# A Novel Approach for Robust Detection and Classification of Valvular Heart Disease Using ProbSparse Self-Attention and Virtual Adversarial Training on Phonocardiography Data

Qijie Feng
*Shanghai United International School*
Shanghai, China
jfeng8035@gmail.com

***Abstract*—Valvular heart disease (VHD) accounts for a significant portion of cardiovascular diseases worldwide. Early-stage diagnosis is crucial for effective treatment, yet traditional diagnostic methods rely on harmful or costly modalities. Moreover, existing datasets for VHD often suffer from data scarcity and low quality. In response, we propose a novel VHD detection and classification method utilizing phonocardiography data. Our approach features a classification model based on ProbSparse self-attention and a training strategy that combines virtual adversarial training with Bayesian optimization to address data scarcity effectively. Evaluated on the corresponding public dataset, our method demonstrated robust performance in VHD classification, achieving state-of-the-art results compared to existing approaches. Additionally, through ablation studies, we validated the influence of the adapted components, confirming the effectiveness of our method.**

***Keywords—valvular heart disease, data scarcity, phonocardiogram, ProbSparse self-attention, virtual adversarial training, Bayesian optimization.***

## I. Introduction

In recent years, cardiac valvular heart disease (VHD) has become a major cause of cardiovascular mortality globally [1]. Common types of VHD include mitral valve prolapse (MVP), aortic stenosis (AS), mitral regurgitation (MR), and mitral stenosis (MS), among others. Established modalities for VHD detection encompass chest radiography, electrocardiography, echocardiography, computed tomography, cardiac MRI, and cardiac catheterization [2]. Nonetheless, these techniques are associated with various limitations, such as invasiveness, radiation exposure, high cost, reliance on specialized equipment, and the need for skilled personnel.

With the development of deep learning algorithms, the last of the preceding issues, i.e., the requirement of skilled personnel or professional training, has been partially solved. Trained with a profusion of data, deep learning models are capable of replicating the process of human decision-making, thus allowing automated classification of the input while preserving a high enough accuracy. Previous works have shown that such methodologies can be effectively applied in the field of clinical medicine [3, 4]. In the context of VHD detection and classification, deep learning methods apply similarly; however, issues of lacking data quality and abundance hinder the way toward generalization. Simultaneously, it is essential to consider the selection of modalities, as different methods produce varying data types that significantly impact model development, including aspects such as model size, complexity, etc.

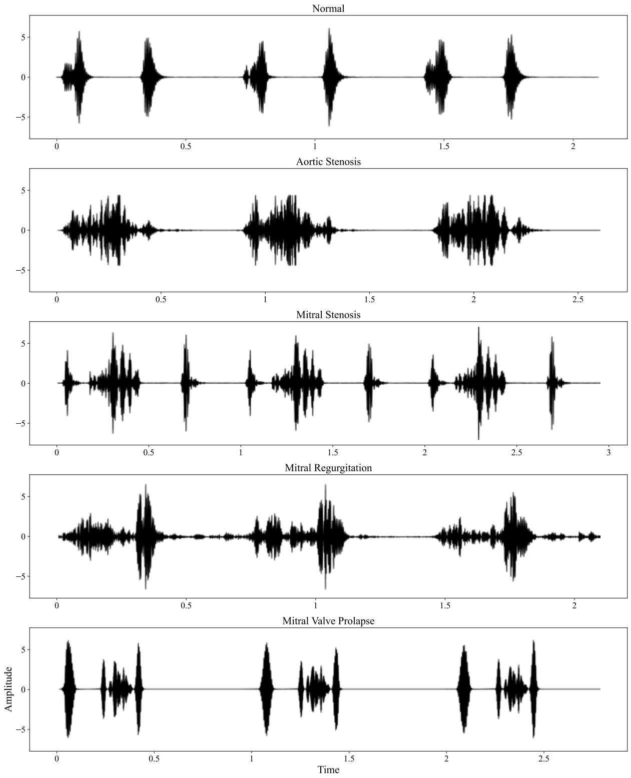

Fig. 1. Examples of PCG recordings (the horizontal axis represents time in seconds; the vertical axis represents relative amplitute after standardization). The top panel shows the waveform of a patient with a healthy heart condition, characterized by regular heartbeats and clear S1 and S2 sounds. The bottom panels illustrates the PCG of patients with VHD, displaying abnormalities such as murmurs and extra heart sounds, indicative of valve abnormalities.

Apart from models built upon the mentioned modalities [5-9], our work proposes a detection and classification model based on phonocardiograms (PCG). The heart produces characteristic sounds during its beating due to the opening and closing of valves. Pathologies such as stenosis, prolapse, or regurgitation of the heart valves cause corresponding changes in these sounds, resulting in abnormal clicks and murmurs. Phonocardiography is a non-invasive, cost-effective, and accessible diagnostic method that detects and records cardiac sounds using specialized sensors placed on the chest. These sensors convert acoustic signals into electrical signals, which are then amplified and filtered. The digitalized signal is subsequently saved or graphed as needed. In this case, the PCG data is passed into a classification model to generate predictions.

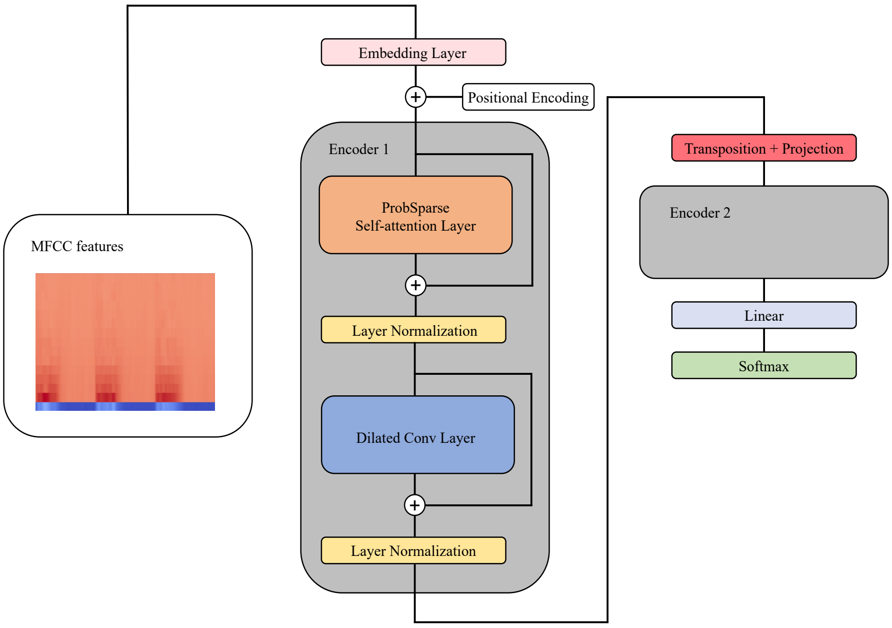

Fig. 2. The overall architecture of the proposed VHD detection and classification model. Encoder 2 in the schemetic has the same structure as Encoder 1.

However, a notable paucity of available PCG data for valvular heart disease poses a challenge. This scarcity limits the representativeness of the dataset, leading to poor generalization and increased risks of overfitting and underfitting. We conjecture a possible correspondence between small dataset size and bias that leads to this suboptimal performance and present a novel training strategy to combat the issue. Our approach aims to mitigate these biases, thereby enhancing the reliability and performance of the VHD detection and classification model.

Our contribution can be summarized as:

- Proposed a novel VHD detection and classification model based on phonocardiography. Trained and evaluated with a public dataset, our model demonstrates robustness under comparison with existing methods.
- Developed a new training strategy to address challenges posed by the lack of data, which, through ablation studies, shows improvement to model performance.

## II. Methodology

In this section, we describe in detail about our proposed method of automated VHD detection and classification, which includes data preprocessing, the classification model, and its training. Experiment setup and results will be presented in the later sections.

### A. Dataset Description and Preprocessing

For training and evaluation of our method, we utilized the public dataset collected by [10], which consists of 5 categories: aortic stenosis (AS), mitral regurgitation (MR), mitral stenosis (MS), mitral valve prolapses (MVP), and the normal (N). To our knowledge, this dataset is currently the only publicly available PCG dataset aimed at VHD classification. The dataset includes 1000 recordings, with 200 recordings per category, each sampled at 8000 Hz, and covers roughly 3 cardiac cycles.

For data preprocessing, all recordings in the dataset are first trimmed or zero-padded to a fixed length of 27500 samples. Then data are normalized with zero-mean, unit-variance, and undergo Mel-frequency cepstral coefficient (MFCC) feature extraction, a widely adopted feature extraction method in audio data preprocessing [11]. We used 30 MFCC coefficients to ensure a comprehensive capturing of features. A plot showing the extracted MFCCs of a recording is shown on the left side of Fig. 2.

### B. Classification Model

We propose a modified transformer encoder [12] model to accomplish VHD classification as a time series classification task. The proposed model utilized ProbSparse self-attention [13] as a regularization to prevent overfitting. The overall structure of our model is shown in Fig. 2. Raw heart sound recording is first feature extracted with MFCC feature extraction. The features, with length $L$, are then passed into an embedding layer where they are linearly projected to an embedding space, denoted $\mathbb{R}^{L\times d_{\text{model}}}$, and positionally encoded [12]. The projected data is passed into two encoders sequentially. The output of the first encoder is transposed and reprojected to $\mathbb{R}^{d_{\text{model}}\times d_{\text{model}}}$ before passing into the second, allowing the second encoder to capture channel-wise dependencies. Each encoder consists of two sub-layers, a ProbSparse self-attention layer and a dilated convolution layer. Both sub-layers have a residual connection connecting between their inputs and outputs and are followed by a layer normalization operation. Finally, the output of the second encoder is passed into a linear layer with softmax activation to produce the classification logits. Below are the detailed descriptions of the key components of the proposed model.

### *1) ProbSparse Self-attention Layer*

First proposed in [13], ProbSparse self-attention is an efficient self-attention mechanism that aims at reducing the computational complexity of time-series prediction with transformer models. In this work, we employ ProbSparse self-attention as an attention mechanism with integrated regularization. The original self-attention [1] utilizes the scaled dot-product attention performed on the queries, keys, and values, denoted $Q$, $K$, and $V$:

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V, \tag{1}$$

where $d_K$ represents dimension of the keys. Denoting the $i$th query, key, value as $q_i$, $k_i$ and $v_i$, this can be rewritten into a kernel smoother [14] as:

$$\mathrm{Attn}(q_i,K,V) = \sum_j \frac{k(q_i,k_j)}{\sum_l k(q_i,k_l)} v_j = \mathbb{E}_{p(k_j,q_i)}[v_j], \tag{2}$$

where $k(q_i,k_j)$ is an asymmetric exponential kernel $\exp\left(q_i {k_j}^T/\sqrt{d_K}\right)$, and $p(k_j,q_i) = k(q_i,k_j)/\sum_l k(q_i,k_l)$. Experiments on the canonical transformer show that the self-attention feature maps tend to form long-tail distributions. Strategies handling the sparsity need to be adopted to obtain dependencies from the data efficiently. To quantify the sparsity, [13] proposed a query sparsity measure which uses KL-divergence to measure the similarity between distribution $p(k_j,q_i)$ and uniform distribution $q(k_j,q_i) = \frac{1}{L_K}$. The $i$th query's sparsity measurement is thus given by:

$$\mathrm{M}(q_i,K) = ln\sum_{j=1}^{L_K} e^{\frac{q_i {k_j}^T}{\sqrt{d_K}}} - \frac{1}{L_K}\sum_{j=1}^{L_K}\frac{q_i {k_j}^T}{\sqrt{d_K}}, \tag{3}$$

which can be approximated with

$$\overline{\mathrm{M}}(q_i,K) = \max_j\left\{\frac{q_i {k_j}^T}{\sqrt{d_K}}\right\} - \frac{1}{L_K}\sum_{j=1}^{L_K}\frac{q_i {k_j}^T}{\sqrt{d_K}} \tag{4}$$

Each time a set of $Q$, $K$, and $V$ passes through the attention, the query sparsity measurements are calculated between the keys and the queries. Then $u$ samples of queries with the top measurements are selected to perform scaled dot-product attention with $K$, and $V$. In our model, $u$ is set to $5\lceil ln(L_Q)\rceil$ as suggested by the original work. To ensure the same output sequence length, places for the abandoned queries are filled with zeros. We also employed the multi-head design proposed by [12], so that the model can simultaneously attend to information from various representation subspaces at different positions. Therefore, with $W$ representing parameter matrices, the ProbSparse self-attention layer is defined as:

$$\mathrm{MultiHead}(X) = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots \mathrm{head}_h)W^O, \tag{5}$$

$$\text{where} \quad head_i = \mathrm{Attn}\left(S(XW_i^Q), XW_i^K, XW_i^V\right). \tag{6}$$

$S(XW_i^Q)$ is the query selection procedure described earlier.

In our work, we use ProbSparse self-attention as a regularization technique. By deleting (setting to zero) queries that are similar to a uniform distribution, ProbSparse self-attention encourages the encoder to focus solely on the most informative features and acts as a noise filter. This prevents the model from overfitting to noise and irrelevant information, enhancing generalization, particularly with audio data such as PCG recordings. An ablation study on the ProbSparse self-attention layer is presented in the results section.

### *2) Dilated Convolution Layer*

Considering PCG's multi-resolution structure, where both large-scale features (e.g., durations of the murmurs) and small-scale features (e.g., frequency components) can be important for the model to determine the existence of abnormality, we incorporated dilated convolution [15] in place of the traditional feed forward network in canonical transformer. The dilated convolution layer consists of a set of dilated convolutions with filter length of {8, 5, 3} and dilation rate of 2, connecting sequentially.

We add residual connections to both sub-layers, followed by a layer normalization operation. Thus, one encoder can be formulated as:

$$\mathrm{Encoder}(X) = \mathrm{Norm}(\mathrm{Conv}(res) + res), \tag{7}$$

$$res = \mathrm{Norm}(\mathrm{MultiHead}(X) + X), \tag{8}$$

where $\mathrm{Norm}(\cdot)$ represents the layer normalization, and $\mathrm{Conv}(\cdot)$ represents dilated convolution.

## *C. Training*

We developed a novel optimization strategy for the training of the proposed model. The training strategy is based on virtual adversarial training (VAT) [16] and Bayesian optimization [17], which when combined, effectively solve the issue of data scarcity.

### *1) Virtual Adversarial Training*

Proposed by [16], VAT is a regularization method that puts an additional *local distributional smoothness* (LDS) regularization term $R_{vadv}$ on the model's objective function. For a model, parameterized by $\theta$, with output distribution $p(y|x,\theta)$, $R_{vadv}$ is defined as:

$$R_{vadv}(B,\theta) = \frac{1}{|B|}\sum_{x\in B} D[p(y|x,\theta), p(y|x + r_{vadv},\theta)], \tag{9}$$

where $B$ represents a training batch, and $D$ represents a non-negative divergence measure. $r_{vadv}$ in (9) is a *virtual adversarial perturbation* given by the maximization:

$$r_{vadv} = \underset{r;\, \|r\|_2 \le \epsilon}{\arg\max}\, D[p(y|x,\theta), p(y|x+r,\theta)], \tag{10}$$

where $\epsilon$ is a norm constraint for perturbation $r$. Thus, the new objective function is defined as:

$$L(B,\theta) = l(B,\theta) + \alpha R_{vadv}(B,\theta), \tag{11}$$

where $l(B,\theta)$ is the original training objective (e.g., cross entropy), and $\alpha$ is a regularization coefficient. Additionally, a fast approximation of $r_{vadv}$ can be achieved via the power iteration method [18] and the finite difference method, yielding:

$$r_{vadv} \approx \frac{g}{\|g\|_2}, \tag{12}$$

$$\text{where } g = \nabla_r D[p(y|x,\theta), p(y|x+r,\theta)]|_{r=\xi d} \tag{13}$$

In (13), $d$ represents a random unit vector and $\xi$ is a small number close to zero.

Optimizing $L(B,\theta)$ improves generalization performance through enhancing isotropic smoothness of the output distribution with the exertion of anisotropic perturbations. *Adversarial* training takes place as the model is optimized to minimize $L(B,\theta)$ while $r_{vadv}$ is obtained to maximize $R_{vadv}(B,\theta)$, forcing the model to have a smooth boundary between different classes.

### 2) Proposed Strategy

We observed that when the model is trained on different subsets of the dataset, but evaluated with the same data, a varying VAT parameter is required to obtain optimal results. This is exemplified in Table I, where the model's loss (in cross entropy) on the test set is recorded for different values of $\epsilon$ and a fixed $\alpha = 1$. The used dataset is from [10], and results from two training sets are compared, both of the equal size 400. Training set 1 comprises the first 80 samples from each category, while training set 2 comprises of samples 81-160 of each category. The testing set is formed with all the remaining samples. For detailed training procedures, please refer to the Experiment section. Table I shows an inconsistency of VAT parameters for the acquisition of optimal performance when the training set varies. Based on the observation, we conjecture that the discrete nature of the dataset, along with the scarcity of data, plays a role in this phenomenon. Specifically, we consider the influence of VAT on model fitting as expanding the margins of categories through exerting adversarial perturbations (maximizing $r_{vadv}$), while ensuring the model's ability to make correct decisions (minimizing $l(B,\theta)$ ). However, incorrect expansion of the margins occurs when penalties are not applied in time due to a lack of certain data points, therefore making data scarcity act as a bias in the dataset.

Since the deterioration of model performance originates from overly or insufficiently expanded margins dominated by VAT, it is reasonable to adjust these parameters to find an optimal balance. In other words, the parameters of VAT give us a degree of freedom to manipulate the category margins, which allows us to mitigate the effect of scarcity-induced bias. Considering the partition invariance of the absence of crucial data points, i.e., the absence of certain crucial data points remains consistent irrespective of how the dataset is divided into subsets, we take a cross validation approach to search for the optimal VAT parameters. In particular, the training set is first split into a larger subset and a smaller one, denoted $T_1$ and $T_2$, with a ratio of 3:1. With initial VAT parameters of $\epsilon = 1$, $\alpha = 1$, the model undergoes training using $T_1$. Then the best model over the validation set (refer to the dataset split in the Experiment section) is selected and evaluated on $T_2$. A 4-fold cross validation is performed, repeating the evaluation across the whole training set, and a generalization measure is calculated by taking the average of the model loss. Next, the generalization is minimized via Bayesian optimization. The Bayesian optimizer will tune the parameters of VAT and make new observations through iteration.

In our research, Bayesian optimization utilizes a Gaussian process with Matérn 5/2 kernel. The kernel parameters are determined and updated by the method of maximizing marginal likelihood [19] during optimization. Noisy expected improvement [20] is used as the acquisition function. The VAT parameters $\epsilon$ and $\alpha$ are both optimized over [0, 5], and a total of 25 optimization iterations are run.

TABLE I. MODEL LOSS ACROSS DIFFERENT VAT PARAMETERS FOR DIFFERENT SUBSETS

| $\epsilon$ | Loss (Cross Entropy, mean±std) | |
|---|---|---|
| | ***Training Set 1*** | ***Training Set 2*** |
| 0.5 | 0.1353±0.0365 | 0.2334±0.0558 |
| 1 | **0.1217±0.0291** | 0.2202±0.0667 |
| 2 | 0.1241±0.0467 | 0.1975±0.0747 |
| 4 | 0.1302±0.0687 | **0.1890±0.0542** |
| 8 | 0.1549±0.0659 | 0.2218±0.0668 |
| 16 | 0.1365±0.0514 | 0.2196±0.0900 |

## III. EXPERIMENTS

### A. Experimental Details

The proposed model is trained using the Adam optimizer [21] with a learning rate of 0.001 and cross-entropy as the loss function. Batch size for stochastic gradient descent is set to 32. A 9:1 train-test split is employed for the dataset. A validation set is formed with the last 100 samples of the training set, which are excluded from training and used to evaluate the model at the end of each epoch. The best-performing model parameters (in terms of loss) over the validation set are evaluated on the testing set and results are reported. The models are trained for a total of 2000 epochs with an early stopping of 100 epochs patience. 10-fold cross validation is adopted for the accuracy of the results.

In terms of model hyperparameter, we used $d_{\text{model}} = 64$ and a head size of 2, resulting in a model size of 204K parameters.

The experiments were conducted on TensorFlow 2.12.0 for model training and BoTorch [22] 0.11.1 for Bayesian optimization, running on Ubuntu 22.04 LTS (64-bit) with a Tesla V100 GPU (32GB memory) and CUDA 11.8.0.

### B. Comparison with State-of-the-art Methods

We compare our method with several strong baselines for VHD classification, including CNN-BiLSTM [23], WaveNet [24], CNN-LSTM [25], CardioXNet [26], TFDDL [27], SVM [10], RF+Multiboost [28], ViT [29], and CNN [30], all are recognized as state-of-the-art approaches in this field. For a fair comparison, all methods were trained and evaluated on the same dataset as the one used by us and utilized k-fold cross validation.

The performance of the baselines and the proposed model are shown in Table II, in which we compared the methods over 4 metrics: accuracy, sensitivity, specificity, and F1 score. Table II shows that the proposed method outperformed all baselines, with an accuracy of 100.0% and an average improvement of 1.09%, indicating that our method has achieved state-of-the-art result in VHD classification via PCG.

### C. Ablation Studies

To reveal the impact of ProbSparse self-attention and VAT, we systematically remove or replace the corresponding components in the model, producing 3 variants: the canonical attention variant without VAT (*Canonical Attention w/o VAT*), the ProbSparse Attention variant without VAT (*ProbSparse Attention w/o VAT*), and the canonical attention variant with VAT (*Canonical Attention w/ VAT*). In the canonical attention variants, the ProbSparse attention mechanism in the proposed model is replaced by the original attention, while in the ProbSparse attention variant, model

TABLE II. Comparison of Performance of Existing VHD Classification Methods

| Methods | Accuracy | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|
| CNN-BiLSTM [23] | 99.32% | 98.30% | 99.58% | 98.33% |
| WaveNet [24] | 97.00% | 92.50% | 98.10% | 92.37% |
| CNN-LSTM [25] | 99.87% | 99.86% | 99.97% | 99.87% |
| CardioXNet [26] | 99.60% | 99.52% | 99.73% | 99.68% |
| TFDDL [27] | 99.48% | 99.48% | 99.48% | 99.73% |
| SVM [10] | 97.90% | 98.20% | 99.40% | 99.70% |
| RF+Multiboost [28] | 98.53% | 96.40% | 99.08% | 96.40% |
| ViT [29] | 99.90% | 99.90% | 99.90% | 99.90% |
| CNN [30] | 98.60% | 98.30% | 98.86% | 98.50% |
| **Proposed** | **100.0%** | **100.0%** | **100.0%** | **100.0%** |

TABLE III. Performance Metrics for Ablation Variants

| Variants | F1 Score | AUROC |
|---|---|---|
| ProbSparse Attention w/ VAT | 100.0% | 100.0% |
| Canonical Attention w/ VAT | 98.00% | 99.91% |
| ProbSparse Attention w/o VAT | 98.59% | 99.90% |
| Canonical Attention w/o VAT | 96.11% | 99.66% |
| Canonical Attention w/ RQS | 96.80% | 99.69% |

structure is kept the same as the description in the Methodology section. Additionally, to reveal the influence of query selection in ProbSparse self-attention, we included a canonical attention variant with random query selection (*Canonical Attention w/ RQS*), where for each attention operation, $5\lceil ln(L_Q)\rceil$ queries are randomly selected and preserved, whereas the unselected queries are set to zero. The performance of each variant is shown in Table III and their corresponding receiver operating characteristic (ROC) curves are shown in Fig. 3. (*ProbSparse Attention w/ VAT* is equivalent to the proposed method). All variants are evaluated by F1 score and area under the receiver operating characteristic curve (AUROC).

Table III shows that implementing ProbSparse attention has an average increase of 2.24% in F1 score in both VAT applied and unapplied models, indicating ProbSparse attention's contribution to model performance. Comparing VAT applied variants to variants without VAT shows that applying VAT induces a 1.65% improvement in F1 score. Finally, comparing the ProbSparse attention variant and the canonical attention variant, both without VAT, to the canonical attention variant with random query selection shows that selecting queries based on the query sparsity measure engenders a more than 3.5 times improvement in F1 score respective to random selection. This suggests that ProbSparse self-attention's improvement to model performance is attributable to its sparsity-based query selection, thus supporting our use of ProbSparse attention as an attention mechanism with information filtering ability.

## IV. Conclusion and Future Work

In this work, we have developed a novel approach for VHD detection and classification. By utilizing a ProbSparse self-attention-based transformer encoder coupled with VAT-based training, our method achieved state-of-the-art results on the public dataset for VHD classification. Through ablation studies, we demonstrated the influence of the model's components and its capability to handle data scarcity. This implies our approach's feasibility in practical clinical scenarios.

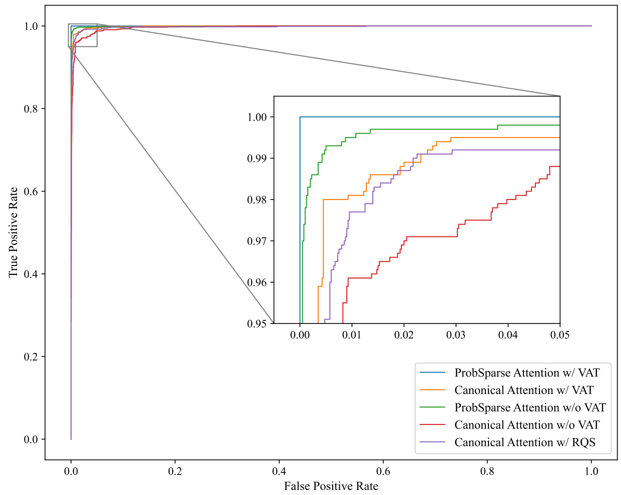

Fig. 3. ROC curves for different ablation variants of the proposed VHD classification method.

However, there remain several areas for future research. Firstly, we will conduct additional investigations to explore the scalability of the proposed method. For example, it is crucial to determine the extent to which the method's ability to resolve data scarcity can be applied to ensure model predictions fall within a domain where safety concerns are minimized. Secondly, we shall perform research on reducing the computational expenses associated with our method. In particular, we aim to integrate Bayesian methods within the VAT framework to eliminate the need for cross-validations, which currently induce a significant computational cost. These future works could enhance clinical practices by facilitating more reliable, efficient, and convenient diagnoses, ultimately improving patient outcomes and reducing resource burdens.

## References

[1] J. S. Aluru, A. Barsouk, K. Saginala, P. Rawla, and A. Barsouk, "Valvular heart disease epidemiology," *Medical Sciences*, vol. 10, no. 2, p. 32, Jun. 2022, doi: 10.3390/medsci10020032.

[2] S. Bhandari, K. Subramanyam, and N. Trehan, "Valvular heart disease: diagnosis and management.," *PubMed*, vol. 55, pp. 575–584, Aug. 2007, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18019800

[3] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, Jun. 2017, doi: 10.1146/annurev-bioeng-071516-044442.

[4] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. V. Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, no. 1, Jun. 2019, doi: 10.1038/s41597-019-0103-9.

[5] F. Yang *et al.*, "Automated analysis of doppler echocardiographic videos as a screening tool for valvular heart diseases," *JACC.*

*Cardiovascular Imaging*, vol. 15, no. 4, pp. 551–563, Apr. 2022, doi: 10.1016/j.jcmg.2021.08.015.

[6] P. Elias *et al.*, "Deep learning electrocardiographic analysis for detection of left-sided valvular heart disease," *Journal of the American College of Cardiology*, vol. 80, no. 6, pp. 613–626, Aug. 2022, doi: 10.1016/j.jacc.2022.05.029.

[7] N. G. Kang, Y. J. Suh, K. Han, Y. J. Kim, and B. W. Choi, "Performance of prediction models for diagnosing severe aortic stenosis based on aortic valve calcium on cardiac computed tomography: incorporation of radiomics and machine learning," *Korean Journal of Radiology/Korean Journal of Radiology*, vol. 22, no. 3, p. 334, Jan. 2021, doi: 10.3348/kjr.2020.0099.

[8] J. A. Fries *et al.*, "Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences," *Nature Communications*, vol. 10, no. 1, Jul. 2019, doi: 10.1038/s41467-019-11012-3.

[9] D. Ueda *et al.*, "Artificial intelligence-based detection of aortic stenosis from chest radiographs," *European Heart Journal. Digital Health*, vol. 3, no. 1, pp. 20–28, Dec. 2021, doi: 10.1093/ehjdh/ztab102.

[10] N. Yaseen, G.-Y. Son, and S. Kwon, "Classification of heart sound signal using multiple features," *Applied Sciences*, vol. 8, no. 12, p. 2344, Nov. 2018, doi: 10.3390/app8122344.

[11] D. Deshwal, P. Sangwan, and D. Kumar, "Feature extraction methods in language identification: a survey," *Wireless Personal Communications*, vol. 107, no. 4, pp. 2071–2103, Apr. 2019, doi: 10.1007/s11277-019-06373-3.

[12] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[13] H. Zhou *et al.*, "Informer: beyond efficient transformer for long sequence time-series forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 11106–11115, May 2021, doi: 10.1609/aaai.v35i12.17325.

[14] Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov, "Transformer dissection: an unified understanding for transformer's attention via the lens of kernel," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4344–4353, Jan. 2019, doi: 10.18653/v1/d19-1443.

[15] F. Yu, V. Koltun and T. Funkhouser, "Dilated residual networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 636-644, doi: 10.1109/CVPR.2017.75.

[16] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019, doi: 10.1109/tpami.2018.2858821.

[17] J. Mockus, "On Bayesian methods for seeking the extremum and their application.," *IFIP Congress*, pp. 195–200, Jan. 1977, [Online]. Available: https://dblp.uni-trier.de/db/conf/ifip/ifip1977.html#Mockus77

[18] G. H. Golub and H. A. Van Der Vorst, "Eigenvalue computation in the 20th century," *Journal of Computational and Applied Mathematics*, vol. 123, no. 1–2, pp. 35–65, Nov. 2000, doi: 10.1016/s0377-0427(00)00413-1.

[19] C. E. Rasmussen and C. K. I. Williams, "Model selection and adaptation of hyperparameters," in *Gaussian Processes for Machine Learning*, MIT Press, 2005, pp.105-128. doi: 10.7551/mitpress/3206.003.0008.

[20] B. Letham, B. Karrer, G. Ottoni, and E. Bakshy, "Constrained Bayesian optimization with noisy experiments," *Bayesian Analysis*, vol. 14, no. 2, Jun. 2019, doi: 10.1214/18-ba1110.

[21] Diederik P. Kingma and Jimmy Ba, "Adam: a method for stochastic optimization," 2017, *arXiv:1412.6980.*

[22] M. Balandat *et al.*, "BoTorch: A framework for efficient Monte-Carlo Bayesian optimization," *Neural Information Processing Systems*, vol. 33, pp. 21524–21538, Jan. 2020, [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/f5b1b89d98b7286673128a5fb112cb9a-Paper.pdf

[23] M. Alkhodari and L. Fraiwan, "Convolutional and recurrent neural networks for the detection of valvular heart diseases in phonocardiogram recordings," *Computer Methods and Programs in Biomedicine*, vol. 200, p. 105940, Mar. 2021, doi: 10.1016/j.cmpb.2021.105940.

[24] S. L. Oh *et al.*, "Classification of heart sound signals using a novel deep WaveNet model," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105604, Nov. 2020, doi: 10.1016/j.cmpb.2020.105604.

[25] Y. Al-Issa and A. M. Alqudah, "A lightweight hybrid deep learning system for cardiac valvular disease classification," *Scientific Reports*, vol. 12, no. 1, Aug. 2022, doi: 10.1038/s41598-022-18293-7.

[26] S. B. Shuvo, S. N. Ali, S. I. Swapnil, M. S. Al-Rakhami, and A. Gumaei, "CardioXNet: a novel lightweight deep learning framework for cardiovascular disease classification using heart sound recordings," *IEEE Access*, vol. 9, pp. 36955–36967, Jan. 2021, doi: 10.1109/access.2021.3063129.

[27] J. Karhade, S. Dash, S. K. Ghosh, D. K. Dash, and R. K. Tripathy, "Time–frequency-domain deep learning framework for the automated detection of heart valve disorders using PCG signals," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, Jan. 2022, doi: 10.1109/tim.2022.3163156.

[28] S. I. Khan, S. M. Qaisar, and R. B. Pachori, "Automated classification of valvular heart diseases using FBSE-EWT and PSR based geometrical features," *Biomedical Signal Processing and Control*, vol. 73, p. 103445, Mar. 2022, doi: 10.1016/j.bspc.2021.103445.

[29] S. Jamil and A. M. Roy, "An efficient and robust phonocardiography (PCG)-based valvular heart diseases (VHD) detection framework using vision transformer (ViT)," *Computers in Biology and Medicine*, vol. 158, p. 106734, May 2023, doi: 10.1016/j.compbiomed.2023.106734.

[30] N. Baghel, M. K. Dutta, and R. Burget, "Automatic diagnosis of multiple cardiac diseases from PCG signals using convolutional neural network," *Computer Methods and Programs in Biomedicine*, vol. 197, p. 105750, Dec. 2020, doi: 10.1016/j.cmpb.2020.105750.