Feature Selection and Extraction in Sequence Labeling for Arrhythmia Detection

Minxiang Ye*†, Vladimir Stankovic*, Lina Stankovic*, Srdjan Lulic‡, Andras Anderla§, Srdjan Sladojevic§
*Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, UK
†Research Center for Intelligent Robot, Zhejiang Lab, Hangzhou 311121, China
‡PanonIT, Laze Nancica 7, 21000 Novi Sad, Serbia
§Department of Industrial Engineering and Management, Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia

***Abstract*—Automated Electrocardiogram (ECG)-based arrhythmia detection methods replace traditional, manual arrhythmia detection reducing the requirement for trained medical staff. Traditionally, ECG-based arrhythmia detection is performed via QRS complex detection followed by feature extraction, based on hand-crafted features, such as RR-intervals, Fast Fourier Transform-based features, wavelet analysis, higher order statistics and Hermite features. After the features are extracted, the ECG segments are classified into pre-defined categories. This study investigates the value of the feature extraction and selection methods for ECG-based arrhythmia detection. That is, with the emerging trend of deep learning methods which are capable of automatic feature extraction and selection, the research question addressed in this paper is if good classification performance can be obtained by feeding the raw ECG sequence directly into robust classifiers or handcrafted feature extraction/selection is necessary. Classification performance across a range of state-of-the-art classification methods indicates that feeding raw signals into the convolution neural network-based classifiers usually leads to the best performance but at the expense of high inference time.**

***Index Terms*—arrhythmia classification, feature selection, sequence labeling, RF, CNN, RNN.**

I. Introduction

Electrocardiogram (ECG) signal analysis is the most effective diagnostic tools to detect cardiac diseases [1]. Numerous methods have been proposed in the literature that replace traditional, manual arrhythmia detection with automated processes that do not require trained medical staff, and hence can be performed remotely, at home. Automated ECG-based arrhythmia detection methods start with QRS complex detection, to split the ECG trace into non-overlapping segments, by identifying Q-, P- and R-wave, that normally appear together and reflect a single ECG trace event. Then, within each segment, feature construction and extraction is performed [2]. The most popular, hand-crafted, features used in the literature include RR-intervals ([3]), that is, the time elapsed between two successive R-waves, Fast Fourier Transform (FFT) based features ([4]), wavelet analysis ([5]), higher order statistics ([6]) and Hermite features ([7]). After the features are extracted, the ECG segments are classified into pre-defined categories, e.g., normal beats, ventricular arrhythmia, supraventricular arrhythmia, etc., using traditional signal classification approaches, such as support vector machines, random decision forests, Hidden Markov models, etc.

Note that the above methods rely on features traditionally used for manual arrhythmia detection, such as the RR-interval, or traditional approaches developed for time-series signal processing analysis, such as FFT and wavelet analysis.

The recent development of deep learning has revolutionised the field of machine learning (ML) based on vast amount of data available to train the models. Deep learning approaches can potentially replace handcrafted feature extraction via an automated, deep feature extraction process relying on the network to recognise and extract the discriminative patterns.

The goal of this paper is to compare the performance of traditional feature extraction and selection methods prior to classification, versus classifying ECG signals by feeding raw signals into deep learning classifiers relying on the machine to select the most appropriate features for the task. We use two popular 'traditional' classifiers, based on support vector machines (SVM) and random decision forests (RF), and three different deep learning architectures. We assess the performance of the classifiers when six different 'traditional' feature maps are compared to the case when we directly feed raw ECG signals into the classifiers. Besides classification accuracy measures, we use inference time to assess practicality of the classification solutions.

Our methodology gives insights into ECG signal analysis, importance of feature selection and extraction for ECG signal analysis, relevance of different features, provides steps towards interpretability and explainability of the results, and the way machine processes these time-series signals.

Our simulation results indicate that feeding raw signals into the convolution neural network (CNN)-based classifiers usually leads to the best performance compared to using expert knowledge to design handcrafted features, but at the expense of high inference time.

The paper is organised as follows. First we describe the setup used for evaluation of the sensitivity of feature selection and extraction in classical sequence labeling machine learning approaches and deep learning based approaches with regard to the classification performance and computational efficiency. Then, we provide our results and discussion. The final section concludes the paper.

This is a peer-reviewed, accepted author manuscript of the following paper: Ye, M., Stankovic, V., Stankovic, L., Lulic, S., Anderla, A., & Sladojevic, S. (Accepted/In press). Feature selection and extraction in sequence labeling for arrhythmia detection. Paper presented at Fourth International Balkan Conference on Communications and Networking, Novi Sad, Serbia.

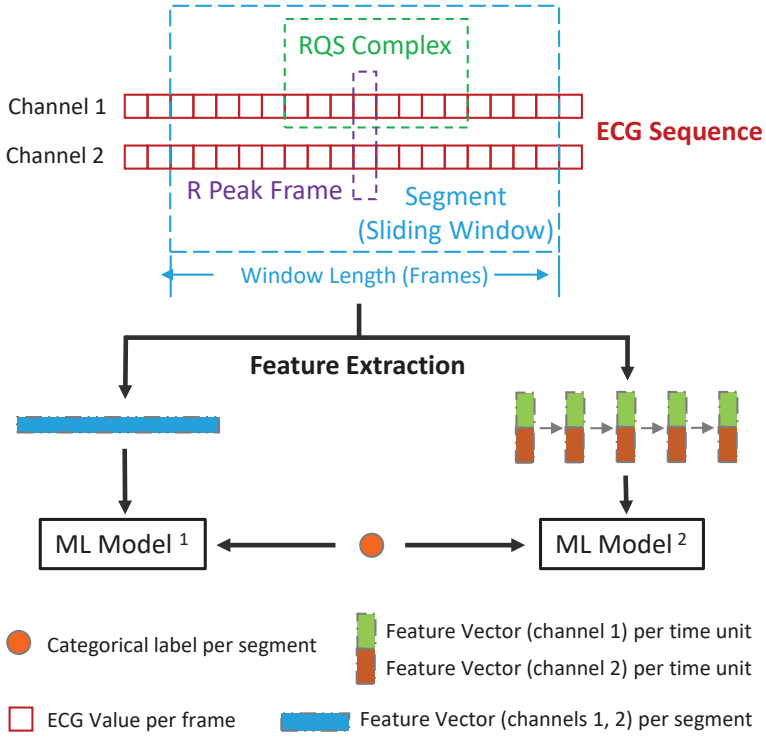


Fig. 1: Many-to-One Architecture: One label is assigned to each ECG window. A frame refers to two samples, one from each channel. ML Model [1] refers to SVM, RF, MLP, or CNN classifier. ML Model [2] refers to Bidirectional GRU classifier. The time unit of time-series feature vectors that are feed into ML Model [2] depends on the feature extraction algorithms, i.e., time resolution in wavelet transform.

II. Background and Evaluation Setup

Sequence labeling or sequence classification assigns categorical/classification labels to a sequence of data points based either on a set of previous timestamped observations or based on a specific time segment of the signal. Sequence labelling is a pattern recognition task that in general can be: (1) one-to-one: predict an individual label in a sequence using a previous observation. (2) one-to-many: predict a sequence of (multiple) labels by retaining states from labels' previous outputs and observations as inputs for the next one. (3) many-to-one: predict a data label for a set of observations from multiple previous observations. (4) many-to-many: assign labels to each member of a subsequence given observations over a period of time. The arrhythmia detection sequence labelling falls into the many-to-one category as we classify each segment of ECG signals into one beat class after QRS complex detection [2].

In the literature, two types of approaches are commonly investigated: (1) Model-based labeling based on a probabilistic model that finds the best matched label via statistical inference. The most commonly used model sequence labeling relies on Markov assumption, that is, the target label is dependent only on the immediately adjacent labels/observations, such as Hidden Markov Model (HMM) and Conditional Random Fields (CRF). (2) Feature-based Labeling, which is a classical classification mapping problem that relates the input observations (features) to a categorical label, using well-known classifiers, such as support vector machines, random decision forest, deep neural networks comprising multi-layer perceptron (MLP), convolutional neural networks and recurrent neural networks (RNN).

Regardless of the classifier used, feature selection is needed to facilitate dimensionality reduction, by selecting the most important features, removing the irrelevant features to make interpretation easier and enable training a simple model with less time, and finally, reducing the risk of over-fitting and hence improving the generalization trading-off the bias and variance. The task of feature selection is to select a subset of the original features to maximise performance based on set criteria, such as, QRS complex detection rate [2] performed to detect the relevant ECG segments of interest.

Feature extraction is usually used to extract the most discriminatory information, transforming the original constructed features into an optimal set of new features. It improves the final predictive performance and is conducted either before applying a machine learning algorithm or within the algorithm during feature engineering, such as linear and nonlinear features used in [3], [4], [6], [8], [9].

For time-series features, sliding window is usually used to represent the sequential pattern [10]. In the many-to-one architecture model, the sliding window method reconstructs the input as a window of observations around a reference frame (R peak frame, extracted by QRS detection) as shown in

Fig. 1 for the case with two ECG channels. The length of the window controls the dimensionality of the input observations, thus directly affecting the performance. Since ECG signals are regular time-series signals of more-or-less fixed duration, we apply a fixed window length and evaluate its resulting predictive performance.

To evaluate the sensitivity of the feature engineering approaches for ECG arrhythmia classification using different sequence labeling classifiers, we select the widely-used open-access MIT-BIH Arrhythmia Dataset [11] to analyse the sensitivity of different state-of-the-art features using classical sequence labeling approaches and deep learning based approaches. The dataset contains ECG recordings from two channels sampled at rate of 360 samples/sec/channel. We separate the whole dataset into training, validation, testing sets with proportions 0.5, 0.15, 0.35, respectively. We select a fixed window length of 180 frames around the extracted segment. The simulation setup is listed in Table I. All evaluations are done in a server with an Intel i7-7800 3.5GHz CPU and a NVIDIA Titan Xp GPU. For each evaluated method, 30 experiments are done to calculate the mean classification error rate and inference speed using different random seeds.

TABLE I: Simulation Setup. N, V, Q, S, F refer to the normal, ventricular ectopic, supraventricular ectopic, fusion, and unknown beats, respectively. The feature map parameters used are all as in the referenced literature. The right column in the last six rows shows the number of real attributes for each feature selection method.

Configuration	Description	
Raw Feature	2 ECG Channels × 180 (Frames) per Detected QRS Segment	
Data Scale	126 Records (613 Strides) with 73854 Segments	
Data Distribution	Train: 44179, Valid: 13242, Test: 31002	
Class Distribution	N: 74683, V: 6605, Q: 3880, S: 2469, F: 786	
Feature Map (Real Attributes per Segment)	RR-Intervals (RRI) [3]	4
	Fast Fourier Transform (FFT) [4]	35 × 2
	Wavelet	23 × 2
	Wavelet with Uniform Local Binary Pattern (WaveletULBP) [5]	59 × 2
	Higher order statistics (HOS) [6]	4 × 5 × 2
	Hermite [7]	22 × 2

A. Classifier Setup

We evaluate the sensitivity of different feature selection methods listed in Table I via Bidirectional GRU (Bi-GRU), SVM, RF, MLP, CNN classifiers. Note that, in the paper, we denote by "MLP" a deep learning based classifier that only uses fully connected layers, and does not use standard convolutional layers and bidirectional GRUs.

For the RF-based classifier, we employ 30 decision trees with depth of 20. The deep learning architectures used in this research are shown in Figs. 2–4 for Bi-GRU, MLP, CNN classifiers, respectively, with respect to the dimensions of input observations. The hyper-parameters used for the MLP classifier are listed in Table II.

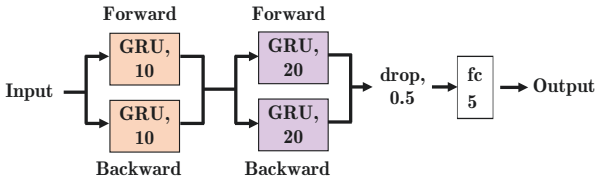


Fig. 2: Deep learning architecture for bidirectional GRU based classifier. 'GRU' refers to the gated recurrent unit (GRU) based recurrent neural network layer with parameters of hidden state size. 'drop' refers to the dropout layer with keeping ratio. 'fc' means the fully connected layer with number of neurons.

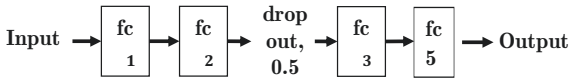
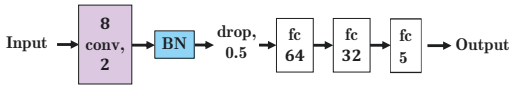


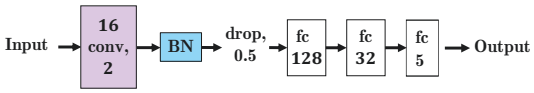
Fig. 3: Deep learning architecture for the MLP classifier. The number of neurons $\upsilon_1, \upsilon_2, \upsilon_3$ in fully-connected (fc) layers vary dependent on the dimension of the input data (see details in Table II).

TABLE II: Hyper-parameters used in the MLP classifier: The number of neurons $\upsilon_1, \upsilon_2, \upsilon_3$ in fully-connected (fc) layers.

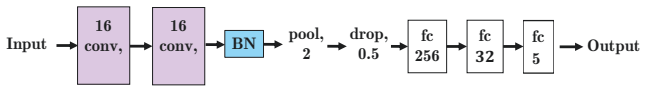
Dimensions of Observation	υ_1	υ_2	υ_3
<8	16	32	None
≥ 8 and <16	32	64	32
≥ 16 and <32	64	128	32
≥ 32 and <64	128	256	64
≥ 64 and <128	256	512	64
≥ 128	512	1024	128



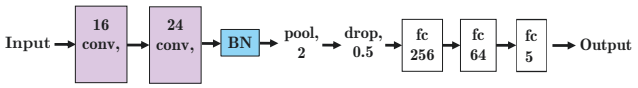
(a) Block diagram of the architecture when number of features <8



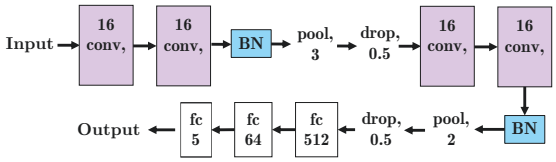
(b) Block diagram of the architecture when number of features ≥8 and <16



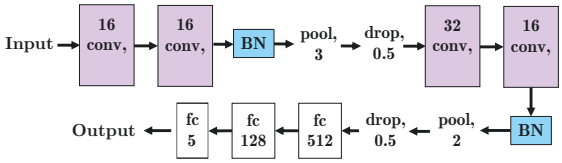
(c) Block diagram of the architecture when number of features ≥16 and <32



(d) Block diagram of the architecture when number of features ≥32 and <64



(e) Block diagram of the architecture when number of features ≥64 and <128



(f) Block diagram of the architecture when number of features ≥128

Fig. 4: Deep learning architecture for the CNN-based classifier. 'conv' refers to the 1D convolutional layer with parameters of kernel size, filter number and constant one stride size. 'BN' refers to the batch normalization layer. Hyper-parameters of different layers vary dependent on the dimensions of the input observation.

III. Results and Discussion

In this section, we first compare the predictive performance of different classifiers using the state-of-the-art feature selection/extraction methods. We discard the RRI, FFT, WaveketULBP, HOS, Hermite features for Bidirectional GRU, since they are designed to explore the relations across multiple frames in the sequence.

We show the classification error rates and F-measure for each class in Tabs. III and IV, respectively. It is clear that 'raw' zero-mean standardized ECG signals fed directly into the classifiers outperform feature extractions across all benchmarked classifiers.

TABLE III: Classification Error Rate (%) for each of five classes (N, S, V, F, and Q. * refers to deep learning based feature extraction approach. Bold numbers show the best performance for each classifier.

Method	Feature Map	Arrhythmia Classes					
		N	S	V	F	Q	Avg
Bi-GRU*	Raw	0.6	25.6	5.6	25.2	0.5	11.5
	Wavelet	0.5	33.8	5.1	24.7	0.6	13.0
SVM	Raw	5.1	12.1	3.8	8.9	0.4	**6.0**
	RRI	1.3	98.6	92.4	85.1	56.8	66.8
	FFT	0.0	100	100	100	100	80.0
	Wavelet	0.0	60.9	96.5	96.8	97.6	70.4
	WaveletULBP	35.8	37.2	39.1	27.7	19.9	31.9
	HOS	1.0	15.7	3.3	22.7	4.0	9.3
	Hermite	36.6	15.5	27.2	21.6	5.4	21.3
RF	Raw	0.1	40.2	6.1	38.8	1.0	17.3
	RRI	1.6	31.1	20.4	77.0	14.2	28.9
	FFT	0.6	59.8	51.8	96.3	84.7	58.6
	Wavelet	0.1	38.9	5.7	37.1	0.9	16.5
	WaveletULBP	0.4	58.4	32.9	48.0	28.9	33.7
	HOS	0.1	35.1	5.7	35.6	2.2	15.7
	Hermite	0.3	36.6	6.6	30.1	0.73	14.8
MLP*	Raw	0.4	12.3	1.9	19.1	0.4	**6.8**
	RRI	0.4	95.2	74.5	100	100	74.0
	FFT	1.7	66.2	54.3	86.4	70.5	55.8
	Wavelet	0.4	17.3	3.9	19.7	0.4	8.4
	WaveletULBP	1.4	45.3	21.9	39.9	15.7	24.8
	HOS	0.5	18.3	2.8	20.1	0.4	8.4
	Hermite	0.6	35.7	5.1	32.5	1.2	15.0
CNN*	Raw	0.3	10.6	3.0	16.7	0.4	**6.8**
	RRI	3.2	56.5	39.9	100	60.4	52.0
	FFT	1.4	57.7	56.7	83.1	82.1	56.2
	Wavelet	0.4	13.4	2.7	17.7	0.4	**6.9**
	WaveletULBP	1.1	56.3	46.8	41.2	24.3	33.9
	HOS	0.4	12.1	2.7	16.0	0.3	**6.3**
	Hermite	0.6	24.6	4.4	29.6	0.6	12.0

It can be seen from the two tables that both SVM classifier and deep learning-based classifiers achieve high predictive performance when raw data is fed directly into the classifier. Deep learning-based classifiers that automatically extract deep features achieve comparable mean classification error rate and F-measure using Wavelet and HOS feature extraction methods. This indicates that investing time and effort on hand-crafting the features following traditional approaches is redundant.

Note that, the inference speed rapidly increases as the model size grows with the increasing dimensionality of input observations/features, that is, increasing the window size. Especially for deep learning based classifiers, a deeper architecture that automatically extracts deeper features suffers from a signifi-

TABLE IV: F-Measure for each of five classes. * refers to deep learning based feature extraction approach. Bold numbers show the best performance for each classifier.

Method	Feature Map	Arrhythmia Classes					
		N	S	V	F	Q	Avg
Bi-GRU*	Raw	0.99	0.80	0.95	0.80	0.99	0.91
	Wavelet	0.99	0.77	0.95	0.81	0.99	0.90
SVM	Raw	0.97	0.57	0.95	0.74	0.99	0.84
	RRI	0.92	0.03	0.14	0.17	0.56	0.36
	FFT	0.92	0.00	0.00	0.00	0.00	0.18
	Wavelet	0.92	0.56	0.07	0.06	0.05	0.33
	WaveletULBP	0.77	0.29	0.44	0.10	0.51	0.42
	HOS	0.99	0.85	0.94	0.84	0.98	0.92
	Hermite	0.77	0.16	0.80	0.15	0.97	0.57
RF	Raw	0.99	0.75	0.96	0.74	0.99	0.89
	RRI	0.97	0.79	0.82	0.32	0.86	0.75
	FFT	0.94	0.57	0.61	0.07	0.26	0.49
	Wavelet	0.99	0.76	0.96	0.76	0.99	0.89
	WaveletULBP	0.97	0.58	0.77	0.66	0.80	0.76
	HOS	0.99	0.78	0.96	0.77	0.99	0.90
	Hermite	0.99	0.77	0.95	0.80	**1.00**	0.90
MLP*	Raw	0.99	0.90	**0.98**	0.86	**1.00**	**0.95**
	RRI	0.92	0.09	0.38	0.00	0.00	0.28
	FFT	0.94	0.47	0.57	0.21	0.41	0.52
	Wavelet	0.99	0.86	0.97	0.86	0.99	0.94
	WaveletULBP	0.97	0.67	0.82	0.68	0.86	0.80
	HOS	0.99	0.86	0.97	0.86	1.00	0.94
	Hermite	0.99	0.75	0.95	0.79	0.99	0.89
CNN*	Raw	**1.00**	**0.91**	0.97	**0.87**	**1.00**	**0.95**
	RRI	0.94	0.59	0.67	0.00	0.44	0.53
	FFT	0.94	0.55	0.55	0.24	0.28	0.51
	Wavelet	0.99	0.89	0.97	**0.87**	0.99	**0.95**
	WaveletULBP	0.96	0.59	0.67	0.59	0.79	0.72
	HOS	0.99	0.89	0.97	**0.87**	**1.00**	**0.95**
	Hermite	0.99	0.81	0.96	0.80	0.99	0.91

TABLE V: Inference Speed (microseconds/frame). * refers to automatic feature extraction approach.

Feature Map	Method				
	Bi-GRU*	SVM	RF	MLP*	CNN*
Raw	33259	8275	8	838	1366
RRI	—	2803	4	429	556
FFT	—	5376	4	514	936
Wavelet	3673	2811	4	506	833
WaveletULBP	—	6633	4	575	1133
HOS	—	540	4	537	854
Hermite	—	2944	4	568	943

cantly longer inference time shown in Tab. V, limited by the size of model and computation capacity.

IV. Conclusion

This paper first reviews the state-of-the-art feature selection and extraction methods for arrhythmia detection. Then, classification error rates and inference speed are measured to analyze the sensitivity of these methods, either with or without domain knowledge. Experimental results indicate that the feature selection/extraction methods, namely FFT and Wavelet transforms, are not as critical to the arrhythmia sequence labeling classification task compared to raw and statistical features, such as HOS. Given sufficient training data, the deep learning-based feature extraction approaches, i.e., Bidirectional GRU, MLP and CNN, achieve better generalization and classification performance. However, they often operate slower at inference, thus require to be accelerated by modern computing devices.

V. Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 734331. The University of Strathclyde gratefully acknowledges the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] S. Gayathri, M. Suchetha, and V. Latha, "Ecg arrhythmia detection and classification using relevance vector machine," *Procedia engineering*, vol. 38, pp. 1333–1339, 2012.

[2] A. Ligtenberg and M. Kunt, "A robust-digital qrs-detection algorithm for arrhythmia monitoring," *Computers and Biomedical Research*, vol. 16, no. 3, pp. 273–286, 1983.

[3] P. De Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ecg morphology and heartbeat interval features," *IEEE transactions on biomedical engineering*, vol. 51, no. 7, pp. 1196–1206, 2004.

[4] E. R. Adams and A. Choi, "Using neural networks to predict cardiac arrhythmias," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 10 2012, pp. 402–407.

[5] S. Nikan, F. Gwadry-Sridhar, and M. Bauer, "Pattern recognition application in ecg arrhythmia classification," *BIOSTEC 2017*, p. 48, 2017.

[6] Y. Kaya, H. Pehlivan, and M. E. Tenekeci, "Effective ecg beat classification using higher order statistic features and genetic feature selection." *Biomedical Research*, vol. 28, no. 17, pp. 7594–7603, 2017.

[7] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sornmo, "Clustering ecg complexes using hermite functions and self-organizing maps," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 838–848, 2000.

[8] F. A. Elhaj, N. Salim, A. R. Harris, T. T. Swee, and T. Ahmed, "Arrhythmia recognition and classification using combined linear and nonlinear features of ecg signals," *Computer Methods and Programs in Biomedicine*, vol. 127, no. C, pp. 52–63, 2016.

[9] T. Teijeiro, P. Félix, J. Presedo, and D. Castro, "Heartbeat classification using abstract features from the abductive interpretation of the ecg," *IEEE journal of biomedical and health informatics*, vol. 22, no. 2, pp. 409–420, 2016.

[10] T. G. Dietterich, "Machine learning for sequential data: A review," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, 2002, pp. 15–30.

[11] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, May 2001.