Ensemble Transformer-Based Neural Networks Detect Heart Murmur In Phonocardiogram Recordings

Mohanad Alkhodari[1], Syafiq Kamarul Azman[2], Leontios J. Hadjileontiadis[1,3], Ahsan H. Khandoker[1]

[1]Healthcare Engineering Innovation Center (HEIC), Department of Biomedical Engineering, Khalifa University, Abu Dhabi, United Arab Emirates
[2]AIQ, ADNOC H.Q., Abu Dhabi, United Arab Emirates
[3]Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece

Abstract

Cardiac auscultation through phonocardiogram (PCG) is still the most commonly used approach for evaluating the mechanical functionality of the heart when diagnosing congenital heart disease. Despite of its time- and cost-effectiveness, it is still limited due to the extensive need for clinical expertise for interpretation. In this study, we propose the use of ensemble transformer-based neural networks to aid in the detection of heart murmur in PCG recordings and for the prediction of clinical outcomes of patients as part of George B. Moody PhysioNet 2022 Challenge. Our team, Care4MyHeart, developed an approach that transforms the raw PCG recordings into wavelet power features signals for the use within the proposed deep learning models. We have achieved a maximum accuracy of 0.855, 0.761, and 0.757 for murmur detection in the training, hidden validation, and hidden testing datasets, respectively. In addition, we had an overall clinical outcome cost of 9980, 11490, and 14410 for the three datasets, respectively. Our team was ranked 6th/40 for murmur detection and 29th/40 for clinical outcome predictions. We had the lowest clinical outcome cost on the validation set of 9737 with a murmur detection score of 0.730 when reducing the number of features used to train the models.

1. Introduction

Congenital heart disease is a birth defect in the structure of the heart that affects nearly 1.2% of newborns worldwide [1]. Early assessment through non-invasive cardiac auscultation devices, i.e., phonocardiogram (PCG), can unveil fundamental information on the mechanical malfunctioning caused by this disease [2, 3]. Although PCG is a time- and cost-effective tool, it still lacks proper interpretation owing to the need for experienced practitioners, thus, it has a limited diagnostic sensitivity [4].

To address this concern, we propose a simple, yet effective, deep learning approach based on attention transformers to detect cardiac malfunction in PCG recordings, which are usually represented as heart murmurs or abnormal sound patterns. We used patient data from The George B. Moody PhysioNet 2022 Challenge [5, 6] with an objective to provide a prediction to whether a patient has absent, present, or unknown murmurs plus identifying the clinical outcomes as either abnormal or normal.

2. Methods

2.1. Data preparation

The challenge data included 1568 patients from the pediatric population, out of which 942 were released as training while the remaining were hidden for validation/testing. Each patient had one or more PCG recordings from four auscultation locations, i.e., aortic valve (AV), mitral valve (MV), pulmonary valve (PV), tricuspid valve (TV), and other (Phc). All recordings were taken sequentially and not simultaneously.

We started our approach by arranging all patient data to include four channels arranged as AV, MV, PV, and TV. If a patient had less than four channels, the previous channel was duplicated once or multiple times as needed. Then, we selected the first 40 seconds from each recording. Many signals were recorded for shorter duration, therefore, we padded them by duplicating the same signal to reach up to 40 seconds. At the end, we apply z-score normalization for the four-channel recording.

2.2. Transformation to wavelet features

Initially, we transformed each PCG signal into wavelet transform-based power features (Fig. 1). For each signal, we used a 32-sample wide sliding window with a shifting interval of 32 samples to split the signal into 5000 segments. The selection of the segment length was based on

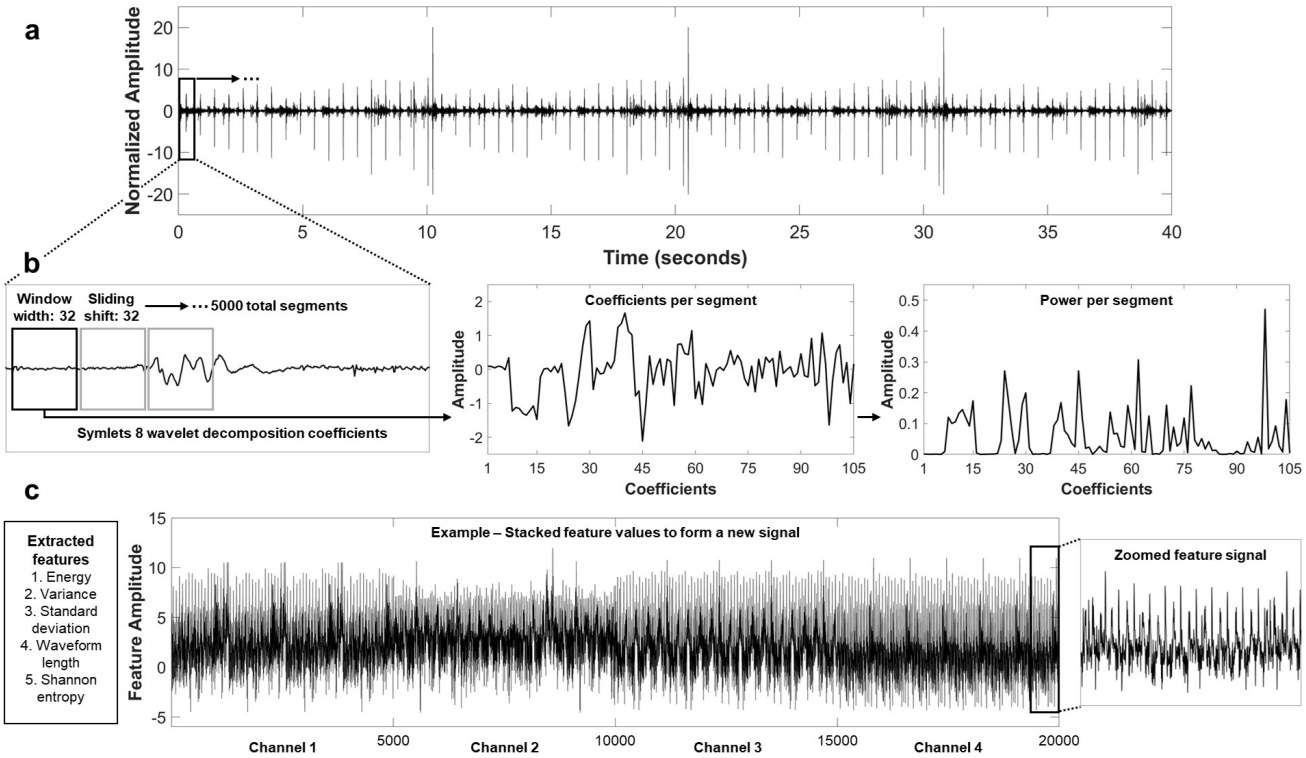

Figure 1. Transformation of the four-channel phonocardiogram (PCG) recording into 30 stacked power features using wavelet transform decomposition. a) An example of a selected channel from the 40-second PCG recording, b) Transformation of the signal to power using sliding windows and Symlets wavelet decomposition, and c) An example of the stacked feature values to form a new signal per extracted feature.

trial and error to ensure better transformation. For each 32-sample segment, we applied wavelet transform decomposition using the Symlets 8 (Sym8) wavelet to decompose each segment into 6 levels. Then, we converted the decomposed signal into 105 concatenated approximation coefficients. Using these coefficients, we calculated the power by taking the square of the absolute value of each coefficients vector. Lastly, we calculated five features from the coefficients vector, namely the energy (summation of values), variance, standard deviation, waveform length, and Shannon entropy yielding a total of 30 features per 32-sample segment. Since we had four-channel PCG recording, we concatenate features of each channel to form a wavelet power transformation of the raw PCG recording with 30 features (6 levels times 5 features) and overall length of 20000 (4 channels time 5000 32-sample segments) to form the deep learning input.

2.3. Transformer-based neural network

We designed our neural network architecture based on the most recent attention transformers, which are usually used in advanced natural language processing (NLP) applications [7]. Our proposed model comprises of four main elements, namely the feature encoder, positional encoder, transformer unit, and decoder (Fig. 2).

The network starts by encoding the features using two sets of one-dimensional (1D) convolutions followed by batch normalization, Gaussian error linear unit (GeLU), and max pooling layers. The parameters for all layers, if any, are provided in Fig. 2. These sets encode the input data (30x20000) by extracting important features and reducing its dimension to 30x1250. Next, a positional encoder assigns a unique representation for each sequence using the regular structure of sequential sin and cosine functions before entering the transformer unit. In the transformer, the multi-head attention (key (K), query (Q), and value (V)) with a scaled dot product mechanism is applied on the input. For simplicity, we used a single transformer unit and two heads of features with a 20% dropout rate. The extracted attention vector obtained from the transformer passes through the decoder, which includes a set of two fully connected layers followed by two dropout layers of 5% rate and a rectified linear unit (ReLU). Lastly, we used a global average pooling layer to pool over the temporal sequence and obtain a single value per vector.

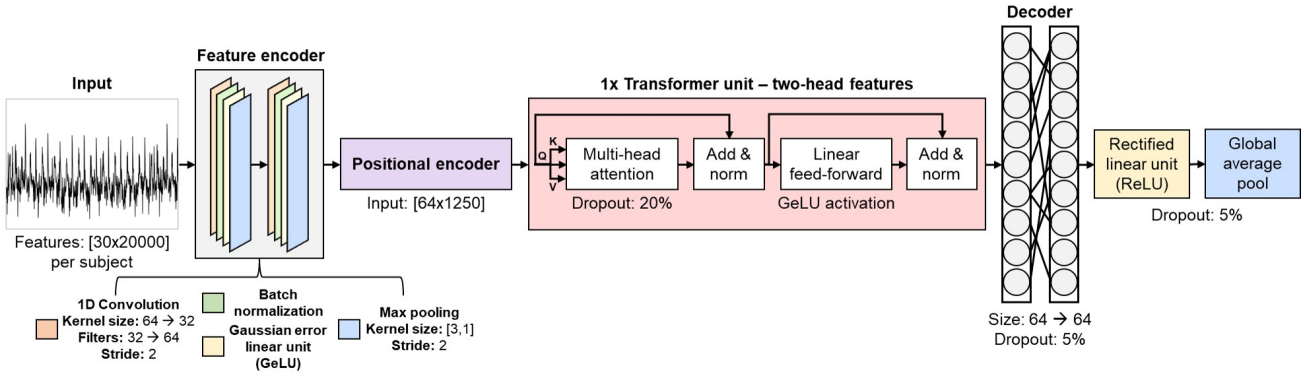

Figure 2. The proposed transformer-based neural network architecture including the four main elements, namely the feature encoder, positional encoder, transformer unit, and decoder.

2.4. Training and classification

2.4.1. Detection of heart murmur

The training data for the detection of heart murmur had a severe imbalance between absent (73.8%), present (19.0%), and unknown (7.2%) classes. Therefore, we applied a weighted cross-entropy loss function during training which was calculated empirically for each class. We optimized the training using the adaptive moment estimation (ADAM) solver with an initial learning rate of 0.001 and a learning rate drop of 0.1% at two-thirds the whole training interval of 90 epochs.

2.4.2. Identification of clinical outcome

We trained the model for another time to predict the clinical outcome of patients (abnormal or normal). We used the same imbalance handling method, however, it was slightly unbalanced with 48.4% for abnormal and 51.6% for normal. We trained the model for a maximum of 60 epochs with a 0.1% drop in learning rate at the 40th epoch.

2.4.3. Additional neural network

As an additional approach to handle the imbalance in the training dataset, we extracted the trained features from the last layer of the transformer network, i.e., the global average pooling features. Then, we applied the synthetic minority over-sampling technique (SMOTE) [8] to oversample the small classes in the murmur detection task (present: 1500, unknown: 1500) and both classes in the clinical outcome task (abnormal: 3500, normal: 3500) using a safe-level mechanism. Next, we trained a simple neural network with the over-sampled classes using the same transformer network's training settings.

For the murmur detection task, the model was trained for 200 epochs and with a learning rate drop of 1% at epoch 180. On the other hand, the clinical outcome model was trained using the features extracted from the transformer network of the murmur task for 120 epochs with a similar learning rate drop at the 100th epoch.

2.4.4. Ensemble classification

We followed ensemble classification as a voting mechanism to enhance the performance for both tasks. We used four scenarios using both the transformer-based and SMOTE-based neural networks as follows,

- **Scenario 1:** 1 transformer – 1 SMOTE networks
- **Scenario 2:** 3 transformers – 3 SMOTE networks
- **Scenario 3:** 5 transformers – 5 SMOTE networks
- **Scenario 4:** 10 transformers – 10 SMOTE networks

The classification task was based on a voting mechanism, where we averaged all scores using all networks at every scenario to generate the final score for every class for every subject.

3. Results

We evaluated our proposed approach first using the training dataset through a 10-fold cross-validation scheme. Then, the scenarios were tested as entries submitted for evaluation on the hidden validation dataset during the official phase.

Table 1 shows the performance of our approach for the four scenarios. The accuracy in murmur detection increased steadily from scenario 1 to scenario 4, that is by adding more trained networks in the ensemble (voters). A similar trend was also observed for the clinical outcome in the form of a decreased cost. The highest achieved accuracy was 0.855 and 0.761 for the training and validation datasets, respectively, in murmur detection. On the other hand, the lowest costs in clinical outcome prediction was 9980 and 11490 in both datasets, respectively.

Scenarios	Training		Validation	
	Murmur	Outcome	Murmur	Outcome
1	0.821	12785	0.734	15696
2	0.835	10890	0.747	12027
3	0.847	10341	0.754	11569
4 (Best Entry)	**0.855**	**9980**	**0.761**	**11490**
1 (Less features)	0.811	**9530**	0.730	**9737**

Table 1. Challenge scores for our team's (Care4MyHeart) entries on the training set and hidden validation set.

	Mumur	Outcome
Score	0.757	14410
Rank	6th/40	29th/40

Table 2. Final challenge scores and ranks for our team's (Care4MyHeart) best entry (scenario 4) on the testing set.

It is worth noting that we have achieved our lowest clinical outcome prediction score, 9737, for the validation dataset in one of the entries after reduction in input features (from 30 to 10) in models of scenario 1. The reduction of features was based on the chi-squared (χ^2) test. However, it had a low murmur detection accuracy of 0.730.

In the final testing phase (Table 2), our team achieved a murmur detection score of 0.757 and a clinical outcome prediction score of 14410. We were ranked the 6th/40 and the 29th/40 in both tasks, respectively.

4. Discussion and Conclusions

In this study, we evaluated the utilization of transformer-based neural networks for the detection of heart murmur and clinical outcome prediction. We have achieved a high level of performance when using ensemble classifiers, that is by using more voters. This approach acts as different experienced doctors evaluating the case with smoother decisions by averaging their diagnosis assessment to obtain as accurate conclusions as possible.

Although we have had high murmur detection accuracy using 10 transformers and 10 SMOTE networks, one would prefer trading off the high accuracy for reducing model complexity and computational demands, although not considered high for this task. This comes with the knowledge based on this study that there was a slight error between the scenarios (almost 1%).

From the findings using the reduced 10 features, the clinical outcome prediction was the lowest, which could suggest that a simpler model could predict better the clinical outcomes. We did not test this case using the four scenarios, which is going to be one of the future works. In addition, the learned parameters during training the models should be further investigated, especially with the ability to extract the attention from the transformer unit. This kind of interpretation is needed to add more explainability to the proposed approach when using it in clinical practice.

Acknowledgments

This work was supported by the Healthcare Engineering Innovation Center (HEIC), Khalifa Univeristy, UAE (Grant: 8474000132).

References

[1] Wu, Weiliang and He, Jinxian and Shao, Xiaobo. Incidence and Mortality Trend of Congenital Heart Disease at the Global, Regional, and National Level, 1990–2017. Medicine 2020;99(23).

[2] Alkhodari, Mohanad and Fraiwan, Luay. Convolutional and Recurrent Neural Networks for the Detection of Valvular Heart Diseases in Phonocardiogram Recordings. Computer Methods and Programs in Biomedicine 2021;200:105940.

[3] Alkhodari, Mohanad and Jelinek, Herbert F et al. Estimating Left Ventricle Ejection Fraction Levels Using Circadian Heart Rate Variability Features and Support Vector Regression Models. IEEE Journal of Biomedical and Health Informatics 2020;25(3):746–754.

[4] Chizner, Michael A. Cardiac Auscultation: Rediscovering the Lost Art. Current Problems in Cardiology 2008; 33(7):326–408.

[5] Oliveira, Jorge and Renna, Francesco et al. The CirCor DigiScope Dataset: From Murmur Detection to Murmur Classification. IEEE Journal of Biomedical and Health Informatics 2021;26(6):2524–2535.

[6] Reyna, Matthew A and Kiarashi, Yashar et al. Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022. medRxiv 2022;URL https://doi.org/10.1101/2022.08.11.22278688.

[7] Vaswani, Ashish and Shazeer, Noam et al. Attention Is All You Need. Advances In Neural Information Processing Systems 2017;30.

[8] Chawla, Nitesh V and Bowyer, Kevin W et al. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research 2002;16:321–357.

Address for correspondence:

Ahsan H. Khandoker
Department of Biomedical Engineering, Khalifa University
PO Box 127788, Abu Dhabi, UAE
ahsan.khandoker@ku.ac.ae