's d for class separability**
Envelope mean	0.70
Envelope variance	1.06
Envelope dominant frequency	0.29
Peak-to-peak	1.06
RMS	0.84
Signal entropy	0.99
Number of peaks in envelope	0.41
Bandpower ratio	0.71
Bandwidth	0.54
Dominant frequency	0.77
Spectral entropy	1.21

Figure 5: Feature Histograms (Healthy vs Murmur) on Training Set

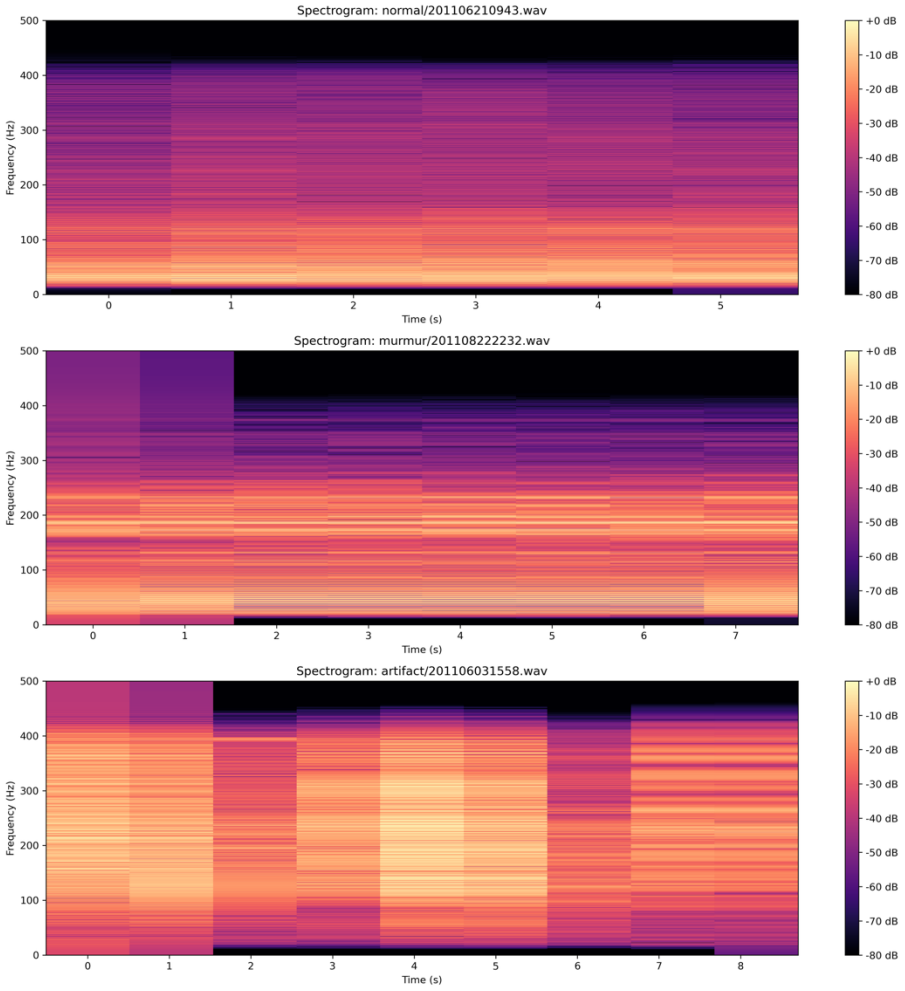


4. Potential improvements and future work

- **Robust evaluation of testing performance:** Model performance may vary across different train-test splits, especially with limited data. While stratified cross-validation was used, further evaluation across multiple random seeds would provide a more robust performance estimate. This was not performed due to time constraints.
- **Deep learning exploration:** future work could convert segments to spectrograms and train CNNs to learn time-frequency dynamics [3]. Current trial count may be limited for deep learning, but initial spectrogram observations show class-specific differences (Figure 6):
    - Normal: Energy concentrated in lower frequency (< 150 Hz)
    - Murmur: Energy diffused into higher frequency components (150-300 Hz)

    - Artifact: Energy diffused in broad frequency bands, up to 400 Hz.
- **Feature engineering enhancements:** it seems envelope variance, peak-to-peak, signal entropy, spectral entropy are discriminant features. These can motivate **development of related features**, such as inter-peak intervals from the envelope.
- **Error analysis**: analyzing trials that were misclassified can inform feature and model improvements.

Figure 6: Spectrogram of a recording from each class



References

[1] Dornbush, S., & Turnquest, A. E. (2019). Physiology, heart sounds.

[2] Yaseen, Son, G. Y., & Kwon, S. (2018). Classification of heart sound signal using multiple features. *Applied Sciences*, *8*(12), 2344.

[3] Dwivedi, A. K., Imtiaz, S. A., & Rodriguez-Villegas, E. (2018). Algorithms for automatic analysis and classification of heart sounds–a systematic review. *IEEE Access*, *7*, 8316-8345.

[4] Singh, M., & Cheema, A. (2013). Heart sounds classification using feature extraction of phonocardiography signal. *International Journal of Computer Applications*, *77*(4).