HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 1 (2023)

The Role of Morphology and Heart Rate Variability Features in Detecting Arrhythmias from Short ECGs

Alice Hu[1] and Shadi Ghiasi[#]

[1]West Windsor-Plainsboro High School North, Plainsboro Township, NJ, USA
[#]Advisor

ABSTRACT

Detecting heart arrhythmias from short Electrocardiogram (ECG) recordings remains challenging since recordings are short and contaminated by noise. ECG morphology features and Heart Rate Variability (HRV) time and frequency domain features are widely used for classifying short ECG recordings. Here we investigate the relative roles of ECG morphology features and HRV time and frequency domain features in classifying short ECG recordings provided by the 2017 PhysioNet/Computing in Cardiology Challenge. The classification is performed separately by four machine learning models: Logistic Regression, Decision Tree, K Nearest Neighbors, and Convolutional Neural Network (CNN). Our best classification score is obtained using the deep learning 1-dimensional CNN model trained on HRV time domain features combined with ECG morphology features. It gives an overall F1 score of 0.70 and 0.73 for the cross validation and hidden test respectively when considering the average classification performance over all 4 categories: Atrial Fibrillation (AF), normal, other arrhythmias, and noisy signal. We found that HRV time domain features play an important role in detecting AF, normal, and other classes, whereas ECG morphology features play a key role in detecting the noisy class. When HRV frequency domain features are combined with HRV time domain features, they do not improve and often degrade classifications of short ECG recordings compared to classifications using only HRV time domain features. Combining ECG morphology features with HRV time domain features leads to a better classification performance for short ECG recordings. Feature-based deep learning could serve as a viable and less expensive approach for ECG classifications.

Introduction

Atrial Fibrillation (AF) is the most common form of heart arrythmias. It occurs when the normal sinus node is unable to control the heart rate because of activity in the upper chambers, or atria, of the heart (Nattel, 2002). Victims of AF experience complications like stroke, heart failure, and coronary artery disease (Clifford et al. 2017). Over the past few decades, AF induced and related deaths have become significantly more common. Electrocardiograms (ECGs) are often useful for AF detection, since AF is characterized by a lack of a P wave and inconsistent Heart Rate Variability (HRV) in ECGs (Da Silva-Filarder & Marzbanrad, 2017). And yet, it is not always easy to detect. Other arrythmias may have irregular heartbeats similar to those of AF on an ECG. Furthermore, AF may be episodic.

Past studies often focused solely on differentiating AF from normal ECGs. They were also limited by the datasets they used (Clifford et al. 2017). Conventionally, AF is detected using atrial activity analysis (by observing P waves and F waves) or ventricular response analysis (Yazdani et al. 2017). Other previous studies have extracted features from both the ECG spectral measures and the HRV (Billeci et al. 2017). The ECG morphology features and HRV time domain and frequency domain features are widely used for detecting heart

arrhythmias from short ECG recordings (Clifford et al. 2017; Coppola et al. 2017; Datta et al. 2017; Goodfellow et al. 2017; Zabihi et al. 2017). However, the relative role of each type of features in the classification performance of short ECG recordings is unclear. The short ECG recordings such as those provided by the 2017 PhysioNet/Computing in Cardiology Challenge, ranging from 9 seconds to 61 seconds, are not long enough to accurately reveal the HRV low frequency components (Shaffer & Ginsberg, 2017). Hence it is unknown whether the widely used HRV frequency domain features for classifying such short ECG recordings (Clifford et al. 2017; Coppola et al. 2017; Datta et al. 2017; Goodfellow et al. 2017; Zabihi et al. 2017) would be necessary. Meanwhile, ECG morphology features might be effective in distinguishing the Noisy class from other ECG classes (Goodfellow et al. 2017; Ghiasi et al. 2017).

To classify ECGs, researchers have used both feature-based conventional Machine Learning (ML) models (Clifford et al. 2017; Goodfellow et al. 2017; Smoleń, 2017) an data-driven Deep Learning (DL) models (Andreotti et al. 2017; Chandra et al. 2017; Hsieh et al. 2020; Warrick & Homsi, 2017; Xiong et al. 2017; Zihlmann et al. 2017; Van Zaen et al. 2019; Weimann & Conrad, 2021), with some studies combining both methods (Datta et al. 2017; Andreotti et al. 2017; Ghiasi et al. 2017). However, the complicated DL models with millions of model parameters are computational expensive to train (Hsieh et al. 2020; Zihlmann et al. 2017). Additionally, the classification score for the minority classes (e.g. Noisy class) may degrade over training time when using complicated DL models such as a combination of Convolutional Neural Networks (CNNs) and a stack of the Long Short-Term Memory layers (Warrick & Homsi, 2017). It is detrimental to make classifications with CNNs of above 7 convolutional layers (Van Zaen et al. 2019).

The purpose of this study is to investigate the role of different types and combination of features, i.e. ECG morphology features, HRV time domain features, and HRV frequency domain features, in classifying short ECG recordings in terms of four categories: AF, Normal rhythm, Other arrhythmias, and Noisy signal. Four different ML classification models (Logistic Regression, Decision Tree, K Nearest Neighbors (KNN), and CNN are applied to the features extracted from the short ECG recordings provided by the 2017 PhysioNet/Computing in Cardiology Challenge (Clifford et al. 2017). Here the CNN model is employed to test the performance of the extracted features from ECG and HRV integrated with the DL approach, since fewer previous studies were focused on the much simpler feature-based methods integrated with DL models, which would cost much less computing time.

Materials and Methods

The flowchart representing the pipeline for classifying the short ECG recordings used in this study is shown in Figure 1. Throughout this section each block in this flowchart will be discussed sequentially.

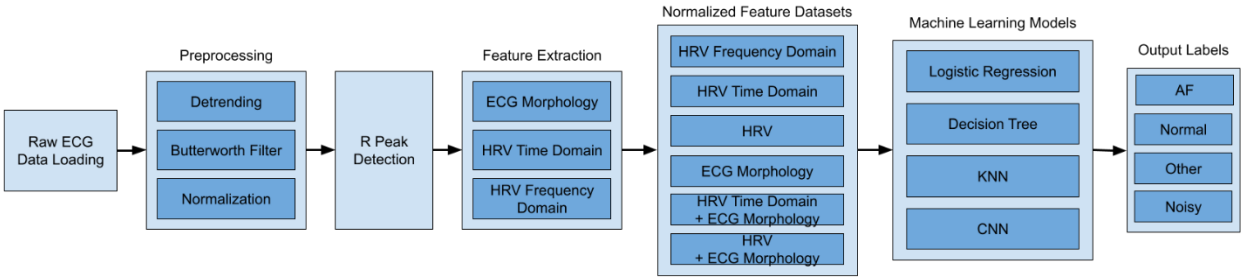

Figure 1. Flowchart of the classification of short ECG recordings into four categories.

Data and Preprocessing

The dataset used to conduct this study is downloaded from the 2017 PhysioNet/Computing in Cardiology Challenge (Clifford et al. 2017). It contains 8528 single lead ECG signals in four different classes: 758 AF, 5076 Normal, 2415 Other rhythm, and 279 Noisy recordings, ranging in length from 9 seconds to 61 seconds at a sampling frequency of 300 Hz. We use version 3 (V3), the latest version of the ECG class labels, which has substantial corrections compared to the original labels posted for the Physionet Challenge 2017 for the Noisy class (Clifford et al. 2017). For the data preprocessing, first the ECG data is detrended, then a butterworth filter (0.05 – 50 Hz) is applied to the detrended ECG data to remove the baseline wander (Datta et al. 2017). Finally, each ECG data is normalized with the mean removed and the signal's difference from the mean is divided by its standard deviation.

R Peak Detection and ECG/HRV Feature Extraction

The typical ECG heartbeat has a pronounced peak of the R wave (the so-called R peak). To extract features, R peak detection is performed on each preprocessed ECG data using the Hamilton-Tompkins algorithm (Hamilton & Tompkins, 1986) through the biosppy.signal.ecg.ecg function provided by the Biosignal Processing in Python (BioSPPy) library (Carreiras et al. 2015). This function also extracts the corresponding heartbeat template associated with each R peak in the preprocessed ECG data. The first 2 seconds of each preprocessed ECG are excluded to avoid the impact of noises on R peak detection. Figure 2 shows sample ECGs and detected R peaks from each class: AF, Normal, Other, and Noisy. The corresponding median and all individual templates from the above four sample ECGs are also shown in Figure 3.

From the R peaks, three types of features are extracted: ECG morphology features, HRV time domain features, and HRV frequency domain features. The ECG morphology features include the following 5 features: (1) Count of negative peaks, i.e. the number of templates within each ECG where the absolute value of the maximum of the template is smaller than the absolute value of the minimum of that template. (2) Normalized P wave amplitude, i.e. the maximum value between -0.2s and -0.1s of the median template (P wave peak) normalized by the maximum of the median template of each ECG (Figure 3a). (3) The median of all correlations between consecutive templates of each ECG. (4) Median R peak amplitude of each ECG. (5) The ratio of the maximum absolute value of each ECG over the corresponding median R peak amplitude. The boxplots of example ECG morphology features are shown in Figure 4. The above ECG morphology features are mainly used to distinguish the Noisy class from AF, Normal, and Other classes (Figure 4), because the Noisy class contains more inverted peaks, lower correlations between templates (Ghiasi et al. 2017; Goodfellow et al. 2017), and smaller R peak amplitudes. The normalized P wave amplitude feature is mainly used to detect the AF signal, since AF is characterized by the absence of a P wave.

We calculate the median R peak - R peak (RR) intervals from the detected R peaks of each ECG as the first HRV time domain feature. Additionally, 5 HRV time domain (i.e. SDNN, RMSSD, SDSD, pNN50, and the triangular index) and 12 frequency domain (i.e. very low frequency (VLF) peak, low frequency (LF) peak, high frequency (HF) peak, LF/HF power ratio, VLF relative power, LF relative power, HF relative power, VLF absolute power, LF absolute power, HF absolute power, LF normalized unit, and HF normalized unit) features (Shaffer & Ginsberg, 2017) are extracted from the calculated RR intervals using pyHRV, an open source python toolbox for HRV (Gomes et al. 2019). The above feature terms are defined in Table 1. The boxplots of example HRV time domain and frequency domain features are shown in Figure 5 and Figure 6 respectively.

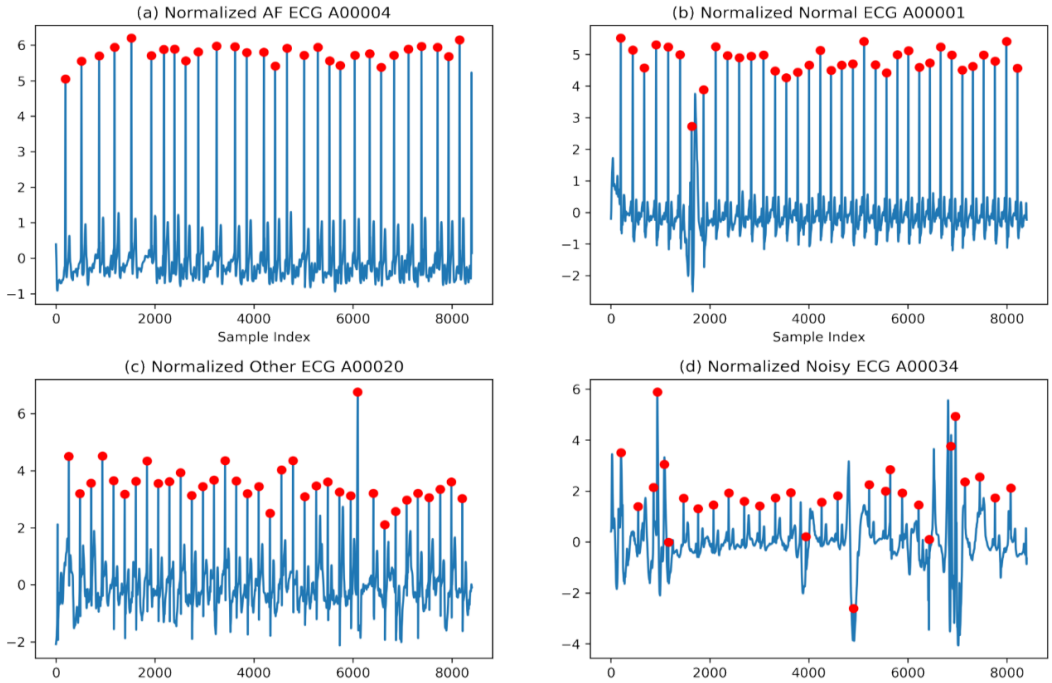

Figure 2. Sample ECGs and detected R peaks from each class. (a) AF, (b) Normal, (c) Other, (d) Noisy.

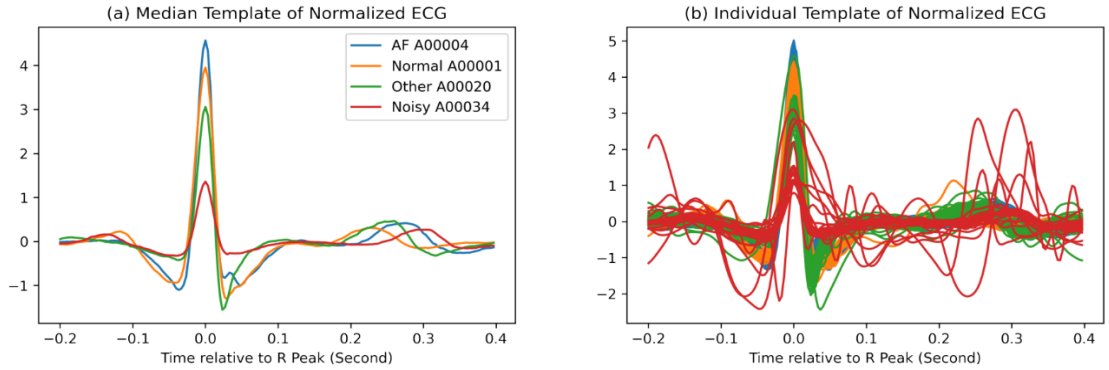

Figure 3. Median (a) and all individual (b) heartbeat templates from the sample ECGs shown in Figure 2.

Table 1 lists and describes all 23 features extracted in this study. A heatmap is created from all 23 extracted features to find the correlation between each of the features (Figure 7). Once all the features are extracted, six distinct feature datasets are created: (1) HRV_F: The 12 HRV frequency domain features. (2) HRV_T: The 6 HRV time domain features. (3) *HRV*: All 18 HRV features (both time domain and frequency domain). (4) *ECG*: The 5 ECG morphology features. (5) HRV_T + *ECG*: The 6 HRV time domain features plus 5 ECG morphology features. (6) *HRV* + *ECG*: All 18 HRV features (both time domain and frequency domain) plus 5 ECG morphology features.

Table 1. Description of the ECG and HRV based features used in this study.

Features (related reference)	Unit	Description
ECG Morphology Feature		
Count of Negative Peaks	None	Number of templates where the absolute of the maximum is less than the absolute of the minimum

Normalized P Wave Amplitude (Goodfellow et al. 2017)	None	P wave amplitude of the median template normalized by the median template maximum
Median Template Correlation (Ghiasi et al. 2017; Goodfellow et al. 2017)	None	Median of correlations between consecutive templates
Median R Peak Amplitude (Goodfellow et al. 2017)	None	Median amplitude of all detected R Peaks
Max/Rpeak Ratio	None	Ratio between the absolute maximum of the preprocessed ECG and the median R peak amplitude
HRV Time Domain Feature (Shaffer & Ginsberg, 2017)		
Median RR Intervals	ms	Median R peak - R peak (RR) intervals
SDNN	ms	Standard deviation of RR intervals
RMSSD	ms	Root mean square of successive RR interval differences
SDSD	ms	Standard deviation of differences between adjacent RR
pNN50	%	Percentage of successive RR intervals that differ by more than 50 ms
Triangular Index	None	Integral of the density of the RR interval histogram divided by its height
HRV Frequency Domain Feature (Shaffer & Ginsberg, 2017)		
VLF Peak	Hz	Peak frequency of very-low-frequency band
LF Peak	Hz	Peak frequency of low-frequency band
HF Peak	Hz	Peak frequency of high-frequency band
LF/HF Power Ratio	None	Ratio between LF power and HF power
VLF Relative Power	%	Relative power of very-low-frequency band
LF Relative Power	%	Relative power of low-frequency band
HF Relative Power	%	Relative power of high-frequency band
VLF Absolute Power	ms^2	Absolute power of very-low-frequency band
LF Absolute Power	ms^2	Absolute power of low-frequency band
HF Absolute Power	ms^2	Absolute power of high-frequency band
LFnu	None	Low-frequency relative power in normalized units
HFnu	None	High-frequency relative power in normalized units

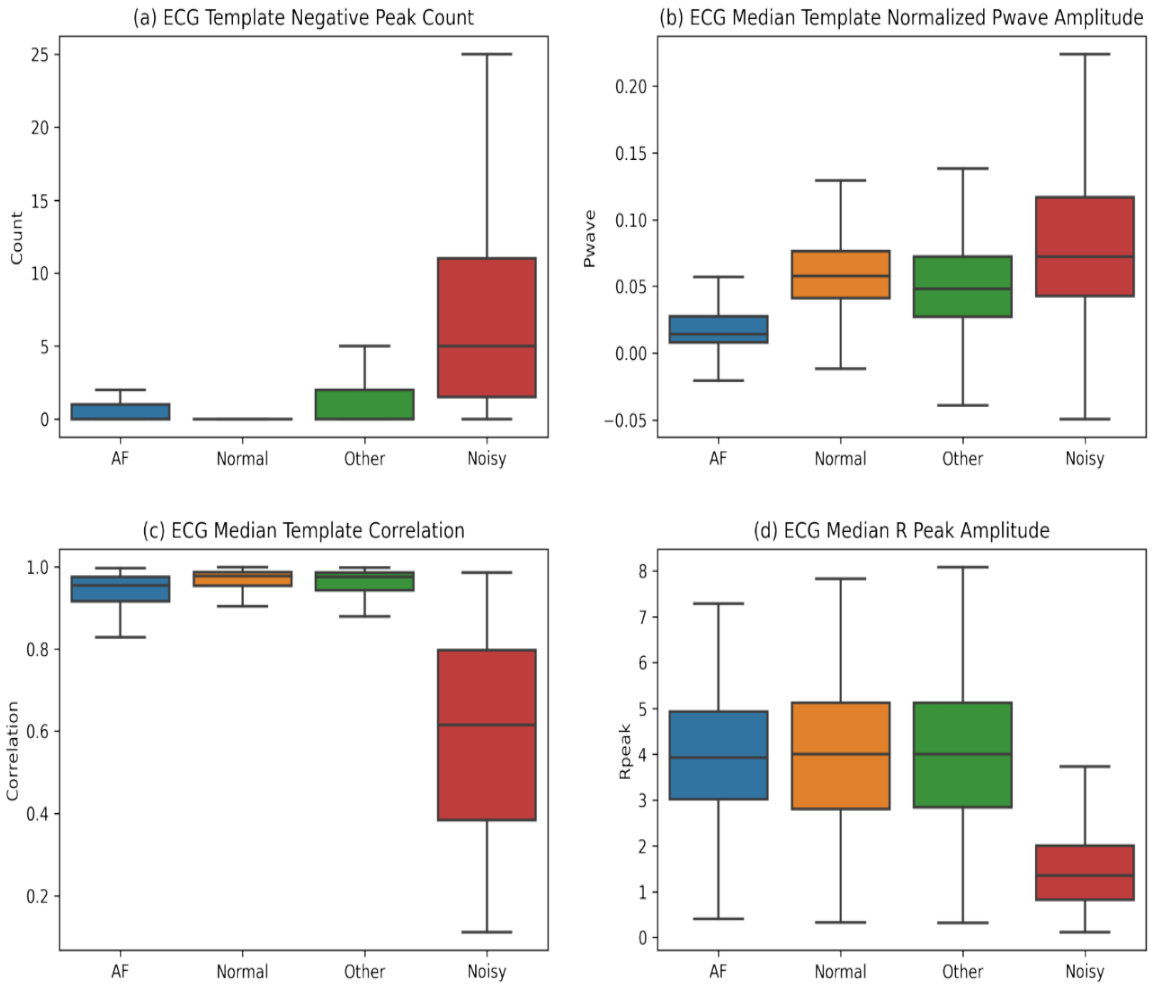

Figure 4. Boxplots of ECG morphology features for each category of heart arrhythmias. (a) Count of negative peaks, (b) Normalized P wave amplitude, (c) Median template correlations, (d) Median R peak amplitude.

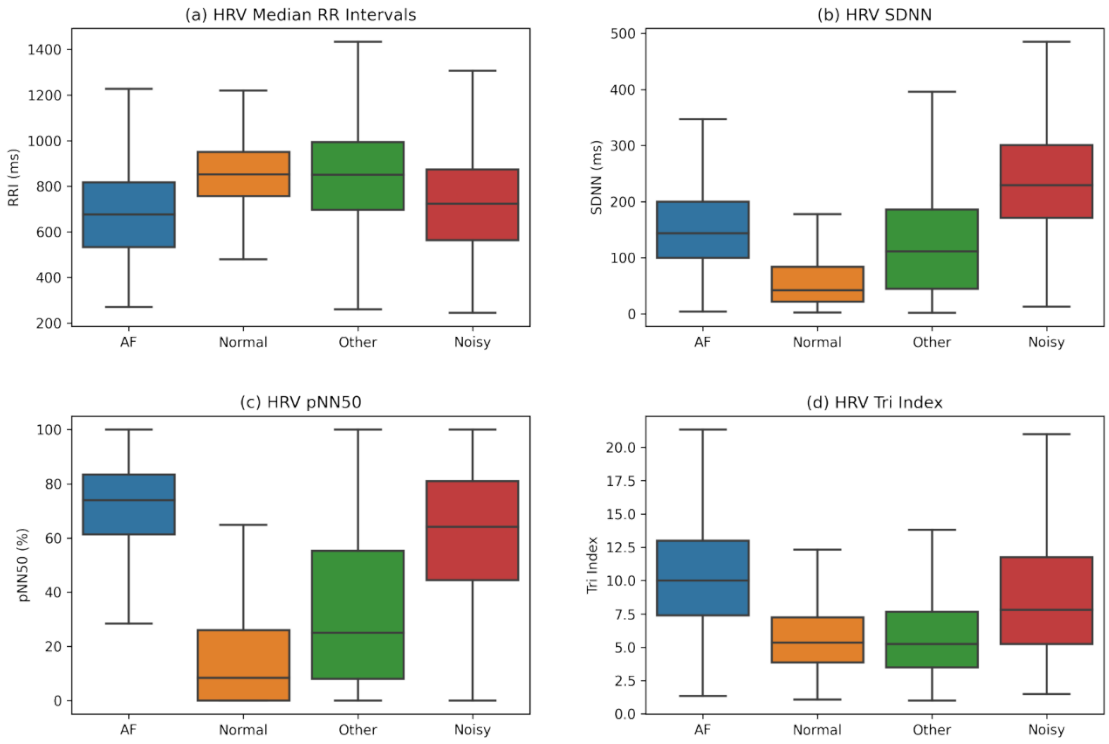

Figure 5. Boxplots of HRV time domain features for each category of heart arrhythmias. (a) Median RR intervals, (b) SDNN, (c) pNN50, (d) Triangular index.

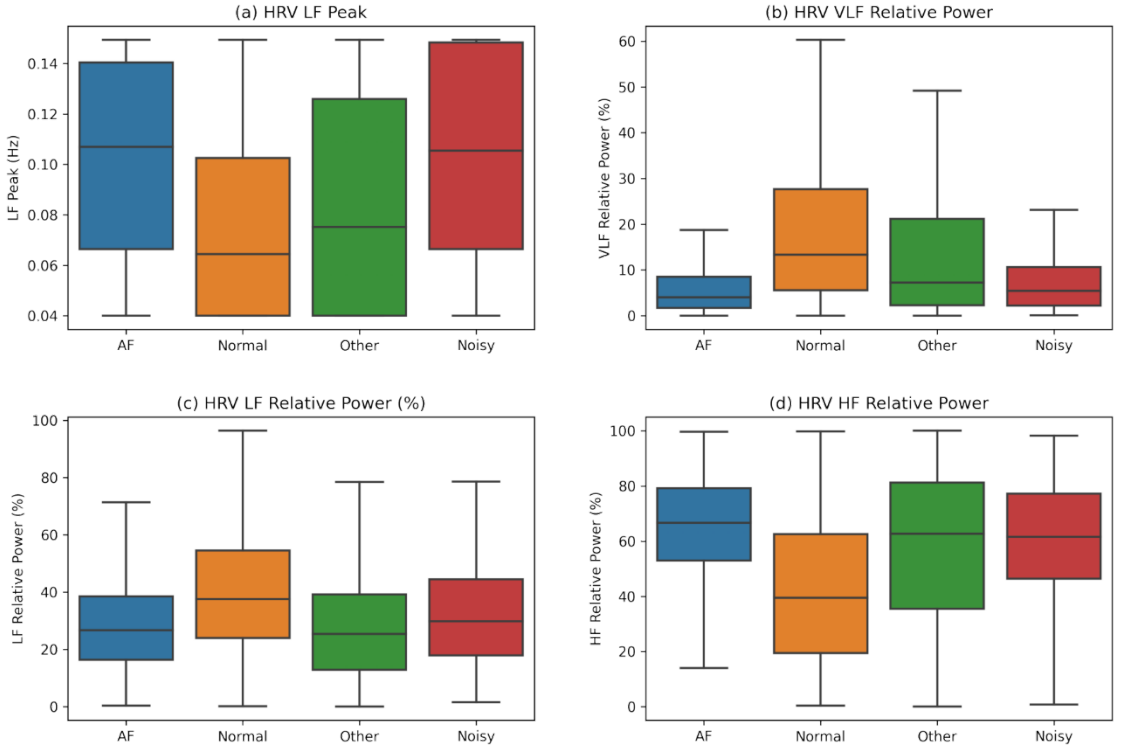

Figure 6. Boxplots of HRV frequency domain features for each category of heart arrhythmias. (a) LF peak, (b) VLF relative power, (c) LF relative power, (d) HF relative power.

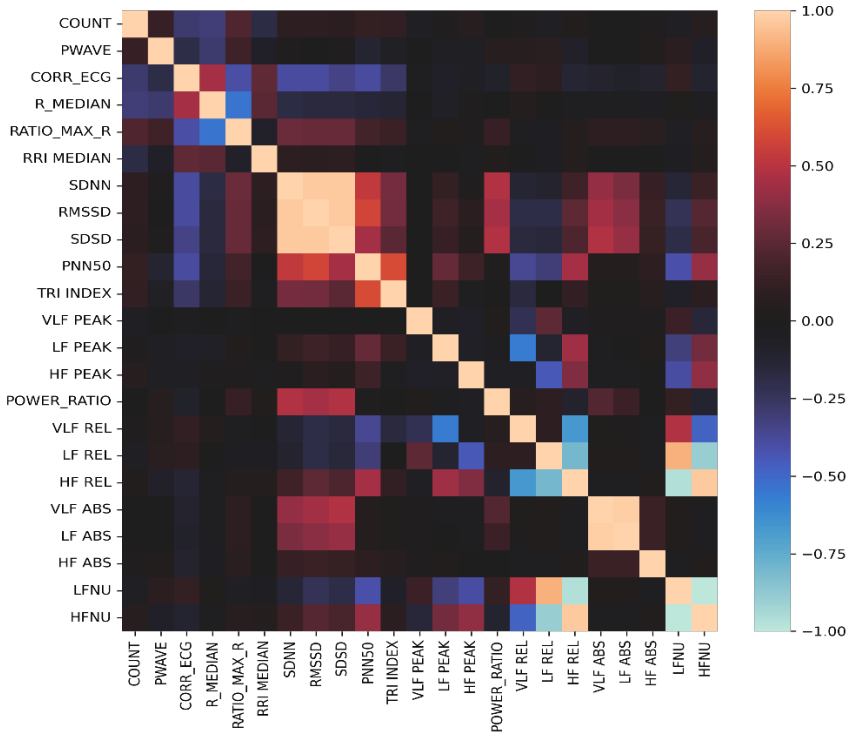

Figure 7. Correlation heatmap between all the extracted features in this study.

ML Models and Classification

In this study, the ECG classification is performed separately by four different ML models: Logistic Regression, Decision Tree, KNN, and CNN. Logistic Regression is the baseline model for classifications, usually used to predict the probability of an event occurring. Decision Tree is a non-parametric model that classifies using decision rules that increase in complexity with the tree's depth. We use a depth of 9 in this study. The KNN model classifies an unknown point according to the categories of the point's k (where k is any integer) nearest neighbors (Fix & Hodges, 1989). The number k of nearest neighbors varies with the dataset. For this paper, the number of nearest neighbors used in the KNN model is 13. The CNN model is a DL model that uses a kernel to detect patterns across an array. Because of its pattern detection capabilities, CNN is often used for image classifications (LeCun & Bengio, 1995). In this study, the 1-dimensional (1D) CNN model has a kernel size of 8 along with 32 filters. The 1D CNN layer is connected to a 1D global average pooling layer, which is then connected to a dense output layer with a softmax activation. The loss function used for the CNN model training is categorical cross entropy and the optimizer used is adam. The CNN model is trained for 1000 epochs with a batch size of 40.

Each of the six feature datasets is used to train the above four ML models to compare the impact of different types and combinations of features. The ML models' classification output (i.e. labels) include all 4 ECG categories (AF, Normal, Other, and Noisy). The formula used to calculate the F1 scores of the classification (for each category and for the overall F1 score) is provided by the PhysioNet Challenge 2017 website (https://physionet.org/content/challenge-2017/1.0.0/). The overall F1 score is the average of the F1 scores of all 4 categories. Because of the labeling issue for the Noisy class at the time of the PhysioNet Challenge 2017 (Clifford et al. 2017), the overall F1 score was only averaged over 3 categories (AF, Normal, Other) in many

previous studies participated the PhysioNet Challenge 2017. Here by using the corrected latest version, version 3 (V3) of the ECG class labels, which was only available after the PhysioNet Challenge 2017, the overall F1 score is averaged over all 4 categories as defined in the PhysioNet Challenge 2017 website.

Each of the six feature datasets is normalized using the Standard Scaler method before being given to the four ML models for classifications. Then it is randomly split into 90% training data and 10% hidden test data across all 8528 ECGs using the stratified Sklearn train test split function, which keeps the same relative distribution of the 4 classes in both the training and the hidden test data (Pedregosa et al. 2011). Cross validation is also used in this study. It is the practice of splitting the training dataset further into different combinations of training and validation sets to gauge how splitting the training dataset differently affects the model's performance. One common way to implement cross validation is through k-folds, which divides the training set into k sections, and selects one of these sections to become the validation set (Hastie et al. 2008). Since the ECG recordings used in this study are heavily imbalanced, with most of the recordings containing normal ECGs and few AF and noisy ECGs, stratified 10-fold cross validation is chosen so that the relative distribution of the 4 classes is the same in each fold as in the full dataset. Finally, each model is trained on the entire training data, which is considered the final model and then applied to the hidden test data for the classification. The cross validation and hidden test F1 scores for each model are calculated for each of the 4 ECG categories and averaged over all 4 ECG categories. The F1 scores of the four ML models trained on the six feature datasets are compared to each other.

Results

Out of the four ML models used in this study, the DL 1-D CNN model trained on the dataset containing a combination of 6 HRV time domain features and 5 ECG morphology features (HRV_T + ECG) leads to the best results in terms of the overall F1 score (0.70 ± 0.03 for the cross validation and 0.73 for the hidden test), which is the average of the F1 scores of all 4 categories (Table 2 and Figure 8). The feature dataset HRV_T + ECG also scores well in each individual category (Table 2). The Noisy F1 scores are much higher when using the feature dataset (HRV_T + ECG) than using only the HRV time domain features (HRV_T) for each of the four ML models (Table 2). Including the 5 ECG morphology features can also increase the AF F1 scores than using only the 6 HRV time domain features (Table 2), since the absence of a P wave is a key feature of AF. Using the 12 HRV frequency domain features (HRV_F) scores lowest in the AF and Noisy categories as well as the overall F1 scores in all four ML models (Table 2). In fact, adding HRV frequency domain features does not significantly improve the F1 scores and often decreases the performance of the models compared to the results using only the HRV time domain features (Table 2).

Figure 9 shows an example plot of the training and validation losses as a function of epochs for the CNN model using the most performant feature dataset (HRV_T + ECG). In this example, 10% of the training data is randomly selected as the validation data using a stratified split. Both the training and validation losses decrease rapidly within the first 50 epochs and decrease slowly afterwards (Figure 9). ML models with a high variance tend to overfit the data during the training process when the validation loss starts to increase and diverge from the training loss in the learning curve. Here in our case, the validation loss does not increase and remains similar to the training loss as the end of the training process, suggesting that our constructed CNN model does not overfit the data and has a small variance. Figure 10 shows the Confusion Matrix for the hidden test using the best ML model (CNN) with the most performant feature dataset (HRV_T + ECG). Table 3 lists the corresponding mean and standard deviation of the cross validation F1 scores along with the hidden test F1 scores for the CNN model trained on the most performant feature dataset (HRV_T + ECG). For the cross validation, the AF, Normal, Other, and Noisy F1 scores are 0.75 ± 0.03, 0.86 ± 0.01, 0.63 ± 0.03, and 0.54 ± 0.11 respectively. For the hidden test, the AF, Normal, Other, and Noisy F1 scores are 0.77, 0.87, 0.62, and 0.67 respectively.

Table 2. The performance of ML models in terms of F1 score for Cross Validation (CV) and hidden Test set. The
values for CV are expressed as mean and standard deviation along all 10 folds.

	HRV_F		HRV_T		HRV		ECG		$HRV_T + ECG$		$HRV + ECG$	
LogReg	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test
AF F1	0.00 ± 0.00	0.00	0.64 ± 0.04	0.67	0.66 ± 0.04	0.66	0.32 ± 0.07	0.39	0.74 ± 0.03	**0.76**	0.74 ± 0.02	**0.76**
Normal F1	0.77 ± 0.00	0.75	0.83 ± 0.01	0.82	0.83 ± 0.01	0.81	0.77 ± 0.00	0.77	0.84 ± 0.01	**0.83**	0.84 ± 0.01	**0.83**
Other F1	0.38 ± 0.02	0.31	0.47 ± 0.04	0.40	0.48 ± 0.04	0.38	0.09 ± 0.02	0.07	0.50 ± 0.03	**0.47**	0.52 ± 0.03	0.46
Noisy F1	0.03 ± 0.04	0.00	0.13 ± 0.08	0.12	0.14 ± 0.09	0.17	0.59 ± 0.08	0.68	0.57 ± 0.10	0.62	0.59 ± 0.10	**0.70**
Overall F1	0.30 ± 0.01	0.26	0.52 ± 0.03	0.50	0.53 ± 0.03	0.51	0.44 ± 0.02	0.48	0.66 ± 0.03	0.67	0.67 ± 0.02	**0.69**
DecTree	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test
AF F1	0.14 ± 0.07	0.26	0.65 ± 0.05	0.62	0.64 ± 0.06	0.61	0.49 ± 0.06	0.49	0.70 ± 0.05	**0.75**	0.68 ± 0.05	0.71
Normal F1	0.77 ± 0.01	0.76	0.86 ± 0.01	0.84	0.84 ± 0.01	0.83	0.77 ± 0.01	0.79	0.85± 0.01	**0.85**	0.85 ± 0.01	**0.85**
Other F1	0.41 ± 0.03	0.36	0.59 ± 0.04	0.55	0.57 ± 0.04	0.51	0.22 ± 0.03	0.18	0.60 ± 0.04	**0.58**	0.61 ± 0.03	0.52
Noisy F1	0.05 ± 0.05	0.06	0.22 ± 0.08	0.18	0.22 ± 0.13	0.29	0.48 ± 0.10	**0.56**	0.47 ± 0.11	0.54	0.45 ± 0.07	0.40
Overall F1	0 34 ± 0.03	0.36	0.58 ± 0.03	0.55	0.57 ± 0.05	0.56	0.49 ± 0.03	0.50	0.66 ± 0.04	**0.68**	0.65 ± 0.02	0.62
KNN	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test
AF F1	0.09 ± 0.03	0.10	0.69 ± 0.03	0.72	0.64 ± 0.04	0.70	0.49 ± 0.05	0.57	0.75 ± 0.02	**0.80**	0.74 ± 0.02	**0.80**
Normal F1	0.76 ± 0.01	0.73	0.86 ± 0.01	0.85	0.84 ± 0.01	0.83	0.77 ± 0.01	0.79	0.86 ± 0.01	**0.86**	0.84 ± 0.01	0.85
Other F1	0.37 ± 0.02	0.32	0.58 ± 0.03	0.54	0.50 ± 0.03	0.48	0.23 ± 0.02	0.24	0.58 ± 0.04	**0.59**	0.52 ± 0.02	0.51
Noisy F1	0.04 ± 0.06	0.00	0.29 ± 0.10	0.18	0.12 ± 0.06	0.13	0.54 ± 0.09	**0.69**	0.54 ± 0.10	0.55	0.53 ± 0.09	0.53
Overall F1	0.32 ± 0.02	0.29	0.61 ± 0.03	0.57	0.52 ± 0.02	0.53	0.51 ± 0.03	0.57	0.68 ± 0.03	**0.70**	0.66 ± 0.02	0.67
CNN	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test
AF F1	0.01 ± 0.01	0.00	0.70 ± 0.04	0.70	0.64 ± 0.05	0.65	0.53 ± 0.05	0.57	0.75 ± 0.03	**0.77**	0.70 ± 0.04	0.75
Normal F1	0.77 ± 0.01	0.75	0.87 ± 0.01	0.86	0.84 ± 0.01	0.84	0.79 ± 0.01	0.81	0.86 ± 0.01	**0.87**	0.84 ± 0.01	0.84
Other F1	0.37 ± 0.03	0.28	0.63 ± 0.04	0.57	0.55 ± 0.03	0.51	0.21 ± 0.03	0.20	0.63 ± 0.03	**0.62**	0.54 + 0.03	0.52
Noisy F1	0.09 ± 0.06	0.06	0.33 ± 0.11	0.41	0.17 ± 0.11	0.43	0.56 ± 0.08	0.61	0.54 ± 0.11	**0.67**	0.55 ± 0.11	0.66
Overall F1	0.31 ± 0.02	0.27	0.63 ± 0.03	0.63	0.55 ± 0.03	0.61	0.52 ± 0.03	0.55	0.70 ± 0.03	**0.73**	0.66 ± 0.02	0.69

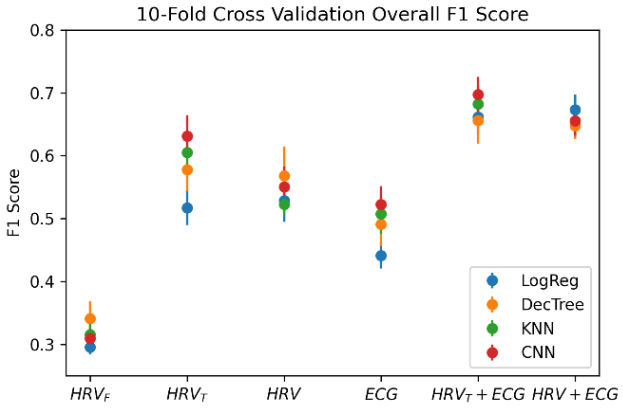

Figure 8. Cross validation overall F1 score for ML models.

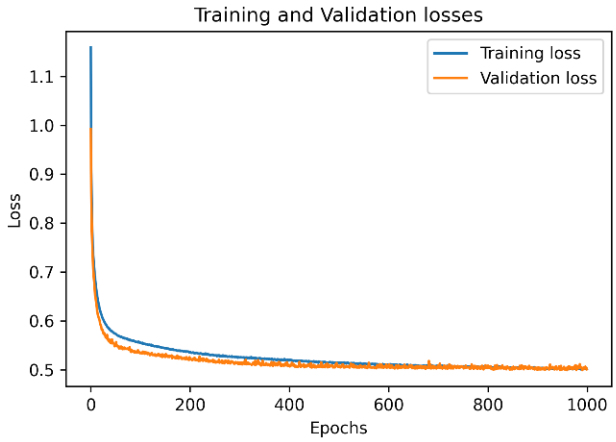

Figure 9. An exemplary plot of training and validation losses as a function of epochs for the CNN model using the most performant feature dataset: HRV time domain features + ECG morphology features.

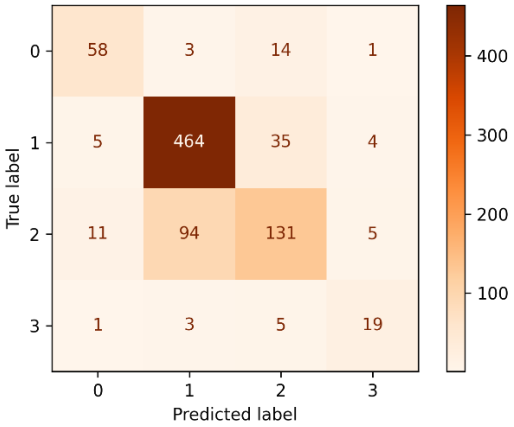

Figure 10. Confusion Matrix for the hidden test using the CNN model with the most performant feature dataset: HRV time domain features + ECG morphology features. Class labels: 0 - AF, 1 - Normal, 2 - Other, 3 – Noisy.

Table 3. F1 Scores using CNN and the most performant feature dataset: HRV time domain features + ECG morphology features. The values for Cross Validation (CV) are expressed as mean and standard deviation along all 10 folds.

CNN	CV	Test
AF F1	0.75 ± 0.02	0.77
Normal F1	0.86 ± 0.01	0.87
Other F1	0.63 ± 0.03	0.62
Noisy F1	0.54 ± 0.11	0.67
Overall F1	0.70 ± 0.03	0.73

Discussion

Our classification results reflect the importance of the ECG morphology features in distinguishing the Noisy class from other classes. As shown in the corresponding boxplots (Figure 4), there is significant difference between the Noisy class and the AF, Normal, Other classes in the ECG morphology features of median template

correlations and R peak amplitudes. The 25th to 75th percentile range for the median template correlation is below 0.8 for the Noisy class and above 0.9 for the AF, Normal, and Other classes. The 25th to 75th percentile range for the median R peak amplitude is below 2.0 for the Noisy class and above 2.7 for the AF, Normal, Other classes.

Meanwhile, the HRV time domain features are important in distinguishing the Normal class from all other classes. For example, the Normal class has the lowest median SDNN and pNN50 values in all classes (Figure 5). From the heatmap of all extracted features (Figure 7), we can see that the SDNN, RMSSD, and SDSD features are highly correlated with each other, and they all reflect the smaller variance of the RR intervals in the Normal class compared to all other classes. In general, the AF and Noisy classes tend to have larger pNN50 values than Normal and Other classes (Figure 5). Hence, the combination of the HRV time domain features with the ECG morphology features could improve the detection of the AF class (Table 2).

The HRV frequency domain features, such as LF relative power which is anticorrelated with HF relative power as seen in the heatmap (Figure 7), are not dramatically different across all 4 ECG classes (Figure 6) and therefore not particularly helpful in detecting AF and Noisy classes. Consistently, the HRV frequency domain feature dataset HRV_F does not contribute much to detecting AF and Noisy classes in all four ML models. For example, for the CNN model trained on the HRV_F feature dataset, the AF and Noisy F1 scores are significantly low for both the cross validation and the hidden test. This is likely due to the limited numbers of heartbeats from the short ECG recordings used in this study for extracting HRV frequency domain features. In particular, the short ECG recording ranging from 9 seconds to 61 seconds is not long enough to resolve the very low frequency (less than 0.04 Hz) and low frequency (0.04 Hz to 0.15 Hz) components of the HRV signal in the frequency domain.

Table 2 shows that the HRV frequency domain feature dataset HRV_F has some contributions in detecting Normal and Other classes in all four ML models, consistent with the boxplots that show that the Normal and Other classes have more VLF relative power (Figure 6). However, in all four ML models, the HRV time domain feature dataset HRV_T alone can lead to higher F1 scores for the Normal, Other, and Noisy classes, and substantially higher AF F1 scores, compared to the F1 scores using the HRV frequency domain feature dataset HRV_F. When the HRV frequency domain feature dataset HRV_F is combined with the HRV time domain feature dataset HRV_T, the F1 scores of each class using the combined dataset *HRV* do not change much for the Logistic Regression and Decision Tree models and even degraded for the KNN and CNN models, compared to those using HRV_T alone. Thus, it is advisable to use HRV time domain features but not HRV frequency domain features if the ECGs recordings are short (less than 1 minute), which costs less time for classifications than using both HRV time and frequency domain features.

On the other hand, when the ECG morphology feature dataset *ECG* is combined with the HRV time domain feature HRV_T, the AF F1 scores are improved and the Noisy F1 scores are much higher in all four ML models for the combined feature dataset HRV_T + *ECG*, compared to those using HRV_T alone. When all 23 features are combined together, i.e. using the combined feature dataset *HRV* + *ECG*, again the F1 scores of each class do not change much for the Logistic Regression and Decision Tree models and even degraded for the KNN and CNN models compared to the case using HRV_T + *ECG* without including the HRV frequency domain feature. In all four ML models, the ECG morphology feature dataset *ECG* dominates the contribution to Noisy F1 scores, and adding the HRV time domain or all HRV (both time and frequency) features to *ECG* do not increase Noisy F1 scores.

For the Logistic Regression model, the best hidden test F1 scores (for each class and for the average of the 4 classes) are obtained with either the feature dataset HRV_T + *ECG* or the feature dataset *HRV* + *ECG* (Table 2), and both feature datasets give similar results. For the Decision Tree and KNN models, the feature dataset HRV_T + *ECG* often leads to the best hidden test F1 scores (Table 2). For the CNN model, the feature dataset HRV_T + *ECG* gives the best hidden test F1 scores for all classes and for the average of the 4 classes

(Table 2). These results again confirm the important role of the combined feature dataset HRV_T + *ECG* in classifying short ECG recordings.

In general, the hidden test F1 score is similar to the corresponding mean cross validation F1 score, and their difference is close to the associated standard deviation of the cross validation F1 scores (Table 2). In all four ML models trained on all six feature datasets, the standard deviation of the cross validation F1 scores is small for the Normal class and large for the Noisy class, which reflects the large sample size for the Normal class and the small sample size for the Noisy class in the short ECG recordings used in this study.

Our study here is limited by the types of extracted features and ML models, and the size and distribution of the ECG data samples used in this study. In future research, the impact of other informative features needs to be explored, and the robustness and generalizability of the results need to be tested with other ML models including more complex DL models and ECG recordings acquired in different settings. The pure feature-based classification methods might be combined with data driven DL models to improve the detection performance. The multiple classification approach used here could also be compared with ensemble of classifiers in future studies.

Conclusion

In this paper, the impact of different types of features on classifying short ECG recordings is investigated, and it is concluded that our DL (1D CNN) model trained on a combination of HRV time domain features and ECG morphology features results in the highest average F1 score over the 4 categories, out of all four ML models used in this study. The results of this study indicate that adding HRV frequency domain features to HRV time domain features does not contribute much improvement to the F1 scores and often degrades the classification performance for short ECG recordings, compared to classifications using only HRV time domain features. Combining the ECG morphology features with the HRV time domain features is helpful in improving the AF detection because AF is characterized by the absence of a P wave. In addition, the ECG morphology features are very useful for detecting the Noisy category specifically, which contains more inverted peaks, lower correlations between templates (Goodfellow et al. 2017; Ghiasi et al. 2017), and smaller R peaks amplitudes than AF, Normal, and Other categories. Hence it is important to combine ECG morphology features with the HRV time domain features in classifying short ECG recordings.

The CNN model used in this study is simple in structure and contains one 1D convolutional layer. DL models are popular because they often do not require extracted features to classify (Andreotti et al. 2017; Warrick & Homsi, 2017; Hsieh et al. 2020). The overall F1 score obtained here using a simple DL (1D CNN) model trained on a combination of 6 HRV time domain and 5 ECG morphology features (Table 3) is comparable to that obtained using some complicated DL models (with millions of model parameters) trained on the raw PhysioNet Challenge 2017 short ECG recordings (Andreotti et al. 2017; Warrick & Homsi, 2017; Hsieh et al. 2020). Our results suggest that feature-based DL could serve as a viable and less expensive approach for classifying short ECG recordings, which costs much less computing time for training.

Acknowledgments

We thank Cambridge Centre for International Research (CCIR, https://www.cambridge-research.org/) One-on-one Research Mentorship Program for providing us this research collaboration opportunity.

References

Andreotti, F., Carr, O., Pimentel, M. A., Mahdi, A., & De Vos, M. (2017). Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ECG. In 2017 Computing in Cardiology (CinC), 1-4, IEEE. https://doi.org/10.22489/CinC.2017.360-239

Billeci, L., Chiarugi, F., Costi, M., Lombardi, D., & Varanini, M. (2017). Detection of AF and other rhythms using RR variability and ECG spectral measures. In 2017 Computing in Cardiology (CinC), 1-4, IEEE. https://doi.org/10.22489/CinC.2017.344-144

Carreiras, C., Alves, A. P., Lourenço, A., Canento, F., Silva, H., & Fred, A. (2015). Biosppy: Biosignal processing in python. https://github.com/PIA-Group/BioSPPy

Chandra, B. S., Sastry, C. S., Jana, S., & Patidar, S. (2017). Atrial fibrillation detection using convolutional neural networks. In 2017 Computing in Cardiology (CinC), 1-4, IEEE. https://doi.org/10.22489/CinC.2017.163-226

Clifford, G. D., Liu, C., Moody, B., Li-wei, H. L., Silva, I., Li, Q., ... & Mark, R. G. (2017). AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017. In 2017 Computing in Cardiology (CinC), 1-4, IEEE. https://doi.org/10.22489/CinC.2017.065-469

Coppola, E. E., Gyawali, P. K., Vanjara, N., Giaime, D., & Wang, L. (2017). Atrial fibrillation classification from a short single lead ECG recording using hierarchical classifier. In 2017 Computing in Cardiology (CinC), 1-4, IEEE. https://doi.org/10.22489/CinC.2017.354-425

Da Silva-Filarder, M., & Marzbanrad, F. (2017). Combining template-based and feature-based classification to detect atrial fibrillation from a short single lead ECG recording. In 2017 Computing in Cardiology (CinC), 1-4, IEEE. https://doi.org/10.22489/CinC.2017.346-357

Datta, S., Puri, C., Mukherjee, A., Banerjee, R., Choudhury, A. D., Singh, R., ... & Khandelwal, S. (2017). Identifying normal, AF and other abnormal ECG rhythms using a cascaded binary classifier. In 2017 Computing in cardiology (cinc), 1-4, IEEE. https://doi.org/10.22489/CinC.2017.173-154

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3), 238-247. https://doi.org/10.2307/1403797

Ghiasi, S., Abdollahpur, M., Madani, N., Kiani, K., & Ghaffari, A. (2017). Atrial fibrillation detection using feature based algorithm and deep convolutional neural network. In 2017 Computing in Cardiology (CinC), 1-4, IEEE. https://doi.org/10.22489/CinC.2017.159-327

Gomes, P., Margaritoff, P., & da Silva H. P, (2019). pyHRV: Development and evaluation of an open-source python toolbox for heart rate variability (HRV), Proc. Int'l Conf. on Electrical, Electronic and Computing Engineering (IcETRAN), 822-828. https://github.com/PGomes92/pyhrv

Goodfellow, S. D., Goodwin, A., Greer, R., Laussen, P. C., Mazwi, M., & Eytan, D. (2017). Classification of atrial fibrillation using multidisciplinary features and gradient boosting. In 2017 Computing in Cardiology (CinC) 44, 1-4, IEEE. https://doi.org/10.22489/CinC.2017.361-352

Hamilton P. S., & Tompkins W. J. (1986) Quantitative Investigation of QRS Detection Rules Using the MIT/BIH Arrhythmia Database. IEEE Trans Eng Biomed Eng. 33, 1157-65. https://doi.org/10.1109/TBME.1986.325695

Hastie, T., Tibshirani, R., & Friedman, J. (2008) The Elements of Statistical Learning; Springer: New York, NY, USA. https://link.springer.com/978-0-387-21606-5

Hsieh, C. H., Li, Y. S., Hwang, B. J., & Hsiao, C. H. (2020). Detection of atrial fibrillation using 1D convolutional neural network. Sensors, 20(7), 2136. https://doi.org/10.3390/s20072136

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10). http://www.iro.umontreal.ca/~lisa/pointeurs/handbook-convo.pdf

Nattel, S. (2002). New ideas about atrial fibrillation 50 years on. Nature, 415(6868), 219-226. https://doi.org/10.1038/415219a

Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python, JMLR 12, 2825-2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com

Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. Frontiers in public health, 258. https://doi.org/10.3389/fpubh.2017.00258

Smoleń, D. (2017). Atrial fibrillation detection using boosting and stacking ensemble. In 2017 Computing in Cardiology (CinC), 1-4. IEEE. https://doi.org/10.22489/CinC.2017.068-247

Van Zaen, J., Delgado-Gonzalo, R., Ferrario, D., & Lemay, M. (2019). Cardiac arrhythmia detection from ECG with convolutional recurrent neural networks. In International Joint Conference on Biomedical Engineering Systems and Technologies, 311-327, Springer, Cham. https://doi.org/10.1007/978-3-030-46970-2_15

Warrick, P., & Homsi, M.N. (2017). Cardiac arrhythmia detection from ECG combining convolutional and long short-term memory networks. In Proceedings of the 2017 Computing in Cardiology (CinC), 44, 1–4, IEEE. https://doi.org/10.22489/CinC.2017.161-460

Weimann, K., & Conrad, T. O. (2021). Transfer learning for ECG classification. Scientific reports, 11(1), 1-12. https://doi.org/10.1038/s41598-021-84374-8

Xiong, Z., Stiles, M. K., & Zhao, J. (2017). Robust ECG signal classification for detection of atrial fibrillation using a novel neural network. In 2017 Computing in Cardiology (CinC), 1-4, IEEE. https://doi.org/10.22489/CinC.2017.066-138

Yazdani, S., Laub, P., Luca, A., & Vesin, J. M. (2017). Heart rhythm classification using short-term ECG atrial and ventricular activity analysis. In 2017 Computing in Cardiology (CinC), 1-4, IEEE. https://doi.org/10.22489/CinC.2017.067-120

Zabihi, M., Rad, A. B., Katsaggelos, A. K., Kiranyaz, S., Narkilahti, S., & Gabbouj, M. (2017). Detection of atrial fibrillation in ECG hand-held devices using a random forest classifier. In 2017 Computing in Cardiology (CinC), 1-4, IEEE. https://doi.org/10.22489/CinC.2017.069-336

Zihlmann, M., Perekrestenko, D., & Tschannen, M. (2017). Convolutional recurrent neural networks for electrocardiogram classification. In 2017 Computing in Cardiology (CinC), 1-4, IEEE. https://doi.org/10.22489/CinC.2017.070-060