Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# A Multi-View Learning Approach to Enhance Automatic 12-Lead ECG Diagnosis Performance

**Jae-Won Choi**
Infomining, Co. Ltd

**Dae-Yong Hong**
Infomining, Co. Ltd

**Chan Jung**
Infomining, Co. Ltd

**Eugene Hwang**
Korea Advanced Institute of Science and Technology

**Sung-Hyuk Park**
sunghyuk.park@kaist.ac.kr

Korea Advanced Institute of Science and Technology

**Seung-Young Roh**
Korea University Medicine Guro hospital

**Article**

**Keywords:**

**Posted Date:** March 29th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-2725046/v1

**License:** 🅭 🅯 This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

**Version of Record:** A version of this preprint was published at Biomedical Signal Processing and Control on July 1st, 2024. See the published version at https://doi.org/10.1016/j.bspc.2024.106214.

# Abstract

The electrocardiogram (ECG) has important clinical value for the early diagnosis of cardiovascular diseases. Recently, the performance of existing diagnosis models based on ECG data has improved with the introduction of deep learning (DL). However, the impact of various combinations of multiple DL components and/or the role of augmentation techniques on the diagnosis have not been sufficiently investigated in this field. In this sense, this study aims to design an integrated model consisting of diverse DL-based modules. Here, an ensemble-based multi-view learning approach with an ECG augmentation technique is proposed to achieve higher performance than traditional automatic 12-lead ECG diagnosis methods. Accordingly, several experiments have been conducted with CPSC2018 dataset for evaluation. The proposed model reports F1 score of 0.840, which outperforms existing state-of-the-art methods in the literature. Thus, this study provides quantitative evidence demonstrating that the multi-view learning approach can be used as a unified algorithmic method in the field of bioinformatics.

# Introduction

Cardiovascular diseases account for one-third of annual deaths worldwide[1]. The electrocardiogram (ECG), which records the electrical activity of the heart as a waveform, is one of the most basic tests available to measure the heart health, which in turn is critical for the management and early diagnosis of cardiovascular diseases. However, high level of expertise is required for the accurate reading of ECG signals. Furthermore, the large number of ECG records, over 300 million cases worldwide every year, places a huge burden on medical staff[2]. To address this issue, the 12-lead ECG device has provided an algorithm-based computerized interpretation of ECG, which has shortened the reading time by 24–28%[3]. Nevertheless, the use of this device has been limited due to its high false-positive rate. For this reason, the need for fast and accurate computer-aided automated ECG interpretation has emerged.

There are three main streams of research on automatic ECG diagnosis. The first is handcrafted feature–based conventional machine learning research[4–7]. However, owing to ECG artifacts and individual differences, these are difficult to generalize and cannot be configured in an end-to-end manner. The second is mathematical modeling–based research. Attempts have been made to model cardiac dynamics by analyzing ordinary differential equations. However, strict formalization is challenging owing to the numerous variables and inherent non-linearity of ECG[8]. Finally, deep learning (DL)–based ECG studies have recently attracted considerable attention. In fact, DL has achieved remarkable success in general tasks over the past decade[9] and its potential as a diagnostic aid in the medical field has been increasing[10–12]. DL components, such as convolutional neural networks (CNNs)[13–17], long short-term memory (LSTM)[18–21], and attention[22], that excel in general tasks within the medical field have been adopted. Considering that ECG has a complex nature of bio-signals, periodic signals, time series, and multi-sensory data, the integration of more varied techniques has been explored[23]. As such, the modern trend of DL-based ECG diagnosis research is to understand ECG more deeply by combining various DL components[24–28].

In previous DL-based studies of ECG diagnosis, end-to-end classification models, which implicitly extract features from raw signals, were developed with good predictive performance. However, most of them restricted the input signal to a fixed length with cleansing procedures. Specifically, zero-padding, cut-out, and/or signal segmentation were utilized to unify the input length. Moreover, filtering techniques such as median filter, wavelet transform, and/or Butterworth filter were implemented to minimize the noises that disturb the interpretation of ECG signals.

In terms of the structure of existing models, He et al.[20] and Luo et al.[21], which achieved F1 scores of 80.6% and 82.2%, respectively, stacked bidirectional LSTM (BiLSTM) layers to a CNN module so that the model could extract both spatial and temporal features. Ye and Lu[24] and Ye et al.[28] attempted to fuse multiple features from the DL components through XGBoost. Wang et al.[16] implemented the attention mechanism on top of a multi-scale CNN layers, reporting a F1 score of 82.8%. Ge et al.[25] utilized Squeeze-and-Excitation (SE) blocks[29] to refine the importance between CNN features. Chen et al.[26] combined CNN, bidirectional RNN (BiRNN), and attention layers to successfully re-scale spatiotemporal features and resulted in a F1 score of 83.7%. To the best of our knowledge, Chen et al. achieved the best F1 score among the comparable studies that used the China Physiological Signal Challenge (CPSC) 2018 database[30].

In this study, we propose a DL model that classifies nine arrhythmia classes with 30 s of 12-lead ECG signals fed as the model input. Given that most previous DL-based studies explore a single way of combining at most two DL components (e.g., CNN, gated recurrent unit (GRU), or attention), we investigated multiple combination of more diverse DL components in the sense of multi-view learning[31]. To evaluate the model performance, candidates of DL components with outstanding practical performance are listed, and their combinations are compared for the investigation of the set that reports better performance. Moreover, because many previous studies[13–15, 18, 26, 27] did not regard the artifacts in ECG signals[32], we propose an ECG-specific data augmentation technique to simulate various ECG artifacts (Fig. 1a). In addition, to learn a more general representation for ECG arrhythmia classification, the degree and number of transformations were fully random, providing diverse data to the model.

The main contributions of this study are as follows: (1) We propose a data augmentation technique specialized for ECG and induce models to perform robust ECG interpretation. (2) With respect to multi-view learning, we investigate various architectures through a combination of DL components, and confirm that by synthesizing them, improved performance can be achieved.

## Results

# Performance Evaluation

A brief overview of the dataset and experiments is described here. The details are described in the *Methods* section. The CPSC2018 dataset contains approximately 10,000 records of 12-lead ECG signal, where each corresponds to one of the nine classes of arrhythmia (AF, I-AVB, LBBB, RBBB, PAC, PVC, STD, STE, and normal). Because only about 70% of the data can be publicly accessed, we split the public dataset into *our training set, our validation set*, and *our test set*. The remaining 30% of the data (i.e., official hidden test set) are marked as *the hidden test set* in this literature. Here, we reported F1 scores to demonstrate comparative result with other challenge participants. Besides the results on our test set, evaluation on the official of the hidden test set was requested to the challenge organizer.

Given an ECG data $X \in \mathbb{R}^{L \times C}$ where *L* and *C* refer to signal length and the number of channels, respectively, our goal is to predict the arrhythmia class $Y \in \mathbb{R}^{K}$of the given ECG input, where *K* refers to the number of arrhythmia classes. Figure 1c illustrates the overview of model training and inference. Herein, *M* independent models are trained based on the preprocessed ECG dataset. All models share the input data, which are segmented into the same length and augmented through a domain-targeted technique. Specifically, the input data are segmented by windowing to deal with varied input lengths. Subsequently, the predictions of *N* models over *S* segments are averaged into the final output. As a result, an ensemble of *N* best-performing models is devised for inference based on these *M* trained models.

## Single Model Results on Our Test Set

The performance of the top 10 models based on the macro average F1 score selected based on our validation set is reported in Table 1. All models except for M10 showed a performance of over 81%, with a difference of up to 0.6%. The best model showed a performance of 84.56%, which was the best for AF, I-AVB, LBBB, and STE. The performance by class exhibited a similar trend in all models. RBBB was the best classified class with a performance of 94% or more in all models, and performances on RBBB, LBBB, and AF were more than 92% in almost all models. The classification performance of PVC, STD, I-AVB, and SNR were more than 80%. In addition, the classification performance of PAC was more than 60%, whereas STE was less than 60%. This largest performance variance in classifying STE among the models may have occurred due to possible noise in labeling STE because the diagnosis of STE is influenced by the experience of doctors and subjectivity[33].

Furthermore, Table 1 lists the combination of the components used in each model. The ResNet-18 backbone and bidirectional gated recurrent unit (BiGRU) with lead-axis unfolding were selected for all models. This suggests that the model complexity of ResNet-34 is too high for ECG signal. We presumed that the GRU was always included for implementation owing to its robust nature in noisy ECG[30]. Lead-axis unfolding may be adopted because it is bidirectional and considers the relationship between leads well. Conversely, the time-axis unfolding seems to have failed in enhancing performance because of the excessive concentration of the local relations. Referring to Chen et al., who stated the different roles of each lead, the lead axis may be successful because inter-lead characteristics are more important than intra-lead characteristics for the

Table 1

Architectural components of top-10 single models and their performances with our test set. *O* and *X* refer to *applied* and *not applied* respectively. *r* indicates the reduction ratio for the SE block. *Lead* stands for lead-axis unfolding. *Element* and *Instance* refer to inter-lead and intra-lead attention respectively. The best F1 score is presented in **bold**.

| Model | Backbone | Augment | SE block | BiGRU | Attention | Activation | F1 |
|---|---|---|---|---|---|---|---|
| M01 | ResNet-18 | O | X | Lead | Element | ELU | **0.8456** |
| M02 | ResNet-18 | O | X | Lead | Element | ReLU | 0.8406 |
| M03 | ResNet-18 | O | X | Lead | Instance | ReLU | 0.8396 |
| M04 | ResNet-18 | O | X | Lead | Instance | ReLU | 0.8331 |
| M05 | ResNet-18 | O | r = 2 | Lead | Element | Leaky ReLU | 0.8321 |
| M06 | ResNet-18 | O | r = 2 | Lead | Instance | Leaky ReLU | 0.8296 |
| M07 | ResNet-18 | O | r = 2 | Lead | Element | ReLU | 0.8259 |
| M08 | ResNet-18 | O | X | Lead | Element | Leaky ReLU | 0.8243 |
| M09 | ResNet-18 | O | r = 2 | Lead | Instance | ELU | 0.8186 |
| M10 | ResNet-18 | O | r = 2 | Lead | Element | ELU | 0.7970 |

Table 2

Top-5 ensemble performances of our test set and with the hidden test set. *Model IDs* indicate the single model number in Table 1. *CPSC2018 Top Models* refer to the performance of previous works posted on the official webpage. Since paper submission was not mandatory, some of the existing models were not published as papers. The best F1 score in each column is written in **bold**. Following the official evaluation of the CPSC2018, all classes were grouped into four types except for *normal*. In this sense, $F_{AF}, F_{Block}, F_{PC}, andF_{ST}$ indicate F1 score for AF, heart blocks (I-AVB, LBBB, and RBBB), premature contractions (PAC and PVC), and ST-segment abnormality (STD and STE), respectively. F1, without any specification, stands for the average score over total of nine classes including sinus rhythm.

| Case | | Model IDs | Our Test Set | Hidden Test Set | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | **F1** | $F_{AF}$ | $F_{Block}$ | $F_{PC}$ | $F_{ST}$ | **F1** |
| Our Models | E01 | 02, 03, 04, 05, 06, 09 | **0.8647** | 0.922 | 0.913 | 0.819 | 0.772 | 0.829 |
| | E02 | 02, 03, 04, 05, 07, 08, 09 | 0.8598 | 0.927 | 0.912 | 0.818 | 0.781 | 0.833 |
| | E03 | 02, 03, 04, 08, 09 | 0.8574 | 0.922 | 0.909 | 0.816 | 0.778 | 0.830 |
| | E04 | 01, 02, 03, 04, 09 | 0.8557 | 0.920 | 0.912 | 0.817 | 0.782 | 0.831 |
| | E05 | 01, 02, 04, 05, 10 | 0.8435 | 0.925 | **0.915** | 0.828 | **0.790** | **0.840** |
| CPSC2018<br>Top Models | Chen et al. | *N/A* | N/A | **0.933** | 0.899 | **0.847** | 0.779 | 0.837 |
| | Cai et al. | *N/A* | N/A | 0.931 | 0.912 | 0.817 | 0.761 | 0.830 |
| | He et al. | *N/A* | N/A | 0.914 | 0.879 | 0.801 | 0.742 | 0.806 |
| | Yu et al. | *N/A* | N/A | 0.918 | 0.890 | 0.789 | 0.718 | 0.802 |
| | Yan et al. | *N/A* | N/A | 0.924 | 0.882 | 0.779 | 0.709 | 0.791 |

analysis of arrhythmia. Moreover, attention was used in all top-10 models, where element attention (6 out of 10) prevailed over instance attention (4 out of 10). In addition, the models reported better performance when squeeze and excitation (SE) blocks were not used (4 of the top-5). This may be because small-scale features were lost as the reduction ratio increased. It can be observed that r = 2 was adopted in all cases when the SE block was used.

# Ensemble Model Results on Our Test Set and the Hidden Test Set

Based on the top-10 models, we compared all ensemble combinations of 1013 cases, where the number of models ranges from two to ten. The performances of our test set and the hidden test set for the top-5 ensembles are presented in Table 2, which further contain the hidden test set results with top-5 participants reported in CPSC2018.

The results of five ensembles evaluated on the hidden test set were 0.5–3.6% lower than our test set. However, the difference between ensembles was reduced from 2.12–1.1%, and all ensemble performances exhibited second to third place performances. This suggests that the multi-view demonstrates a stable inference. The best ensemble was E05, which achieved a state-of-the-art performance of 84%. Specifically, E05 had the lowest average F1 in our test set and was well-generalized as the difference in performance between our test set and the hidden test set was minimal (0.35%).

# Discussion

Multi-view learning trains multiple models to extract different representations of identical data, aiming to derive a good representation[34]. Here, multi-view learning involves two basic principles[35]. The first is consistency, which states that the context should be consistent even if the views are different. In our experiment, each model had the same inference goal for the same task. The second is the complementary principle, which states that each view should be able to complement the other. In this sense, the CNN feature from the proposed model considers the shape of the waveform while the RNN feature considers the temporal context. Accordingly, the latter complements the former with temporal information. Because we focused on the difference in viewpoints between different components, models of various structures were assembled on the same data. To fuse the model outputs, we averaged all probability vectors with the intention of equal contribution of all models. Thus, our model obtained an accurate classification ability, noise immunity, and the ability for rich representation by synthesizing different central features.

As demonstrated in Table 2, the ensemble based on five to seven models exhibited the best performance on our test set (i.e., average F1 of 84.35–86.47%), suggesting that the optimal number of models was approximately six. In fact, the ensemble of all ten models did not show the optimal performance, which is consistent with the law of diminishing returns[37]. Moreover, the performances of E01 and E05 differed by 2.12%, but E05 (84.35%) still exceeded all single models except for M01 (84.56%). This implies that the fusion of classifiers (i.e., ensemble) may have encouraged the models to complement each other's weakness, leading to superior performance. Therefore, the effectiveness of classifier fusion based on multiple views was confirmed.

Furthermore, we analyzed the trade-off within the number of models of an ensemble using our test set (Fig. 2). Intuitively, the number of models and the predictive performance were proportional. Because the interquartile range decreases from 1.32–0.565% as the number of models increases, it can be inferred that the greater the number of models, the more stable it is. Given that the ensemble of three models showed the largest performance increase of 0.28% and the maximum g-mean of 0.7046, we may conclude that the optimal number of models for good predictive performance was three.

The results for our test sets E01–E05 are shown as red dots in Fig. 2. All four ensembles, except E05, outperformed the average of the ensembles with the same number of models. Nevertheless, given that the best performing model on the hidden test set was E05, E01–E04 may have been fitted only to the public data. In fact, according to Bonab et al.[37], an optimal ensemble requires independence between models. This implies that M02 and M04 adopted in all ensembles are independent and complementary. Specifically, M09 was used for E01–E04, whereas M10 was used for E05. This indicates that a set of M02, M04, and M09 and a set of M02, M04, M10 is the optimal for our test set and the hidden test set, respectively. In Fig. 2, the two sets of M02, M04, M09 and M02, M04, M10 are denoted as O_1 and O_2, respectively. Here, the relationship between O_1 and O_2 is similar to that of E01, E04, and E05. Thus, this indicates that the independence and complementary between the models were consistent throughout the experiment.

The limitations of this study are as follows. We reviewed a number of model architectures and their complementary cooperations through ensembles. However, due to the computational limit, not every DL component and combination has been explored. There are three possible further improvements for this literature. First, searching broader hyperparameter space may lead to a more optimal performance. Second, the method of how to fuse features was not investigated in various ways. Other methods such as representation-level fusion may possibly help extract better representations. Lastly, another way of interactions within components or the application of other DL components may allow the model with a more powerful ability to analyze ECGs. Given that we restricted the number of conditions for each component to at most four cases, it would be worth applying new variants or components to observe closer interactions.

## Conclusion

Up to the early 2010s, general DL-based techniques were mainly focused on the difference in performance according to the model architecture or hyperparameters. Their successful results stimulated many researchers to introduce some DL components to the automatic ECG diagnosis and achieved significant performance improvements. However, previous studies have not focused on the difference between the combination of various DL components and separated the ECG artifacts from the model. Nevertheless, recent DL studies for general tasks concentrated on what good representations are and how to obtain them. Specifically, these representation-based techniques commonly focus on synthesizing features from various perspectives, where multi-view learning is one of these concepts. Inspired by these works, we constructed and compared ensembles of models, each of which analyzes ECGs from different perspectives based on multi-view learning. In addition, we suggested augmentation specialized for ECG, allowing the model to observe more diverse data. Accordingly, experimental results demonstrated that the integration of models outperform state-of-the-art performance. Overall, the results of this study demonstrate the possibility of further improvements in our approach by incorporating a richer combination of DL components.

# Methods

## Data Augmentation for ECG

Because data augmentation must preserve labels, it is difficult to apply computer vision techniques, such as vertical flipping. Given that ECG data is similar to the three-channel bio-signal data of Parkinson’s disease, we propose an ECG augmentation technique (Fig. 1a) by referring to physiological signal augmentation techniques for Parkinson’s disease data[38] and for ECG signal[39]. The proposed random augmentation method is based on the following six operations.

- Jittering is applied by additive high-frequency noise and represents power line interference (PLI).
- Scaling introduces electromyography (EMG) noise by adjusting the amplitude of the signal.
- Magnitude warping simulates baseline wander by distorting the magnitude of an arbitrary location.
- Time warping provides small changes to each point of the ECG along the x-axis; therefore, the interval characteristics of signal components such as PR, RR, or QT intervals.
- Permutation prevents position memorization by reversing positions between arbitrary sections within the ECG.
- RandomSample resamples based on non-uniform random points. The sampling rate was lowered by up to 20% from the original, representing different devices or environments. This may lead to a signal quality degradation by losing a small number of signal details

The parameters of each transformation were experimentally set according to the characteristics of ECG signal. Parameters are not specified as constants but randomly assigned for each mini-batch within a heuristically defined parameter space. The number of transformations to be applied is also randomly determined.

## Components and Their Roles in ECG Diagnosis

## Characteristics of ECG

In automatic ECG diagnosis, multiple factors should be considered. Morphological information is extracted from the shape of the waveform (i.e., P, QRS, and T). Temporal information is obtained by referencing the intervals between the waveforms or peaks. The dependency within the points in each lead or between the leads must be considered because certain patterns of ECG, which may be recorded within the same lead or different leads, specify both the lesion and the region of the heart it originates from. Furthermore, three characteristics of ECG signal including integrity, diversity, and periodicity were highlighted by Liu et al.[23]. Therefore, various types of information should be considered in the automatic ECG diagnosis.

## Combination of DL Components Based on Multi-View Learning

Multi-view learning can be broadly classified into three types, data-level, classifier-level, and representation-level integration, according to the integration point[36]. The data-level learning integrates different types of data in the embedding process and uses techniques such as principal component analysis (PCA). The classifier-level learning integrates the predictions of the models trained using different views. Ensemble is a representative example of a classifier-level scheme. The representation-level learning incorporates latent vectors and is often used in multimodal tasks. In this study, to utilize ECG augmentation and the representation ability of the model, an ensemble was constructed based on the second category. Nonetheless, it is also related to the third category in that it induces fusion between representations from diverse leads through interactions between components. As a result, this configuration of layers aimed to facilitate information exchange between the signal segments, enabling higher-level inference.

# CNN

CNNs efficiently extract spatial features and are parameter-efficient. In this study, a CNN was adopted as the basic component (Fig. 1b1). Here, 1D convolution and depthwise separable convolution[40] were used to separate the features for each channel, increasing the computational efficiency. Referring to the effect of the residual connection in ECG[20, 41], ResNet was chosen as the default architecture.

# RNN

RNN is effective in reflecting the sequential characteristics of time-series data. In this sense, this study utilized GRU[42] for computational efficiency. The GRU operates as follows:

$$r_t = sigmoid\left(W_r x_t + U_r h_{t-1}\right)$$

1

$$h_t = z_t h_{t-1} + \left(1 - z_t\right) \tilde{h}_t,$$

2

where $z_t = sigmoid(W_z x_t + U_z h_{t-1})$ and $\tilde{h}_t = tanh(W x_t + U(r_t * h_{t-1}))$. Here, equations 1 and 2 represent a reset gate and an update gate, respectively, where $W_r, W_z, W, U_r, U_z$ and $U$ are trainable parameters. Considering that ECG does not have unidirectional dependency, BiGRU was used as a component (Fig. 1b2). Moreover, because the 12-lead ECG consists of a temporal axis and a spatial axis, we considered t cases of time-axis unfolding and lead-axis unfolding.

# Attention

Attention introduces weights to the encoding process in the neural machine translation (NMT) task[43], emphasizing key information. The experiment was performed as shown in Fig. 1b3 with reference to the method in Chen et al[26] (instance attention). In addition, to consider the inter-lead difference, a point-wise attention (element attention) method was added. Both types of attention layers share the operation as follows:

$$u = tanh\left(W_1 X^T\right) \tag{3}$$

$$a = softmax\left(W_2 u\right) \tag{4}$$

$$z = \sum_{i=1}^{L} a_i^T \bullet x_i, \tag{5}$$

where $X \in \mathbb{R}^{L\times C}$, $W_1 \in \mathbb{R}^{C\times C}$, and $W_2 \in \mathbb{R}^{L\times C} or \mathbb{R}^{C\times C}$. For simplicity, the bias was omitted. Here, $X, a, z, L,$and $C$ refers to the input, weight to be applied, output, input length, and the number of channels, respectively. $W_1$ and $W_2$ are parameters where $W_2 \in \mathbb{R}^{1\times C}$ for instance attention and $W_2 \in \mathbb{R}^{C\times C}$ for element attention.

## SE Block

The SE network[29] rescales the channel-wise feature response based on its importance to the task. The SE operations, which summarize the feature map and importance, respectively, can achieve a high-performance improvement with a small parameter increase (Fig. 1b4). Ge et al.[25] introduced an SE block for ECGs and confirmed that the degree to which each character is reflected in the target diagnosis can be well readjusted.

## Experimental Setup

## CPSC2018 Database

Prior to the 1990s, researchers constructed databases such as MIT-BIH[44], PTB[45], and INCART[46] for automatic ECG diagnosis. However, most were small-scale and contained fewer than 12 leads. With the recent advent of the DL approach, large-scale databases such as CPSC2018 and PTB-XL[47] have emerged. We experimented with the CPSC2018 dataset using varied-length ECGs with more abundant target diagnoses. CPSC2018 contains 6877 public ECG records and 2954 hidden ECG records. The database consists of 12-lead ECGs with a sampling rate of 500 Hz and includes nine arrhythmia classes (AF, I-AVB, LBBB, RBBB, PAC, PVC, STD, STE, and normal). Specifically, the public dataset contains 476 multi-label records, where the challenge's official evaluation defines the classification of multi-label data as sufficiently correct if the classified outcome is one of its labels. Inappropriate data containing NaN were removed and only z-score normalization was applied for preprocessing data.

For the experiment, the public dataset was split with a ratio of 8:1:1, namely our training set, our validation set, and our test set, respectively. Here, ECG data with multiple labels were included in the test during the data split. The length of the signals in the public dataset ranges from 6 s to 60 s, which were segmented into 30 s. If the length is shorter than 30 s, zero-padding was applied to unify the length of all data. The hidden test set in this literature indicates the ECG records that are not disclosed by the challenge organizer. Every model was trained using our training set, whereas the hyperparameters were tuned based on our validation set. Our test set was used for internal evaluation, and the final top-5 ensembles were evaluated by the challenge organizer with the hidden test set for external validation.

# Implementation Details

Experiments were performed using TensorFlow 2.5, on an Intel Xeon Gold 5218, GEFORCE 3090 24GB X 4, 192GB RAM, CentOS. The overall composition of the experiment is as follows. First, the architecture was constructed using a combination of various components. Because all possible scenarios cannot be tested, the combinations and parameter ranges were heuristically limited. Thus, we defined six components of augmentations, backbone, SE blocks, activation functions, BiGRU, and attention.

As described in the *Data Augmentation for ECG* section, the augmentation consists of six operations. For all cases, the ECG-specific augmentation was applied. Here, the number, types, or parameters of operations are fully random for each mini-batch. ResNet-18 and − 34 were tested as the CNN backbone, both with 1D depthwise separable convolution operation. Because ECG has a lower dimension than images, the number of channels in all convolutional blocks is reduced by half. Specifically, 16 kernels (reduced by half again) were used instead of 32 channels in the first layer, considering that the ECG has 12 channels. All kernel sizes were set as three. Moreover, SE block, BiGRU, and/or attention components were stacked in addition to the convolutional feature extractor. The reduction ratio *r* of the SE block was set as 2, 4, and 8 to consider the trade-off between the complexity and performance. BiGRU and attention layers were set with time- or lead-axis unfolding, and instance or element attention, respectively. Specifically, BiGRU was added as a single layer with a dropout rate of 0.2, tanh activation, and 256-length input and output, where lead-axis unfolding (Fig. 1b2) was implemented by transposing the input. The attention component was applied following the BiGRU layer. Because both include weighted summation (Eq. 5), the global average pooling is added only when attention is not used. Lastly, a softmax dense classifier was applied at the end of the model architecture, whereas the activation functions of linear, ReLU, ELU, and Leaky ReLU for other layers were compared. As a result, 216 architectures in total were reviewed under these conditions.

Model training was carried out for 100 epochs with a batch size of 256. The initial learning rate was set as 0.001 or 0.01, which decreased according to the cosine decay rule. Adam was used as the optimizer. For each case, the optimal hyperparameters were tuned based on our validation set in the search space. After the convergence, the candidate set of ten optimal performances were composed for ensembles. All possible ensembles were constructed based on a random number of models in the candidate set, and the top-5 ensembles were selected. In the inference step, the final inference was derived by averaging the probabilities for each inferred 30 s segment. The evaluation was based on the challenge’s official metric. We reported the performances of our test set and the hidden test set.

# Declarations

## Data availability

The CPSC2018 dataset is publicly available at the official challenge page http://2018.icbeb.org/Challenge.html.

## Author contributions statement

J.C., C.J., and D.H. conceived and conducted the experiments, analyzed results, and wrote the first draft. E.H., S.P., and S.R. revised the subsequent versions. All authors reviewed and approved the submitted version of the manuscript.

## Competing interests

The authors declare no competing interests.

## References

1. World Health Organization. Regional Office for South-East Asia. WHO south-east asia journal of public health, volume 8, issue 2, september 2019. Place: New Delhi Publisher: World Health Organization. Regional Office for South-East Asia.
2. Hedén, B. et al. Detection of frequently overlooked electrocardiographic lead reversals using artificial neural networks. 78, 600–604, DOI: 10.1016/S0002-9149(96)00377-3.
3. Smulyan, H. The computerized ECG: Friend and foe. 132, 153–160, DOI: 10.1016/j.amjmed.2018.08.025.
4. Jambukia, S., Dabhi, V. & Prajapati, H. Classification of ECG signals using machine learning techniques: A survey.
5. Macfarlane, P., Devine, B. & Clark, E. The university of glasgow (uni-g) ECG analysis program. In Computers in Cardiology, 2005, 451–454, DOI: 10.1109/CIC.2005.1588134 (IEEE).
6. de Chazal, P., Celler, B. & Reilly, R. Using wavelet coefficients for the classification of the electrocardiogram. In Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No.00CH37143), vol. 1, 64–67, DOI: 10.1109/IEMBS.2000.900669 (IEEE).
7. Aziz, S., Khan, M. U., Choudhry, Z., Aymin, A. & Usman, A. ECG-based Biometric Authentication using Empirical Mode Decomposition and Support Vector Machines. Pages: 0912.
8. Zewdie, G. & Xiong, M. Fully automated myocardial infarction classification using ordinary differential equations. 1410.6984.

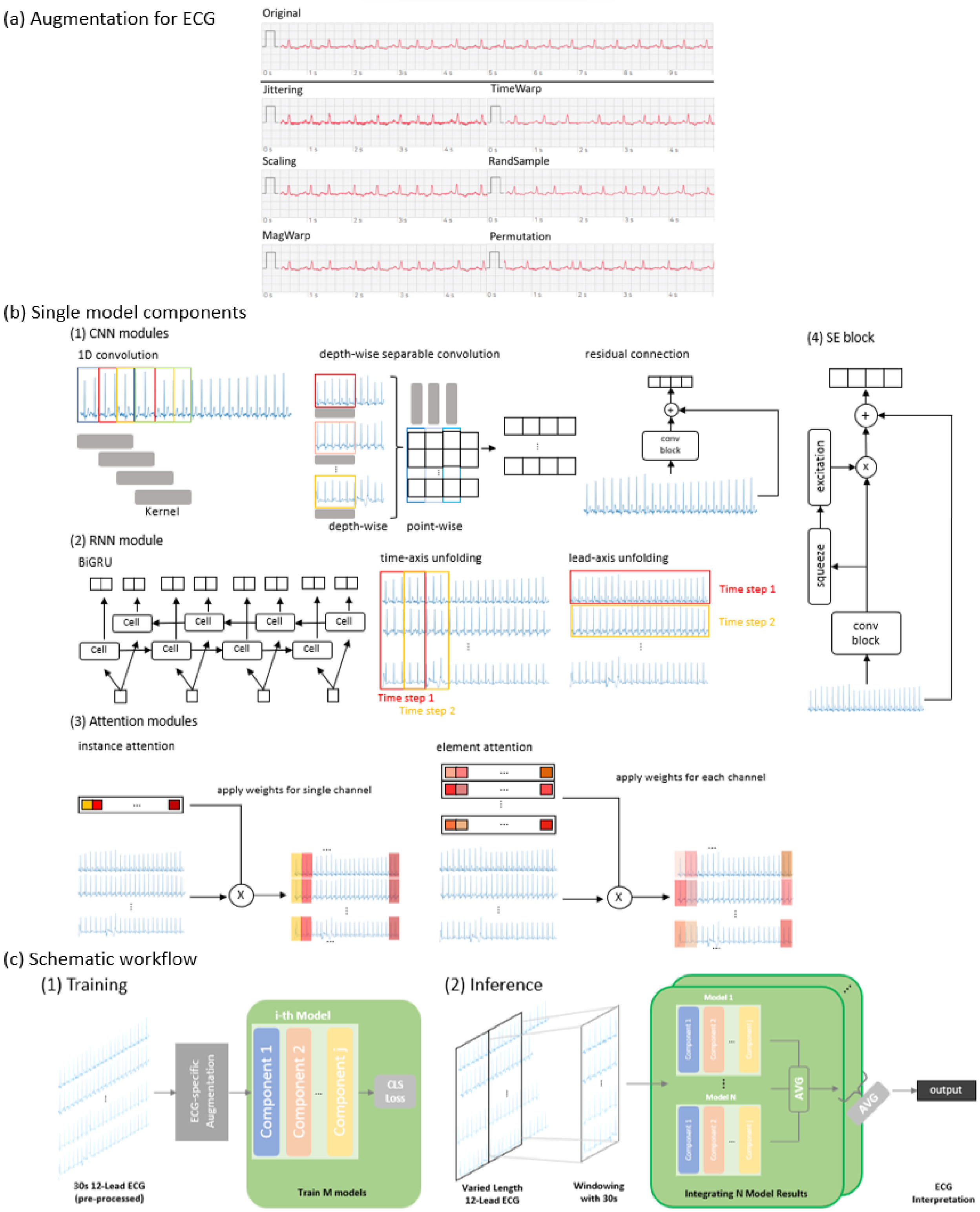
9. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. 521, 436–444, DOI: 10.1038/nature14539. Number: 7553 Publisher: Nature Publishing Group.
10. Wang, D., Khosla, A., Gargeya, R., Irshad, H. & Beck, A. H. Deep learning for identifying metastatic breast cancer. 1606.05718.
11. Esteva, A. et al. Dermatologist–level classification of skin cancer with deep neural networks. 542, 115–118, DOI: 10.1038/nature21056.
12. Li, R. et al. Deep learning based imaging data completion for improved brain disease diagnosis. In Golland, P., Hata, N., Barillot, C., Hornegger, J. & Howe, R. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, vol. 8675, 305–312, DOI: 10.1007/978-3-319-10443-0_39 (Springer International Publishing). Series Title: Lecture Notes in Computer Science.
13. Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C. & Ng, A. Y. Cardiologist-level arrhythmia detection with convolutional neural networks. 1707.01836.
14. Acharya, U. R. et al. Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. 415, DOI: 10.1016/j.ins.2017.06.027.
15. Liu, Z., Meng, X., Cui, J., Huang, Z. & Wu, J. Automatic identification of abnormalities in 12-lead ECGs using expert features and convolutional neural networks. In 2018 International Conference on Sensor Networks and Signal Processing (SNSP), 163–167, DOI: 10.1109/SNSP.2018.00038.
16. Wang, R., Fan, J. & Li, Y. Deep multi-scale fusion neural network for multi-class arrhythmia detection. 24, 2461–2472, DOI: 10.1109/JBHI.2020.2981526. Conference Name: IEEE Journal of Biomedical and Health Informatics.
17. Zhang, D., Yang, S., Yuan, X. & Zhang, P. Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. 24, 102373, DOI: 10.1016/j.isci.2021.102373.
18. Chauhan, S. & Vig, L. Anomaly detection in ECG time signals via deep long short-term memory networks. Pages: 7.
19. Mostayed, A., Luo, J., Shu, X. & Wee, W. Classification of 12-lead ECG signals with bi-directional LSTM network. 16.
20. He, R. et al. Automatic cardiac arrhythmia classification using combination of deep residual network and bidirectional LSTM. 7, 102119–102135, DOI: 10.1109/ACCESS.2019.2931500. Conference Name: IEEE Access.
21. Luo, C., Jiang, H., Li, Q. & Rao, N. Multi-label classification of abnormalities in 12-lead ECG using 1d CNN and LSTM. In Liao, H. et al. (eds.) Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting, Lecture Notes in Computer Science, 55–63, DOI: 10.1007/978-3-030-33327-0_7 (Springer International Publishing).
22. Yao, Q., Wang, R., Fan, X., Liu, J. & Li, Y. Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network. 53, 174–182, DOI: 10.1016/j.inffus.2019.06.024.
23. Liu, W. et al. Real-time multilead convolutional neural network for myocardial infarction detection. 22, 1434–1444, DOI: 10.1109/JBHI.2017.2771768. Conference Name: IEEE Journal of Biomedical and

Health Informatics.

24. Ye, X. & Lu, Q. Automatic classification of 12-lead ECG based on model fusion. In 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 733–738, DOI: 10.1109/ CISP-BMEI51763.2020.9263559.
25. Ge, R. et al. Convolutional squeeze-and-excitation network for ECG arrhythmia detection. 121, 102181, DOI: 10.1016/j. artmed.2021.102181.
26. Chen, T.-M., Huang, C.-H., Shih, E. S. C., Hu, Y.-F. & Hwang, M.-J. Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model. 23, 100886, DOI: 10.1016/j.isci.2020.100886.
27. Hannun, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. 25, 65–69, DOI: 10.1038/s41591-018-0268-3.
28. Ye, X., Huang, Y. & Lu, Q. Explainable prediction of cardiac arrhythmia using machine learning. In 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 1–5, DOI: 10.1109/CISP-BMEI53629.2021.9624213.
29. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. 10.
30. Liu, F. et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. 8, 1368–1373, DOI: 10.1166/jmihi.2018.2442.
31. Yan, X., Hu, S., Mao, Y., Ye, Y. & Yu, H. Deep multi-view learning methods: A review. 448, 106–129, DOI: 10.1016/j.neucom.2021.03.090.
32. Velayudhan, A. & Peter, S. Noise analysis and different denoising techniques of ECG signal - a survey. 5.
33. McCabe, J. M. et al. Physician accuracy in interpreting potential ST-segment elevation myocardial infarction electrocardiograms. 2, e000268, DOI: 10.1161/JAHA.113.000268. Publisher: American Heart Association.
34. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. 1206.5538.
35. Liu, J., Jiang, Y., Li, Z., Zhou, Z.-H. & Lu, H. Partially shared latent factor learning with multiview data. 26, 1233–1246, DOI: 10.1109/TNNLS.2014.2335234. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.
36. Ramachandram, D. & Taylor, G. W. Deep multimodal learning: A survey on recent advances and trends. 34, 96–108, DOI: 10.1109/MSP.2017.2738401. Conference Name: IEEE Signal Processing Magazine.
37. Bonab, H. & Can, F. Less is more: A comprehensive framework for the number of components of ensemble classifiers. 30, 2735–2745, DOI: 10.1109/TNNLS.2018.2886341. Conference Name: IEEE Transactions on Neural Networks andLearning Systems.
38. Um, T. T. et al. Data augmentation of wearable sensor data for parkinson’s disease monitoring using convolutional neural networks. 216–220, DOI: 10.1145/3136755.3136817.1706.00527.

39. Pan, Q., Li, X. & Fang, L. Data augmentation for deep learning-based ECG analysis. In Liu, C. & Li, J. (eds.) Feature Engineering and Computational Intelligence in ECG Monitoring, 91–111, DOI: 10.1007/978-981-15-3824-7_6 (Springer).
40. Howard, A. G. et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. 1704.04861.
41. Zhu, H. et al. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. 2, e348–e357, DOI: 10.1016/S2589-7500(20)30107-2.
42. Cho, K. et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 1406.1078.
43. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. 1409.0473.
44. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet. 101, e215–e220, DOI: 10.1161/01.CIR.101.23.e215. Publisher: American Heart Association.
45. Bousseljot, R., Kreiseler, D. & Schnabel, A. Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das internet. 40, 317–318, DOI: 10.1515/bmte.1995.40.s1.317. Publisher: De Gruyter Section: Biomedical Engineering / Biomedizinische Technik.
46. Tihonenko, V., Khaustov, A., Ivanov, S. & Rivin, A. St.-petersburg institute of cardiological technics 12-lead arrhythmia database, DOI: 10.13026/C2V88N. Type: dataset.
47. Wagner, P. et al. PTB-XL, a large publicly available electrocardiography dataset. 7, 154, DOI: 10.1038/s41597-020-0495-6. Number: 1 Publisher: Nature Publishing Group.

# Figures



## Figure 1

(a) Augmentation for ECG. The six operations of the ECG-specific augmentation are shown. (b) The components of single models including CNN, RNN, attention, and SE block. (c) The schematic workflow for training and inference stage

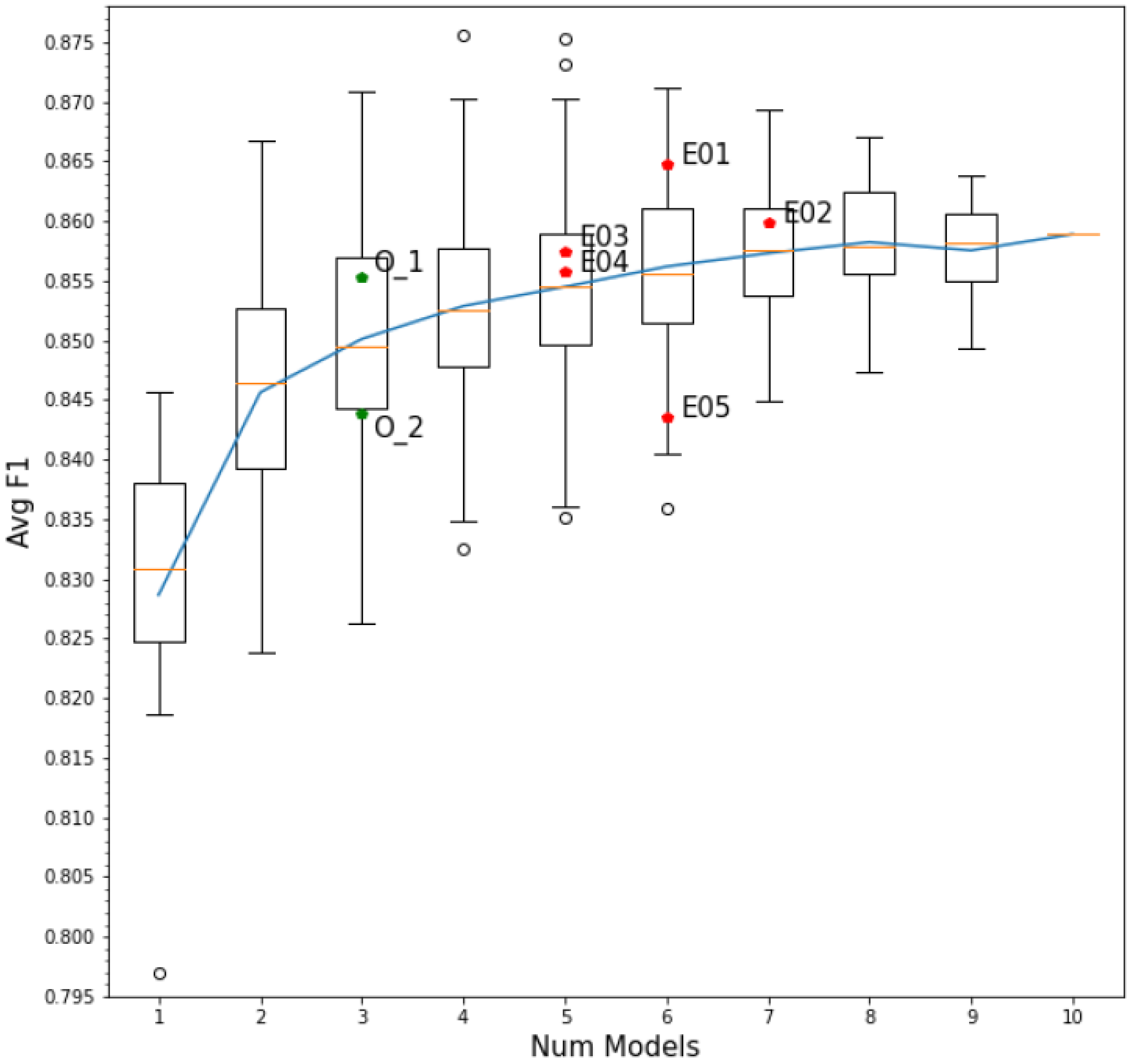


**Figure 2**

F1 score on our test set with respect to the number of models in ensembles. Ensemble performance for E01–E05 in Table 2 is marked as red dots. O1 and O2 indicate ensembles of #models=3, presumed mutually independent.