Group #4 Final Report: Classification of Heart Sounds

INFO 5375 - Machine Learning for Health

Jacob Ioffe (ji97)
Gerhard Nordemann (gn233)
William J. Reid (wjr83)

Cornell Tech — May 11, 2024

1 Introduction

Cardiovascular diseases are a leading cause of mortality worldwide, necessitating timely and accurate diagnostic methods [1, 2, 3]. Cardiac auscultation, despite being a cost-effective tool, requires extensive training and expertise. Computer-aided decision systems based on auscultation could help address this challenge by assisting healthcare professionals in diagnosing cardiac conditions. However, the lack of large, publicly available datasets with detailed annotations has hindered the development of effective recommendation systems in clinical trials. Cardiac auscultation involves listening to heart sounds using a stethoscope, with specific sounds indicating normal or abnormal cardiac function. Murmurs, clicks, and snaps are additional sounds associated with turbulent blood flow, often indicating underlying cardiac conditions [4]. Understanding the characteristics of these sounds is crucial for accurate diagnosis and referral of patients.

Accurate detection and classification of heart murmurs in pediatric patients are critical for early diagnosis and appropriate clinical intervention. Traditional methods often rely on manual feature extraction and heuristic algorithms, which may not effectively capture the complexity of heart sound patterns. In contrast, machine learning, particularly deep learning, offers a promising approach for automating this analysis. Leveraging newly available datasets such as the CirCor DigiScope Phonocardiogram dataset [5], machine learning models can automatically learn discriminative features from raw data, enhancing diagnostic accuracy and clinical utility. These models can be trained to distinguish between normal and abnormal heart sounds, classify different types of murmurs, and assist clinicians in making more accurate and timely diagnoses. Beyond comparative effectiveness, this machine learning investigation lays the groundwork for future advancements in pediatric heart sound analysis, including medical training and the development of stethoscopes with sound classification capabilities.

The major contribution of this work is twofold. First, we demonstrate the importance and effect of preprocessing on eventual model performance. By exploring various sound preprocessing techniques and medical literature, we showcase performance improvements on a model-by-model basis. In doing so, we show that, especially with this small dataset, employing the right preprocessing techniques such as the identification and segmentation of the S1 and S2 heart sounds, noise reduction and MFCC conversion enables simpler models to perform as effectively as much more complex machine learning models like neural networks, which are commonly used in existing research. Furthermore, the use of simpler models like Random Forest not only matches the performance of complex neural networks but also offers significant advantages in terms of scalability and ease of deployment. These attributes make simpler models highly suitable for implementation in diverse healthcare settings, including hospitals with limited computational resources, and for integration into portable devices such as wearables, enhancing their practical utility in real-time health monitoring and diagnostics.

Second, we introduce a potential workflow for community healthcare workers to aid in accessing the murmurs of children in remote areas using a stethoscope. Though not perfect, we hope the introduced interface represents a first step in assisting healthcare workers to more reliably screen and triage the care of children with heart murmurs. We achieve this by providing community healthcare workers with three insights: we output a processed audio file containing only the segments of the audio signal where murmurs can occur, we provide a score indicating the likelihood that the sound contains a murmur, and if so, we offer the clinical details that characterize the potential murmur detected.

2 Related Work

Following the PRISMA guidelines, we conducted a systematic literature review focusing on the use of machine learning in heart sound classification. Our search spanned several major databases, including PubMed, Google Scholar, and Physionet challenges using key terms such as "heart sound classification", "PCG analysis", "machine learning for cardiac auscultation", and "heart murmur detection". From an initial pool of approximately 250 articles, careful screening based on titles and abstracts reduced the pool to 80 articles. This initial reduction was primarily guided by the relevance of the studies to machine learning applications in heart sound classification, methodological rigor, and the date of publication, prioritizing works that were no older than 10 years unless they were seminal articles. We also focused on filtering out non-research articles, such as editorials and reviews, to ensure a focus on empirical research findings.

The second stage of reduction to 30 critically assessed articles involved a deeper evaluation of each study's contribution to the field, focusing on articles that introduced novel methodologies or significant advancements in diagnostic accuracy. We also prioritized studies with clear clinical applications, those that provided comprehensive datasets, or demonstrated potential enhancements in clinical practice. Further considerations included the quality and impact of the research, such as citation count and the journal's impact factor, to ensure inclusion of influential and high-quality studies that offer significant insights and implications for further research and practical applications.

Recent advancements in heart sound classification have demonstrated significant progress, primarily driven by contributions such as those from McDonald et al.[6], who developed a model combining recurrent neural networks with hidden semi-Markov models for effective murmur detection. Similarly, Lu et al.[7] proposed a lightweight CNN combined with a random forest model, focusing on computational efficiency for real-time processing in clinical settings. Additionally, Chang et al.[8] introduced a multi-task learning framework that leverages both time-domain and frequency-domain features to predict the presence of murmurs and clinical outcomes, illustrating the potential of integrated learning systems to enhance diagnostic processes.

Inspired by the Busono et al.[9] study, we recognized the importance of medically informed data preprocessing and segmentation. Busono's work, which highlighted the challenges of digital auscultation and the necessity for precise feature extraction to classify heart sounds accurately, led us to refine our approach to audio data preprocessing. By adopting their insights on how to effectively isolate heart sound components and emphasize the regions most likely to contain murmurs, we enhanced our segmentation process to ensure that our models learn from the most clinically relevant and high-quality data sections. This has significantly increased the detection accuracy of our models, particularly in distinguishing between 'Murmur Audible' and 'Murmur Not Audible' categories.

Building on these insights, our research addresses several critical gaps. Inspired by the robust feature extraction and model interpretability in McDonald et al.'s work, we have developed streamlined computational models that adapt more readily to diverse clinical settings without sacrificing accuracy. From Lu et al., we derived the value of creating lightweight models, leading us to innovate efficient preprocessing techniques to ensure high-quality data inputs from varied recording conditions, addressing issues of recording quality variance and inconsistent labels.

By integrating these advanced techniques, our study builds a model that significantly improves upon the accuracy and utility of existing diagnostic tools for cardiac auscultation, providing more reliable, accurate, and clinically useful tools that can ultimately lead to better patient outcomes.

3 Methods

3.1 Our Approach

This study adopts a comparative approach to evaluate the performance of traditional machine learning models and deep learning architectures in the classification of heart sounds from phonocardiogram (PCG) recordings. Initially, we plan to apply classical signal processing techniques, including Fourier analysis, wavelet transform, and Mel-frequency cepstral coefficients (MFCCs), to extract salient features that capture the spectral and temporal characteristics of PCG signals. These features will then be utilized as inputs for various machine learning classifiers, such as Support Vector Machines (SVMs), Random Forests, K-Nearest Neighbors (KNN), Decision Trees, Logistic Regression, and Gaussian Naive Bayes. In parallel, we will explore deep learning models, specifically Convolutional Neural Networks (CNNs), to identify spatial and temporal patterns within the heart sounds.

3.2 Dataset Introduction

The dataset used is described in the paper *The CirCor DigiScope Dataset: From Murmur Detection to Murmur Classification* and made freely available on PhysioNet [5]. It comprises 5230 unique heart sound recordings, derived from the four primary auscultation sites (PV, AV, TV, MV) of 1568 subjects aged 0 to 21 years (average age 6.1 ± 4.3 years), with durations ranging from 4.8 to 80.4 seconds (average duration 22.9 ± 7.4 seconds), summing up to over 33.5 hours of audio data. This dataset includes three main target variables: presence of a murmur [*Absent*, *Present*, *Unknown*], whether or not a murmur is audible at a specific recording location [*AV*, *MV*, *PV*, *TV*], and the outcome after expert consultation [*Normal*, *Abnormal*].

The data is organized into four primary types of files for each subject:

- **Wave Recording Files (.wav):** Binary files containing heart sound data from various auscultation locations.
- **Header Files (.hea):** Text files providing metadata about the .wav files in standard WFDB format.
- **Segmentation Data Files (.tsv):** Text files with segmentation information marking the start and end points of the fundamental heart sounds S1 and S2, for each auscultation location.
- **Subject Description Files (.txt):** Text files offering detailed demographic data (e.g., weight, height, sex, age group, pregnancy status) and descriptions of murmur events for each subject.

The filename is systematic, reflecting the subject ID and auscultation location, followed by an integer index for multiple recordings at the same location (e.g., `ABCDE_XY.wav` where `ABCDE` is the subject identifier and `XY` corresponds to specific auscultation points such as PV, TV, AV, MV, Phc).

Murmurs are meticulously classified based on timing, shape, pitch, quality, and grade, providing a granular view of heart sound anomalies. These classifications, along with the precise segmentation of heart sounds into systolic and diastolic periods as delineated in the .tsv files, aid in the detailed analysis of these audio recordings. Each .tsv file tags key periods in the sound files that are significant for clinical analysis, enhancing the dataset's utility for automated and detailed heart sound classification.

Figure 1 presents a summary of murmur occurrences among the patients whose sound recordings were available, illustrating the significant skew toward the absence of murmurs within the dataset.

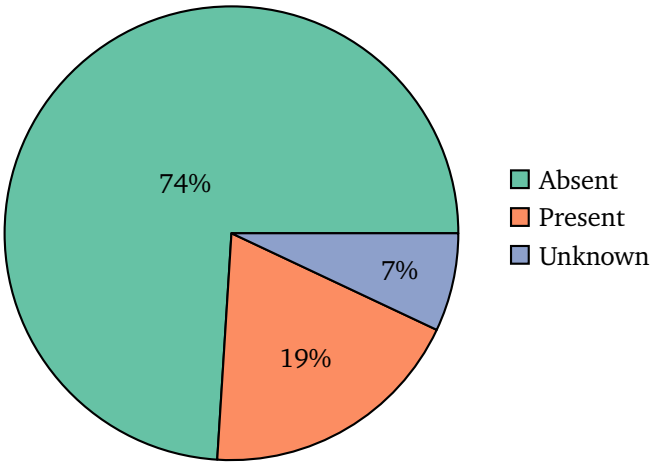

Figure 1: Distribution of Murmur Categories Among Patients in the Training Dataset. The total number of patients in the Training Dataset is 942

Upon further examination, we observe that a significant portion of cases initially identified as absent of murmur ultimately yielded abnormal conclusions (37.8% within the absent group), shown in Table 1. This stark discrepancy influenced our decision to define the target variable for this study more definitively, focusing on whether a murmur was audibly detected by an expert at specific recording locations [AV, MV, PV, TV].

Moreover, when examining the sound files in ".wav" format within the database, we noticed that not all patients have recordings from all different locations. As shown in Table 2, variations in the number of recording types are evident. In Table 3 we indicated the number of recordings in the training, validation, and test sets, after removing duplicates.

Murmur	Normal	Abnormal
Absent	432	263
Present	29	150
Unknown	25	43

Table 1: Presence of Murmurs per Clinical Expert Diagnosis (Ground Truth of Outcome)

Category	Number of Recordings	Percentage
PV	766	24.23%
TV	732	23.14%
AV	800	25.30%
MV	861	27.26%
Phc	4	0.13%
Total	3163	100.00%

Table 2: Distribution of Auscultation Locations in the Training Data Heart Sound Recordings

Dataset	Number of Unique Recordings	Percentage of Total
Training	3141	60%
Validation	480	10%
Test Set	1609	30%
Total	5230	100%

Table 3: Distribution of Unique Recordings Across the Training, Validation, and Test Sets

Figure 2 illustrates the distribution of the murmur audible locations per recording location for the training, validation, and test sets. For the training dataset, we investigated the locations where murmurs were detected, and the findings are presented in Figure 3. As anticipated, murmurs can manifest in multiple locations simultaneously. Notably, the most audible locations include PV, TV, and MV.

Upon further stratifying the data into distinct age groups, it becomes apparent that a majority of patients fall into the "Child" category. Additionally, it is noteworthy that over one-third of the cases where murmurs were initially *absent* were subsequently labeled as *abnormal*. This observation is illustrated in Table 4. Additional analysis revealed that the predominant type of murmur observed is Systolic Murmur,

Age Group	Outcome	Murmur Absent	Murmur Present	Murmur Unknown	Total
Unknown	Normal	56	3	2	61
	Abnormal	11	2	0	13
Neonate	Normal	3	0	0	3
	Abnormal	1	1	1	3
Infant	Normal	37	4	6	47
	Abnormal	39	21	19	79
Child	Normal	305	22	16	343
	Abnormal	190	110	21	321
Adolescent	Normal	31	0	1	32
	Abnormal	22	16	2	40
Total		695	179	68	942

Table 4: Training Data: Distribution by Age Group, Outcome, and Presence of Murmur

as depicted in Table 5. This finding is consistent with established medical knowledge, as systolic murmurs are typically more prevalent in both pediatric and adult populations [10]. We processed the header files

Murmur Type	Number of Recordings
Systolic	178
Diastolic	5

Table 5: Training Data: Summary of Systolic and Diastolic Murmur Recordings

to extract metadata from each audio file. This metadata includes information about patient demographics and murmur characteristics. All metadata is provided in Table 6. In addition to the metadata, the dataset owners utilized an advanced model to tag specific periods within the audio recordings of heart sounds. These tags are stored in a *.tsv* file format for every audio file is illustrated in Table 7. The *.tsv* files delineate

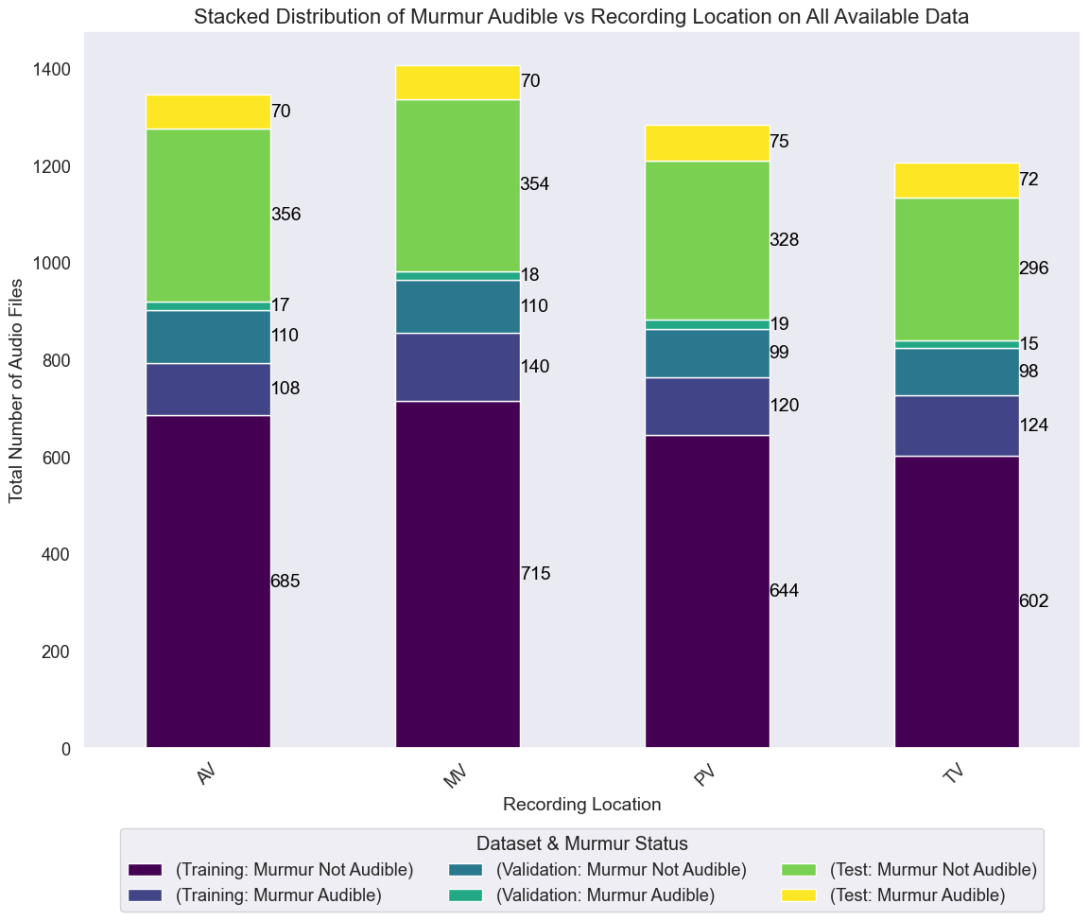

Figure 2: Distribution of murmur audible per recording location on the training, validation, and test data

Feature Category	**Features**
Patient Info	Patient ID, Campaign, Additional ID
Recording Info	Recording Locations, Frequency (Hz)
Patient Demographics	Age, Sex, Height, Weight, Pregnancy Status
Murmur Characteristics	Murmur (Present/Absent), Murmur Audible, Most Audible Location
Systolic Murmur Details	Systolic Murmur Timing, Systolic Murmur Shape, Systolic Murmur Grading, Systolic Murmur Pitch, Systolic Murmur Quality
Diastolic Murmur Details	Diastolic Murmur Timing, Diastolic MurmurShape, Diastolic Murmur Grading, Diastolic Murmur Pitch, Diastolic Murmur Quality
Expert Diagnosis	Outcome

Table 6: Metadata Extracted and Compiled from Audio Files

distinct segments in each recording as follows:

- **Period 0:** Noise — ambient or non-cardiac sounds.
- **Period 1:** The S1 Region — corresponds to the first heart sound.
- **Period 2:** The Systolic Region — crucial for identifying potential systolic murmurs.
- **Period 3:** The S2 Region — corresponds to the second heart sound.
- **Period 4:** The Diastolic Region — another key period for detecting potential diastolic murmurs.

3.3 Experimentation Plan

In this initial part of our research, we aim to classify murmurs based on the information provided, including audio recordings and metadata. As observed in Table 5, diastolic murmurs are scarcely present. Consequently, we have decided to focus this research solely on systolic murmurs. Figure 9 illustrates how we set up our training, validation, and testing phases. Below, we will discuss each step in greater detail.

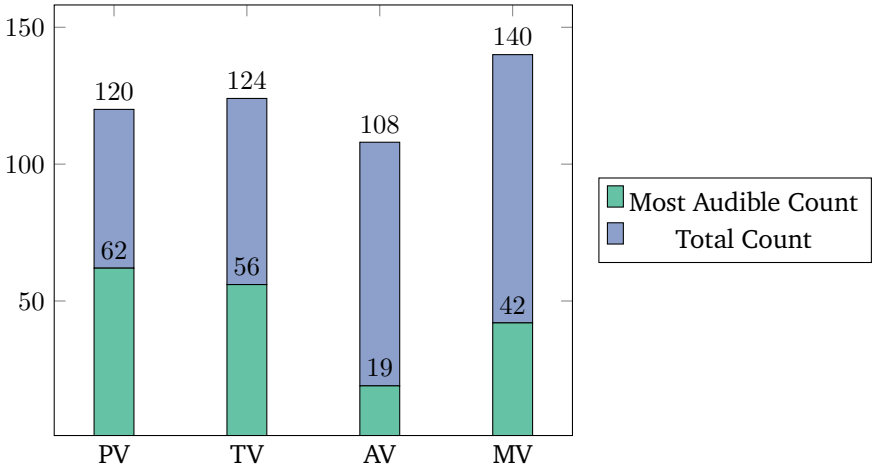

Figure 3: Summary of Murmur Locations and Most Audible Locations in the Training Dataset

Begin (s)	End (s)	Period
0.0000	0.6780	0
0.6780	0.8202	4
0.8202	0.9602	1
0.9602	1.0802	2
1.0802	1.1802	3
1.1802	1.4402	4
...	...	...

Table 7: Sample Data from TSV File

In the development of our methodology for classifying heart murmurs using phonocardiograms, several advanced techniques were explored but ultimately not adopted due to various challenges in achieving superior results. Below, we provide a detailed account of these explorations, the rationale for their initial consideration, and the reasons for their eventual inclusion or exclusion from the final methodology that was implemented.

Inputs/Outputs
In the training and selection phase of this study, several machine learning models were carefully tuned using both the training and validation datasets, while the test dataset was preserved for final evaluation. Initially, the training set comprised 60% of the data, and the validation set held 10%.

1. Splitting
The segments between the S1 and S2 heart sounds, where murmurs are predominantly present, were isolated for detailed analysis. Through empirical testing, different audio segment lengths were tested, including a segment length of 392 milliseconds which captured the critical period between the S1 and S2 heart sounds (e.g., where murmurs, if present, will occur) for all patients. To avoid inadvertently excluding any murmurs or merging them with other noise in this period, a small amount of padding was added around these segments. The audio files were processed to retain only segments from Period 2, where systolic murmurs are most likely to occur. These segments were also padded to standardize the length across all samples, ensuring consistency in analysis and modeling. An example of a final preprocessed audio file can be seen in Figure 4.

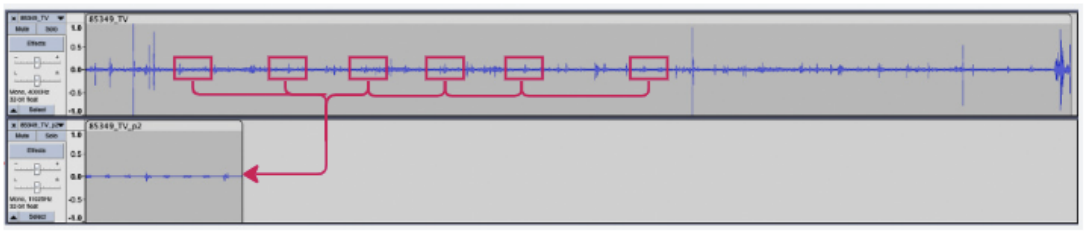

Figure 4: Segmentation is done based on the pre-tagged period 2 locations in the .tsv files

2. Noise Reduction
In clinical diagnostics, four main heart sounds — S1, S2, S3, and S4 — were distinguished through this methodology. S1, registering between 10 and 200 Hz, and S2, between 20 and 250 Hz, were essential markers for detecting heart murmurs. S3 and S4, present in lower frequencies, indicated potential heart failure or abnormal rhythms. The noise reduction preprocessing involved applying a low pass filter with a threshold of 500Hz, which helped attenuate high-frequency noise components that are typically not relevant to the heart sound analysis. Additionally, sounds below 20 decibels were set to zero. These thresholds were determined empirically to concentrate the analysis on the most significant audio components of the heart sounds. This approach not only reduced the processing time but also enhanced the accuracy of the analysis in subsequent stages.

3. Train Test Validation Split
In the training and selection phase of this study, several machine learning models were carefully tuned using both the training and validation datasets, while the test dataset was preserved for final evaluation. Initially, the training set comprised 60% of the data, and the validation set held 10%. To improve hyper-parameter optimization, iterative cross-validation was applied using the validation set alone. This is displayed in Figure 9.

4. Feature Extraction
Initially, our work incorporated feature engineering techniques like Mel-spectrograms and wavelet features. Mel-spectrograms were pivotal for visualizing the spectral patterns over time, capturing the textural nuances of heart sounds which are essential for identifying murmurs. Wavelet features were employed to grasp both frequency and location information, which is crucial for the analysis of non-stationary signals like heart sounds. However, despite their theoretical advantages, these features alone did not sufficiently enhance the model's performance in preliminary tests.

In parallel, we delved into more advanced signal analysis techniques such as auto-correlation to detect repeating patterns and periodic behavior within the audio data, indicative of regular heart rhythms and potential murmurs. Auto-correlation, supplemented with band-pass filtering to minimize noise interference, aimed to isolate significant patterns and identify periodic events like heartbeats and murmurs. This approach, while sound in theory, faced practical challenges in effectively isolating the nuanced sounds of heart murmurs from other cardiac noises in the highly variable clinical data.

Therefore, in terms of the feature extraction phase of our project, we settled on using only Mel Frequency Cepstral Coefficients (MFCCs) to analyze the audio data from heart sounds, focusing exclusively on the Period 2 segments—identified as the systolic region. This selection is rooted in medical literature, as these segments are where potential systolic murmurs are most likely to be present [11]. Murmurs differ from normal heart sounds in their softness and frequency profile, presenting a significant challenge for acoustic analysis. These variations are not only subtle but also contaminated with background noise and artifacts that further obscure the murmurs. Additionally, murmurs vary acoustically depending on their type, location, and severity; for example, systolic and diastolic murmurs exhibit different spectral signatures. This variability necessitates sophisticated feature extraction and machine learning techniques for effective identification and classification of heart murmurs.

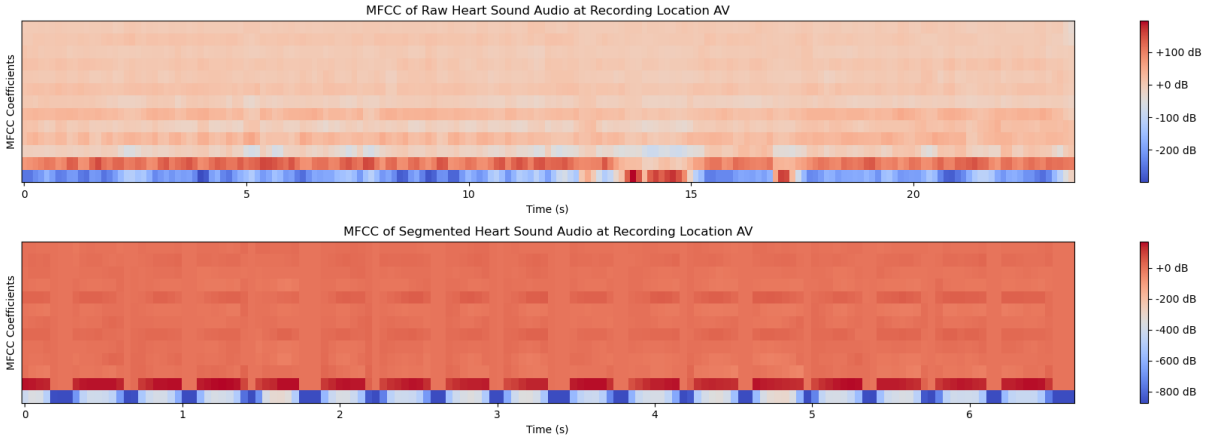

Figure 5: MFCC features extracted from a heart sound recording audio file at the atrial valve (AV) recording location for both the raw heart sound audio file and its corresponding segmented heart sound audio file.

By employing 40 MFCC features, we effectively capture essential auditory characteristics that mimic human perception, such as loudness and pitch, within this critical period. Figure 5 and Figure 6 provide a

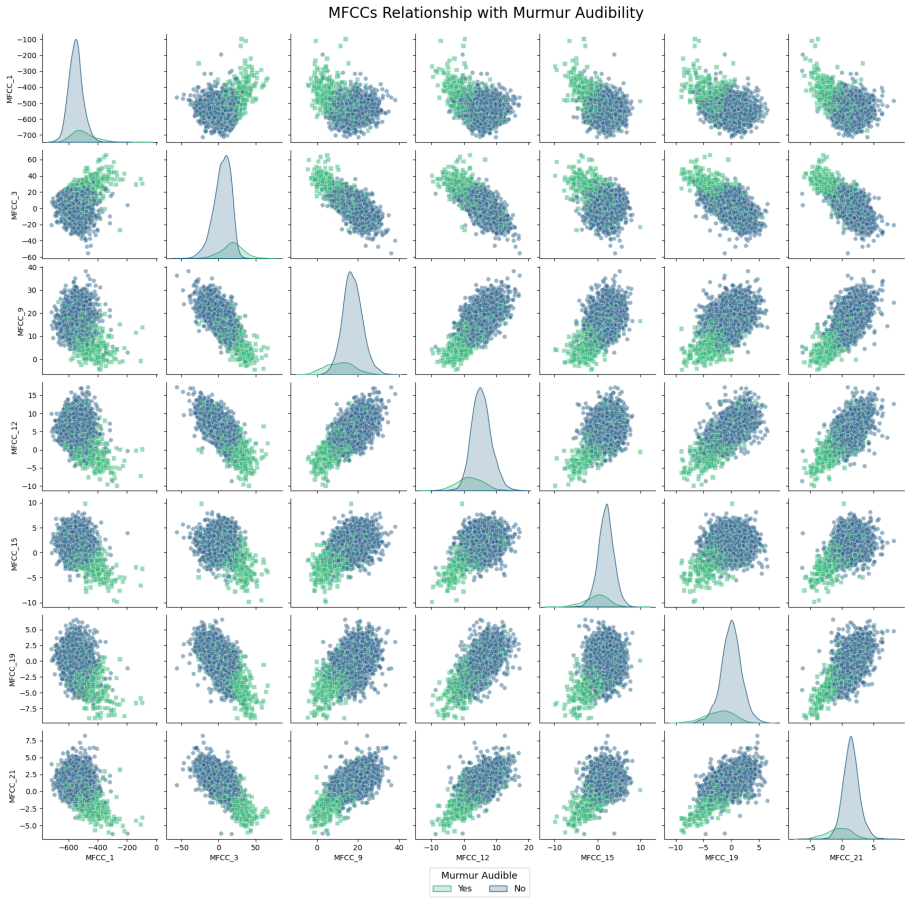

Figure 6: Pair-plot showing 7 of the 40 MFCC features extracted from all heart sound segmented audio recordings in the training dataset.

detailed visual representation of the Mel Frequency Cepstral Coefficients (MFCCs) utilized in our analysis of heart sounds. Figure 5 displays the MFCC features extracted from both raw and segmented heart sound recordings at the atrial valve (AV) location, illustrating the transformation of audio data through preprocessing. Meanwhile, Figure 6 presents a pair-plot of seven selected MFCC features from the training dataset, showcasing their interactions and the distinctions between murmurs and normal heart sounds. These figures underscore the critical role of MFCCs in enhancing our understanding and detection of heart murmurs. This targeted approach allowed our machine learning models to better capture the nuances of systolic heart sounds, facilitating a more precise analysis and identification of the murmur conditions. It is also worth mentioning that in our pursuit to optimize the performance of the machine learning models for heart murmur detection, a critical observation emerged regarding the variability in model efficacy when training and evaluating models based on specific recording locations versus aggregating data across all locations. This distinction became a focal point in our study, as it significantly influenced the detection accuracy and reliability of the models.

During the initial phases of our research, models were trained on the entirety of the dataset without distinction between the different recording locations (PV, AV, TV, MV). This approach, while holistic, did not account for the unique acoustic signatures that murmurs might exhibit depending on where on the chest the recording was taken. Heart murmurs can have distinct sounds based on their anatomical origin, influenced by factors such as the proximity to the heart valves and the direction of blood flow. Recognizing this, we shifted our strategy to develop location-specific models. By segregating the data according to the recording site and tailoring the models to these subsets, we aimed to capture the unique characteristics of the heart sounds from each location. This methodological refinement led to a noticeable improvement in model performance, and therefore, feature extraction was segregated per recording location.

5. Scaling

After extracting the MFCC features from the Period 2 segments, we applied scaling to standardize the data, normalizing each feature to have a mean of 0 and a standard deviation of 1, ensuring uniformity in the data for optimal model performance.

6. Metadata Location Split

To further refine our analysis and enhance the model's accuracy, we segmented the training dataset based

on the measurement locations indicated in the metadata—Aortic Valve (AV), Mitral Valve (MV), Pulmonic Valve (PV), and Tricuspid Valve (TV). This was an empirical decision detailed in section 4. Each location represents a specific area where heart sounds are recorded. In our dataset, we have roughly the same number of audio files for each measurement location. This segmentation allows for more precise model training and higher diagnostic accuracy across different heart valve areas.

7. Training and Validating ML models

Our methodology, as outlined in section 3.1, involves a dual-path exploration of both traditional machine learning models and more complex deep learning frameworks. To reiterate, each model has been selected based on a comprehensive review of existing literature and the promising results reported in related works (section 2).

Hyperparameter tuning was performed exclusively on the training and validation datasets, with different models optimized for their distinct architectures and parameters. During the optimization process, cross-validation was crucial in finding the best hyperparameters for each classifier. For instance, grid searches were employed to explore optimal combinations for Random Forests (number of estimators), Support Vector Machines (kernels and regularization), K-Nearest Neighbors (number of neighbors), and Logistic Regression (penalty parameters). Iterative cross-validation on the validation set alone ensured the models were not over-fitting on unseen data.

The Random Forest classifier was configured with 100 trees, striking a balance between computational cost and predictive performance, with the 'sqrt' option for maximum features to promote diversity among the trees and mitigate over-fitting. The Support Vector Machine (SVM) utilized a linear kernel due to its effectiveness with high-dimensional data, paired with a regularization parameter (C) of 1.0 to adequately balance margin maximization and error minimization. For the K-Nearest Neighbors (KNN) model, we chose 5 neighbors to achieve a beneficial mix of locality and noise reduction, employing the Euclidean distance metric for its simplicity and effectiveness in high-dimensional space. Decision Trees were limited to a maximum depth of 10 to prevent over-fitting, using Gini impurity as the criterion for its robustness in handling multi-class classification problems. Logistic Regression was implemented with a regularization parameter (C) of 1.0 to optimize the balance between fitting and over-fitting, utilizing the 'lbfgs' solver for its efficiency, particularly in smaller datasets. Lastly, the Naive Bayes (GaussianNB) model was fine-tuned with a variance smoothing parameter set to 1e-9, ensuring numerical stability in scenarios where feature variances are minimal, thus safeguarding the model's performance and stability.

Extensive experiments were also performed with various CNN architectures incorporating multiple convolutional and pooling layers to leverage their potential in capturing spatial hierarchies in spectrogram data. Techniques such as dropout and L2 regularization were integrated to mitigate overfitting concerns, especially given the limited size and imbalanced nature of our dataset. Despite their sophistication, CNN models required extensive computational resources and tuning, and they did not yield the expected improvement over traditional machine learning models when evaluated strictly on heart sound classification tasks. This was partly because the added model complexity did not translate into better generalization across the diverse acoustic profiles present in the dataset.

Similarly, to further refine the ability to handle outliers (e.g., murmurs) and enhance a model's attention on relevant patterns indicative of murmurs, anomaly detection techniques such as Isolation Forest and One-Class SVM were explored. By focusing on 'normal' data during training (which consists of 80% of the available data), the model could better learn the characteristic features of typical heart sounds and murmurs, thereby improving its generalization performance on unseen data. Despite their potential benefits, the techniques of anomaly detection, majority voting, and handling class imbalance through synthetic data augmentation and class weighting were ultimately not incorporated into the final methodology as we began to notice the complexity and subtle distinctions in heart murmur detection required a simpler and more transparent approach. Therefore, we opted to use the following six traditional machine-learning models with carefully tuned hyperparameters.

9. Model Evaluation

We primarily focused on metrics such as precision, recall (sensitivity), F1-score, and accuracy, which are critical for assessing the performance of models in the context of medical diagnostics. Recall was particularly emphasized, as minimizing false negatives is paramount in clinical settings to avoid missing any potential diagnoses of heart murmurs. Each model's performance was rigorously tested using a dedicated test set, which constituted 30% of the entire dataset, ensuring that the models were evaluated on unseen data. Additionally, the Receiver Operating Characteristic (ROC) curves and the corresponding Area Under the Curve (AUC) were computed for each model across all recording locations (PV, AV, TV, MV). To address potential class imbalance, which is common in medical datasets, we utilized stratified sampling during dataset splitting.

10. Model Interpretation

Interpreting machine learning models, especially in a clinical context, involves understanding how model decisions are made, which can be challenging with complex models. For the traditional machine learning models employed (e.g., Random Forest, SVM, Decision Tree, KNN, Logistic Regression, and Naive Bayes), we leveraged several techniques to enhance interpretability. For models like Random Forest and Decision Trees, we extracted feature importance scores, which indicate how valuable each feature is in making predictions. This helps in understanding which characteristics of the heart sounds are most indicative of murmurs. For simpler models such as Logistic Regression and SVM with a linear kernel, we examined the decision boundaries to understand how different features influence the prediction outcome. This can be particularly insightful when the models are linear or near-linear. For all models, confusion matrices were used to visually assess model performance concerning false positives, false negatives, true positives, and true negatives (these are summarized in tables to save space). This allows clinicians to understand the types of errors the models are making and consider these in the context of clinical decision-making.

11. Novelty

Our project aims to demonstrate a significant advancement in heart sound analysis by showing that simpler machine learning models using Mel-frequency cepstral coefficients (MFCCs) with meticulous preprocessing can outperform more complex deep learning architectures. For healthcare professionals, the ability to comprehend how diagnostic predictions are made is essential for trust and acceptance, facilitating easier integration into clinical workflows and supporting more informed decision-making. As part of the novelty of this research, we focused on making heart sound classification more explainable and potentially applicable in the real world. We developed a Streamlit interface to integrate our best-performing model to classify murmur characteristics. Healthcare workers can upload audio files into the interface. The audio data undergoes preprocessing steps such as segmenting, noise reduction, and feature extraction, precisely those used in model training. The processed audio is analyzed to detect murmurs (outputting the result) and uses SHAP values to identify which MFCC features were most influential. If a murmur is detected, the system analyzes specific murmur characteristics using the secondary model. These characteristics are displayed to provide detailed insights, enhancing clinical decision-making. The overall process flow is illustrated in Figure 7. The significance and implications of these capabilities will be discussed further in the sections 4 and 5 of our study.

In summary, while our initial experiments with complex features, advanced signal processing, and neural networks provided valuable insights, they underscored the challenges of working with highly variable and noisy clinical data. The streamlined approach described in our final methodology, which emphasizes precise audio segmentation and the use of simpler machine learning models, ultimately proved to be the most effective in enhancing the detection of heart murmurs, for robust clinical applications.

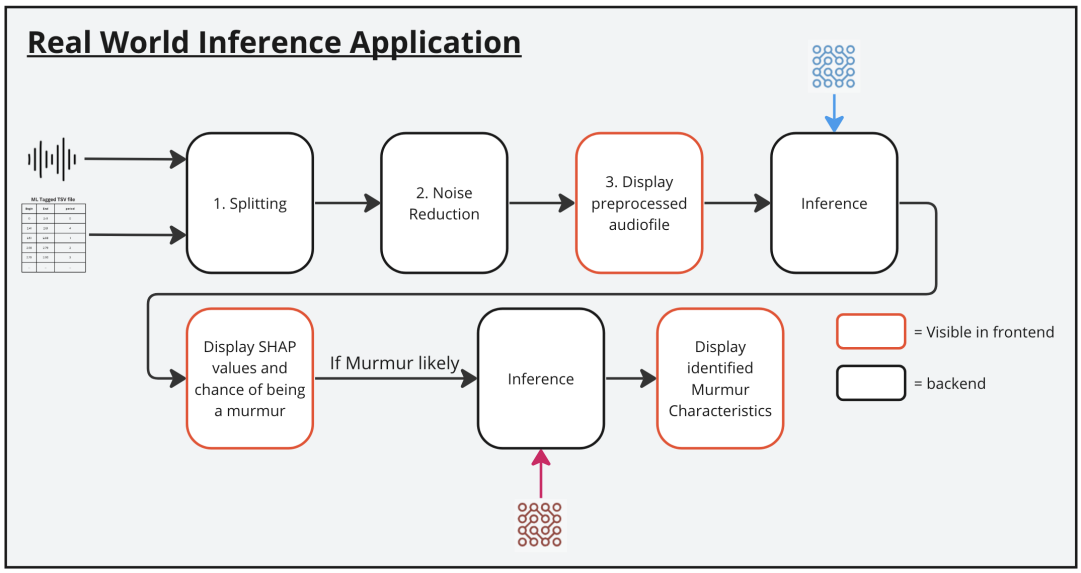

Figure 7: Real World Application: Community Health Care Worker Streamlit Backend Flow

4 Results

4.1 Experimental Results

Table 8 presents the classification results of various models before segmentation and ignoring location data. The performance of each model is evaluated by metrics such as weighted F1-Score, True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Model	F1-Score	TP	TN	FP	FN
Random Forest	81.03%	4	416	1	65
SVM	79.25%	0	417	0	69
KNN	81.08%	13	391	26	56
Decision Tree	76.65%	16	353	64	53
Logistic Regression	84.13%	13	412	5	56
Naive Bayes	77.41%	41	320	97	28

Table 8: Performance without preprocessing and segmentation

Segmenting the data by location shows subtle improvements in model performance as noted in the adjusted, weighted F1-Scores and reduction in False Negatives in Table 9.

Model	F1-Score	TP	TN	FP	FN
Random Forest	82.04%	7	414	3	62
SVM	82.44%	8	415	2	61
KNN	81.23%	11	397	20	58
Decision Tree	78.41%	21	357	60	48
Logistic Regression	83.43%	12	410	7	57
Naive Bayes	76.14%	29	328	89	40

Table 9: Performance without preprocessing and with segmentation

By preprocessing we see improvements across models in terms of the weighted F1-Score increase and reduction in False Negatives displayed in Table 10.

Location	F1-Score	TP	TN	FP	FN
PV	86.24%	13	88	11	6
AV	90.30%	13	101	9	4
TV	80.14%	11	76	22	4
MV	82.02%	13	89	21	5

Table 10: Performance with preprocessing and with segmentation

Final improvements can be made by focusing on optimizing for recall of the "murmur prediction" class. Figure 8 displays the relationship between Target Recall and Specificity as a function of different classification thresholds for our best-performing RF model. It highlights how the Target Recall (True Positive Rate) decreases as the threshold increases from 0.05 to 0.30, indicating a decline in the model's ability to correctly identify positive cases as the threshold becomes more stringent. Conversely, the Specificity (True Negative Rate over Total Positives), represented by the green dashed line, increases with the threshold, indicating an improvement in the model's ability to correctly identify negative cases as false alarms are reduced. Optimizing for high recall (sensitivity) is crucial, particularly when the cost of missing a true positive could have significant health repercussions. Setting the classification threshold at 0.15 is a strategic choice aimed at balancing the need for high recall, ensuring most positive cases are captured while maintaining a reasonable level of specificity to avoid too many false positives. Setting the threshold to 0.15 results in a further reduction in False Negatives and improvements in F1-Scores across different locations, suggesting a tailored approach per location significantly enhances model performance indicated in Table 11.

For the final evaluation, the training and validation datasets were merged for the models to train on 75% of the available data (refer to Table 3 and Figure 9) and evaluate the models on 25% of the available data.

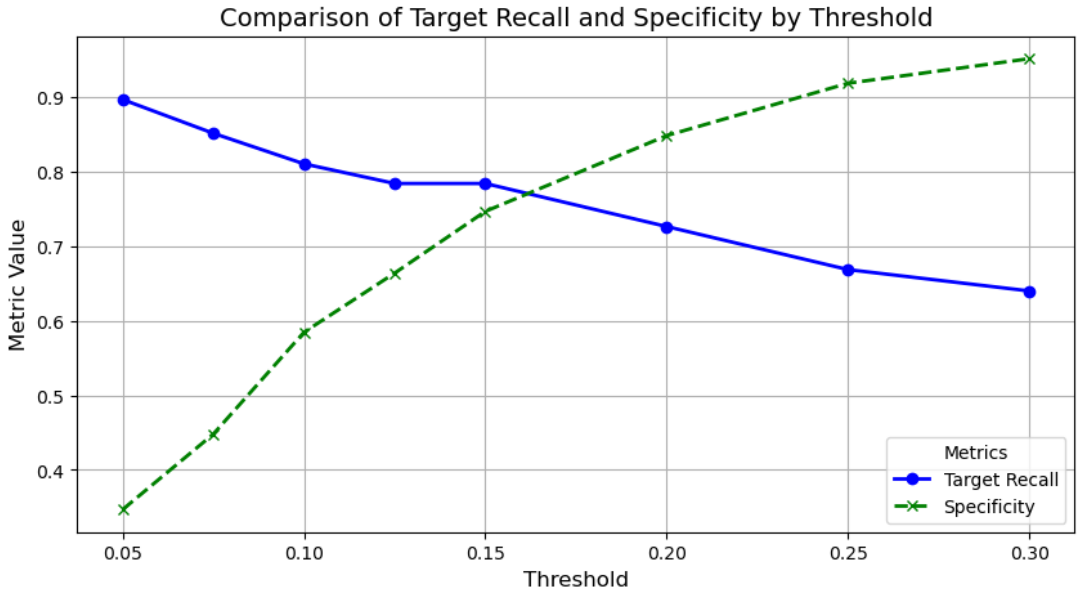

Figure 8: Trade off between recall and specificity for different classification thresholds for the best performing RF model

Model	Avg F1-Score Increase	Avg FN Reduction
Random Forest	7.62%	7.25
SVM	6.87%	6.25
KNN	12.98%	10.00
Decision Tree	6.23%	2.75
Logistic Regression	6.36%	4.75
Naive Bayes	14.86%	11.75

Table 11: Performance on validation set by lowering the murmur classification to a 0.15 threshold

From the analysis of various machine learning classifiers applied to different heart sound recording locations (PV, AV, TV, and MV), each model's performance was evaluated based on precision, recall, F1-score, and the count of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These metrics help identify the most reliable classifiers for clinical application, prioritizing those with minimal FN rates to ensure all potential murmurs are examined. Table 12 through Table 15 synthesize the performance metrics of various machine learning classifiers across four distinct heart sound recording locations — PV, AV, TV, and MV on the test set. Additionally, below each table, the AUC score for the two best-performing models per recording location are shown in Figure 10 through Figure 13. The AUC score is particularly useful in this context because it is not affected by the proportion of patients with or without the condition (class imbalance), which is common in medical datasets including the one concerning this report as shown in Figure 1. This property allows for a fair comparison of model performance in datasets where one class might be significantly less common than the other.

In the PV location (Table 12), the Naive Bayes classifier excels with a balanced performance, notably in detecting murmurs while minimizing false alarms. Although the Random Forest model has lower precision, its high recall rate for "Murmur Audible" makes it valuable for initial screenings where capturing most potential cases is crucial. At the AV location (Table 13), both SVM and Logistic Regression demonstrate exceptional performance, effectively balancing sensitivity and specificity in detecting both murmur and non-murmur cases. These models are particularly adept at minimizing false negatives, ensuring that murmurs are reliably identified. In the TV location (Table 14), Naive Bayes and Logistic Regression maintain strong performance, with Naive Bayes showing high specificity and recall for "Murmur Not Audible," indicating its effectiveness in such settings. The MV location (Table 15) also sees these models perform well, especially in reducing false negatives, a crucial factor in clinical diagnostics. Their consistently high recall rates underline their suitability in environments where missing a condition could pose significant patient risks.

The classification outcomes at each recording location emphasize the critical need to minimize false negatives in clinical diagnostics to prevent missed murmurs and the subsequent risk of severe cardiac conditions. Although a false positive can be reevaluated, a false negative could lead to undiagnosed, progressing

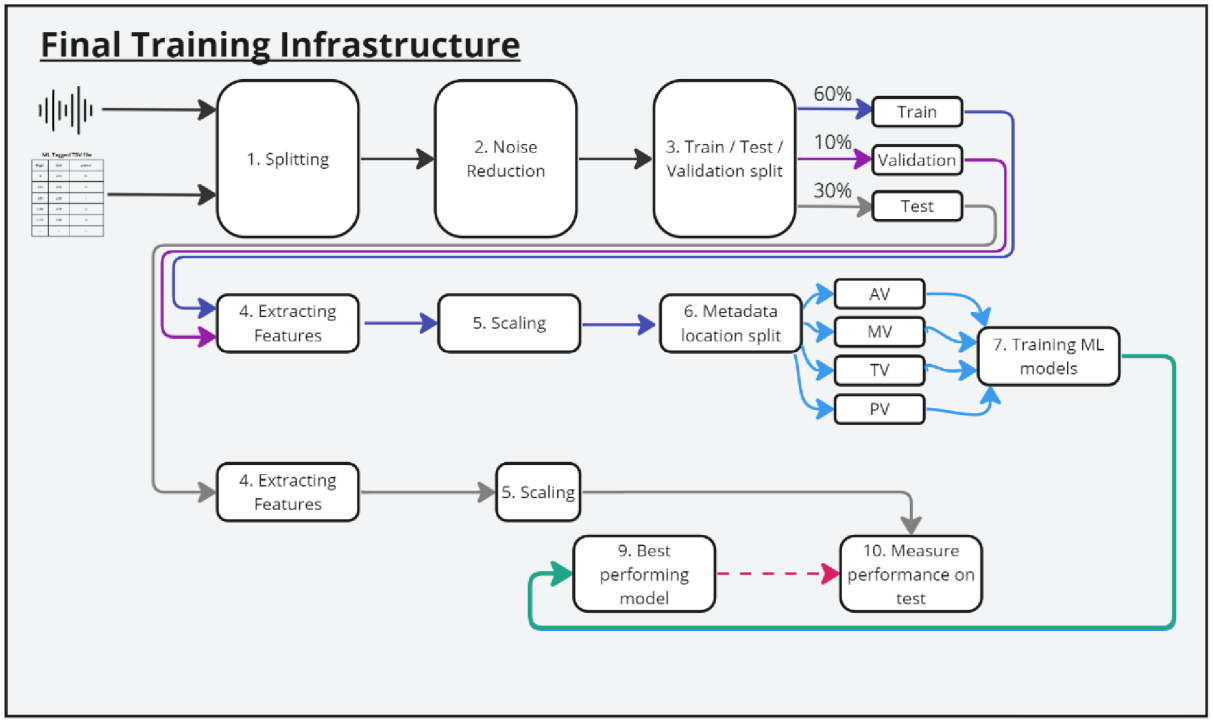

Figure 9: Machine learning pipeline with the training and validation dataset combined.

diseases. For instance, while the Random Forest model at the PV location doesn't achieve the highest precision, its significant recall (86.67% with 65 true positives and only 10 false negatives) proves its efficacy in capturing murmurs that might be missed by other models, making it ideal for preliminary screenings.

Conversely, in situations requiring high sensitivity and specificity—where each false alarm could lead to unnecessary and costly interventions—models like SVM and Naive Bayes are preferable. These models consistently perform well in both the AV and MV locations, making them suitable for comprehensive cardiac assessments where the impact of false positives is also a concern. The results highlight that no single model is universally superior across all locations; however, leveraging each model's strengths based on specific clinical needs and data characteristics can significantly enhance the reliability and effectiveness of automated heart murmur detection systems, providing clinicians with tools that are finely tuned to the diagnostic context.

Classifier	Murmur Audible			Murmur Not Audible			TP	TN	FP	FN
	Precision	Recall	F1-Score	Precision	Recall	F1-Score				
Random Forest	43.05%	86.67%	57.52%	96.03%	73.78%	83.45%	65	242	86	10
SVM	49.09%	72.00%	58.38%	92.83%	82.93%	87.60%	54	272	56	21
KNN	36.91%	73.33%	49.11%	92.13%	71.34%	80.41%	55	234	94	20
Decision Tree	54.55%	56.00%	55.26%	89.88%	89.33%	89.60%	42	293	35	33
Logistic Regression	50.47%	72.00%	59.34%	92.91%	83.84%	88.14%	54	275	53	21
Naive Bayes	72.22%	69.33%	70.75%	93.05%	93.90%	93.47%	52	308	20	23

Table 12: Classification results for the PV location on the test set

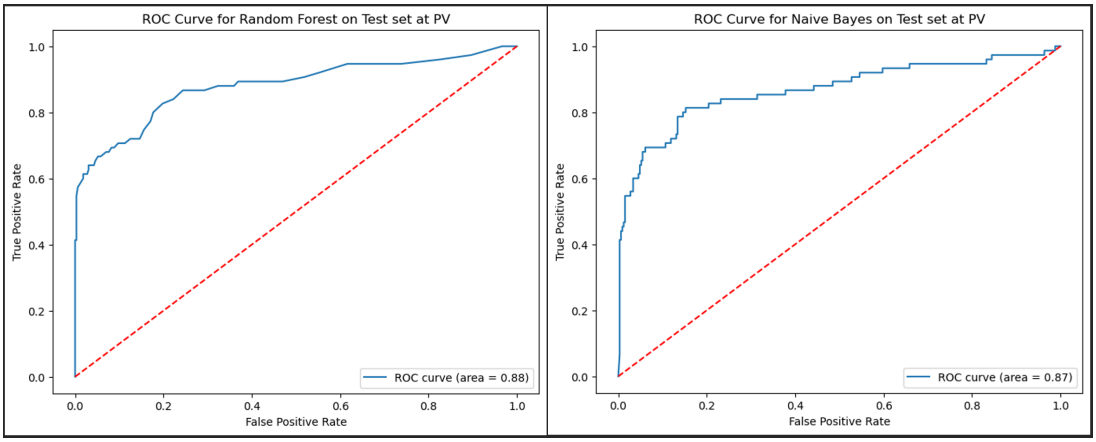

Figure 10: ROC curve for the top two models with the highest AUC score in the PV Location (Random Forest and Naive Bayes)

Classifier	Murmur Audible			Murmur Not Audible			TP	TN	FP	FN
	Precision	Recall	F1-Score	Precision	Recall	F1-Score				
Random Forest	44.54%	75.71%	56.08%	94.46%	81.46%	87.48%	53	290	66	17
SVM	50.00%	78.57%	61.11%	95.25%	84.55%	89.58%	55	301	55	15
KNN	26.23%	68.57%	37.94%	90.95%	62.08%	73.79%	48	221	135	22
Decision Tree	42.62%	37.14%	39.69%	87.95%	90.17%	89.04%	26	321	35	44
Logistic Regression	49.54%	77.14%	60.34%	94.95%	84.55%	89.45%	54	301	55	16
Naive Bayes	67.14%	67.14%	67.14%	93.54%	93.54%	93.54%	47	333	23	23

Table 13: Classification results for the AV location on the test set

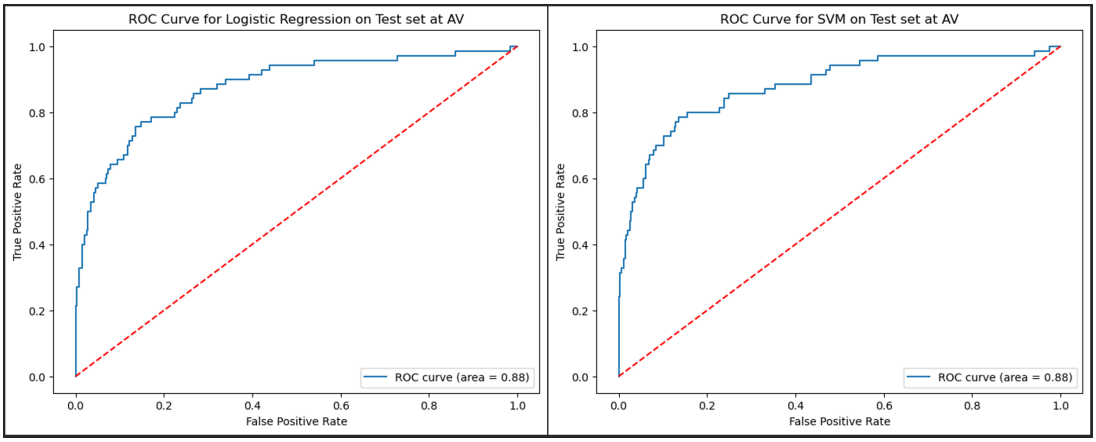

Figure 11: ROC curve for the top two models with the highest AUC score in the AV Location (Logistic Regression and SVM)

Classifier	Murmur Audible			Murmur Not Audible			TP	TN	FP	FN
	Precision	Recall	F1-Score	Precision	Recall	F1-Score				
Random Forest	44.35%	70.83%	54.55%	91.70%	78.38%	84.52%	51	232	64	21
SVM	53.54%	73.61%	61.99%	92.94%	84.46%	88.50%	53	250	46	19
KNN	34.87%	73.61%	47.32%	91.20%	66.55%	76.95%	53	197	99	19
Decision Tree	58.90%	59.72%	59.31%	90.17%	89.86%	90.02%	43	266	30	29
Logistic Regression	54.08%	73.61%	62.35%	92.96%	84.80%	88.69%	53	251	45	19
Naive Bayes	69.23%	62.50%	65.69%	91.09%	93.24%	92.15%	45	276	20	27

Table 14: Classification results for the TV location on the test set

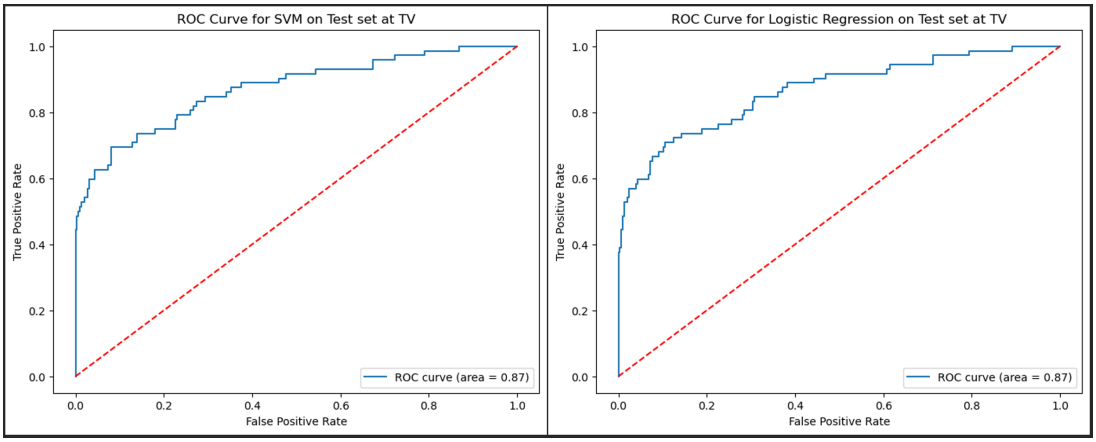

Figure 12: ROC curve for the top two models with the highest AUC score in the TV Location (SVM and Logistic Regression)

Classifier	Murmur Audible			Murmur Not Audible			TP	TN	FP	FN
	Precision	Recall	F1-Score	Precision	Recall	F1-Score				
Random Forest	34.78%	80.00%	48.48%	94.68%	70.34%	80.71%	56	249	105	14
SVM	44.74%	72.86%	55.43%	93.87%	82.20%	87.65%	51	291	63	19
KNN	26.42%	72.86%	38.78%	91.77%	59.89%	72.48%	51	212	142	19
Decision Tree	42.31%	47.14%	44.59%	89.31%	87.29%	88.29%	33	309	45	37
Logistic Regression	47.27%	74.29%	57.78%	94.27%	83.62%	88.62%	52	296	58	18
Naive Bayes	58.75%	67.14%	62.67%	93.31%	90.68%	91.98%	47	321	33	23

Table 15: Classification results for the MV location on the test set

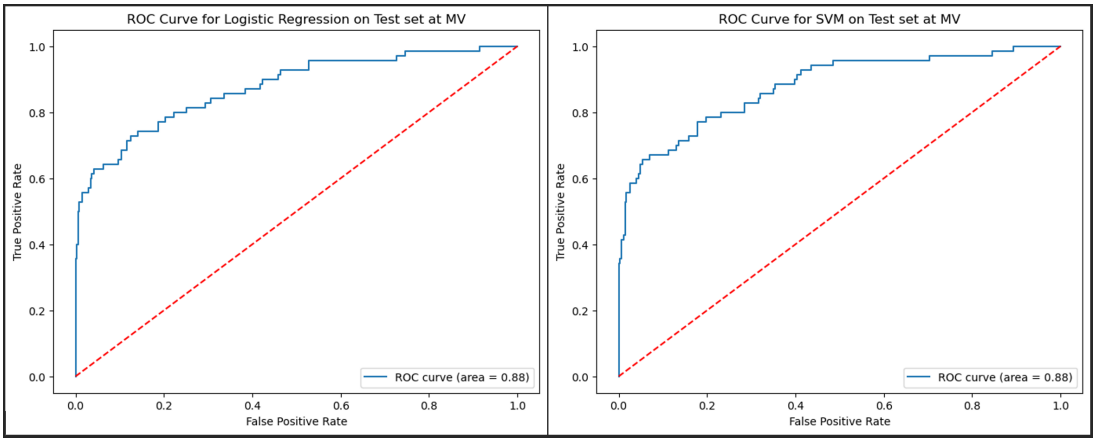

Figure 13: ROC curve for the top two models with the highest AUC score in the MV Location (Logistic Regression and SVM)

4.2 Application Results

As mentioned in section 3.3, we explored how to make heart sound classification more interpretative and how our results can be applied in real life by developing a front-end that utilizes inference from two trained machine learning models. Here is an example workflow designed to potentially assist users in identifying heart murmurs:

Ana, a community health worker in a remote area of Brazil, conducts daily examinations of children using only a digital stethoscope. Diagnosing heart murmurs from sound recordings poses a significant challenge, even for experienced cardiologists. Given her limited resources, Ana must carefully decide which children to send on a long journey to a distant hospital for further treatment.

To aid Ana, our model processes the audio files from the digital stethoscope with noise reduction, audio enhancements, and by stripping out segments where a murmur cannot occur. This preprocessing allows her to listen more attentively to the exact timings at which a murmur might occur, simplifying the complex task of classification. What this looks like in the front-end is shown in Figure 14. Once a potential murmur

is detected, the model delivers a verdict with SHAP explanations on why it reached that conclusion, as shown in Figure 15. If a murmur is confirmed, the model provides additional details on the characteristics of the murmur observed, helping Ana navigate its complex features. An example is shown in Figure 16. This information assists her in prioritizing which children might have the most serious underlying conditions, ensuring that only those most likely to need further medical intervention undertake the long and costly journey for a more thorough examination. This thoughtful application of machine learning thus aims to make Ana's critical judgment calls both more informed and more efficient.

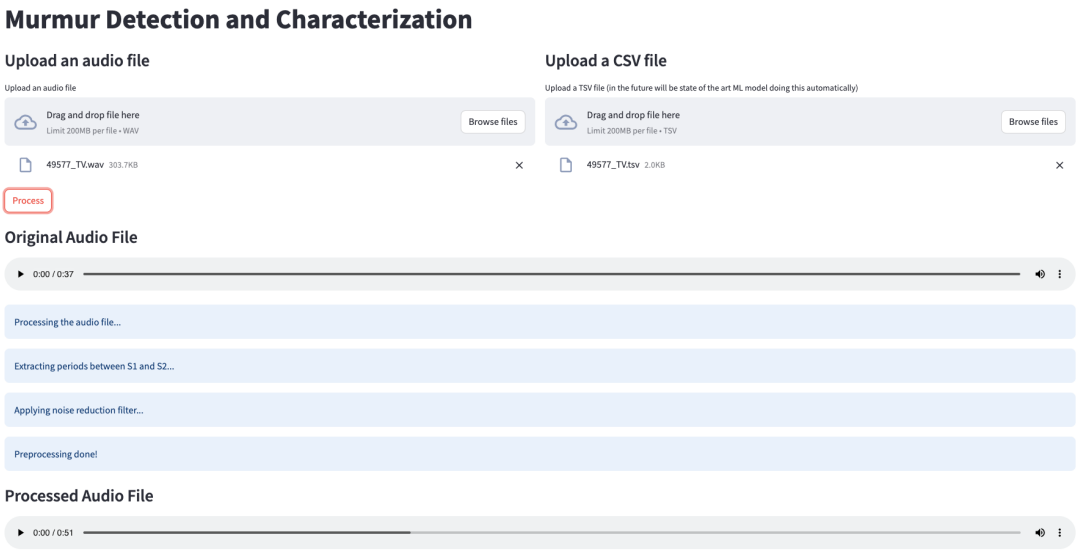

Figure 14: Uploading the .wav file and metadata results in a processed audio file that only contains sounds where a potential murmur can occur

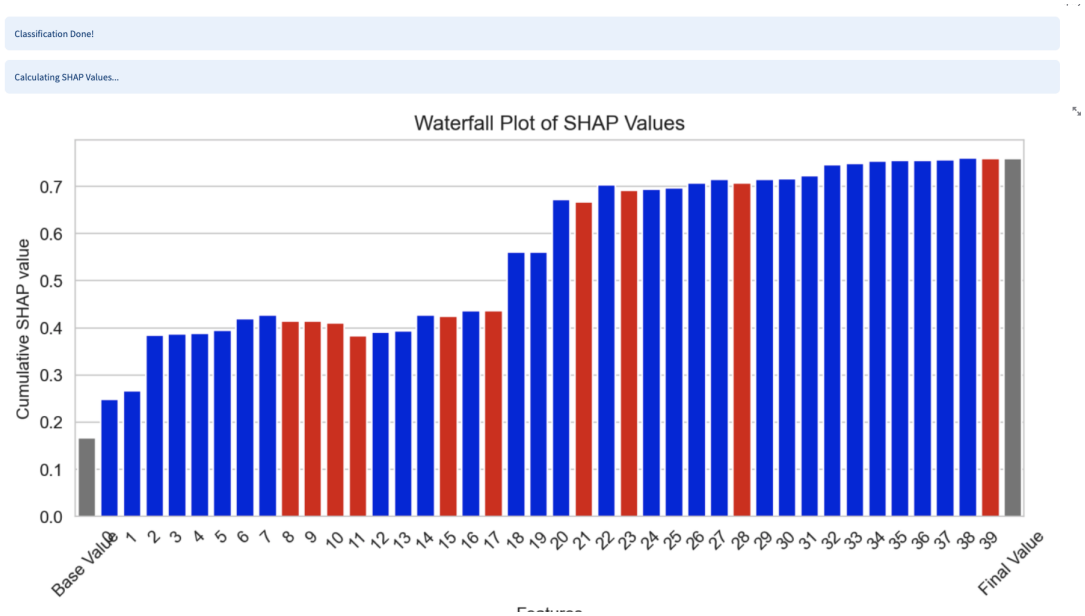

Figure 15: The model provides feedback with SHAP values on whether it thinks it is likely this recording contains a murmur

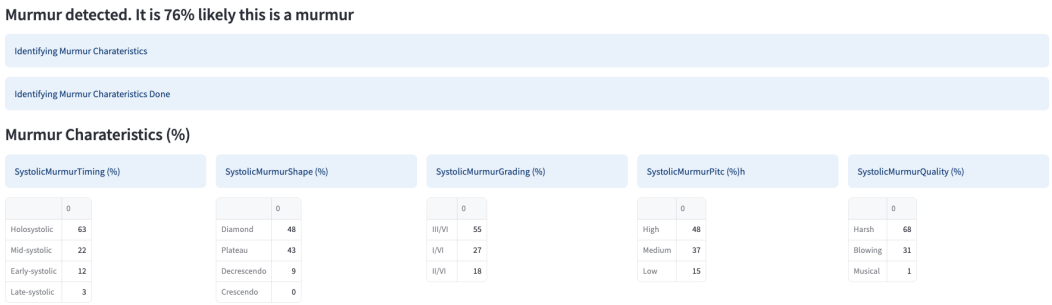

Figure 16: If a murmur is detected, it gives detailed characteristics of the murmur detected

5 Discussion

5.1 Overview

In this study, we have explored various approaches to classify heart sounds and detect murmurs using machine learning techniques. Our findings demonstrate the potential of using simpler models with carefully engineered features, such as segmentation techniques and Mel Frequency Cepstral Coefficients (MFCCs), to achieve performance comparable to complex deep learning architectures. Moreover, we have developed a Streamlit interface that showcases how our models can be integrated into a user-friendly tool to assist community healthcare workers in assessing heart murmurs in resource-constrained settings. This section discusses the main contributions, sub-optimal approaches, real-world applications, methodological nuances, ethical considerations, and future directions of our research.

5.2 Advancing Heart Sound Classification with Novel Approaches

Our work makes significant contributions to the field of heart sound classification by demonstrating the effectiveness of using MFCCs with simpler machine learning models. This approach offers several advantages over complex deep learning architectures, including reduced computational requirements, improved interpretability, and increased accessibility for medical professionals. By leveraging MFCCs to capture the timbral and textural qualities of heart sounds, we have shown that simpler models can achieve performance levels comparable to state-of-the-art deep learning techniques. For example, in the Pulmonic Valve (PV) location, the Random Forest model achieved a recall of 86.67% for 'Murmur Audible' classifications and a precision of 96.03% for 'Murmur Not Audible' classifications, underscoring the capability of these models to provide reliable diagnostic insights. Similarly, in the Mitral Valve (MV) location, the Random Forest model demonstrated a recall of 80.00% for detecting audible murmurs, which is crucial for ensuring that fewer cases of potential abnormalities go unnoticed. This finding has important implications for the broader adoption of heart sound analysis technologies, as it lowers the barriers to entry and enables the deployment of these tools in resource-limited settings, such as remote clinics or on mobile devices like smartphones and digital stethoscopes.

5.3 Lessons from Sub-optimal Approaches

Throughout our research, we explored several techniques that ultimately did not yield optimal results. These suboptimal approaches, however, provide valuable insights and lessons for future research in heart sound classification. For instance, our experiments with wavelet transforms highlighted the importance of balancing feature richness with model interpretability. Similarly, the use of Mel spectrograms as inputs for CNNs underscored the significance of retaining critical time-domain information when analyzing heart sounds. Our attempts at segmenting audio files based on individual heartbeats and applying anomaly detection techniques further emphasized the challenges associated with the inconsistent occurrence of murmurs and the nuanced nature of distinguishing between normal and abnormal heart sounds. By sharing these findings, we aim to guide future researchers in their choice of techniques and help them avoid potential pitfalls.

5.4 Empowering Community Healthcare Workers

One of the key contributions of our work is the development of a Streamlit interface that demonstrates how our heart sound classification models can be integrated into a user-friendly tool for community healthcare workers. This interface allows users to upload heart sound recordings and receive immediate feedback on the likelihood of a murmur being present, along with detailed characteristics of the detected murmur. By providing a processed audio file that focuses on the segments where murmurs are most likely to occur, our tool enables healthcare workers to make more informed decisions about referring patients for further

evaluation. The potential benefits of such a system are particularly significant in resource-constrained environments, where access to specialized medical expertise may be limited. However, it is crucial to recognize that this is just the first step in a longer journey to ensure the effectiveness and reliability of our models in real-world settings. Comprehensive testing in diverse environments, further refinement of the models, and the development of targeted training programs for healthcare workers are essential to realize the full potential of this technology in enhancing healthcare delivery in underprivileged areas.

5.5 Methodological Nuances

In our study, we made a strategic choice to adjust the classification threshold to 0.15 to optimize for high recall (sensitivity). This decision was driven by the clinical need to minimize false negatives and ensure that most positive cases are captured, even at the cost of a slightly reduced specificity. By prioritizing recall, we aim to align our models with the real-world requirements of heart sound classification, where missing a potentially abnormal case could have serious consequences for patient outcomes. This methodological nuance highlights the importance of considering the specific context and objectives of the application when designing and evaluating machine learning models in healthcare settings.

5.6 Ethical Considerations and Privacy Implications

The use of machine learning models in clinical settings raises important ethical considerations and privacy implications that must be carefully addressed. Ensuring the reliability, fairness, and interpretability of our models is of utmost importance to maintain trust and avoid potential harm arising from misclassification. Regular validation and auditing of the models are necessary to identify and mitigate any biases or errors that may arise over time. Moreover, the handling of sensitive patient data requires strict adherence to relevant regulations and ethical guidelines. Robust data management practices, including anonymization and secure storage, are essential to protect patient privacy and maintain confidentiality. As we move towards the real-world deployment of our heart-sound classification tools, it is crucial to engage with healthcare professionals, ethicists, and policymakers to develop appropriate governance frameworks and ensure the responsible use of these technologies.

5.7 Future Directions and Collaborations

Our research lays the foundation for several exciting future directions in the field of heart sound classification. Firstly, we plan to focus on advanced audio preprocessing techniques and the expansion of our datasets with diverse, well-labeled recordings. By collaborating with healthcare institutions and leveraging standardized data collection protocols, we aim to improve the accuracy and reliability of our models across a wide range of clinical settings. Secondly, we intend to explore sophisticated machine learning architectures that integrate deep learning with traditional signal processing techniques. These hybrid models have the potential to enhance feature extraction capabilities while maintaining computational efficiency. Additionally, we will investigate the use of unsupervised and semi-supervised learning methods to address the challenge of limited labeled data in medical domains. Lastly, we will work closely with healthcare professionals and technology experts to ensure the seamless integration of our heart sound classification tools into existing clinical workflows. By assessing technical feasibility, navigating regulatory landscapes, and developing user-friendly interfaces, we aim to create a technology that truly augments the capabilities of medical professionals and improves patient care.

5.8 Conclusion

In conclusion, our research demonstrates the potential of using simpler machine learning models with carefully engineered features, such as MFCCs, for heart sound classification and murmur detection. By developing a user-friendly Streamlit interface, we have shown how these models can be integrated into tools that empower community healthcare workers in resource-constrained settings. However, our work also highlights the importance of considering methodological nuances, ethical implications, and the need for comprehensive testing and refinement before deploying these tools in real-world clinical environments. As we move forward, we remain committed to collaborating with healthcare professionals, researchers, and technology experts to advance the field of heart sound classification and develop innovative solutions that can transform the delivery of healthcare in underserved communities worldwide.

References

[1] A. Tandon, S. Sengupta, V. Shukla, and S. Danda. Risk factors for congenital heart disease (chd) in vellore, india. *Current Research Journal of Biological Sciences*, 2(4):253–258, 2010.

[2] M. D. Seckeler and T. R. Hoke. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. *Clinical Epidemiology*, 3:67–84, 2011.

[3] A. Gheorghe, U. Griffiths, A. Murphy, H. Legido-Quigley, P. Lamptey, and P. Perel. The economic burden of cardiovascular disease and hypertension in low-and middle-income countries: a systematic review. *BMC Public Health*, 18(1), 2018.

[4] P. Libby, R. Bonow, D. Mann, and D. Zipes. *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*. Elsevier Science, 8th edition, 2007.

[5] Jorge Oliveira, Francesco Renna, Paulo Dias Costa, Marcelo Nogueira, Cristina Oliveira, Carlos Ferreira, Alípio Jorge, Sandra Mattos, Thamine Hatem, Thiago Tavares, Andoni Elola, Ali Bahrami Rad, Reza Sameni, Gari D. Clifford, and Miguel T. Coimbra. The circor digiscope dataset: From murmur detection to murmur classification. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2524–2535, 2022.

[6] Andrew. McDonald, Mark. JF. Gales, and Anurag Agarwal. "detection of heart murmurs in phonocardiograms with parallel hidden semi-markov models". https://physionet.org/files/challenge-2022/1.0.0/papers/index.html.

[7] Hui Lu, Julia Beatriz Yip, Tobias Steigleder, Stefan Grießhammer, Maria Heckel, Naga Venkata Sai Jitin Jami, Bjoern Eskofier, Christoph Ostgathe, and Alexander Koelpin. A lightweight robust approach for automatic heart murmurs and clinical outcomes classification from phonocardiogram recordings. `https://physionet.org/files/challenge-2022/1.0.0/papers/index.html`.

[8] Yale Chang, Luoluo Liu, and Corneliu Antonescu. Multi-task prediction of murmur and outcome from heart sound recordings. `https://physionet.org/files/challenge-2022/1.0.0/papers/index.html`.

[9] P Busono, S Karim, A Kamaruddin, and IPA Yogiswara. Heart sound signal analysis for digital auscultation. In *Journal of Physics: Conference Series*, volume 2377, page 012024. IOP Publishing, 2022.

[10] Eugene Braunwald, Douglas P. Zipes, Peter Libby, Robert O. Bonow, and Douglas L. Mann. *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*. Elsevier Science, 8th edition, 2007.

[11] A. Aly and S. Dasgupta. The cardiac cycle. `https://www.utmb.edu/pedi_ed/corev2/cardiologypart1/cardiologypart12.html`, August 2022. Accessed: 2024-05-10.