2021 IEEE International Conference on Data Mining (ICDM)

2021 IEEE International Conference on Data Mining (ICDM) | 978-1-6654-2398-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICDM51629.2021.00043

BaT: Beat-aligned Transformer for Electrocardiogram Classification

Xiaoyu Li[*,‖], Chen Li[†], Yuhua Wei[*], Yuyao Sun[§], Jishang Wei[‡], Xiang Li[§], Buyue Qian[¶]
[*]*School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, China*
Email: {xiaoyuli, weiyuhua}@stu.xjtu.edu.cn
[†]*National Engineering Lab for Big Data Analytics, Xi'an Jiaotong University, Xi'an, Shaanxi China*
Email: cli@xjtu.edu.cn
[‡]*HP Labs, 1501 Page Mill Rd, Palo Alto, CA 94304, USA*
Email: weijishang@gmail.com
[§]*Ping An Healthcare Technology, Beijing, China*
Email: {sunyuyao188, lixiang453}@pingan.com.cn
[¶]*The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, China*
Email: qianbuyue@xjtu.edu.cn

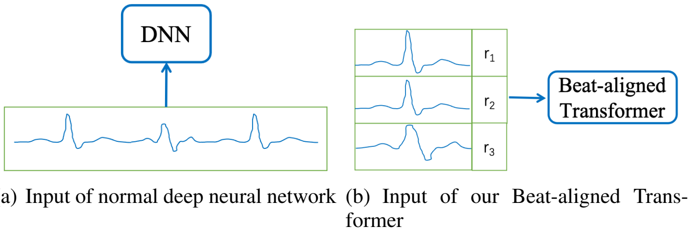
***Abstract*—Electrocardiogram (ECG) is one of the critical diagnostic tools in healthcare. Various deep learning models, except Transformers, have been explored and applied to map ECG patterns to heart abnormalities. Transformer models have been adopted from natural language processing to computer vision with advanced features. Most recently, vision transformers show exceptional performances, even on moderate-scale datasets. However, naively applying vision transformers on electrocardiogram datasets leads to poor results. In this paper, we propose a novel network called Beat-aligned Transformer (*BaT*), a hierarchical Transformer that sufficiently exploits the cyclicity of ECG. We organize and treat an input ECG as multiple aligned beats instead of a single time series. In the *BaT*, shifted-window-based Transformer blocks (SW Block) are adopted to learn the representation for each beat, and aggregation blocks are designed to exchange information among the beat representations. Nested SW Blocks and aggregation blocks form a beat-aware hierarchical structure of *BaT*. In this way, the new data format and the *BaT* hierarchical structure boost Transformer performance on ECG classification. From the experiments on public ECG datasets, we observe *BaT* outperforms other Transformer-based models and achieves competitive performance compared with other state-of-the-art methods.**

***Keywords*-ECG Classification, Transformer, Deep Learning**

I. Introduction

Electrocardiogram (ECG) is one of the most common tools for monitoring heart activities. In medicine and healthcare, ECG is widely used to diagnose different kinds of cardiac arrhythmia. Besides, ECG can also be used for many applications, including emotion recognition, sleep staging and human identification. Traditional methods based on hand-crafted features achieve considerable performance on small data. Along with the rapid development of wearable devices, the amount of ECG data keeps growing. In recent years, various deep learning methods from natural language process (NLP) and computer vision (CV) have been adopted and applied to ECG data .

In NLP, Transformers are the dominant model choice because of their computational efficiency and scalability. Most recently, Transformer models have been applied to images and show promising results. ECG data share similar nature with images, including the high resolution of data points (pixels) and the large variations in the scale of peaks and waves (objects). Thus, in this paper, we investigate how to apply and boost a vision Transformer on ECG data.



(a) Input of normal deep neural network (b) Input of our Beat-aligned Transformer

Figure 1. Comparison of Input Format. (a) Previous studies feed a deep neural network with the whole ECG time series. (b)In our method, we send aligned beats and the corresponding resampling ratios to the Beat-aligned Transformer.

We observe poor performance when naively applying Vision Transformer on ECG data leads. This is mainly because of the lack of inductive bias, such as translation equivariance and locality. In the successful application of previous vision Transformer, *ViT* [1] requires a large-scale dataset for pretraining, e.g. ImageNet-21k and JFT-300M (303M). However, the size of ECG dataset is usually much smaller, e.g. PTB-XL (∼22K) [2] and Physionet Challenge2020 (∼44K) [3]. Vision Transformers with hierarchical structures alleviate the inductive bias problem. Hierarchical Transformers apply local self-attention on the part of input without overlap. Among the hierarchical Transformers, Swin Transformer [4] is based on a shifted-window mechanism, and Nested Transformer (*NesT*) [5] is based

[‖] Xiaoyu Li contributed to this work as a research intern at Ping An Healthcare Technology.

978-1-6654-2398-4/21/$31.00 ©2021 IEEE
DOI 10.1109/ICDM51629.2021.00043

on a block aggregation function. The key lies in how to organize the hierarchical structure. When we directly apply the above hierarchical Transformers on ECG data, there is still a performance gap between Transformer and widely used networks such as ResNet. To further boost the performance of the Transformer on ECG data, we consider the nature of ECG to set the hierarchies well, and propose a novel hierarchical Transformer, Beat-aligned Transformer, to learn the representation inside and between aligned beats.

We propose to sufficiently utilize the cyclicity of ECG. The heartbeat cycle consists of diastole and systole, and an ECG usually consists of multiple heartbeat cycles. Previous deep learning methods seldom consider this nature and always feed deep neural network with the whole ECG, as shown in Figure 1.(a). The ECG is treated as 1-D image, and the deep neural networks are supposed to automatically learns the cardiac cyclicity. In our setting, we try to adopt the inductive bias of the cardiac cyclicity into ECG data format and network structure. We segment ECG into beats and resample the beats with ratios rs for alignment. Then, we send the beats and ratios to the proposed Beat-aligned Transformer, as shown in Figure 1.(b). Inside the Beat-aligned Transformer, each beat is concatenated with a resampling ratio after a linear projection layer. Then, shifted-window blocks from Swin-T are adopted to learn the representation for each beat. Inspired by *NesT*, a beat aggregation block, is designed for the information exchange among the beat representations.

We use Challenge2020 and PTB-XL for experiments. Ablation studies are also conducted to check the effectiveness of each component in the Beat-aligned Transformer. The empirical results prove that the Beat-aligned Transformer well models the cardiac cyclicity and boosts the performance of the Transformer on ECG data from scratch.

In summary, our contributions can be concluded as follows:

- We propose a novel model called Beat-aligned Transformer for ECG classification. We design the hierarchical structure of *BaT* based on the cardiac cyclicity of ECG.
- We propose a novel format of ECG, which is efficient to utilize the cyclicity of ECG data. Benefited from such a format, we boost Transformer performance with local self-attention and hierarchy structure.
- Experiments show Beat-aligned Transformer outperforms other Transformer-based networks and is also competitive compared with other state-of-the-art methods for ECG classification.

II. Related work

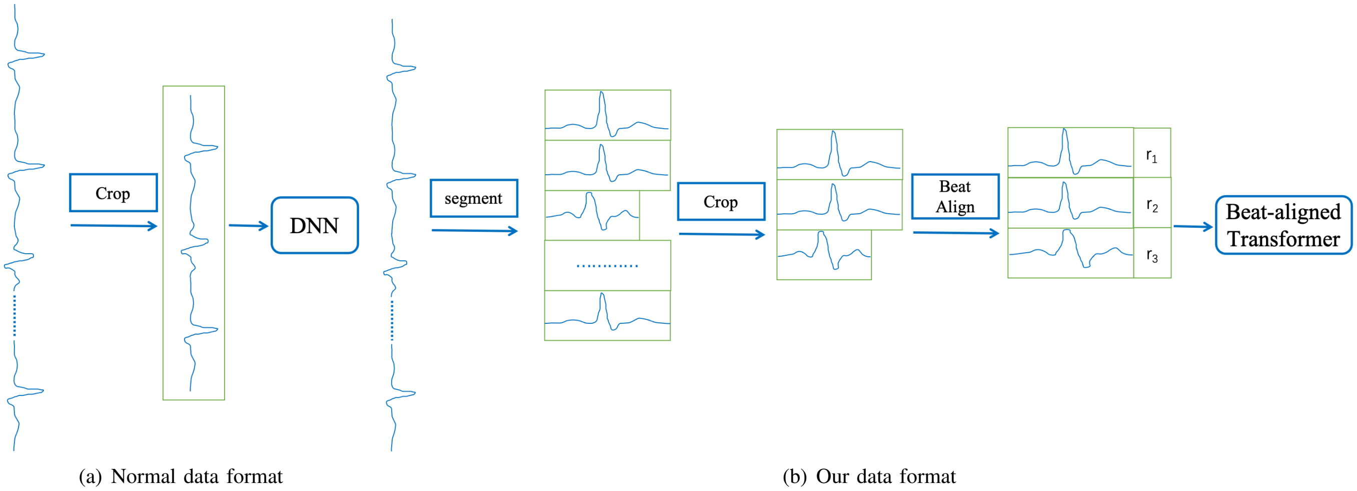
The Beat-aligned Transformer is based on vision Transformer and designed for ECG classification. Thus, in this section, we introduce related work in vision Transformer and methods for ECG classification.

A. Vision Transformer

After [6] proposes Transformer network, Transformer-based methods achieve state-of-the-art performance in many NLP tasks. Later, large pre-trained Transformer-based models, such as *BERT* and *GPT*, show the computational efficiency and scalability of the Transformer. Inspired by the great success of Transformer in NLP, researchers also try to adopt similar structures into CV. In *ViT* [1], an image is split into fixed-size patches (e.g. 16×16), and the patches are treated as tokens in the original Transformer after linear projection. Order of patches in the original image is used as position embedding. *ViT* outperforms ResNet when there are large-scale datasets for pretraining, such as ImageNet-21k and JFT-300M (303M), but yields lower accuracy on “mid-size datasets” such as ImageNet (1.3M). As [1] explains, it is because Transformers lack some inductive bias compared with CNN, such as “translation equivariance” and “locality”. To alleviate such lack, in *DeiT* [7], Transformer learns from a pre-trained deep convolutional network, with less dependency on external datasets. However, *ViT* and *DeiT* are still limited by the quadratic cost in the number of pixels. Recently, hierarchical Transformers with local self-attention have been proposed and getting popular. Swin Transformer [4] are proposed with a shifted-window mechanism. For each transformer block, a fixed length of the window is set to apply local self-attention. In each even window-based attention layer, the windows are shifted to enhance modeling power. An efficient batch computation approach for the cyclic shift is also proposed. Swin Transformer constructs the hierarchical structure from bottom to up with windows, while Nested Transformer [5] organizes the hierarchy from up to bottom with a block aggregation function. Cross-block non-local information communication only happens in the aggregation function where a MaxPooling layer is applied. *SwinT* and *NesT* provide promising practice to apply local self-attention. In our Beat-aligned Transformer, we adopt the shifted-window mechanism to help learn beat-wise representation and the aggregation function to help exchange information among beats. Compared with other vision Transformer, our *BaT* is specially designed for ECG data and is the first application for vision Transformer on ECG classification.

B. ECG Classification

Conventional methods for ECG classification are mainly based on feature engineering [8]. In these methods, time series features are usually automatically generated and selected, including features based on entropy and morphology. Then, these features are used to train classifiers such as SVM [9], Random Forest [10], [11] and XGBoost [12]. Additionally, especially on small data, hand-crafted features based on domain knowledge are proved to be effective for some cardiac arrhythmia such as atrial fibrillation [13]–[15].



(a) Normal data format

(b) Our data format

Figure 2. Comparison for Preprocessing ECG Data

With the rapid growth of the amount of ECG data, end-to-end deep learning methods emerge and become dominant choice [16]–[18]. Deep neural networks from natural language processing and computer vision are adopted for ECG classification. Considering the temporality of ECG, some studies treat ECG as sequential data, and RNN-based models (LSTM, GRU) are usually applied [19]–[22], especially for heartbeat classification where ECG data is relatively short [23]–[26]. In most of other work, ECG is usually treated as 1-D image data, and CNN-based models achieve impressive performance for ECG classification [27]–[30]. These CNN-based models vary in network structure, including 1-D FCN, ResNet and Inception [31]–[34]. ResNet-based models have become the de-facto standard for most ECG classification tasks. Among the above models, transformer-based methods are seldomly explored. [35] fuses Transformer model with temporal features for heartbeat classification. Original transformer structure is capable of dealing with a short sequence of one single heartbeat. While an ECG usually consists of multiple heartbeats, and original Transformer is not enough to handle the ECG. The top team in the Physionet Challenge2020 [10] replaces LSTM with Transformer behind ResNet. However, the Transformer doesn't help improve the final Challenge Score compared with LSTM, as reported. Besides, these Transformers are directly adopted from [6], not utilizing the new features from the above vision Transformers. In this paper, we aim to apply and boost a vision Transformer on ECG. Thus, we propose Beat-aligned Transformer which exploits cardiac cyclicity sufficiently.

III. Method

A. Input Format

Most ECG related tasks can be formulated as classification tasks, including arrhythmia diagnosis and emotion recognition. We aim to apply and boost Transformer for ECG classification. Our solution is to explicitly encode cardiac cyclicity into data format and model structure.

Suppose $X \in \mathbb{R}^{b\times l\times t}$ is a batch of ECG data, where b is the batch size, l is the lead number, and t is the length of the ECG sequence. We take l as 1 to explain our data format. In most of the previous studies, an ECG of the certain length is usually cropped to be fed into a deep neural network directly, shown in Figure 2.(a). In contrast, in our setting, we apply a heartbeat segmentation method [36] on the ECG and organize the ECG into multiple heartbeats. Then, we crop a certain number of beats. The heartbeats are usually not the same length, and we need to feed the network with a vector. Thus, we resample each beat to the same length with a resampling ratio r_i, and get a new ECG signal as $X \in \mathbb{R}^{b\times l\times h\times f}$, where h is the number of beats cropped, and f is the length of each beat. Then, we feed the beats and resampling ratios into the Beat-aligned Transformer. Note that there inevitably might be information loss when we resample the beats. It is needed to set f in a proper range. Considering that the heart rate of an adult usually ranges from 60 $\sim$ 100 times per minute, e.g., each beat lasts about 0.6 $\sim$ 1 seconds, we set f as 0.8 times of *sampling rate* in our experiments.

Folling [36], we segment ECG signals and keep R peaks in the middle of the beats. An intuitive alternative is to segment ECG signals directly with R peaks. Then, each beat starts with a half of R peak and end with the other, which also models the cyclicity of heartbeats. However, such treatment splits the R peaks, and the R peaks are critical for many tasks such as atrial fibrillation detection. We validate our choice in the ablation study.

B. Network Structure

Besides the aligned beats, input also includes the corresponding resampling ratios. The Transformer structure is

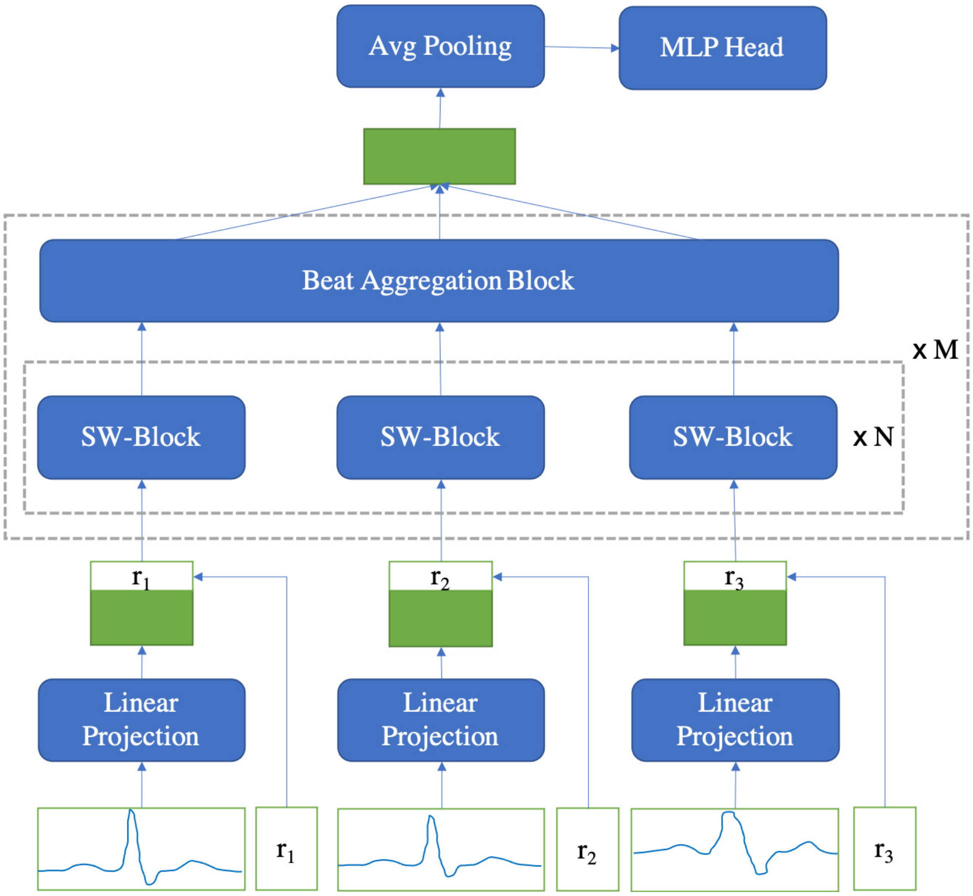


Figure 3. The structure of the Beat-aligned Transformer. Green rectangles denote different vectors, and blue rectangles denote different layers or blocks. We take a cropped ECG with three beats as an example.

flexible compared with CNN, which makes the Transformer naturally match such input format. The network structure of the Beat-aligned Transformer is shown in Figure 3. In Figure 3, green rectangles denote different vectors, and blue rectangles denote different layers or blocks. We take a cropped ECG with three beats as an example. First, the three beats are linearly projected respectively and concatenated with corresponding resampling ratios. Then, we get the representation for each beat feature. The linear projection can be a dense layer or a convolution layer with a kernel size of 1. With the linear projection, each beat is separated into small patches. Empirically, smaller patches bring more modeling power but cost more memory quadratically. We do not add absolute position encoding here. Instead, we add a relative bias to all of self-attention layers later following [37]–[40].

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}} + B)V \qquad (1)$$

where Q, K, V are *query*, *key* and *value* matrices in self-attention, d is the dimension of the small patch, and B is the relative bias.

Then, there are two nested blocks called shifted-window-based Transformer blocks (SW Block) and beat aggregation blocks. We explain details inside the two kinds of blocks in the next subsections. The SW Blocks are adopted from [4]. Each SW Block is applied to the corresponding beat feature. The SW Block is designed to focus on learning the representation inside each beat. After several SW Blocks, beat features are forward to a beat aggregation block. Because that the context of the beat can also affect the current beat itself, which is similar to the context in natural language processing. We design the beat aggregation block to share information among the beats. Every time after several SW Blocks, there is a beat aggregation block. By the nested structure of the two kinds of blocks, the network learns local beat-level features and global features in the meantime.

After the last beat aggregation block, the representations from each beat are concatenated as the final representation for the whole cropped ECG. An average pooling layer and a multi-layer perceptron head are applied to the final representation to classify the ECG data.

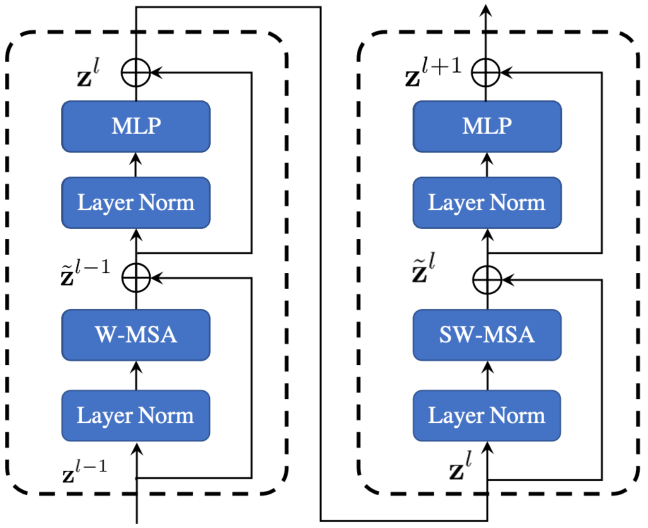


Figure 4. The Structure of SW Block. The SW Block is similar to two normal self-attention layers with additional window-shift mechanism. The SW Block is adopted to learn the representation of each beat.

C. SW Block

SW Block is first proposed in [4]. We introduce SW Block here and explain how to apply SW Block on ECG data. In the SW Blocks, local self-attention is applied based on a window mechanism. Such layer is called the window-based multi-head self-attention layer (W-MSA). Then, the W-MSA layer is used to replace the original multi-head self-attention layer in Transformer. Considering that there are no connections across the windows, the window in each even W-MSA layer is shifted by the half length of window size. The W-MSA with shifted window is called SW-MAS. Then, the SW Block works as below:

$$\begin{aligned} \tilde{z}^{l-1} &= \textit{W-MSA}(LN(z^{l-1})) + z^{l-1} \\ z^{l} &= \textit{MLP}(LN(\tilde{z}^{l-1})) + \tilde{z}^{l-1} \end{aligned} \tag{2}$$

$$\begin{aligned} \tilde{z}^{l} &= \textit{SW-MSA}(LN(z^{l})) + z^{l} \\ z^{l+1} &= \textit{MLP}(LN(\tilde{z}^{l})) + \tilde{z}^{l} \end{aligned} \tag{3}$$

where LN denotes layer normalization, z^l is the representation in the layer l, $\tilde{z}$ is the middle variable for temporal notation. Figure 4 shows the structure of SW Block.

When we apply SW Block on ECG data, we mainly consider two aspects. The first is whether to apply SW Block on the whole cropped ECG, in which way we treat ECG as 1-D images. A model with such a setting outperforms *ViT* in our experiment. Then, trying to boost this model, we split these windows in the corresponding beat. Thus, information of one beat is not to be sent to another beat, e.g., each SW Block focuses on learning the representation of the corresponding beat. Secondly, after separating the windows, we need to consider the window size for the beat instead of the whole ECG sequence. When the window size is too long, there might be few windows in a beat, which limits modeling power. When the window size is too small, the required memory explodes. Considering a kernel size of 5 and 7 are usually set for CNN-based networks for previous studies on ECG data, we set window size as 5 in our experiments.

D. Beat Aggregation Block

In the above SW Blocks, window-based local self-attention is limited inside each beat, which means information is blocked across beats. Inspired by [5], we apply the Beat Aggregation Blocks to encourage information communication among different parts. The Beat Aggregation Block is shown in Figure 5. The block takes beat representations as inputs and concatenates them together as a representation for the whole cropped ECG. Then, a convolutional layer, a layer-normalization layer and a max-pooling layer are applied sequentially. We set the kernel size of the convolutional layer and max-pooling layer as 7 for the sample of 500 Hz. The kernel size, which is larger than 1, means information can be fused across beats. The stride of the max-pooling layer is also set as 2. Then, the representation is still split into parts corresponding to different beats.

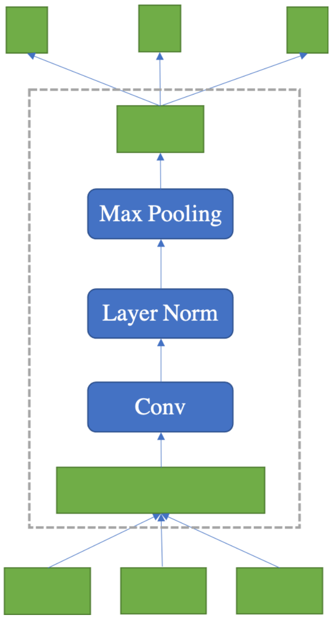


Figure 5. The Structure of Beat Aggregation Block.

We organize the Beat-aligned Transformer with several stages. Each stage begins with some SW Blocks and ends with a Beat Aggregation Block.

IV. Experiments

A. Datasets

We choose large datasets in recent research, PhysioNet Challenge2020 and PTB-XL, to conduct experiments.

Table I

Statistical Information of Challenge2020, including mean duration of recordings in years, mean age of patients in recordings in seconds, and sex (male/female) of patients in recordings for each training set.

Dataset	Duration	Age	Sex
CPSC	15.9	60.2	53.8%/46.2%
CPSC-Extra	15.9	63.7	53.4%/46.6%
INCART	1800	56.0	54.1%/45.9%
PTB	110.8	56.3	73.1%/26.7%
PTB-XL	10.0	59.8	52.1%/47.9%
G12EC	10.0	60.5	53.9%/46.1%

The PhysioNet/Computing in Cardiology Challenge 2020 [3] is a competition to identify clinical diagnoses from 12-lead ECG recordings, which assembled multiple databases across the world. Each database contains recordings with diagnoses and demographic data. 66,405 recordings were sourced from hospital systems from four distinct countries and annotated with clinical diagnoses, including 43,101 annotated recordings that were posted publicly. The first source is the public (CPSC Database) and unused data (CPSC-Extra Database) from the China Physiological Signal Challenge in 2018 (CPSC2018). This training set consists of two sets of 6,877 and 3,453 of 12-ECG recordings lasting from 6 seconds to 60 seconds. Each recording was sampled at 500 Hz. The second source set is the public dataset from St Petersburg INCART 12-lead Arrhythmia Database. This database consists of 75 annotated recordings extracted from 32 Holter records. Each record is 30 minutes long and contains 12 standard leads, each sampled at 257 Hz. The third source from the Physikalisch Technische Bundesanstalt (PTB) comprises two public databases: the PTB Diagnostic ECG Database and the PTB-XL. The first PTB database contains 549 records. Each recording was sampled at 1000 Hz. The PTB-XL contains 21,837 clinical 12-lead ECGs of 10-second length with a sampling frequency of 500 Hz. The fourth source is a Georgia database which represents a unique demographic of the Southeastern United States. This training set contains 10,344 12-lead ECGs (male: 5,551, female: 4,793) of 10-second length with a sampling frequency of 500 Hz. Each annotated ECG recording contained ECG signal data and demographic information, including age, sex, and a diagnosis, i.e., the labels for the Challenge data. The statistical information of Challenge2020 is shown in Table I. We conduct experiments on the PTB-XL dataset and on the whole Challenge 2020 datasets, respectively, because different sources show different data and label distribution.

B. Task and Metric

PTB-XL comes with a variety of labels. We conduct experiments on the multi-label classification task that are directly related to ECG statements. To address the performance of our model, we report the macro-averaged area under the receiver operating characteristic curve (AUC), which is obtained by averaging class-wise AUCs over all classes.

In Challenge2020, we also conduct experiments on the multi-label classification task that are directly related to 24 diagnoses. To test the performance of our model, we report the macro G-beta , macro F-beta measures and a new scoring metric called Challenge Score, which awards partial credit to misdiagnoses that result in similar treatments or outcomes as the true diagnosis as judged by our cardiologists. The macro G-beta and macro F-beta are formally defined as below:

$$\begin{aligned} F_\beta &= (1+\beta^2)\cdot\frac{Precision\cdot Recall}{\beta^2\cdot Precision + Recall} \\ &= (1+\beta^2)\cdot\frac{TP}{(1+\beta^2)TP+FP+\beta^2 FN} \end{aligned} \tag{4}$$

$$G_\beta = \frac{TP}{TP+FP+\beta FN} \tag{5}$$

where β is set as 2.

The challenge score is the final evaluation metric for Challenge2020. It is officially defined as follows:

Let $C = [c_i]$ be a collection of diagnoses. We compute a multi-class confusion matrix $A = [a_{ij}]$, where a_{ij} is the number of recordings in a database that were classified as belonging to class c_i but actually belonging to class c_j. We assign different weights $W = [w_{ij}]$ to different entries in this matrix based on the similarity of treatments or differences in risks. The score s is given by $s = \sum_{ij} w_{ij} a_{ij}$, which is a generalized version of the traditional accuracy metric. The score s is then normalized so that a classifier that always outputs the true class(es) receives a score of 1 and an inactive classifier that always outputs the normal class receives a score of 0.

The scoring metric is designed to award full credit to correct diagnoses and partial credit to misdiagnoses with similar risks or outcomes as the true diagnosis. Therefore, true positives are rewarded, false negatives are partially rewarded, and false positives are effectively penalized by receiving no credit at all, or, equivalently, by reducing the credit for true positives and false negatives. (True negatives are technically neither rewarded nor penalized by this metric.) A classifier that returns only positive outputs should now receive a negative score, i.e., a lower score than a classifier that returns only negative outputs.

C. Data Preprocess

PTB-XL provides waveform files stored in WaveForm DataBase (WFDB) format, with 16-bit precision at a sampling frequency of 500Hz and also release downsampled versions of the waveform data at a sampling frequency of 100Hz. We choose the records of 500Hz sampling rate as input data because our method splits the ECG record into several heartbeats, and convert the heartbeats to high dimensional representation, which needs a higher resolution. Normally, Each 10-second ECG recording contains about

Table II
Performance Comparison on Challenge2020. The mean and variance values from 10-fold cross validation and 5 random seeds are reported. The rows of vision Transformers are grayed. The best results of vision Transformers and the best results of other baselines are bold.

Model	F2	G2	Challenge Score
LSTM	0.4323±0.0024	0.2742±0.0052	0.4372±0.0073
CNN	0.4519±0.0070	0.2862±0.0083	0.4542±0.0076
ResNet	**0.5088±0.0021**	**0.3278±0.0088**	**0.5158±0.0041**
ViT	0.3263±0.0054	0.1970±0.0037	0.3197±0.0078
Swin Transformer	0.4812±0.0042	0.3045±0.0020	0.4811±0.0068
Beat-aligned Transformer	**0.5011±0.0034**	**0.3125±0.0036**	**0.4958±0.0041**

Table III
Performance Comparison on PTB-XL. The mean and variance values from 5 random seeds are reported. The rows of vision Transformers are grayed. The best results of vision Transformers and the best results of other baselines are bold.

Model	AUC
LSTM	0.8896±0.0062
CNN	0.8968±0.0041
ResNet	**0.9030±0.0052**
ViT	0.8622±0.0058
Swin Transformer	0.8958±0.0038
Beat-aligned Transformer	**0.9053±0.0049**

9∼13 heartbeats. We split the ECG recordings into several heartbeats, and choose 5 sequential heartbeats randomly, then concatenate them in another dimension.

We preprocess Challenge2020 in a similar manner with PTB-XL, except for the following differences. For Challenge2020, we need to resample ECG from different databases to 500Hz. We randomly select 10 sequential beats. We also follow [41] to only select 8 leads from 12 leads ECG because the other 4 leads contain no incremental information.

D. Training and Evaluation

We use the recommended train-test splits provided by PTB-XL benchmarks [42]. PTB-XL proposes 10-fold train-test splits obtained via stratified sampling. All recordings of a particular patient were assigned to the same fold. Recordings in fold 9 and 10 receive at least one human evaluation and are therefore of particularly high label quality. Therefore, we use folds 1-8 as the training set, fold 9 as the validation set and fold 10 as the test set. We set AdamW optimizer with a learning rate of 0.001 for Transformer-based models. We follow the PTB-XL benchmarks for the other training and evaluation settings.

For Challenge2020, we use 5 random train-validation-test splits (8:1:1) obtained via a stratified sampling method for multi-label data [43]. For all models, we set batch size as 128, decay the learning rate by 10 after every 20 epochs, add early stopping with patience of 10, and constantly monitor the max of Challenge Score on the validation set. For Transformer-based models, we set AdamW optimizer with a learning rate of 0.0001, following the setting in the Swin Transformer for image classification. We set Adam optimizer with a learning rate of 0.003 for the other models. After training stops, we reload the best weight according to the Challenge Score on the validation set. Test-time augmentation is utilized during evaluation. We report F_2, G_2 and Challenge Score on Challenge2020, and AUC on PTB-XL.

E. Baselines

We evaluate adaptations of a range of different algorithms from the literature that can be broadly categorized as convolutional neural networks, recurrent neural networks, especially Transformer architecture, which has been popular in CV and NLP recently. In detail, besides *BaT*, we evaluate Vision Transformer, Swin Transformer, and other deep learning benchmarks on PTB-XL, such as LSTM, CNN and ResNet. The 34-layers ResNet is widely used in ECG classification [44] and serves as a strong baseline. We set hyperparameters using grid search. Since vision Transformers have not been applied on ECG data, we share our settings for Challenge2020 as below, and our code will be released on https://github.com/lixiaoyu0575/Beat-aligned_Transformer.

- ***ViT***. We set patch size as 25, embedding dimension as 768, number of attention head as 12, and the depth of Transformer as 12.
- **Swin Transformer**. We construct 4 Swin Transformer blocks. The depths of the blocks are 2, 2, 6, 2, the numbers of attention heads in different blocks are 8, 16, 16, 24, the window size is set as 7, the patch size is set as 4, and the patch embedding dimension is 96.
- **Beat-aligned Transformer**. We set the numbers of SW-blocks as 2, 2, 6, 2, 2, the numbers of attention heads in different layers as 8, 16, 32, 64, 128. Patch size and window size are set as 5. The patch embedding dimension is also set as 96.

F. Performance Comparison

The experimental results on Challenge2020 are shown in Table.II. The mean and variance values from 5-fold cross validation and 5 random seeds are reported. From the table, we can observe that, among the Transformer-based methods, Beat-aligned Tranasformer achieves the best performance,

Table IV
ABLATION STUDY RESULTS ON CHALLENGE2020. THE BEST RESULTS ARE BOLD.

Model	F2	G2	Challenge Score
Beat-aligned Transformer	**0.5011±0.0034**	**0.3125±0.0036**	**0.4958±0.0039**
To Segment ECG with R peaks	0.4352±0.0039	0.2659±0.0074	0.4293±0.0065
To replace aggregation block with Patch Merging	0.4342±0.0066	0.2538±0.0067	0.4472±0.0052
To keep windows unshifted	0.4966±0.0031	0.3063±0.0029	**0.4935±0.0028**
To remove resampling ratios	0.4809±0.0051	0.2844±0.0064	0.4793±0.0035

outperforming *ViT* by a large margin. If we compare all of the models, although ResNet still performs best, Beat-aligned Transformer also achieves competitive performance.

The experimental results on PTB-XL are listed in the Table.III. The mean and variance values from 5 random seeds are reported. The results are similar to Challenge2020. The performance of Beat-aligned Transformer keeps the best among vision Transformers, and ResNet is still the best among the other baselines. Beat-aligned Transformer is more competitive to ResNet. There is no statistical performance gap between ResNet and Beat-aligned Transformer under this setting.

Overall, on both datasets, Beat-aligned Transformer consistently outperforms other vision Transformers. Compared with other baseline models, Beat-aligned Transformer also shows competitive modeling power. Note that all of the performances are achieved by training from scratch without pretraining. In computer vision, compared with ResNet, better performances of vision Transformer are only reported on the large datasets with millions of samples.

G. Ablation Study

We also conduct ablation studies on Challenge2020 to check the effectiveness of each component in the Beat-aligned Transformer. The results are shown in Table.IV. The first model is the original Beat-aligned Transformer on ECG classification. The rest models are modified from the *BaT*.

- **To Segment ECG with R-peaks**. In this setting, ECG are segmented into beats with R-peaks, and the model performance drops more than 6% on Challenge Score. In this way, R-peaks are split. However, R-peak is usually the most important and explicit features in ECG.
- **To replace aggregation block with Patch Merging**. In this setting, we remove aggregation block to check whether the information exchange among beats is important. Patch Merging is a block from Swin Transformer which is used to downsample input. The performance drop of this variant model means that the aggregation block plays an important role in the *BaT*.
- **To keep windows unshifted**. In this setting, we aim to investigate the shifted-window mechanism. It is interesting to observe that the Challnege score of this model is very close to the original *BaT*. It can be explain that, since we utilize convolution layer and maxpooling layer with a kernel size larger than 1, information from different windows can also be exchanged, not only the information from different beats. From this perspective, aggregation block shares similar function with SW Block. However, G_2 and F_2 are still lower than *BaT*, which indicates that SW Block still enhances the modeling power.
- **To remove resampling ratios**. In this setting, we try to confirm the importance of resampling ratio. The performance drop explains the informatiness of the resampling information.

Above ablation studies prove the effectiveness of each part in the *BaT*. The SW Block might share a similar function with Beat Aggregation Block, which shows a direction to simplify the *BaT* further.

V. CONCLUSION

In this paper, we investigate how to apply and boost a vision Transformer on ECG data. To fully utilize the nature of cardiac cyclicity, we propose a new ECG data format as aligned beats and a novel network called Beat-aligned Transformer (*BaT*). In the *BaT*, a hierarchical structure is specially designed to exploit the cyclicity of ECG data. Extensive experiments show that *BaT* not only outperforms *ViT* and Swin Transformer for ECG classification, but also achieves competitive performance compared with other state-of-the-art methods. Ablation studies prove the effectiveness of our hierarchical design. Our work demonstrates a new solution for ECG classification, which is promising with the rapid growth of data amount. In future work, besides collecting large external datasets, we will also explore self-supervised learning with our *BaT* on ECG data.

ACKNOWLEDGMENT

This work has been supported by National Natural Science Foundation of China (61772409); The consulting research project of the Chinese Academy of Engineering (The Online and Offline Mixed Educational Service System for “The Belt and Road” Training in MOOC China); Project of China Knowledge Centre for Engineering Science and Technology; The innovation team from the Ministry of Education (IRT_17R86); and the Innovative Research Group of the National Natural Science Foundation of China (61721002).

REFERENCES

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[2] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "Ptb-xl, a large publicly available electrocardiography dataset," *Scientific data*, vol. 7, no. 1, pp. 1–15, 2020.

[3] E. A. P. Alday, A. Gu, A. J. Shah, C. Robichaux, A.-K. I. Wong, C. Liu, F. Liu, A. B. Rad, A. Elola, S. Seyedi *et al.*, "Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020," *Physiological measurement*, vol. 41, no. 12, p. 124003, 2020.

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[5] Z. Zhang, H. Zhang, L. Zhao, T. Chen, and T. Pfister, "Aggregating nested transformers," *arXiv preprint arXiv:2105.12723*, 2021.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[7] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *arXiv preprint arXiv:2012.12877*, 2020.

[8] H. Li, X. Wang, C. Liu, Y. Wang, P. Li, H. Tang, L. Yao, and H. Zhang, "Dual-input neural network integrating feature extraction and deep learning for coronary artery disease detection using electrocardiogram and phonocardiogram," *IEEE Access*, vol. 7, pp. 146 457–146 469, 2019.

[9] Z. Li, X. Feng, Z. Wu, C. Yang, B. Bai, and Q. Yang, "Classification of atrial fibrillation recurrence based on a convolution neural network with svm architecture," *IEEE Access*, vol. 7, pp. 77 849–77 856, 2019.

[10] A. Natarajan, Y. Chang, S. Mariani, A. Rahman, G. Boverman, S. Vij, and J. Rubin, "A wide and deep transformer neural network for 12-lead ecg classification," in *2020 Computing in Cardiology*, 2020, pp. 1–4.

[11] S. Wang, C. Aggarwal, and H. Liu, "Using a random forest to inspire a neural network and improving on it," in *Proceedings of the 2017 SIAM international conference on data mining*. SIAM, 2017, pp. 1–9.

[12] S. Hong, M. Wu, Y. Zhou, Q. Wang, J. Shang, H. Li, and J. Xie, "Encase: An ensemble classifier for ecg classification using expert features and deep neural networks," in *2017 Computing in cardiology (cinc)*. IEEE, 2017, pp. 1–4.

[13] S. Datta, C. Puri, A. Mukherjee, R. Banerjee, A. D. Choudhury, R. Singh, A. Ukil, S. Bandyopadhyay, A. Pal, and S. Khandelwal, "Identifying normal, af and other abnormal ecg rhythms using a cascaded binary classifier," in *Computing in Cardiology Conference*, 2017.

[14] T. Teijeiro, C. A. García, D. Castro, and P. Félix, "Arrhythmia classification from the abductive interpretation of short single-lead ecg records," 2017.

[15] M. Zabihi, A. B. Rad, A. K. Katsaggelos, S. Kiranyaz, S. Narkilahti, M. Gabbouj, M. Zabihi, A. B. Rad, A. K. Katsaggelos, and S. Kiranyaz, "Detection of atrial fibrillation in ecg hand-held devices using a random forest classifier," in *Computing in Cardiology*, 2017.

[16] S. Goto, M. Kimura, Y. Katsumata, S. Goto, T. Kamatani, G. Ichihara, S. Ko, J. Sasaki, K. Fukuda, and M. Sano, "Artificial intelligence to predict needs for urgent revascularization from 12-leads electrocardiography in emergency patients," *PloS one*, vol. 14, no. 1, p. e0210103, 2019.

[17] R. Xiao, Y. Xu, M. M. Pelter, D. W. Mortara, and X. Hu, "A deep learning approach to examine ischemic st changes in ambulatory ecg recordings," *AMIA Summits on Translational Science Proceedings*, vol. 2018, p. 256, 2018.

[18] Y. Xia, N. Wulan, K. Wang, and H. Zhang, "Atrial fibrillation detection using stationary wavelet transform and deep learning," in *2017 Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.

[19] S. L. Oh, E. Y. Ng, R. San Tan, and U. R. Acharya, "Automated diagnosis of arrhythmia using combination of cnn and lstm techniques with variable length heart beats," *Computers in biology and medicine*, vol. 102, pp. 278–287, 2018.

[20] O. Yildirim, U. B. Baloglu, R.-S. Tan, E. J. Ciaccio, and U. R. Acharya, "A new approach for arrhythmia classification using deep coded features and lstm networks," *Computer methods and programs in biomedicine*, vol. 176, pp. 121–133, 2019.

[21] B. Yuen, X. Dong, and T. Lu, "Inter-patient cnn-lstm for qrs complex detection in noisy ecg signals," *IEEE Access*, vol. 7, pp. 169 359–169 370, 2019.

[22] F. Zhu, F. Ye, Y. Fu, Q. Liu, and B. Shen, "Electrocardiogram generation with a bidirectional lstm-cnn generative adversarial network," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[23] Y. Xu, S. Biswal, S. R. Deshpande, K. O. Maher, and J. Sun, "Raim: Recurrent attentive and intensive model of multimodal patient monitoring data," in *Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, 2018, pp. 2565–2573.

[24] O. Yildirim, U. B. Baloglu, R.-S. Tan, E. J. Ciaccio, and U. R. Acharya, "A new approach for arrhythmia classification using deep coded features and lstm networks," *Computer methods and programs in biomedicine*, vol. 176, pp. 121–133, 2019.

[25] S. P. Shashikumar, A. J. Shah, G. D. Clifford, and S. Nemati, "Detection of paroxysmal atrial fibrillation using attention-based bidirectional recurrent neural networks," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 715–723.

[26] S. Saadatnejad, M. Oveisi, and M. Hashemi, "Lstm-based ecg classification for continuous monitoring on personal wearable devices," *IEEE journal of biomedical and health informatics*, vol. 24, no. 2, pp. 515–523, 2019.

[27] A. Habib, C. Karmakar, and J. Yearwood, "Impact of ecg dataset diversity on generalization of cnn model for detecting qrs complex," *IEEE access*, vol. 7, pp. 93 275–93 285, 2019.

[28] R. Rahim, S. Nadeem *et al.*, "End-to-end trained cnn encoder-decoder networks for image steganography," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[29] E. Fotiadou, T. Konopczyński, J. Hesser, and R. Vullings, "Deep convolutional encoder-decoder framework for fetal ecg signal denoising," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[30] B. Zhou, S. Liu, B. Hooi, X. Cheng, and J. Ye, "Beatgan: Anomalous rhythm detection using adversarially generated time series." in *IJCAI*, 2019, pp. 4433–4439.

[31] P. A. Warrick and M. N. Homsi, "Ensembling convolutional and long short-term memory networks for electrocardiogram arrhythmia detection," *Physiological measurement*, vol. 39, no. 11, p. 114002, 2018.

[32] J. H. Tan, Y. Hagiwara, W. Pang, I. Lim, S. L. Oh, M. Adam, R. San Tan, M. Chen, and U. R. Acharya, "Application of stacked convolutional and long short-term memory network for accurate identification of cad ecg signals," *Computers in biology and medicine*, vol. 94, pp. 19–26, 2018.

[33] D. Jia, W. Zhao, Z. Li, J. Hu, C. Yan, H. Wang, and T. You, "An electrocardiogram delineator via deep segmentation network," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 1913–1916.

[34] X.-C. Cao, B. Yao, and B.-Q. Chen, "Atrial fibrillation detection using an improved multi-scale decomposition enhanced residual convolutional neural network," *IEEE Access*, vol. 7, pp. 89 152–89 161, 2019.

[35] G. Yan, S. Liang, Y. Zhang, and F. Liu, "Fusing transformer model with temporal features for ecg heartbeat classification," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019, pp. 898–905.

[36] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "Neurokit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, Feb 2021. [Online]. Available: https://doi.org/10.3758/s13428-020-01516-y

[37] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.

[38] H. Bao, L. Dong, F. Wei, W. Wang, N. Yang, X. Liu, Y. Wang, J. Gao, S. Piao, M. Zhou *et al.*, "Unilmv2: Pseudo-masked language models for unified language model pre-training," in *International Conference on Machine Learning*. PMLR, 2020, pp. 642–652.

[39] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3588–3597.

[40] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3464–3473.

[41] Z. I. Attia, P. A. Noseworthy, F. Lopez-Jimenez, S. J. Asirvatham, A. J. Deshmukh, B. J. Gersh, R. E. Carter, X. Yao, A. A. Rabinstein, B. J. Erickson *et al.*, "An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction," *The Lancet*, vol. 394, no. 10201, pp. 861–867, 2019.

[42] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ecg analysis: Benchmarks and insights from ptb-xl," *arXiv preprint arXiv:2004.13701*, 2020.

[43] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2011, pp. 145–158.

[44] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature medicine*, vol. 25, no. 1, pp. 65–69, 2019.

Appendix

- Hardware and software for experiments on Challenge2020:
  The experiments were run on the same machine with Intel(R) Core(TM) i9-7900X CPU @ 3.30GHz, 20 cores, 128 GB of RAM and 4 GPUs (all NVIDIA TITAN V, 12GB of video memory) The OS of the server is Ubuntu, 16.04.5LTS. The Python version is 3.5, and Pytorch 1.5.0 is used as the deep learning framework. CUDA version is 10.1, and cuDNN version is 7.4. The version of torch is 1.5.0+cu101. The version of torchvision is 0.6.0+cu101. The version of numpy is 1.18.3.
- Hardware and software for experiments on PTB-XL:
  The experiments were run on the same machine with Intel Core Processor (Broadwell) CPU@2.3GHz, 16 cores, 128G RAM and 2 GPUs(Nvidia Tesla P100-PCIE, 16G of video memory) Software The OS of the server is Centos 7.3. The python version is 3.6.5, and pytorch 1.6.0 and fastai 1.0.61 are used as deep learning framework. Cuda version is 11.1.