Towards Real-Time Heart Health Monitoring in Firefighting Using Convolutional Neural Networks

Jiajia Li[a,†], Christopher Brown[a,†], Dillon J. Dzikowicz[b], Mary G. Carey[b], Wai Cheong Tam[a,*], Michael Xuelin Huang[c]

[a]Fire Research Division, National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, MD 20899, USA, {jiajia.li, christopher.brown, waicheong.tam}@nist.gov

[b]School of Nursing, University of Rochester, 255 Crittenden Blvd, Rochester, NY 14642, USA, {dillon_dzikowicz, mary_carey}@urmc.rochester.edu

[c]Google Inc., Mountain View, CA, USA, mxhuang@google.com

[†]Joint first authors, [*]Corresponding author

Highlights:

- A deep-learning model was developed to determine ECG cardiac rhythms in real-time.
- 24-hour ECGs from 112 career on-duty firefighters were used for training.
- The model predicted normal, abnormal, and noisy ECG with an error of < 6 %.
- Using non-firefighters' ECG datasets led to substantial errors (~ 40 %).

Abstract:

A machine learning-based heart health monitoring model, named H2M, was developed. Twenty-four-hour electrocardiogram (ECG) data from 112 career firefighters were used to train the proposed model. The model used carefully designed multi-layer convolution neural networks with maximum pooling, dropout, and global maximum pooling to effectively learn the indicative ECG characteristics. H2M was benchmarked against three existing state-of-the-art machine learning models. Results showed the proposed model was robust and had an overall accuracy of approximately 94.3 %. A parametric study was conducted to demonstrate the effectiveness of key model components. An additional data study was also carried out, and it was shown that using non-firefighters' ECG data to train the H2M model led to a substantial error of ~ 40 %. The contribution of this work is to provide firefighters on-demand, real-time status of heart health status to enhance their situational awareness and safety. This can help reduce firefighters' injuries and deaths caused by sudden cardiac events.

Keywords: Abnormal heartbeat detection; machine learning; on-duty ECG signals; sudden cardiac death prevention; smart firefighting

1. Introduction

Sudden cardiac death (SCD) has been the leading killer for U.S. firefighters. Over the past 10 years, SCD consistently accounted for more than 40 % of on-duty fatalities [1]. In the year of 2021 alone, it resulted in 31 firefighter fatalities. In the same study [1], statistics showed that firefighters aged 50 years and over accounted for roughly two-thirds of the total number of SCD. Moreover, the incidence of SCD among firefighters was about twice that of police officers and four times higher than other emergency responders [2]. In terms of injuries, cardiac events led to about 13 % of the severe injuries during fireground operations between 2010 and 2014 [3]. From the studies carried out by the National Fire Protection Association (NFPA) [4-6], there was an annual average of about 831 instances due to cardiac related events for on-duty firefighters between 2015 and 2020. Based on these statistics, research is needed to prevent future firefighter deaths and injuries.

The National Institute for Occupational Safety and Health (NIOSH) conducts independent investigations of on-duty firefighter deaths through the Fire Fighter Fatality Investigation and Prevention Program. Currently, there are about 700 completed investigation reports [7]. These reports are useful because they provide a detailed timeline of the cardiac event. From the recent reports [8-12], there are two consistent observations before the fatal cardiac event occurs: 1) the firefighter feels physical discomfort and 2) their fellow firefighters notice unusual symptoms. Important notes from one investigation [12] are provided here. A 44-year-old female firefighter (FF) was dispatched as the driver of a rescue unit at 1022 hours. Although the light-duty work, her fellow firefighter noticed she was diaphoretic (Moment 1). When questioned, the FF indicated that she completed a physical test in the morning and she was just tired (Moment 2). The second dispatch took place at 1125 hours. While returning to the fire station, the FF complained about a burning sensation in her throat (Moment 3) but insisted that she was physically healthy and the unusual feeling was attributed to breathing cold air during the morning physical test. At around noon after arriving at the fire station, the FF indicated the symptoms were getting worse. The FF began to experience chest pain and complained that she could not breathe. Shortly after, the FF went into seizure-like activity and had a cardiac arrest. The FF's heart rhythm was shown to be ventricular fibrillation. At about 1215 hours, the FF was unresponsive and pulseless. Based on the details provided from [12], if the FF had understood her cardiac status at any one of the three moments, she could have sought immediate medical attention and this fatal event could have been avoided.

Three NFPA standards help firefighters prevent heart attacks and/or other cardiac related issues. Firstly, NFPA 1500 [13] addresses firefighter safety with general guidance on operations, health and wellness, equipment, fitness assessments, and rehabilitation. Secondly, NFPA 1582 [14] provides guidance for medical testing, minimum performance, and specific testing criteria. Finally, NFPA 1583 [15] provides guidance on fitness and wellness programs. However, there are two potential problems. The first problem is that compliance with the NFPA standards is voluntary [16] and the second problem is that all firefighter victims from the NIOSH reports [8-12] had received medical clearance for their duties and there were no major concerns noted in their medical evaluations. This is a major concern because the medical evaluations aimed at protecting firefighters fail to accurately acquire the true physiological demands of firefighting; as such, firefighters are incorrectly classified as fit yet suffer SCD. Additional efforts are needed to understand the relationship between emergency duties and SCD among firefighters specifically in the real world environment.

Contributions from the fire and medical research communities provide a better understanding of cardiovascular risk factors. These studies investigated the effect of firefighter's age [17,18], sex [19], fitness [20,21], career path [22,23], and roles [24]. Research findings indicated that there was a prevalence of overweight and obesity within a cohort of male career firefighters. This was alarming because obesity was found to be highly correlated with increased cardiovascular risk. In addition, a great deal of efforts has been made to understand firefighter's physiological responses in various emergency duties and firefighting environments. For example, early studies examined the effect from various simulated firefighting activities such as a response to a fire alarm [25], training [26], fire suppression [27], high-rise building operation [28], and recovery [29]. More recently, several research groups, such as those in references [30-32], expanded the studies to accommodate real emergency and fire responses. It was found that strains due to strenuous work, dangerous environments, and heavy protective equipment, which include attack and suppression, search and rescue, climbing stairs, extreme temperatures, toxic gases, low visibility, increased metabolic work, decreased heat dissipation, and restrictive body movement, contributed as cardiac stressors that may trigger sudden cardiac events. In [32], the study showed that firefighters, who did not have any underlying cardiac diseases and had completed NPFA 1582, do experience at least one non-sustained cardiac arrhythmia (supraventricular and/or ventricular) in the 24-hour shift. However, none of these cardiac reports were available to the firefighters in real-time and none of the firefighters noticed any of these events during their 24-hour shifts. Indeed, the traditional approaches in the fire and/or medical research communities are limited to offline analysis of physiological signals. Therefore, a robust approach is required to transfer fundamental knowledge into practical applications and to provide on-demand, real-time heart health status to the firefighters.

Deep learning algorithms have achieved great success in electrocardiogram (ECG) classification tasks. The current state of the art models can provide cardiologist-level detection of ECG waveforms [33], heart beats [34], artifacts [35] and classification of abnormal heart rhythms [36]. The performance of these models is promising, and the model accuracy for heart rate [34] and abnormal cardiac ECG rhythm [36] detection can be nearly 99 % and at least > 80 %, respectively. However, there are three major problems. Firstly, ECG data obtained from hospital patients were used for model development [33-36]. Secondly, the existing models generally rely on multi-lead, lengthy ECG sequences for predictions. Finally, none of these models has been validated against any ECG recordings obtained from on-duty firefighters where these models may not be reliable because the models have not learned sufficient ECG characteristics (i.e., more noise and higher heart rate) from career firefighters and their unique activities. In this paper, the development of a lightweight, domain specific, deep learning-based heart rhythm classification model is presented. The proposed model only requires the use of single-lead, six-second, ECG segments and is trained using the ECG recordings obtained from career on-duty firefighters. It is expected that the proposed model can provide firefighters on-demand, real-time, heart health status to enhance their situational awareness and safety and to help reduce firefighters' deaths and injuries due to sudden cardiac events.

This paper is organized as follows. Section 2 describes the on-duty firefighters' ECG data covering baseline information about the firefighters, data collection and annotation procedure, data behaviors and potential challenges, and data processing. Section 3 presents the development of the heart health monitoring (H2M) model. Then, Section 4 provides the model performance of the H2M model, benchmark results against the current state-of-the-art models, and model comparison with hospitalized ECG datasets. Finally, Section 5 presents the conclusions of the study.

2. On-Duty Firefighters' ECG Data

Data is one of the most important elements for the development of a reliable machine learning model. In contrast to [33-36], this study utilizes realistic firefighters' ECG data collected from on-duty firefighters [31]. This dataset is unique because it accounts for a diverse population of career firefighters and includes various dynamic on-duty activities. Thus, the proposed model is expected to be used in emergency response and firefighting contexts.

2.1 Firefighters Demographic and Anthropometric Characteristics

ECG data from one-hundred and twelve (112) career firefighters mainly from metro fire stations in the Western New York area were used. Of the 107 male firefighters and 5 female firefighters, 91 were White, 15 were Black, and the remaining were considered as Others. The average age of the firefighters was (43.6 ± 7.7) years old and about 47 % were ≥ 45 years old. It should be noted that the age significance was attributed to the fact that more than 75 % of on-duty fatalities in the US were older than 45 years old [37]. The mean length of fire service experience was about 15.5 years with a standard deviation of about 7 years.

Anthropometric data were measured before the study started. Based on the body mass index (BMI), almost half of the firefighters (~ 49 %) were overweight, and more than 40 % were obese with the BMI ≥ 30 kg/m^2. In the group of obese firefighters, about 55 % had a waist circumference larger than 100 cm. For blood pressure, the systolic and diastolic readings were (129.3 ± 14.9) mmHg and (81.8 ± 10.6) mmHg, respectively. Hypertension was observed in 35 firefighters. Past medical history from the firefighters was also collected. It showed that about 13 % were active smokers, 3 % had a history of coronary artery disease, and 9 % had respiratory disease (i.e., asthma, chronic obstructive pulmonary disease, or sleep apnea). This baseline information provided important characteristics about the firefighter data which was crucial to understanding the model capabilities.

2.2 Data Collection and Annotations [31]

Portable ambulatory recorders (H12+ Holter V3.12[1]) were used to obtain the 12-lead ECG data from the firefighters. In order to optimize signal quality, the contact areas were prepped. For example, skin hair was removed and the skin was cleaned with alcohol wipes. Electrodes were applied utilizing the Mason-Likar lead configuration [38] under the firefighters' uniformed t-shirts and the Holter was secured to the uniform belt. Fig. 1a shows the corresponding placement locations of the 10 electrodes, and Fig. 1b presents an overview of normal 12-lead ECGs in a resting state. Each ECG had different temporal characteristics because each ECG lead corresponded to electrical activity of the heart muscle at different locations.

Twenty-four-hour Holter ECG recordings were collected from all 112 firefighters. The 24-hour recordings consisted of data from 16-hour on-duty shifts and the following 8-hour post-duty shifts. Various activities were engaged by the firefighters during the 16-hour shifts and grouped into six different categories: i) fire calls, ii) medical calls and non-emergency categories, iii) physical activities (i.e., trainings, exercises, etc.), iv) sitting/talking (i.e., shift reports, administration, instruction, etc.), v) meals, and vi) rest/sleep. Post-duty activities were also grouped into the same non-emergency categories.

[1] Disclaimer: any mention of commercial products by NIST authors is for information only; it does not imply recommendation or endorsement by NIST

The ECG recordings were downloaded for annotations. First, each beat of the ECG recordings was annotated by a computer software. There were seven different classes: 1) normal beat, 2) supraventricular premature beat (SVPB), 3) ventricular premature beat (VPB), 4) paced rhythm, 5) atrial fibrillation (AF), 6) R on T, and 7) artifact due to movement. These classes were selected based on expert knowledge and previous studies from [31,32] that suggested the irregular heart rhythms from Class 2 to Class 6 were most indicative to potentially trigger SCD for on-duty firefighters. Then, all ECGs and the corresponding annotations were reviewed by an expert investigator with over 15 years of experience in electrocardiography. In general, the ECG dataset from the 112 firefighters during a 24-hour shift had a total number of 9 588 015 beats. Table 1 provides detailed beat counts for each class.

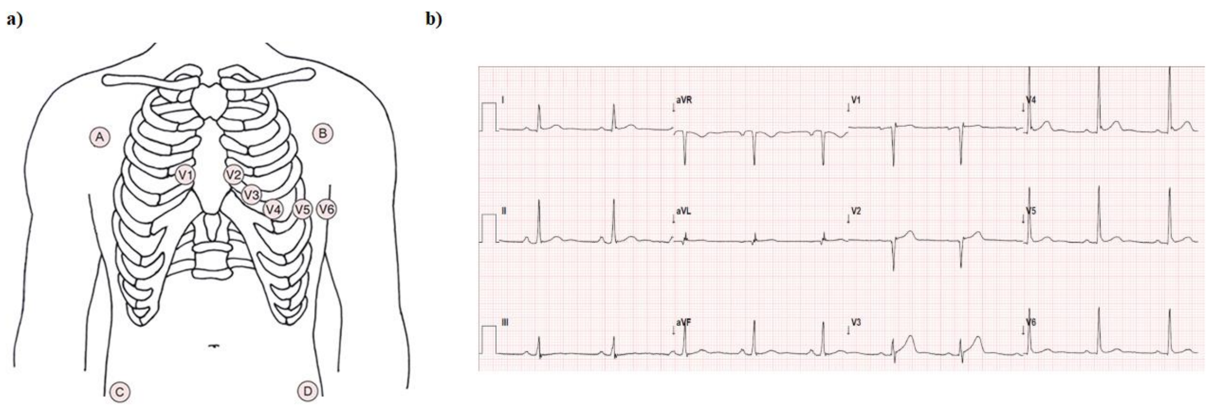

Fig. 1. a) A diagram of the 10 electrode placements [38] and b) an example of normal 12-lead ECG signals [31].

Table 1. Total beat counts for 7 different classes.

	Class 1 Normal	Class 2 SVPB	Class 3 VPB	Class 4 Paced	Class 5 AF	Class 6 R on T	Class 7 Artifact
Counts	9 393 057	21 746	45 437	1 128	9 502	192	116 953

2.3 ECG Characteristics and Potential Challenges

Understanding the characteristics from normal and abnormal ECG rhythms was vital to the design of a robust model. Fig. 2a shows an overview of a complete cardiac cycle. It consists of a P-wave, a QRS complex (Q-wave, R-wave, and S-wave), and a T-wave. In principle, the P-wave, QRS complex, and T-wave correspond to the atria contraction, ventricular depolarization, and ventricular relaxation, respectively. To determine the rhythm normality, cardiologists compare consecutive cardiac cycles and examine the length, relative difference in magnitude, and the shape of each wave. Fig. 2b depicts a 6-second normal sinus rhythm (NSR) obtained at lead position V6 (see Fig. 1) from Firefighter-2 (FF-2), and there are six complete cardiac cycles. As shown in the figure, the overall shape of the ECG rhythms from each cycle is consistent and the relative difference of the length and magnitude of each wave is negligible. Fig. 2b also shows that the heart rate of FF-2 (obtained from measuring the R-to-R intervals) is increasing over time because the firefighter was moving while performing on-duty tasks. It is important to note that this kind of normal ECG characteristics (i.e., monotonic increasing or decreasing R-to-R intervals over time) were not available from the ECG datasets being used in [33-36] because those ECG datasets were taken from hospital patients who were lying on beds.

Three abnormal ECG recordings obtained at lead position V6 are shown in Fig. 3. These rhythms are selected to demonstrate various information associated with abnormal ECGs. Fig. 3a shows the 6-second ECG recording with a SVPB (see the red arrow in the figure). As compared to the preceding cardiac cycles, the 4th cardiac cycle begins about 0.5 s earlier, and there is a significant discrepancy in the TP segment between the 3rd and the 4th cycle (the duration is less than 0.2 s). Fig. 3b shows the ECG recording with a VPB. Comparing each of the cycles, the start and the duration of different waveforms are relatively consistent. However, during the 4th cycle, there is an elevated R-wave and a missing positive S-wave. The expected S-wave is replaced with a large, inverted wave pattern. Fig. 3c shows the ECG recording with AF rhythms. Unlike SVPB and VPB, there is a significant change in the R-to-R interval. The deviation is rather random and the R-to-R interval varies from ~ 1 s to ~ 1.5 s. Given the observed ECG data characteristics, the model needs to capture indicative features at different magnitudes and time scales. In Section 3, a sensitivity study on model structure is presented to understand the effect of different modeling components.

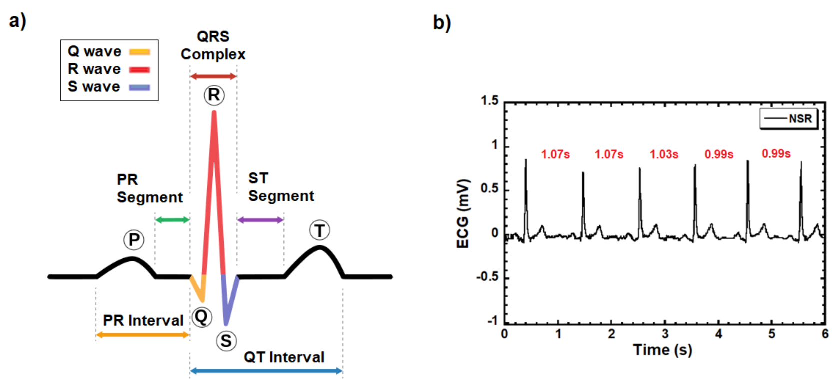

Fig. 2. a) An overview of a complete cardiac cycle and b) 6-second normal sinus rhythm (NSR) obtained at lead position V6.

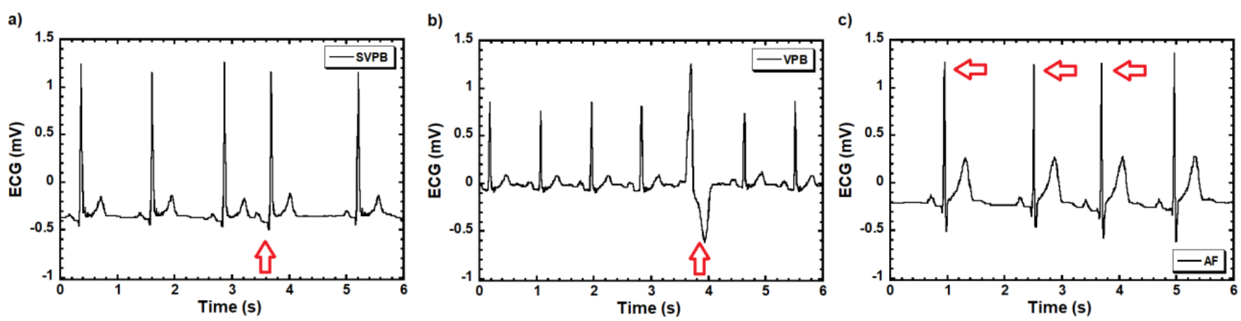

Fig. 3. Abnormal ECG due to a) SVPB, b) VPB, and c) AF at lead position V6 from FF-2, FF-3, and FF-93, respectively.

2.4 Data Preprocessing

Four additional steps were taken to prepare the final dataset. 1) The ECG dataset was re-organized from seven classes into three major classes: a) normal, b) abnormal, and c) noisy ECGs. The number of classes was reduced to provide simple actionable information to enhance firefighters' awareness of their heart health. The normal (class 1) and noisy (class 7) ECG data remained the same. The abnormal data now consisted of ECGs with SVPB, VPB, paced rhythm, AF, and R on T (classes 2 through 6). With that, there were 9 393 057 samples, 30 864 samples, and 116 953 samples for normal, abnormal, and noisy beats, respectively. The dataset is obviously imbalanced at this stage, so further processing is needed. 2) Therefore, data balancing was conducted to help avoid prediction bias and the modified dataset only contained 30 864 selected samples for each of the classes. During the selecting process, the normal, abnormal, and noisy samples were forced to select from the same firefighter. By doing so, the dataset was optimized to make use of all available abnormal ECG data, to maximize data diversity, and to capture well-balanced data characteristics from each firefighter. In total, the modified dataset contained 92 592 samples (30 864 + 30 864 + 30 864 for normal, abnormal, and noisy beats, respectively). 3) The modified dataset was then split into different subsets using a fixed ratio. Approximately 60 %, 20 %, and 20 % of data were assigned to the training, validation, and testing subsets, respectively. 4) Data normalization was carried out and the z-score normalization method [39] was used to maintain the data from each subset in a specific range. The normalization helped to improve the training stability and to expedite the learning process. The final training, validation, and testing subsets were used to facilitate the machine learning (ML) model development.

3. Development of the Heart Health Monitoring Model

The Heart Health Monitoring (H2M) model was developed using a convolutional neural network [40] (CNN) which is a class of deep learning algorithms. There were three reasons why CNN was selected: a) CNN has unique operations, such as convolution and pooling, that automatically and adaptively learn temporal hierarchies (i.e., from local to global and from low level to high level) of features. These operations were important to help the model to accurately capture the abnormal ECG characteristics mentioned in Sec. 2.3; b) the size of the ECG dataset being used in this study was sufficiently large so the model had adequate data to distinguish indicative features and ignore irrelevant information, such as high frequency noise, for the classification task; c) CNN can be finetuned to have robust model architecture to facilitate training (i.e., less computational time) and to be relatively lightweight (i.e., less memory). These benefits are favorable for practical engineering applications, including this present study.

3.1 Model Structure

Fig. 4 shows the overall model structure of the H2M model. The network took an array of ECG sequences with a dimension of $(X_1, X_2, 1)$ as inputs. X_1 and X_2 were the number of training samples and the sequence length of each sample and they were taken to be 55 555 (60% of 92 592) and 1800 (12 s ECG signals with a sampling frequency of 150 Hz), respectively. In terms of prediction, the model provided an output every 1 second.

As shown in Fig. 4, the model consisted of 8 layers of convolution blocks. For each convolution layer, 1-D convolution were applied. There were three hyperparameters, namely kernel size, stride, and number of filters, to modify the 1-D convolution (conv). For each conv, the kernel/filter size was 3 and the stride was 1. In principle, this convolution configuration allowed the model to extract temporal features from three neighboring input representations. The 1st conv was set to have 8 different kernels/filters and the number of filters was increased by a factor of 2 in every two convolution blocks. Each convolution operation was then followed by a batch normalization (BN) and an activation function using ReLU (Rectified Linear Unit). The BN normalized output features from conv to improve training stability [40] and the use of ReLU provided nonlinearity to activate useful features [40].

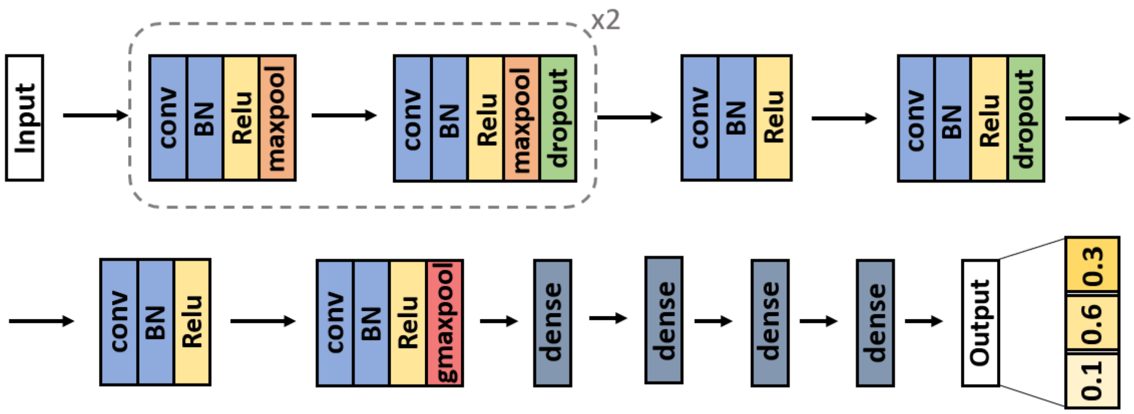

Fig. 4. Overview of the H2M model structure.

To allow the model to learn indicative features from a larger time scale, maximum pooling (maxpool) was used. There were 4 maxpool operations, and they were added after the ReLU activation function in the first 4 convolution blocks. Using a pool size of 2, the model selected the feature with the highest activation values from every 2 temporal features. In addition, dropout was also utilized, and they were added to the 2nd and the 4th convolution blocks. The dropout rate was taken to be 0.1. Physically, this dropout operation forces the model to randomly retain 90 % of the features. The use of maxpool and dropout helped the model to extract better features and avoid overfitting [40]. The learned features from the final conv were passed into a global maximum pooling (gmaxpool) operation in which the gmaxpool took the strongest activation to separate the different classes of ECGs. The selection of the exact number, locations, and the size/rate for both maxpool and dropout, was based on the observation from the data characteristics made in Sec. 2.3, and the model was optimized based on numerical experiments. Table 2 provides a summary of the important layer parameters.

Table 2. A summary table of the H2M layer parameters.

Layer	Type	Output Size	Kernel/ Pooling Size	Stride	Layer	Type	Output Size	Kernel/ Pooling Size	Stride
1	conv	(, 1798, 8)	3	1	**6**	conv	(, 106, 24)	3	1
	maxpool	(, 899, 8)	2	–	**7**	conv	(, 104, 48)	3	1
2	conv	(, 897, 8)	3	1	**8**	conv	(, 102, 48)	3	1
	maxpool	(, 448, 8)	2	–		gmaxpool	(, 48)	–	–
3	conv	(, 446, 16)	3	1	**9**	dense	(, 128)	–	–
	maxpool	(, 223, 16)	2	–			(, 64)	–	–
4	conv	(, 221, 16)	3	1			(, 32)	–	–
	maxpool	(, 110, 16)	2	–			(, 8)	–	–

5	conv	(, 108, 24)	3	1	**10**	softmax	(, 3)	–	–

Besides the convolution blocks, the H2M model also had 4 fully connected layers (denoted as dense). Differing from convolution blocks in which they were used to extract features, the dense layers were utilized to combine the high-level features to make classifications. The dense layers had a nonlinear activation function (ReLU) with decreasing numbers of neurons, which reinforced dimension reduction. Finally, there was an output layer with a dimension of 3 for three different prediction classes: normal, abnormal, and noisy ECGs. Softmax was used as the activation function because the outputs were expected to range from 0 to 1. Given the ECG sequences, the H2M model was optimized by solving the cross-entropy objective or the loss function ($\mathcal{L}$):

$$\mathcal{L}(X, r) = \frac{1}{n}\sum_{i=1}^{n} \log p(R = r_i \mid X) \tag{1}$$

where X was the ECG sequences, r was the corresponding labels of the ECG signals, $p(\cdot)$ was the probability the model assigned to the i-th output taking on the value r_i, and n was 3.

3.2 Training and Testing

The proposed CNN-based H2M model was trained on a PC workstation with a Nvidia Quadro RTX 5000 and an Intel Xeon 3.70GHz (W-2145). Tensorflow-GPU 2.0 with CUDA 10.0 and cuDNN 7.4.1 was used as a backbone to enable parallel computing. Adam optimizer [40] with an initial learning rate of 5e-4 was used to update trainable parameters during the training model. The H2M model size was lightweight with only 31 298 parameters. The model convergence was monitored using the validation subset. Fig. 5 shows the validation loss and accuracy for the optimized H2M model. When the validation loss did not improve for 10 consecutive epochs, the learning rate was decreased by a factor of 2 to stabilize the training. Early-stopping with a patience number of 25 was used to avoid overfitting. The training stopped at epoch 171, leaving the best model saved at epoch 146. The total training time was about 1371.3 s. The best model was applied to a testing subset to evaluate its model performance.

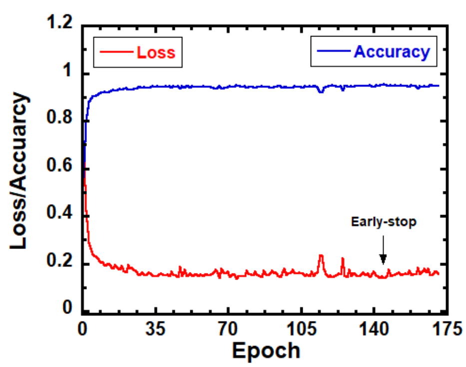

Fig. 5. Validation loss and accuracy for the H2M model.

4. Results and Discussion

Table 3 shows the model performance for predicting the normal, abnormal, and noisy ECGs from the testing subset. There were a total number of 18 519 samples for the testing set and these samples are evenly distributed over the three different ECG classes. The three ECG classes, namely normal, abnormal, and noisy ECGs, were denoted as C1, C2, and C3, respectively. The proposed model, H2M, was benchmarked against three state-of-the-art ECG rhythm classification models. The baseline models include i) MLP – a feedforward multiple-layer perceptron [41], ii) LSTM – a three-layer long short-term memory [42], and iii) ResNet – a 12-layer residual neural network [43]. Each model was fine-tuned to obtain optimal model performance without overfitting. The following metrics: accuracy, precision, and recall, were used to evaluate the model performance. The mathematical expressions were given as:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2a}$$

$$Precision = \frac{TP}{TP + FP} \tag{2b}$$

$$Recall = \frac{TP}{TP + FN} \tag{2c}$$

where TP, TN, FP, and FN were true positive, true negative, false positive, and false negative, respectively. Since the classification task involved three different classes, it yielded a 3-by-3 confusion matrix. The determination of TP, TN, FP, and FN from the 3-by-3 confusion matrix is trivial, and readers who are not familiar with this calculation method can refer to [44] for the details.

Table 3. Model performance of the H2M model against three different ML algorithms.

Method		**Predictions**			**Acc.**	**Prec.**	**Recall**	**Testing Time**	**Param.**
		C1	C2	C3					
MLP	C1	5133	311	729	89.2 %	84.4 %	83.2 %	2.4 s	38 974
	C2	444	4278	1451	81.5 %	73.7 %	69.3 %		
	C3	507	1216	4450	78.9 %	67.1 %	72.1 %		
LSTM	C1	5131	534	508	68.9 %	52.1 %	83.1 %	199.6 s	41 387
	C2	2803	1625	1745	67.2 %	51.6 %	26.3 %		
	C3	1907	990	3276	72.2 %	59.3 %	53.1 %		
ResNet	C1	5498	149	526	94.7 %	94.9 %	89.1 %	10.8 s	944 659
	C2	130	5401	642	93.7 %	93.1 %	87.5 %		
	C3	168	252	5753	91.8 %	83.1 %	93.2 %		
H2M	C1	5909	84	180	96.4 %	93.7 %	95.7 %	6.2 s	31 298
	C2	94	5914	165	97.1 %	95.4 %	95.8 %		
	C3	306	203	5664	95.4 %	94.3 %	91.8 %		

As shown in Table 3, H2M outperformed the existing ML-based prediction models and achieved a better overall accuracy of about 94.3 %. MLP and LSTM have an overall accuracy of ~ 74.9 % and ~ 52.0 %, respectively. ResNet had a similar model performance (~ 89.9 %) as compared to H2M. In terms of total testing time, H2M needed about 6.2 s to provide predictions for 18 519 samples. This yields only 3.3 x 10^{-4} s for a single prediction. For that, the proposed model is numerically suitable for real-time applications. Also, the precision and recall scores suggest that H2M was a more well-balanced model minimizing the false positives and the false negatives. The recall score is a more important evaluation metric for the current application because a high number of abnormal misclassifications (i.e., low recall score) might put firefighters into dangerous situations. In general, the main reason why H2M tended to perform better was that the model was designed carefully to capture the important ECG characteristics at different timescales. In the later section, results from a parametric study are provided to highlight the effect of each modeling component for H2M.

Fig. 6 shows examples of three correct prediction cases selected from the testing subset: a) normal, b) abnormal, and c) noisy ECGs. Two observations are worth noting. Firstly, H2M was capable of differentiating noise due to powerline interference and minor muscular activities (i.e., the blue-arrow region in Fig. 6a) and noise due to movement artifacts (i.e., the blue-arrow regions in Fig. 6c). Secondly, the model effectively recognized ECG abnormalities (see the red-arrow regions from Fig. 6b) and ignored motion induced ECG peaks shown in Fig. 6c (see the red-arrow regions). These example cases demonstrate that H2M does learn indicative patterns that can be used to separate different classes of ECGs. Another interesting note is that the output probabilities for these cases were high. The output probabilities (normal, abnormal, noise) for Case a, Case b, and Case c, are (0.99, 0.01, 0.00), (0.00, 1.00, 0.00), and (0.00, 0.01, 0.99), respectively. These results show that the model has more than a 99 % confidence level for its predictions.

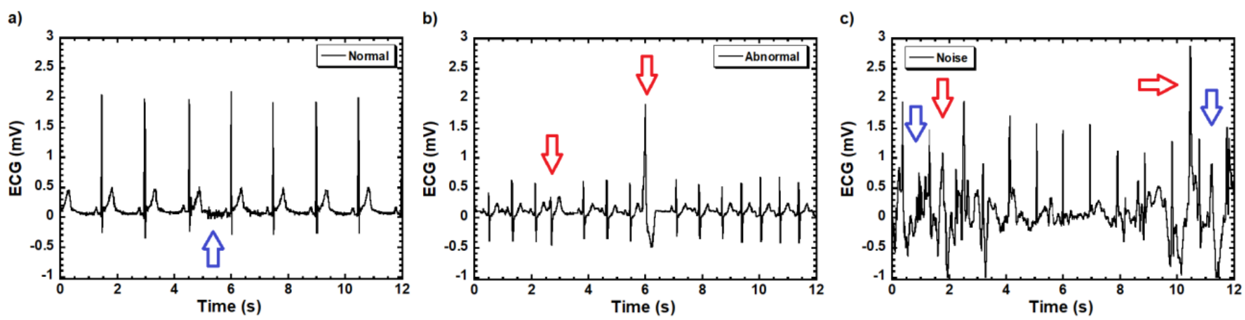

Fig. 6. Correct prediction of a) normal, b) abnormal, and c) noisy ECGs with the confidence level of approximately 0.99, 1.00, 0.99, respectively.

Fig. 7 presents two selected misclassification cases. The model prediction and the corresponding ground-truth are shown in the figures. There are two reasons why these example cases are being discussed. The first reason is that the classification task becomes challenging when the ECGs have various dynamic effects from emergency response/firefighting related activities. For example, there existed unusual peaks in the P-wave and other minor noise in both Fig. 7a and 7b. However, the ground-truths for these cases were completely different: one was an abnormal ECG and the other one was a noisy ECG. The second reason is that the model has relatively low confidence level in its predictions. For Case a, the output probability was (0.03, 0.46, 0.51), and the output probability for Case b was (0.13, 0.49, 0.38). These examples indicate that the model is likely to be more reliable if it can omit or disregard predictions that have relatively low confidence.

Fig. 8a shows the adjusted accuracy for seven sensitivity tests in which the classification threshold varies from 0.3 to 0.99. Given a 12-second ECG sample, if the model output probability was lower than the classification threshold, the ECG sample was omitted. For example, if the classification threshold was 0.4 and if the output probability was (0.33, 0.34, 0.33), the model prediction was disregarded. Fig. 8b shows the histogram for the number of omitted cases, misclassification cases, and correctly predicted cases for seven different sensitivity tests. As the classification threshold increased, the number of omitted cases increased and the number of misclassification cases decreased. When the classification threshold became 0.99, Fig. 8a shows a corresponding adjusted accuracy of ~ 99.7 %. A drawback was that approximately 7000 cases were disregarded. Yet, depending on the application requirements, the classification threshold can be modified.

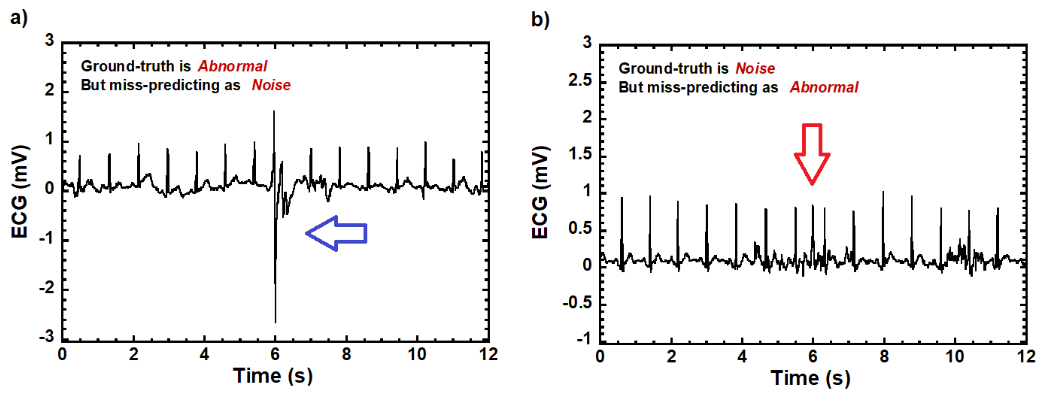

Fig. 7. Misclassification of a) abnormal and b) noisy ECGs with the confidence level of approximately 0.51 and 0.49, respectively.

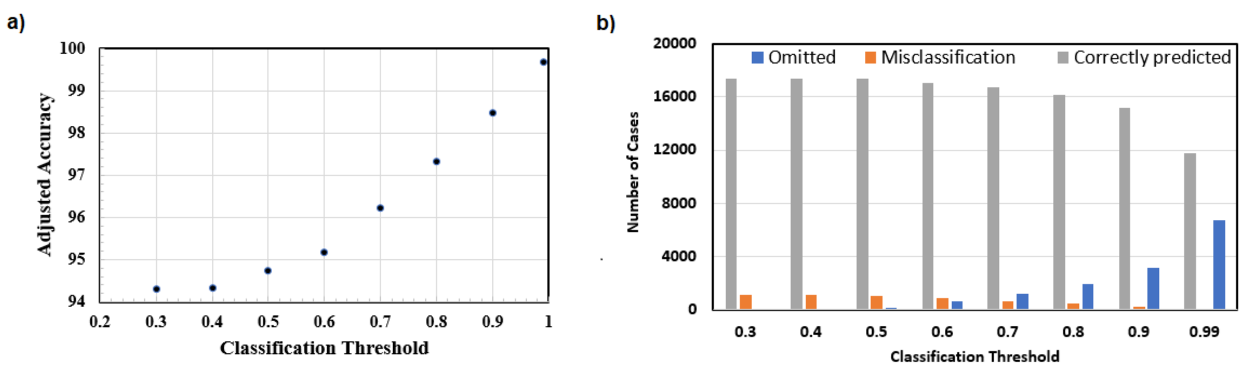

Fig. 8. a) Adjusted accuracy and b) histogram for seven sensitivity tests.

A parametric study was conducted to examine the effectiveness of key components that contributed to the improved outcomes for H2M. The full model of H2M was compared with four model variations: i) w/o gmaxpool – H2M without global maximum pooling and it was replaced by a flatten layer, ii) w/o dropout – taking out dropout and all convolution layers were fully connected, iii) w/o maxpool – all maximum pooling operations were removed, and iv) plain CNN – all global maximum pooling, dropout, and maximum pooling operations were removed.

Table 4 shows the accuracy, precision, and recall scores for each of the models. The inclusion of maximum pooling improved the model performance the most as it allowed the model to learn the ECG characteristics from larger timescales. The effect from using global dropout and maximum pooling and dropout was evident. As shown in Table 4, when all of these modeling components were removed, the overall accuracy of the model dropped to about 89.8 % and each of these components helped the model through the training process to learn useful data patterns for classifications.

Table 4. Parametric study for H2M.

	H2M	w/o gmaxpool	w/o dropout	w/o maxpool	Plain CNN
Accuracy	96.3 %	95.8 %	92.6 %	91.7 %	89.8 %
Precision	94.1 %	93.9 %	88.9 %	87.6 %	85.6 %
Recall	94.7 %	93.7 %	88.2 %	87.5 %	84.7 %

4.1 Effect of the ECG Dataset

In order to examine the contextual importance from firefighters' ECG data, a cross validation was carried out. Two public datasets from the 2021 Computing in Cardiology Challenge [45] were selected. The datasets were from the Chapman University and Ningbo First Hospital with about 10 247 and 34 905 ECG recordings, respectively. Both datasets were obtained from anonymous patients and contained normal and more than 100 different abnormal ECG rhythms. The ECG recordings were prepared accordingly. They were divided into 10 second segments at 150 Hz and each ECG recording had a sequence annotation. Unlike the firefighters' data, the ECGs from the public datasets did not contain any noisy data. For that, the cross validation can only be done with binary classifications with normal and abnormal classes. Also, the public dataset did not contain any ECG characteristics due to movements, emergency response, and/or firefighting related activities because they were solely obtained for medical diagnostic purposes.

Table 5. Cross-validation results from public and firefighter datasets.

Train on	**Ningbo**	**Chapman**	**Ningbo**	**Chapman**	**Ningbo**	**Chapman**	**Combine**
Test on	**Ningbo**	**Chapman**	**Chapman**	**Ningbo**	**Firefighter**	**Firefighter**	**Firefighter**
Accuracy	86.5 %	96.5 %	92.9 %	87.5 %	62.7 %	66.0 %	71.5 %
Precision	87.9 %	96.0 %	91.1 %	87.4 %	69.5 %	62.7 %	66.9 %
Recall	87.3 %	97.4 %	94.7 %	88.1 %	45.2 %	78.9 %	85.0 %

Table 5 shows results from the seven cross-validation tests. Each test was trained on dataset-A and was tested on dataset-B. The subset assignment was the same where 60 %, 20 %, and 20 % of the data were assigned to the training, validation, and testing subsets. The testing subset from a dataset was identical to have a consistent comparison. Three metrics, namely accuracy, precision, and recall, were used to assess the data effects. As shown in Table 5, when the public datasets were used for training and testing (i.e., train on Ningbo and test on Ningbo, or train on Chapman and test on Chapman), the model performance has an overall accuracy of > 86 %. The same model performance was also observed for two special cases in which the model was trained on Ningbo (or Chapman) and was tested on Chapman (or Ningbo). However, the model performance dropped significantly when the trained model used either one or both (denoted as 'Combine') of the public datasets then tested on the firefighter dataset. An error of more than 37 % with a recall score of only 45.2 % was observed. Even when both public datasets were used, the best model accuracy was only about 71.5 %. The results from these cross-validation tests suggest that the data characteristics were substantially different. Therefore, in order to develop a robust heart health monitoring model for emergency response and/or firefighting related activities, firefighters' ECG data is essential. The use of non-firefighters' data is likely to lead to substantial errors.

5. Conclusions

This paper presents the development of a deep learning-based heart health monitoring model that can provide firefighters real-time, on-demand, beat-by-beat classifications of normal, abnormal, and noisy ECG rhythms. The heart health monitoring (H2M) model utilized 24-hour ECG recordings from 112 career firefighters. This dataset had approximately 92 592 samples and was unique from public ECG datasets because it contained firefighters' beat-to-beat ECGs from various emergency response and/or firefighting related activities. H2M was designed carefully to learn indicative ECG characteristics. Model comparison against three current-state-of-the-art ECG prediction models showed that H2M offered convincing performance with an overall accuracy of about 94.3 % with a relatively lightweight model structure that required only 31,298 trainable parameters. Results from the parametric study demonstrated the effectiveness of each model component. Using the multi-layer CNN structures with maximum pooling, dropout, and global maximum pooling, H2M effectively captured ECG behaviors at different timescales. Examples for correctly predicted cases and misclassification cases were discussed. A sensitivity study on prediction thresholds showed an extremely high model reliability with an accuracy of about 99.7 % if low-level confidence predictions were omitted. Results from cross-validation tests were presented. The importance of firefighters' ECG data was demonstrated when non-firefighters' ECG data were used to train the heart health monitoring model for firefighters and resulted in a substantial error of about 40 %. Therefore, on-duty firefighters' data was crucial to develop a robust and reliable model. The outcome of this work is expected to enhance firefighters' situational awareness and safety about their heart health and to help reduce firefighters' deaths and injuries due to sudden cardiac events.

References

1. Fahy, R.F. and Petrillo, J.T., 2022. Firefighter Fatalities in the US in 2021. National Fire Protection Association. Quincy, Massachusetts.
2. Maguire, B.J., Hunting, K.L., Guidotti, T.L. and Smith, G.S., 2005. Occupational injuries among emergency medical services personnel. Prehospital Emergency Care, 9(4), 405-411.
3. Campbell, R., 2018. US firefighter injuries on the fireground, 2010–2014. Fire Technology, 54(2), 461-477.
4. Haynes, H. and Molis, J., 2016. U.S. Firefighter Injuries – 2015. National Fire Protection Association. Quincy, Massachusetts.
5. Campbell, R., Evarts, B., and Molis, J., 2019. United States Firefighter Injury Report 2018. National Fire Protection Association. Quincy, Massachusetts.
6. Campbell, R. and Evarts, B., 2021. United States Firefighter Injuries in 2020. National Fire Protection Association. Quincy, Massachusetts.
7. NIOSH. https://wwwn.cdc.gov/NIOSH-fire-fighter-face (accessed 21 January 2023).
8. F2021-04. https://www.cdc.gov/niosh/fire/pdfs/face202104.pdf (accessed 21 January 2023).
9. F2019-15. https://www.cdc.gov/niosh/fire/pdfs/face201915.pdf (accessed 21 January 2023).
10. F2019-08. https://www.cdc.gov/niosh/fire/pdfs/face201908.pdf (accessed 21 January 2023).

11. F2018-14. https://www.cdc.gov/niosh/fire/pdfs/face201814.pdf (accessed 21 January 2023).
12. F2018-05. https://www.cdc.gov/niosh/fire/pdfs/face201805.pdf (accessed 21 January 2023).
13. NPFA 1500, 2021. Standard on Fire Department Occupational Safety, Health, and Wellness Program. National Fire Protection Association. Quincy, Massachusetts.
14. NPFA 1582, 2022. Standard on Comprehensive Occupational Medical Program for Fire Departments. National Fire Protection Association. Quincy, Massachusetts.
15. NPFA 1583, 2022. Standard on Health-Related Fitness Programs for Fire Department Members. National Fire Protection Association. Quincy, Massachusetts.
16. NIOSH, 2007. Preventing fire fighter fatalities due to heart attacks and other sudden cardiovascular events. Department of Health and Human Services. Cincinnati, OH, p. 32.
17. Yang, J., Teehan, D., Farioli, A., Baur, D. M., Smith, D., & Kales, S. N. (2013). Sudden cardiac death among firefighters ≤ 45 years of age in the United States. The American Journal of Cardiology, 112(12), 1962-1967.
18. Farioli, A., Christophi, C. A., Quarta, C. C., & Kales, S. N. (2015). Incidence of sudden cardiac death in a young active population. Journal of the American Heart Association, 4(6), e001818.
19. Li, K., Lipsey, T., Leach, H. J., & Nelson, T. L. (2017). Cardiac health and fitness of Colorado male/female firefighters. Occupational Medicine, 67(4), 268-273.
20. Staley, J. A., Weiner, B., & Linnan, L. (2011). Firefighter fitness, coronary heart disease, and sudden cardiac death risk. American Journal of Health Behavior, 35(5), 603-617.
21. Tsismenakis, A. J., Christophi, C. A., Burress, J. W., Kinney, A. M., Kim, M., & Kales, S. N. (2009). The obesity epidemic and future emergency responders. Obesity, 17(8), 1648-1650.
22. Sen, S., Palmieri, T., & Greenhalgh, D. (2016). Cardiac fatalities in firefighters: An analysis of the US fire administration database. Journal of Burn Care & Research, 37(3), 191-195.
23. Dzikowicz, D. J., & Carey, M. G. (2021). Severity of Myocardial Ischemia Is Related to Career Length Rather Than Age Among Professional Firefighters. Workplace Health & Safety, 69(4), 168-173.
24. Eglin, C. M., & Tipton, M. J. (2005). Can firefighter instructors perform a simulated rescue after a live fire training exercise? European Journal of Applied Physiology, 95(4), 327-334.
25. Kuorinka, I., & Korhonen, O. (1981). Firefighters' reaction to alarm, an ECG and heart rate study. Journal of Occupational Medicine, 23(11), 762-766.
26. Lannon, C. M., & Milke, J. A. (2014). Evaluation of Fire Service Training Fires. Fire Protection Research Foundation.
27. Al-Zaiti, S., Rittenberger, J. C., Reis, S. E., & Hostler, D. (2015). Electrocardiographic responses during fire suppression and recovery among experienced firefighters. Journal of Occupational and Environmental Medicine, 57(9), 938-942.
28. Smith, D. L., Haller, J. M., Benedict, R., & Moore-Merrell, L. (2015). Cardiac strain associated with high-rise firefighting. Journal of Occupational and Environmental Hygiene, 12(4), 213-221.
29. Yang, Y. C., Dzikowicz, D., Al-Zaiti, S. S., & Carey, M. G. (2019). Heart Rate Recovery, Blood Pressure Recovery, and 24-hour Heart Rate among Firefighters. Journal of Electrocardiology, 57, S117.
30. Kerber, S. (2013). Analysis of one and two-story single family home fire dynamics and the impact of firefighter horizontal ventilation. Fire Technology, 49(4), 857-889.
31. Al-Zaiti, S. S., & Carey, M. G. (2015). The prevalence of clinical and electrocardiographic risk factors of cardiovascular death among on-duty professional firefighters. The Journal of Cardiovascular Nursing, 30(5), 440.

32. Smith, D. L., Horn, G. P., Fernhall, B., Kesler, R. M., Fent, K. W., Kerber, S., & Rowland, T. W. (2019). Electrocardiographic responses following live-fire firefighting drills. Journal of Occupational and Environmental Medicine, 61(12), 1030.
33. Peimankar, A., & Puthusserypady, S. (2021). DENS-ECG: A deep learning approach for ECG signal delineation. Expert Systems with Applications, 165, 113911.
34. Murat, F., Yildirim, O., Talo, M., Baloglu, U. B., Demir, Y., & Acharya, U. R. (2020). Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review. Computers in Biology and Medicine, 120, 103726.
35. Bashar, S. K., Ding, E., Walkey, A. J., McManus, D. D., & Chon, K. H. (2019). Noise detection in electrocardiogram signals for intensive care unit patients. IEEE Access, 7, 88357-88368.
36. Baloglu, U. B., Talo, M., Yildirim, O., San Tan, R., & Acharya, U. R. (2019). Classification of myocardial infarction with multi-lead ECG signals and deep CNN. Pattern Recognition Letters, 122, 23-30.
37. Mbanu, I., Wellenius, G. A., Mittleman, M. A., Peeples, L., Stallings, L. A., & Kales, S. N. (2007). Seasonality and coronary heart disease deaths in United States firefighters. Chronobiology International, 24(4), 715-726.
38. Khan, G. M. (2015). A new electrode placement method for obtaining 12-lead ECGs. Open Heart, 2(1), e000226.
39. Al Shalabi, L., & Shaaban, Z. (2006, May). Normalization as a preprocessing engine for data mining and the approach of preference matrix. In 2006 International Conference on Dependability of Computer Systems (pp. 207-214). IEEE.
40. Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Transactions on Neural Networks and Learning Systems.
41. Li, K., Pan, W., Li, Y., Jiang, Q., & Liu, G. (2018). A method to detect sleep apnea based on deep neural network and hidden Markov model using single-lead ECG signal. Neurocomputing, 294, 94-101.
42. Sun, L., Wang, Y., He, J., Li, H., Peng, D., & Wang, Y. (2020). A stacked LSTM for atrial fibrillation prediction based on multivariate ECGs. Health Information Science and Systems, 8(1), 1-7.
43. Han, C., & Shi, L. (2020). ML–ResNet: A novel network to detect and locate myocardial infarction using 12 leads ECG. Computer Methods and Programs in Biomedicine, 185, 105138.
44. Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.
45. Reyna, M. A., Sadr, N., Alday, E. A. P., Gu, A., Shah, A. J., Robichaux, C., & Clifford, G. D. (2021, September). Will two do? Varying dimensions in electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021. In 2021 Computing in Cardiology (CinC) (Vol. 48, pp. 1-4). IEEE.

Figure captions

Fig. 1. a) A diagram of the 10 electrode placements [38] and b) an example of normal 12-lead ECG signals [31].

Fig. 2. a) An overview of a complete cardiac cycle and b) 6-second normal sinus rhythm (NSR) obtained at lead position V6.

Fig. 3. Abnormal ECG due to a) SVPB, b) VPB, and c) AF at lead position V6 from FF-2, FF-3, and FF-93, respectively.

Fig. 4. Overview of the H2M model structure.

Fig. 5. Validation loss and accuracy for the H2M model.

Fig. 6. Correct prediction of a) normal, b) abnormal, and c) noisy ECGs with the confidence level of approximately 0.99, 1.00, 0.99, respectively.

Fig. 7. Miss-classification of a) abnormal and b) noisy ECGs with the confidence level of approximately 0.51 and 0.49, respectively.

Fig. 8. a) Adjusted accuracy and b) histogram for seven sensitivity tests.