NOISE-ROBUST CONTRASTIVE LEARNING WITH AN MFCC-CONFORMER FOR CORONARY ARTERY DISEASE DETECTION

Milan Marocchi, Matthew Fynn, Yue Rong

Curtin University, Bentley 6102, WA, Australia

ABSTRACT

Cardiovascular diseases (CVD) are the leading cause of death worldwide, with coronary artery disease (CAD) comprising the largest subcategory of CVDs. Recently, there has been an increased focus on the detection of CAD using phonocardiogram (PCG) signals, with high success in clinical environments with low noise and optimal sensor placement. Multichannel techniques have been found to be more robust to noise, however, there are still challenges in achieving robust performance and practical real-world performance. This work utilises a novel energy-based noisy-segment rejection algorithm for the discarding of audio segments with large amounts of non-stationary noise, before training a deep learning classifier. This conformer-based classifier takes mel-frequency cepstral coefficients (MFCCs) from multiple channels, further helping improve the models noise robustness. The proposed method achieved 78.4% accuracy and 78.2% balanced accuracy on 297 subjects, representing improvements of 4.1% and 4.3%, respectively, compared to training without noise-segment rejection.

***Index Terms*—** Contrastive Learning, Noise-robust audio classification, Deep Learning, Phonocardiograms analysis, Coronary artery disease detection

1. INTRODUCTION

Cardiovascular disease (CVD) result in 31% of deaths annually around the globe [1]. Coronary artery disease (CAD) is the largest subtype. CAD requires prompt diagnosis to help manage the disease before it progresses. However, auscultation yields relatively low diagnostic accuracy, partly because heart sounds often lie near the threshold of human hearing [2, 3, 4]. With expensive and highly invasive angiograms being the gold-standard diagnostic tool [5], there is a gap in prescreening tools. Recently deep learning aided phonocardiogram (PCG) methods have been employed to accuractely pre-screen CAD [6, 7]. It has been found that multichannel PCG signals help improve CAD classification performance in the presence of background hospital noise [8]. The use of linear frequency cepstral coefficients (LFCCs) and mel-frequency cepstral coefficients (MFCCs) have further been found to improve the noise robustness of deep learning models [6]. Modern strong performing transformer-based architectures such as conformers have seen success in speech but have not yet been evaluated in PCG signals [9]. They have been found to lead to state-of-the-art (SOTA) performance for speech in clean and noisy conditions, showing promise for use on noisy PCG data. However, there is currently a lack of work utilising all these techniques, along with noisy segment removal, based on external and internal noise, on a real-world noisy dataset to validate all of these approaches [10]. This study makes use of data collected in a noisy hospital environment with unoptimal sensor placement, to evaluate the performance of noise-segment rejection, along with a conformer-based model that uses MFCCs as input.

This work's novel contributions to the field are:

- An energy-based noisy segment rejection algorithm leveraging multichannel PCG data and built-in noise reference microphones
- Use of contrastive learning for an MFCC conformer-based model to improve balanced performance with noisy data

2. MATERIALS

All data processing and model training were conducted using a Ryzen 7 3800X CPU and an Nvidia RTX 3090 (24 GB), with Python 3.11 and PyTorch 2.1.2.

2.1. Data Aquistion

A wearable vest embedded with multiple PCG sensors was used to acquire synchronised multichannel PCG data from participating subjects [6]. Each stethoscope incorporated two microphones: one positioned beneath the diaphragm (Heart mic – HM) and another on the rear of the stethoscope (Noise mic – NM). The vest can be fitted easily, requiring less than a minute. This work made use of channels 1, 2, 3 and 4 of the seven PCG channels.

2.2. Dataset

The wearable vest collected data from 297 male subjects at Fortis Hospital, Kolkata, across three separate rounds: May–June 2023, January–February 2024, and February 2025.

Of these, 155 subjects were diagnosed with CAD, defined as having greater than 50% stenosis in the right coronary artery, left coronary artery, or left circumflex artery, confirmed through angiography. The remaining 142 subjects were classified as normal, including 32 subjects under 35 years of age, assumed to be free of CAD as the risk is significantly higher in males above 45 years [11]. Data was collected in a clinical environment, and thus typical hospital background noise was present, including talking, privacy curtains closing, and doors slamming. Subjects were instructed to sit comfortably on a chair and breathe normally during recording. Between one and three 60-second recordings were acquired from each subject.

3. METHOD

Segments of audio from the PCG signals are extracted and preprocessed before being used to train a conformer-based classifier with a contrastive loss. The methods will first detail the novel energy-based noisy segment rejection approach, preprocessing, and feature extraction before detailing the model training and inference.

3.1. Preprocessing

The PCG signals first are concatenated so that there is one contiguous recording for each subject. This will ensure that there is no data leakage. The regions around the joins are then discarded when segmenting the signal, which is further discussed in Section 3.3. Following this, the signals undergo noisy segment rejection, which will mark noisy segments so that they will not be included in any fragments. Following this, the signals undergo spike removal [12] and then are bandpassed using a second-order Butterworth filter between 25Hz and 450Hz. The signals are then k-peak mean normalised [7], before being used to extract features. Then the signals are segmented into fragments to be used for training.

3.2. Noisy Segment Rejection

Both HM and NM signals were utilised for noisy segment rejection. We propose an algorithm that identifies and mitigates impulsive and movement noise within the recordings. The algorithm outputs the set of indices deemed corrupted by impulse noise, enabling clean signal segments to be used in downstream training and inference. The algorithm takes a signal (either from HM or NM) as input and divides it into frames of fixed length. For each frame, the energy is computed as the sum of squared samples. The median frame energy (excluding the first and last frames) is then calculated. Each frame whose energy exceeds the product of the median value and the given threshold is flagged, and the corresponding start and end indices of that frame are stored in a variable. Additionally, the first and last seconds of each recording are flagged as noisy, ensuring boundaries between concatenated signals are excluded from downstream tasks, including filtering. Algorithm 1 describes this process of highlighting noisy segment indices. Sources of identifiable noise included sudden bursts of external voices and door slams, while patient movement introduced friction noise between the diaphragm and the skin. As the NM signals from all stethoscopes were highly correlated, only channel 4 was used to detect noisy indices, whereas each HM was processed separately. The frame length was tailored to the dominant noise source of each channel: for HM signals, it was set to 2.5 s to capture longer-duration friction noise, while for NM signals, it was set to 0.25 s to detect brief impulsive events such as door slams or speech bursts. For both signal types, the threshold was fixed at 2.5 times the median frame energy, chosen to balance sensitivity to noise events against robustness to natural signal variability. Figure 1 illustrates an example HM and NM signal with noisy indices highlighted in red. The resulting indices from all HM and NM signals were then combined to form a final vector of noise-corrupted segments across the concatenated recording, accounting for overlapping indices highlighted by separate channels. From this combined output, the complementary indices corresponding to noise-free segments were extracted and applied uniformly across the entire multichannel recording.

Algorithm 1 Noisey Segment Identification

Require: Signal $x[0 \dots L-1]$, sampling rate f_s, frame length (s) T_f, threshold τ
Ensure: List of index intervals $\mathcal{I}$ containing frames flagged as noise

$N \leftarrow \lfloor L/(T_f \cdot f_s) \rfloor$ ▷ number of full frames
$F \leftarrow T_f \cdot f_s$ ▷ samples per frame
$\mathbf{E} \leftarrow \mathbf{0}_{1\times N}$ ▷ frame energies
for $i = 0$ to $N-1$ **do**
  $s \leftarrow iF; \quad e \leftarrow (i+1)F - 1$
  $\mathbf{E}[i] \leftarrow \sum_{n=s}^{e} x[n]^2$ ▷ sum of squares (energy)
end for
$m \leftarrow \mathrm{median}\big(\mathbf{E}[1{:}N-2]\big)$ ▷ exclude first/last frame
$\mathcal{I} \leftarrow [\,]$ ▷ empty list of (start,end) indices
for $i = 1$ to N **do**
  if $\mathbf{E}[i] > \tau \cdot m$ **then**
    $s \leftarrow iF; \quad e \leftarrow (i+1)F - 1$
    append (s, e) to $\mathcal{I}$
  end if
end for
return $\mathcal{I}$

3.3. Segmentation

Following denoising, the noise-free indices of each subject's recording were used to extract clean segments, whose lengths varied according to the distribution of noise. Segments shorter than four seconds were discarded. From the remaining segments, fixed-length fragments of four seconds were extracted. To ensure class balance during training, a base number of fragments F_{base} was specified per class, with additional fragments drawn from the underrepresented class

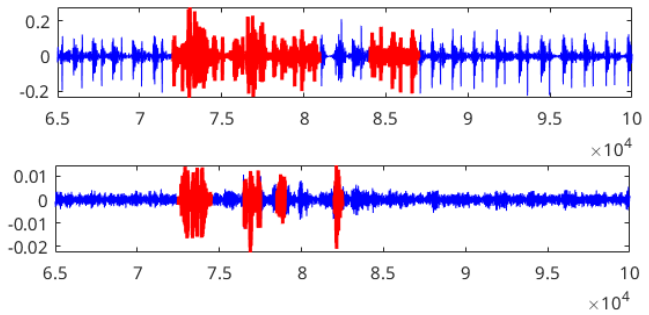

Fig. 1. Zoomed-in HM (top) and NM (bottom) signals with noise-corrupted segments highlighted in red. These segments were discarded from all channels during downstream training and inference.

until equal counts were achieved. During validation and testing, the target number of fragments was fixed across classes to avoid bias.

For a subject in class c, the assigned number of fragments F_{class} was distributed across its noise-free segments in proportion to segment duration. Specifically, for the i^{th} noise-free segment,

$$F_i = \left\lfloor F_{\text{class}} \cdot \frac{L_i}{\sum_j L_j} \right\rfloor, \tag{1}$$

where L_i is the length of the i^{th} segment and $\sum_j L_j$ is the total length of all valid segments for that subject. Any remaining fragments were allocated to the longest segments to preserve proportionality (i.e $\sum_i F_i = F_{\text{class}}$). Fragments were then extracted with variable overlap, determined by both F_i and the length of the corresponding segment.

3.4. Feature Extraction

MFCCs were extracted from each recording segment following amplitude normalisation to mitigate inter-recording variability. The normalisation conducted was k-peak normalisation, which has been shown to be more effective for PCG signals [7]. For each segment, the short-time Fourier transform (STFT) was computed, mapped to the mel scale, and subsequently transformed into cepstral coefficients. There were 128 MFCCs extracted between 25Hz and 450Hz with a window length of 512 samples, hop length of 160 samples for the STFT computation. The MFCC vectors obtained from individual segments for each channel were then concatenated along the time channel axis to create a unified representation of all features from each channel.

3.5. Models

The proposed network is a conformer-style encoder operating on MFCC sequences. After per-segment MFCC preprocessing (single- or multi-channel; features concatenated across channels), a linear projection maps the input F-dimensional frames to the model width D. The encoder comprises B stacked conformer blocks, each following a layered topology: a pre-normalized feed-forward sublayer (scaled by 0.5), multi-head self-attention with H heads, a convolutional module, and a second pre-normalised feed-forward sublayer of dimension M (again scaled by 0.5), with residual connections throughout. The convolutional module employs a pointwise expansion with gated linear units (GLU), depthwise convolution (kernel size k, which was fixed to 31), batch normalisation, SiLU activation, and a pointwise projection back to D. Layer normalisation is applied before attention and convolutional sublayers, and dropout is used in the feed-forward paths. A final layer normalisation precedes temporal aggregation via adaptive average pooling to produce an embedding, which is fed to a shallow MLP classifier (one hidden layer with ReLU and dropout) to predict one of the two classes.

Parameter	Value	Parameter	Value
batch size (N_b)	256	α	0.7235
learning rate	2.97e-06	β	0.9807
weight decay	5.71e-05	temperature	0.8050
s	2	λ_c	0.00281
γ	0.2903	D	1024
H	8	M	128
B	3	dropout	0.2903

Table 1. Hyperparameter values.

3.6. Contrastive Learning

3.6.1. Hybrid-Contrastive Loss

A hybrid-contrastive loss is utilised to best shape the embedding space to ensure more robust classifications, especially in the case of noise, which is typical within this dataset. It is a supervised loss function to best make use of the data being fully labelled. The training objective combines three components: a supervised contrastive loss, a classification loss, and an optional center loss. Given a batch of feature embeddings $\mathbf{z}_i \in R^d$ with corresponding class labels y_i, the total loss is defined as

$$\mathcal{L} = \beta\, \mathcal{L}_{\text{contr}}(\mathbf{z}, \mathbf{y}; \tau) + \alpha\, \mathcal{L}_{\text{CE}}(\mathbf{p}, \mathbf{y}) + \lambda_c\, \mathcal{L}_{\text{center}}(\mathbf{z}, \mathbf{y}), \tag{2}$$

where $\mathbf{p}$ are the classifier logits, α and β weight the classification and contrastive terms, respectively, and λ_c controls the influence of the center loss. The hyperparameter τ denotes the temperature. A standard cross-entropy objective is used for the classification loss ($\mathcal{L}_{\text{CE}}$).

3.6.2. Supervised Contrastive Loss

We normalise all embeddings and compute a cosine similarity matrix $\mathbf{S} = \hat{\mathbf{z}}\hat{\mathbf{z}}^\top$. The contrastive loss encourages embeddings of the same class to be close, and embeddings of different classes to be pushed apart:

$$\mathcal{L}_{\text{contr}} = -\frac{1}{N_{mb}} \sum_{i=1}^{N_{mb}} \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{\exp(\mathbf{S}_{ij}/\tau)}{\sum_{k=1}^{N_{mb}} \exp(\mathbf{S}_{ik}/\tau)}, \tag{3}$$

Table 2. Model performance at the fragment and subject level - Base number of fragments is 61

Method	Acc	UAR	TPR	TNR	$F1^+$	$F1^-$	MCC
			Fragment Level				
Noisy MFCC Conformer	71.2±0.05%	70.9±0.05%	77.5±0.07%	64.2±0.17%	73.9±0.02%	67.5±0.09%	0.425±0.009
Denoised MFCC Conformer	**73.9±0.35%**	**73.7±0.37%**	76.8±0.42%	**70.7±0.76%**	75.5±0.32%	**71.9±0.44%**	**0.478±0.072**
Noisy Wav2Vec 2.0 [13]	70.7±0.21%	68.2±0.18%	**78.9±0.46%**	57.6±0.30%	**76.8±0.22%**	59.9±0.22%	0.372±0.004
			Subject Level				
Noisy MFCC Conformer	74.3±0.09%	73.9±0.10%	80.9±0.11%	66.9±0.30%	76.8±0.06%	70.6±0.15%	0.490±0.019
Denoised MFCC Conformer	**78.4±0.29%**	**78.2±0.32%**	81.9±0.49%	**74.5±0.97%**	79.9±0.20%	**76.4±0.48%**	**0.570±0.058**
Noisy Wav2Vec 2.0 [13]	77.1±1.50%	74.3±1.73%	**86.5±1.30%**	62.0±2.76%	**82.3±1.10%**	67.1±2.56%	0.510±0.035

where $\mathcal{P}(i)$ is the set of positive indices (samples with $y_j = y_i$, excluding i itself), and N_{mb} is the number of samples within the mini batch.

3.6.3. Center Loss

For each class c, a learnable center vector $\mathbf{c}_c \in R^d$ is maintained. The center loss penalises the distance between feature vectors and their class centers:

$$\mathcal{L}_{\text{center}} = \frac{1}{N_{mb}} \sum_{i=1}^{N_{mb}} \|\mathbf{z}_i - \mathbf{c}_{y_i}\|_2^2. \quad (4)$$

3.7. Model Training

Training is done on a fragment level to optimise the fragment-level metrics. The model was trained using the AdamW optimiser [14]. An exponential decay learning rate scheduler was also utilised parameterised by the step size (s) and the decay rate (γ). The model makes use of gradient accumulation with a mini batch size (N_{mb}) and a batch size (N_b), where the gradients from each mini batch update are accumulated until the number of samples sum to the N_b. The models are trained for 10 epochs, with the best model from training being selected as the one with the best weighted average Matthew's Correlation Coefficient (MCC) between the training and validation set; a scaling factor of 0.9 to the validation MCC and 0.1 to the training MCC. The MCC metric provides a single measure that captures all aspects of model performance [15]. For tuning the hyperparameters and the architecture of the model, a Bayesian optimisation was conducted using the Optuna library [16]. Table 1 contains the parameters included in this optimisation. Each trial was repeated three times and was optimised over the average validation MCC score to ensure a less noisy value being used.

3.8. Model Inference

Before being used for inference, the MLP is removed and replaced with a support vector machine (SVM) with a radial basis function (RBF) kernel. The subject-level predictions are then taken by majority vote of each of the fragment-level predictions. The accuracy, unweighted average recall (UAR), true positive rate (TPR), true negative rate (TNR), F1 scores, and MCC are reported.

4. RESULTS AND DISCUSSION

Table 2 displays the fragment and subject performance which compares the baseline with no noise-segment rejection to a model that was trained with the contrastive loss and the signals denoised. These results are presented as average±standard deviation, where the models are averaged over the five folds and run three times to account for the stochasticity of the training of the neural networks. The table also contained a comparison to a previous method on this same vest data, with the best method for each metric being highlighted in bold.

The proposed noise-segment rejection algorithm, when applied in conjunction with the conformer-based model, yielded subject-level improvements of 4.1%, 4.3%, and 0.08 in accuracy, UAR, and MCC, respectively. The model's performance was also significantly more balanced between TPR and TNR with the use of the noise-segment rejection, highlighting the importance of removing segments heavily contaminated with non-stationary noise. Comparing this method to another work utilising data from the same vest, it is seen that this method results in a more performant and noise robust model, with subject-level increases of 1.3%, 3.9% and 0.06 in the accuracy, UAR and MCC, respectively. This confirms that the use of these techniques to deal with a noisy real-world dataset help to improve performance, whilst also providing a model which is signifigicantly smaller than the one in [13]. It is much smaller as an early feature fusion is employed, as opposed to a late feature fusion.

5. CONCLUSION AND FURTHER WORK

This work presented a CAD classifier that is more robust to the noise of a real-world dataset by employing a noise-segment rejection algorithm, the use of MFCC features and a conformer architecture trained with a hybrid contrastive loss. This model performed better than a previous method which utilised multiple Wav2Vec 2.0 feature extractors. Further work includes the use of augmentations coupled with the contrastive learning method to improve the robustness of the model, along with testing on out-of-distribution datasets to assess how the preprocessing and model generalise to other datasets that the model is not trained on.

Ethics approval: — This study received approval from the ethics committee of Fortis Hospital, Kolkata, India, where the data collection took place (ECR/240/Inst/WB/2013/RR-19, Date of approval: 13/01/2023) in accoradance to the Helsinki Declaration. Informed consent was obtained from all subjects.

Acknowledgment: — We would like to thank Ticking Heart Pty Ltd for providing the wearable vest design, and Fortis Hospital Kolkata and Dr. Mandana for their support in data collection.

6. REFERENCES

[1] WHO, "Cardiovascular Diseases (CVDs)". *Geneva, Switzerland: WHO*, 2021.

[2] M. A. Chizner, "Cardiac auscultation: Rediscovering the lost art," *Current Problems in Cardiology*, vol. 33, no. 7, pp. 326–408, Jul. 2008.

[3] C. A. Feddock, "The Lost Art of clinical skills," *The American Journal of Medicine*, vol. 120, no. 4, pp. 374–378, Apr. 2007.

[4] Q.-M. Zhao, C. Niu, F. Liu, L. Wu, X.-J. Ma, and G.-Y. Huang, "Accuracy of cardiac auscultation in detection of neonatal congenital heart disease by general paediatricians," *Cardiology in the Young*, vol. 29, no. 5, pp. 679–683, May 2019.

[5] R. J. Gibbons, K. Chatterjee, J. Daley, J. S. Douglas, S. D. Fihn, J. M. Gardin, M. A. Grunwald, D. Levy, B. W. Lytle, R. A. O'Rourke, W. P. Schafer, S. V. Williams, J. L. Ritchie, R. J. Gibbons, M. D. Cheitlin, K. A. Eagle, T. J. Gardner, A. Garson, R. O. Russell, T. J. Ryan, and S. C. Smith, "Acc/aha/acp-asim guidelines for the management of patients with chronic stable angina1: A report of the american college of cardiology/american heart association task force on practice guidelines (committee on management of patients with chronic stable angina)," *Journal of the American College of Cardiology*, vol. 33, no. 7, pp. 2092–2197, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0735109799001503

[6] M. Fynn, K. Mandana, J. Rashid, S. Nordholm, Y. Rong, and G. Saha, "Practicality meets precision: Wearable vest with integrated multi-channel pcg sensors for effective coronary artery disease pre-screening," *Computers in Biology and Medicine*, vol. 189, p. 109904, 2025.

[7] A. Maity and G. Saha, "Enhancing cross-domain robustness in phonocardiogram signal classification using domain-invariant preprocessing and transfer learning," *Computer Methods and Programs in Biomedicine*, vol. 257, p. 108462, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260724004553

[8] A. Pathak, P. Samanta, K. Mandana, and G. Saha, "An improved method to detect coronary artery disease using phonocardiogram signals in noisy environment," *Applied Acoustics*, vol. 164, p. 107242, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X19305742

[9] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020. [Online]. Available: https://arxiv.org/abs/2005.08100

[10] Z. Ren, Y. Chang, T. T. Nguyen, Y. Tan, K. Qian, and B. W. Schuller, "A comprehensive survey on heart sound analysis in the deep learning era," 2023.

[11] R. Hajar, "Risk factors for coronary artery disease: historical perspectives," *Heart views*, vol. 18, no. 3, pp. 109–114, 2017.

[12] S. E. Schmidt, C. Holst-Hansen, J. Hansen, E. Toft, and J. J. Struijk, "Acoustic features for the identification of coronary artery disease," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 11, pp. 2611–2619, Nov. 2015.

[13] M. Marocchi, M. Fynn, K. Mandana, and Y. Rong, "Scaling to multimodal and multichannel heart sound classification: Fine-tuning wav2vec 2.0 with synthetic and augmented biosignals," 2025. [Online]. Available: https://arxiv.org/abs/2509.11606

[14] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[15] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 6, 2020.

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.