Shape Analysis of Heart Murmurs: Adding Explainability to Murmur Detection

Chandan Pandey[1] , **Pinaki Dey**[2] , **Anirban Dutta Choudhury**[1] , **Naman Mohan Paul**[3] , **Purvesh Sanjeev Doud**[3] , **Riddhi Jain**[3] , **Aniruddha Sinha**[1]

[1]TCS Research, Kolkata, India , [2]IIT Dhanbad, India , [3]TCS Research, Pune, India

c.pandey1@tcs.com, 21je0647@iitism.ac.in, anirban.duttachoudhury, naman.paul, purvesh.doud, jain.riddhi1, aniruddha.s@tcs.com

Abstract

Heart murmurs, caused by aberrant blood flow due to valvular defects, are often early indicators of cardiovascular diseases. Detecting and analyzing the shapes of these murmurs in Phonocardiogram (PCG) signals can provide critical insight into the underlying heart conditions. This paper explores unsupervised and supervised techniques to detect Plateau, Decrescendo, and Diamond shapes in heart murmurs. We introduce a novel unsupervised pipeline that identifies shape characteristics by computing peak patterns in the windowed power envelope of murmur segments and comparing them to a template bank consisting of representative murmur shape profiles. his template-based approach enables interpretable classification, aligning with clinical expectations. n parallel, we develop and evaluate supervised learning models that leverage deep and shallow learning paradigms. These include a Convolutional Neural Network (CNN), a Long Short-Term Memory (LSTM) network, and a k-Nearest Neighbors (k-NN) classifier, all trained on the envelope features of the systolic murmur segments. Empirical results demonstrate that the best-performing supervised pipeline achieves a mean f1-score of 66.2%, outperforming the best unsupervised approach, which yields a mean f1-score of 60.9%. Despite the superior quantitative performance of the supervised methods, the unsupervised technique offers a compelling advantage in explainability through its reliance on a transparent and clinically meaningful template bank.

1 Introduction

Cardiovascular disease (CVD) remains the foremost cause of mortality globally, accounting for an estimated 19 million deaths in the year 2020 alone [Tsao *et al.*, 2023]. CVD encompasses a wide range of disorders affecting the heart and vascular system, including coronary artery disease (CAD), congenital heart disease (CHD), and valvular heart disease (VHD) [Peter, 2007]. Among these, VHD is characterized by structural and functional abnormalities of the heart valves which leads to the generation of atypical heart sounds known as murmurs. These murmurs result from turbulent or altered blood flow across defective valves and are often early indicators of underlying cardiac pathology.

Traditionally, cardiac auscultation using a stethoscope remains the primary screening method for murmur detection in clinical practice. However, accurate interpretation of heart sounds is heavily dependent on the skill and experience of the clinician, resulting in high inter-observer variability [Gardezi *et al.*, 2018]. This limitation has led to increased interest in leveraging digital phonocardiogram (PCG) signals capable of recordings of cardiac activity for the automated detection and analysis of heart murmurs.

Earlier approaches primarily utilized conventional machine learning techniques. For instance, [El Badlaoui *et al.*, 2020] extracted features from PCG signals using Principal Component Analysis (PCA), and applied Support Vector Machines (SVMs) and k-Nearest Neighbors (k-NN) classifiers to distinguish between normal and abnormal heart sounds. In a similar effort, [Abduh *et al.*, 2020] extracted Mel-frequency cepstral coefficients (MFCCs) from segmented cardiac cycles and utilized a combination of k-NN, SVM, and ensemble classifiers for murmur classification.

The recent proliferation of publicly available PCG datasets [Oliveira *et al.*, 2021], [Liu *et al.*, 2016], [Spadaccini and Beritelli, 2013] along with George B. Moody PhysioNet Challenge 2022 [Reyna *et al.*, 2023] have significantly accelerated research with DL-based algorithms for murmur detection and clinical outcome prediction. For example, [Lu *et al.*, 2022] extracted mel spectrograms along with features such as spectral bandwidth and demographic information to train a dual-branch lightweight Convolutional Neural Network (CNN). [McDonald *et al.*, 2023] computed log-spectrograms and trained a bidirectional Recurrent Neural Network (RNN) to identify heart sound states (S1, S2, systole, diastole, murmur), with further segmentation achieved using hidden semi-Markov models (HSMMs). Similarly, [Prince *et al.*, 2023] utilized a mel-spectrogram input to train a Residual Network (ResNet) for murmur detection, and [Elola *et al.*, 2023] further expanded this approach by grading murmur intensity into three levels: absent, soft, and loud.

Despite the success of these deep learning models in improving detection performance, a major shortcoming remains in their lack of explainability. As these models function as "black boxes," it becomes challenging to interpret the ratio-

Characteristic	Representative Values
Shape	Crescendo, Decrescendo, Diamond, Plateau
Pitch	Low, Medium, High
Grading	Grade 1 (barely audible) to Grade 6 (exceptionally loud)
Quality	Blowing, Harsh, Musical

Table 1: Murmur Characteristics

nale behind their predictions. This has motivated the integration of explainable artificial intelligence (xAI) techniques such as LIME [Ribeiro *et al.*, 2016], SHAP [Lundberg and Lee, 2017], and DeepLIFT [Shrikumar *et al.*, 2017]. For example, [Wang *et al.*, 2023] demonstrated the superior performance of the Stockwell Transform for murmur detection, and further used SHAP values to attribute importance to specific regions in the transformed signal. In another study, [Alkhodari *et al.*, 2024] employed gradient-weighted class activation mapping (Grad-CAM) [Selvaraju *et al.*, 2017] to identify critical attention heads in a transformer network, revealing the relative importance of auscultation channels such as the mitral and tricuspid valves.

While these studies primarily focus on murmur detection and its explainability, a significant research gap persists in the characterization of murmurs. Comprehensive murmur characterization spanning multiple features such as location, timing, duration, pitch, intensity, and shape are essential for accurate diagnosis and disease staging. Till date, very few studies have attempted to classify murmur shapes directly. In one such attempt, [Ning *et al.*, 2016] proposed the use of an average magnitude index function to quantify murmur shape over a short temporal segment. However, further efforts in this direction remain scarce.

In this study, we address the challenge of murmur shape classification, with an emphasis on explainability. The principal contributions of our work are as follows:

1. We introduce a template bank containing canonical murmur shapes and develop a novel unsupervised signal processing pipeline that detects peaks in the windowed power of the murmur envelope and aligns them with the most appropriate shape templates.

2. We explore supervised learning approaches using both shallow and deep models, including CNN, LSTM, and k-NN, with systolic murmur envelopes as input features.

3. We perform a comparative analysis, demonstrating that the best supervised model achieves a mean F1-score of 66.2%, while the best unsupervised model achieves 60.9%. Despite this performance gap, the unsupervised method offers significant advantages in clinical explainability through the use of shape-based templates.

This work provides a unique contribution toward bridging the gap between murmur detection and explainable murmur shape characterization, which is vital for the development of clinically useful diagnostic tools.

2 Dataset

In this study, we utilize the CirCor DigiScope dataset [Oliveira *et al.*, 2021], a high-quality collection of phonocardiogram (PCG) recordings acquired from diverse clinical settings. The dataset comprises a total of 3,163 PCG recordings, sampled at 4,000 Hz, collected from 942 pediatric and adult patients through auscultation at various anatomical locations. A subset of approximately 60% of the full dataset has been made publicly available as part of the George B. Moody PhysioNet Challenge 2022 [Reyna *et al.*, 2023], enabling reproducibility and benchmarking of murmur detection algorithms in the research community.

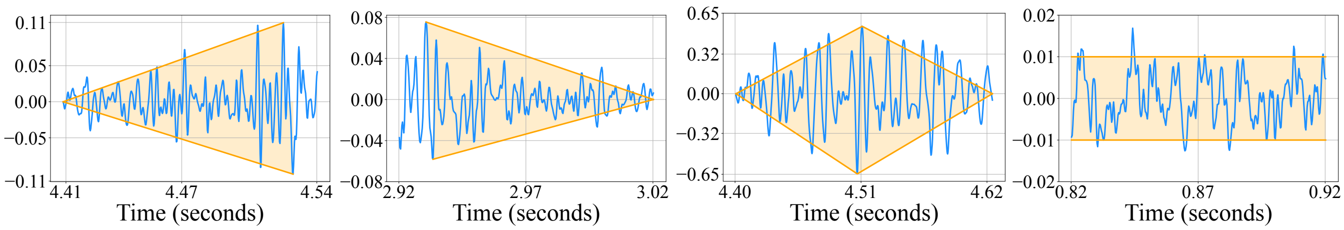
Each PCG recording in the dataset is accompanied by expert-annotated temporal boundaries of the fundamental cardiac phases—namely the first heart sound (S1), systolic interval, second heart sound (S2), and diastolic interval. Additionally, the presence or absence of systolic murmurs is annotated by cardiologists based on auscultatory findings. Table 1 presents an overview of common murmur characteristics. Among these, murmur shape which refers to the temporal evolution of murmur intensity is particularly informative. For instance, a crescendo murmur, characterized by a gradual rise in amplitude, is often associated with mitral valve prolapse; decrescendo murmurs show a decline in intensity and may indicate acute mitral regurgitation; diamond-shaped (crescendo-decrescendo) murmurs are frequently linked to papillary muscle dysfunction; and plateau murmurs, which maintain uniform intensity, are commonly found in holosystolic conditions such as mitral or tricuspid regurgitation and ventricular septal defects [Walker HK, 1990].

Out of the 942 total participants, 179 subjects were identified to have systolic murmurs. However, one subject was excluded from the analysis due to missing murmur shape information. Furthermore, four additional subjects were excluded due to their belonging to murmur shape categories with insufficient class sizes (e.g., holosystolic crescendo, mid-systolic plateau, late-systolic diamond), which were statistically underrepresented and unsuitable for model training. Following this refinement, we retained data from 174 subjects who exhibited one of the three predominant murmur shapes: Plateau, Decrescendo, and Diamond. To ensure high-quality input for the shape classification task, we selected the PCG recording corresponding to the auscultation site with the highest audibility for each subject, as specified in the dataset.

Figure 1 illustrates representative examples of the murmur shapes under investigation: Crescendo, Decrescendo, Diamond, and Plateau. Although the Crescendo class was not included in the final classification task due to data insufficiency, it is presented here for completeness.

3 Methodology

This section presents two distinct pipelines for murmur shape classification: (1) an interpretable, unsupervised signal processing pipeline Fig. 2 and (2) a set of supervised learning models using shallow and deep architectures Fig.3. The unsupervised pipeline emphasizes transparency and explainability, while the supervised approaches prioritize classification performance. Both pipelines are designed to identify murmur



(a) Crescendo (b) Decrescendo (c) Diamond (d) Plateau

Figure 1: Representative murmur shapes based on PCG signal morphology. These temporal patterns are observed in the systolic phase of cardiac cycles and aid in differentiating underlying valvular abnormalities.

shapes—Plateau, Decrescendo, and Diamond—from systolic segments of PCG signals.

3.1 Unsupervised and Explainable Pipeline

Pre-processing

Heart murmurs predominantly occupy the frequency range of 20–500 Hz [McGee, 2018]. To eliminate extraneous noise while preserving murmur content, we apply a second-order Butterworth band-pass filter with cutoff frequencies at 20 Hz and 500 Hz. The PhysioNet 2022 dataset provides detailed annotations demarcating S1, S2, systole, and diastole phases for each cardiac cycle. We isolate the systolic phase of each cycle for further analysis.

To examine murmur intensity, we first extract the Hilbert envelope of the filtered signal. However, the resulting envelopes exhibit excessive local fluctuations (Fig. 4a), complicating shape identification. As a remedy, we introduce an alternative envelope based on windowed power estimation (Fig. 4b). It can be seen that the variation of the windowed power is able to reveal the expected shape (Diamond). This motivated us to to use a set of predefined shape templates and then correlate with the envelope of the windowed power.

Template Bank

We construct a template bank that approximates canonical murmur shapes, enabling correlation-based matching. Each template comprises four synthetic curves, defined according to distinct slope dynamics:

- **Template 1 (T1)**: Linear slopes (constant positive and negative gradients).
- **Template 2 (T2)**: Increasing slopes on both ascending and descending edges.
- **Template 3 (T3)**: Decreasing slopes for both ascent and descent.
- **Template 4 (T4)**: Gaussian-shaped slope variation.

Given that the peak of a Diamond-shaped murmur may not occur centrally, we generate three variants per template with peaks positioned at 25%, 50%, and 75% of the murmur duration. Thus, each template contains four shapes: one Decrescendo and three Diamond variants (Fig. 5). For each murmur segment, we compute Pearson correlation with all template shapes. The shape with the highest correlation is selected. If none exceed a threshold of 0.5, the murmur is classified as a Plateau.

Windowed Power Estimation

To obtain a robust representation of murmur dynamics, we compute the windowed power of each systolic segment (Fig. 6b). Given the variable duration of systole across cardiac cycles, we divide each segment into 30 overlapping windows (50% overlap) and compute signal power within each. The resulting power time series provides a smooth surrogate for shape characterization.

Peak-based Shape Matching

Despite smoothing, abrupt changes remain in the power series, adversely affecting correlation accuracy. To mitigate this, we extract local maxima (peaks) from the power series and construct a peak sequence (Fig. 6c). Pearson correlation is then computed between this peak sequence and each template, improving robustness to minor signal artifacts.

3.2 Supervised Learning Pipeline

To benchmark the unsupervised pipeline, we develop a suite of supervised models using stratified training, validation, and testing protocols. These models are trained to classify murmur shapes based on features extracted from systolic PCG segments.

Data Splitting Protocol

We employ stratified sampling to partition the dataset into training (60%), validation (20%), and testing (20%) sets. This ensures class balance is maintained across all splits.

Pre-processing

The signal is passed through a second-order Butterworth band-pass filter with cutoff frequencies at 20 Hz and 500 Hz. To standardize murmur length across samples, we interpolate all segments to the median murmur length (480 data points) using cubic interpolation. Zero-padding is avoided to preserve shape integrity. The processed dataset yields 1,561 Plateau, 590 Decrescendo, and 396 Diamond murmurs. To mitigate class imbalance, we apply random over- and under-sampling to ensure 600 samples per class.

Envelope features are extracted by identifying local maxima (peaks) and minima (troughs) in the signal. We then compute the upper envelope by interpolating the peaks only, as the lower envelope closely mirrors the inverse of the upper envelope (Fig. 7).

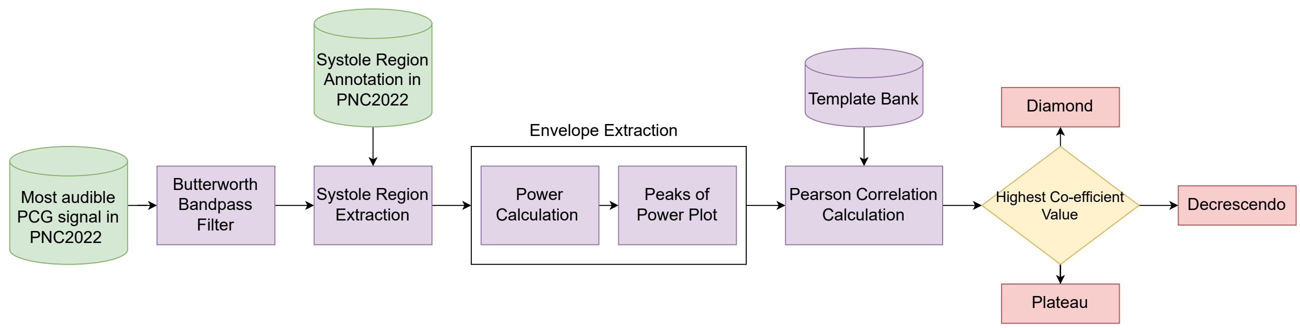


Figure 2: Data flow diagram of the proposed Unsupervised Pipeline

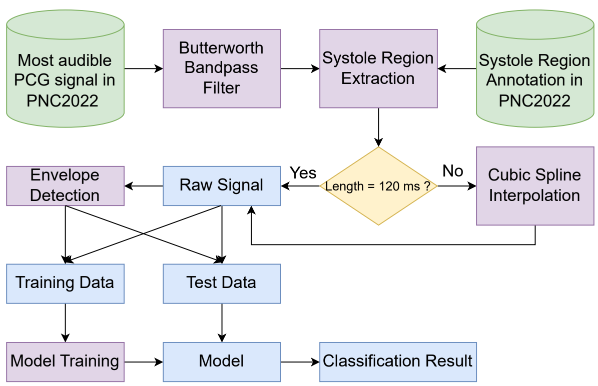
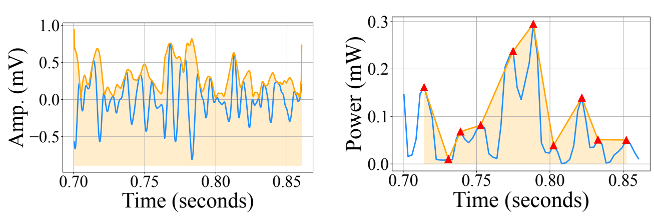


Figure 3: Data flow diagram of the proposed Supervised Pipeline



(a) Envelope of PCG Signal (b) Envelope of Peak Power

Figure 4: Comparison of signal envelope (left) and power peak envelope (right) for systolic murmur. Signal: Cycle 1 of Recording 9979_TV

k-Nearest Neighbors (k-NN)

The k-NN algorithm is applied to both raw interpolated signals and envelope representations. The value of k is empirically selected as 40. This is guided by a heuristic approach [Nadkarni, 2016], where k is approximated as $\sqrt{N}$. Here, $N = 1800$ is the total number of samples across all the three classes.

CNN + LSTM Hybrid Model

We design a five-layer deep learning model composed of convolutional and recurrent units, applied to both raw and envelope data:

1. 1D Convolutional layer with 512 filters (dropout = 0.2)
2. 1D Convolutional layer with 1024 filters (dropout = 0.2)
3. LSTM layer with 512 units (dropout = 0.2)
4. Dense layer with 256 nodes (dropout = 0.3)
5. Output dense layer with 3 nodes (softmax activation)

The model is trained using categorical cross-entropy as the loss function and optimized via the Adam algorithm.

4 Results and Discussion

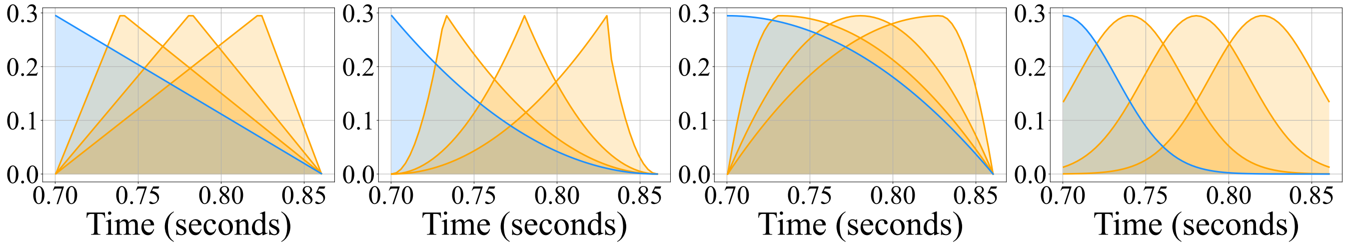
In this section, we present a comprehensive evaluation of the unsupervised and supervised pipelines described in the methodology. The performance of each approach is analyzed in terms of class-wise accuracy and F1-score across the murmur shape categories: Plateau, Decrescendo, and Diamond. Mean metrics are also computed to provide an overall performance measure.

4.1 Unsupervised Pipeline

We begin by analyzing the performance of the unsupervised method using the *windowed power* time series as the basis for shape identification. Table 2 presents the performance of the four template variants (T1–T4). Among these, **Template 1 (T1)** emerges as the top performer, achieving a mean accuracy of 69.5% and a mean F1-score of 53.9%. Notably, T1 also achieves the highest F1-scores across all three murmur classes: Plateau (80.2%), Decrescendo (44.8%), and Diamond (36.8%).

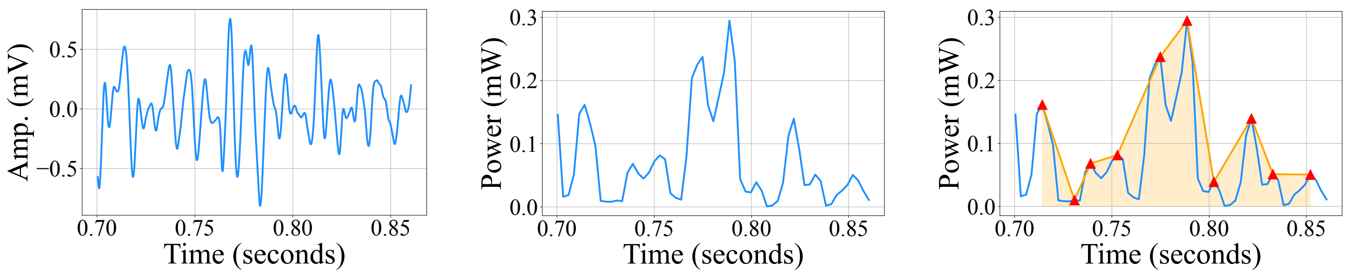
To improve shape resolution, we evaluated the *peaks of the windowed power* instead of the full power signal. Table 3 shows that this refinement leads to a consistent increase in mean F1-score across all templates. T1 again achieves the highest mean F1-score of 60.9%, while T3 achieves the highest mean accuracy of 69.0%. At the class level, the best F1-scores are observed with T3 (Plateau, 78.4%), T2 (Decrescendo, 54.1%), and T1 (Diamond, 56.6%).

In comparing the two strategies, we observe that using the peaks of the windowed power enhances the mean F1-scores by 7% (T1), 8% (T2), 17% (T3), and 2% (T4), respectively. Consequently, the best-performing unsupervised configuration is T1 using peak-based power features, with a mean F1-score of **60.9%**.



(a) Template 1 (b) Template 2 (c) Template 3 (d) Template 4

Figure 5: Templates used for murmur shape matching. Each template includes one Decrescendo (blue) and three Diamond (orange) profiles. Shapes are normalized by murmur duration and amplitude.



(a) Raw Murmur Signal (b) Windowed Power (c) Power Peaks

Figure 6: Windowed power and corresponding peaks for systolic murmur. Signal: Cycle 1 of Recording 9979_TV

	Plateau		Decrescendo		Diamond		Mean	
Templates	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
T1	**91.8**	**80.2**	**38.2**	**44.8**	23.2	**36.8**	**69.5**	**53.9**
T2	90.0	78.9	35.3	40.0	20.0	32.4	67.2	50.4
T3	91.2	79.7	14.7	23.8	13.3	22.9	67.2	42.1
T4	**91.8**	79.8	23.5	40.7	**23.3**	34.1	68.4	51.6

Table 2: Performance using Windowed Power

	Plateau		Decrescendo		Diamond		Mean	
Templates	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
T1	73.6	74.7	58.8	51.3	50.0	**56.6**	66.7	**60.9**
T2	65.6	70.2	**67.6**	**54.1**	50.0	51.7	63.2	58.7
T3	**82.7**	**78.4**	52.9	50.7	36.7	48.9	**69.0**	59.4
T4	66.4	70.5	47.1	48.9	**53.3**	42.7	60.3	53.9

Table 3: Performance using Peaks of Windowed Power

4.2 Supervised Pipeline

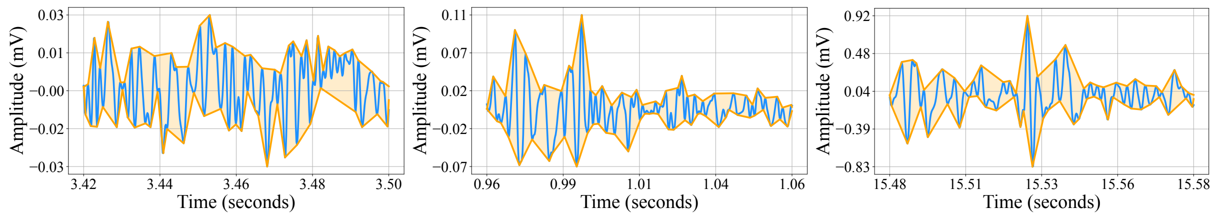
In Table 4 we summarize the results of four supervised classification models. The hybrid CNN+LSTM model performs reasonably well when trained on raw signal data (mean F1-score: 60.6%), but its performance deteriorates significantly when trained on signal envelopes (46.4%).

Conversely, the k-Nearest Neighbors (k-NN) classifier benefits from the use of envelope features, improving from a mean F1-score of 59.1% to a peak value of **66.2%**, making it the best-performing supervised model overall. The k-NN + envelope configuration also yields the highest class-wise F1-scores for Decrescendo (62.5%) and Diamond (62.5%).

4.3 Comparative Analysis and Discussion

To visualize the classification efficacy, Figure 8 compares the confusion matrices of the best unsupervised and supervised models. The supervised pipeline (k-NN with envelope) exhibits a stronger diagonal structure, with per-class accuracies ranging from 64% to 83%. In contrast, the unsupervised method (T1 with peaks of power) shows lower diagonal consistency (50%–73%).

While the supervised pipeline achieves a higher F1-score, it is inherently dependent on training data and subject to vari-

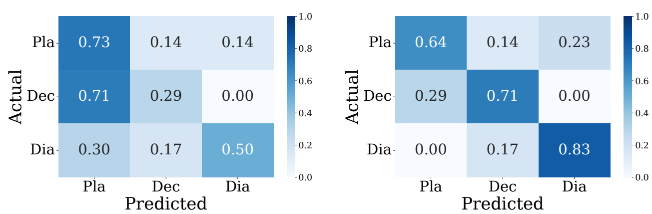


(a) Plateau (b) Decrescendo (c) Diamond

Figure 7: Upper envelopes extracted from systolic murmurs using local peaks and troughs. Representative cycles are shown for each murmur shape class.

	Plateau		Decrescendo		Diamond		Mean	
Learning Techniques	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
CNN LSTM	**68.2**	**75.0**	57.1	53.3	66.7	53.3	65.7	60.6
CNN LSTM + Env	54.5	58.5	28.6	30.8	66.7	50.0	51.4	46.4
kNN	63.6	70.0	57.1	57.1	66.7	50.0	62.9	59.1
kNN + Env	63.6	73.7	**71.4**	**62.5**	**83.3**	**62.5**	**68.6**	**66.2**

Table 4: Performance using Supervised Approaches



(a) Best Unsupervised (T1 + Peaks) (b) Best Supervised (k-NN + Env)

Figure 8: Confusion matrices of best-performing unsupervised and supervised pipelines. The supervised model demonstrates greater classification consistency.

ation with data distribution. In contrast, the unsupervised pipeline offers greater robustness and explainability by directly comparing murmur segments with shape-based templates. This explainability is especially valuable in clinical settings, where understanding the rationale behind predictions is critical for trust and adoption.

In summary, the supervised approach (mean F1: 66.2%) delivers superior performance, while the unsupervised pipeline (mean F1: 60.9%) provides a valuable balance between performance and explainability, making it suitable for scenarios with limited annotated data or where explainability is paramount.

5 Conclusion

Automated detection and classification of heart murmurs from phonocardiogram (PCG) signals hold significant promise for enabling large-scale, early screening of valvular heart diseases. The temporal shape of murmurs—categorized as Crescendo, Decrescendo, Diamond, or Plateau—is a medically validated diagnostic feature. Incorporating murmur shape enhances the clinical relevance and explainability of automated diagnostic tools. This work proposes an unsupervised, explainable pipeline that identifies murmur shapes by correlating pre-defined templates with the envelope of peak power values in the systolic phase. Among supervised models evaluated, the k-nearest neighbors (k-NN) classifier applied to envelope features achieved the highest mean F1-score of 66.2%. However, the unsupervised method, despite a slightly lower F1-score of 60.9%, provides better explainability through direct shape matching. This trade-off between accuracy and explainability highlights the potential of hybrid diagnostic systems that combine the predictive power of deep learning with the interpretive clarity of template-based methods. Furthermore, we plan to improve the performance of the unsupervised approach by integrating additional features such as pitch and spectral characteristics.

References

[Abduh *et al.*, 2020] Zaid Abduh, Ebrahim Ameen Nehary, Manal Abdel Wahed, and Yasser M Kadah. Classification of heart sounds using fractional fourier transform based mel-frequency spectral coefficients and traditional classifiers. *Biomedical Signal Processing and Control*, 57:101788, 2020.

[Alkhodari *et al.*, 2024] Mohanad Alkhodari, Leontios J Hadjileontiadis, and Ahsan H Khandoker. Identification of congenital valvular murmurs in young patients using deep learning-based attention transformers and phonocardiograms. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[El Badlaoui *et al.*, 2020] O El Badlaoui, A Benba, and A Hammouch. Novel pcg analysis method for discriminating between abnormal and normal heart sounds. *Irbm*, 41(4):223–228, 2020.

[Elola *et al.*, 2023] Andoni Elola, Elisabete Aramendi, Jorge Oliveira, Francesco Renna, Miguel T Coimbra, Matthew A Reyna, Reza Sameni, Gari D Clifford, and Ali Bahrami Rad. Beyond heart murmur detection: automatic murmur grading from phonocardiogram. *IEEE Journal of Biomedical and Health Informatics*, 27(8):3856–3866, 2023.

[Gardezi *et al.*, 2018] Syed KM Gardezi, Saul G Myerson, John Chambers, Sean Coffey, Joanna d'Arcy, FD Richard Hobbs, Jonathan Holt, Andrew Kennedy, Margaret Loudon, Anne Prendergast, et al. Cardiac auscultation poorly predicts the presence of valvular heart disease in asymptomatic primary care patients. *Heart*, 104(22):1832–1835, 2018.

[Liu *et al.*, 2016] Chengyu Liu, David Springer, Qiao Li, Benjamin Moody, Ricardo Abad Juan, Francisco J Chorro, Francisco Castells, José Millet Roig, Ikaro Silva, Alistair EW Johnson, et al. An open access database for the evaluation of heart sound algorithms. *Physiological measurement*, 37(12):2181, 2016.

[Lu *et al.*, 2022] 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. In *2022 Computing in Cardiology (CinC)*, volume 498, pages 1–4. IEEE, 2022.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[McDonald *et al.*, 2023] Andrew McDonald, Mark John Francis Gales, and Anurag Agarwal. A recurrent neural network and parallel hidden markov model algorithm to segment and detect heart murmurs in phonocardiograms. *medRxiv*, pages 2023–12, 2023.

[McGee, 2018] Steven McGee. Chapter 39 - auscultation of the heart: General principles. In Steven McGee, editor, *Evidence-Based Physical Diagnosis (Fourth Edition)*, pages 327–332.e1. Elsevier, Philadelphia, fourth edition edition, 2018.

[Nadkarni, 2016] Prakash Nadkarni. Chapter 10 - core technologies: Data mining and "big data". In Prakash Nadkarni, editor, *Clinical Research Computing*, pages 187–204. Academic Press, 2016.

[Ning *et al.*, 2016] Taikang Ning, Lin Cheng, and Kai-Sheng Hsieh. Identification and features extraction of systolic and diastolic murmurs. In *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, pages 271–276. IEEE, 2016.

[Oliveira *et al.*, 2021] Jorge Oliveira, Francesco Renna, Paulo Dias Costa, Marcelo Nogueira, Cristina Oliveira, Carlos Ferreira, Alípio Jorge, Sandra Mattos, Thamine Hatem, Thiago Tavares, et al. The circor digiscope dataset: from murmur detection to murmur classification. *IEEE journal of biomedical and health informatics*, 26(6):2524–2535, 2021.

[Peter, 2007] Libby ROB Peter. Braunwald's heart disease: a textbook of cardiovascular medicine. *(No Title)*, 2007.

[Prince *et al.*, 2023] John Prince, John Maidens, Spencer Kieu, Caroline Currie, Daniel Barbosa, Cody Hitchcock, Adam Saltman, Kambiz Norozi, Philipp Wiesner, Nicholas Slamon, et al. Deep learning algorithms to detect murmurs associated with structural heart disease. *Journal of the American Heart Association*, 12(20):e030377, 2023.

[Reyna *et al.*, 2023] Matthew A Reyna, Yashar Kiarashi, Andoni Elola, Jorge Oliveira, Francesco Renna, Annie Gu, Erick A Perez Alday, Nadi Sadr, Ashish Sharma, Jacques Kpodonu, et al. Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022. *PLOS Digital Health*, 2(9):e0000324, 2023.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Shrikumar *et al.*, 2017] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[Spadaccini and Beritelli, 2013] Andrea Spadaccini and Francesco Beritelli. Performance evaluation of heart sounds biometric systems on an open dataset. In *2013 18th International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2013.

[Tsao *et al.*, 2023] Connie W Tsao, Aaron W Aday, Zaid I Almarzooq, Cheryl AM Anderson, Pankaj Arora, Christy L Avery, Carissa M Baker-Smith, Andrea Z Beaton, Amelia K Boehme, Alfred E Buxton, et al. Heart disease and stroke statistics—2023 update: a report from the american heart association. *Circulation*, 147(8):e93–e621, 2023.

[Walker HK, 1990] Hurst JW Walker HK, Hall WD, editor. *Chapter 26 -Systolic Murmurs*. Boston: Butterworths, 1990.

[Wang *et al.*, 2023] Zhihua Wang, Kun Qian, Houguang Liu, Bin Hu, Björn W Schuller, and Yoshiharu Yamamoto. Exploring interpretable representations for heart sound abnormality detection. *Biomedical Signal Processing and Control*, 82:104569, 2023.