Cardiac Arrhythmias Classification in Kardiovize Population Study

Martin Pesl[1,2,3], Jakub Hejc[2,4], Tomas Kulik[1,2], Tomas Vicar[4], Petra Novotna [4], Marina Ronzhina [4], Juraj Jakubik[2] Pavel Leinveber [1,2], Juan Pablo Gonzalez Rivas [2], Zdenek Starek [1,2]

[1] 1st Department of Internal Medicine, Cardio-Angiology, Faculty of Medicine, Masaryk University, Brno, Czech Republic [2] ICRC, St. Anne's University Hospital, Brno, Czech Republic
[3] Department of Biology, Faculty of Medicine, Masaryk University, Brno, Czech Republic
[4] Department of Biomedical Engineering, Brno University of Technology, Brno, Czech Republic

Abstract

Automatic classification of heart rhythm becomes essential in population studies. There are new options for ECG data handling, such as deep-learning models. In this study, we compare our detector and ELI™ 350 ECG Mortara detector using 12-lead volunteers ECG, from Kardiovize study. The ECGs were analyzed by a self-developed deep-learning arrhythmia detector. The evaluation process was focused on atrial fibrillation (AF), the most common arrhythmia in the Czech population. The training database from publicly available datasets included 43,000 variable records. On the training set, the F1-score of the model reached 0.86 and 0.87 for normal sinus rhythm and AF, respectively. In both categories, false positives occur. One of the reasons for the model misclassification was incorrect expert evaluation used as a predicted model output. In the test phase, no records were assigned to the AF category. On the contrary, the Mortara system classified 6 records as AF. Visual verification confirmed the correctness of the model output. From present pilot study, deep-learning classifier ResNet model outperform currently used system in both main categories reaching Sp of 1.0 for atrial fibrillation, and F_1 of 0.875 for long QT syndrome. Particularly, no false atrial fibrillation detection were indicated in model output. ECG evaluation using a deep-learning model seems to be useful tool for handling population data.

1. Introduction

Number of automated electrocardiogram (ECG) classification methods have been reported over last decade, often evaluated solely on small or homogeneous datasets. Available ECG identification is still sub-optimal, despite advanced software solutions. Atrial fibrillation (AF) is the most frequently encountered cardiac arrhythmia and related to substantial cardiovascular (chronic heart failure) and cerebrovascular (stroke) morbidity and mortality [1]. Stroke patients with the highest mortality rate are more likely to suffer from preexisting AF, which is up to 5 times more likely than the general population [2].

In this paper, we present application of deep learning model for ECG classification, aiming to automatically identify patients with AF and other selected criteria.

2. Material and Methods

2.1. Training dataset

Training data originated in the 2020 Physionet/CinC Challenge (PhCi2020) publicly available dataset. It is composed of 43,101 labeled recordings from 6 different sources and includes 24 scored pathologies [3,4]. Recordings are sampled with various sampling frequencies (257, 500 or 1000 Hz) and resolution settings.

2.2. Kardiovize population dataset

The Kardiovize study is a cross-sectional population-based study. Aiming to evaluate adult population health in Brno, the second-largest city in the Czech Republic, with 373,327 residents [5]. Survey sampling was preformed in January 2013 with technical assistance from the health insurance providers (HIP). Providing a random selection of age– and sex- stratified sample of 6,377 permanent residents from Brno aged 25 to 64 years. HIPs directly mailed invitation letters to selected individuals with a description of the study ensuring confidentiality. The overall response rate was 33.9 %. The Kardiovize was approved by the Ethics Committee of St Anne's University Hospital, Brno, Czech Republic. All participants signed the informed consent. Screened were 397 anonymous ECGs, recorded between May and September 2020 without any information on age and sex.

Original ECG data were recorded with Mortara ELI™ 350 system at sampling frequency 1000 Hz. Each file con-

 ISSN: 2325-887X DOI: 10.22489/CinC.2021.271

tains 12-lead ECG encoded by BASE64 system and one or more diagnostic classes provided by the Mortara/Veritas conventional algorithm automatic evaluation system, Version 7.3.0 (Mortara Instrument, Milwaukee, WI, USA).

2.3. Data preparation

Several preprocessing steps had to be applied in order to standardize input/output of the model for data originating from distinct sources. First, diagnostic labels attached in the Mortara system were filtered out and remapped onto labels that properly matched the ones provided in the Physionet/CinC Challenge dataset. Additionally, within both inference and training phase, the ECG data were re-sampled with fixed sampling frequency of 125 Hz (decimation combined with anti-aliasing FIR filter), possible baseline wander was reduced by extracting moving-averaged signal (5s window) and signal were then smoothed by Savitzky-Golay filter (3 samples long; 2nd order). During the training phase an augmentation pipeline consisted of several randomly controlled transformation methods, which were applied on the ECG signal to prevent the model from over-fitting. More detailed description about used augmentation methods can be found in [6].

2.4. Deep neural model architecture

Baseline model was based on modified convolutional neural network (CNN) previously reported by our team in [6]. Briefly, the architecture consists of 6 residual layers [7] with three 1D convolutional filters in each layer. The model was extended by global skip connection in order to efficiently address vanishing gradient problem. Element-wise addition of an identity mapping and residual block output was replaced by a concatenation operator. Multi-label output of the model was guaranteed by $c = 24$ independent fully connected binary classifiers, which allows arbitrary combination of class labels [8]. Implementation details of proposed model architecture are depicted in Figure 1.

Final class label was given by a majority vote of an ensembles of 3 bootstrap models. Each of them was fitted to a 90 % subset of the training data. The number of ensembles was chosen to balance the benefit and computational complexity. Model ensemble should be able to decrease variance and produce better generalized output [9]. Additional step towards an improvement of performance was made by Bayesian optimization of class specific thresholds mapping raw model scores into class label and application of the Stochastic Weight Averaging algorithm during training phase. In the former method, default threshold of 0.5 was replaced by a value maximizing custom loss function [6] used for training. Main goal of the latter technique was to find flat minimum of the criterion function by averaging

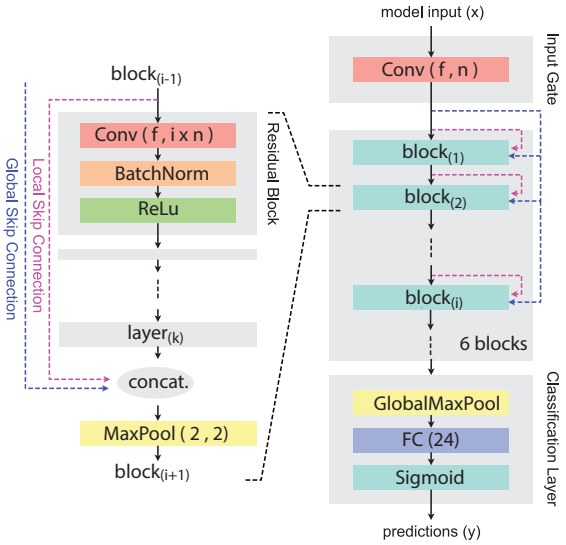

Figure 1: Architecture of proposed model. i – ResNet block number; k – ResNet block layer number; f – kernel size; n – number of filters in first layer.

model parameters captured within various phases of the training [10]. Finally, the model was trained with Adam optimizer and two cycle learning rate strategy. Within the first cycle, weighted cross-entropy was used as a loss function, while in the second cycle custom loss function [6] was used to retrain the model.

2.5. Validation of the Kardiovize dataset

Since Kardiovize screening relies primarily on an automated ECG interpretation done by the Mortara system, we have performed additional evaluation to get a ground truth diagnostic information. Clinical validation was done independently by two cardiac electrophysiology professionals and was primarily focused towards atrial fibrillation and long QT syndrome. Inter-rater agreement was calculated by Krippendorff's alpha coefficient. For the purpose of evaluation algorithms, a ground truth label in case of disagreement between the raters was established as consensus of the two.

3. Results

Model performance was evaluated separately on training (PhCi2020_T), hidden (PhCi2020_H) and our test (Kardiovize) dataset using partial F1 score, sensitivity (Se) and specificity (Sp) for each classification group. Neither confusion matrix nor Se or Sp for PhCi2020_H were publicly available and therefore were not listed in the results. The training performance was previously stated and discussed in detail here [6].

As is shown in Table 1, the Kardiovize contains clinical

statements with high degree of agreement in between cardiology experts with Krippendorrf's alpha of 0.874. Degree of agreement of 1.0 was achieved with regard to the most severe disorders followed here, atrial fibrillation and the long QT syndrome. On the contrary, a compliance between ResNet model and Mortara system was lowered to 0.591, mainly due to misdiagnosed cases with atrial fibrillation and long QT on the side of Mortara and premature ventricular beats on the side of ResNet Model.

Table 1: Inter-rater agreement given as pair-wise Krippendorff's alpha score. Diagnostic classes included in the analysis were: SR, AF, LQT, PAC and PVC.

Raters	α
Human rater A/Human rater B	0.874
Ours/ELI^{TM} 350	0.591

Our main subject of interest was to identify patients with atrial fibrillation. Here, training F1 score was 0.829. On the hidden part of the PhCi2020 database F1 score was actually 0.871. For the normal sinus rhythm F1 score reached 0.864 and 0.633 on the training and hidden set, respectively.

Table 2: Confusions of the ResNet model and Mortara system considering atrial fibrillation and long QT syndrome. AF – atrial fibrillation; PAC – premature atrial contraction; PVC – premature ventricular contraction; LQT – long QT syndrome; SR – sinus rhythm.

ID	ELI^{TM} 350	Our model	Reference
171	AF	PAC	PAC
193	AF	PAC	PAC
21	AF	PAC	PAC
297	AF	PAC, PVC, 1AVB	PAC, 1AVB
329	AF	PAC, PVC	PAC
82	AF	PAC	PAC
54	LQT	LQT	LQT
104	LQT	SR	LQT
148	SR	LQT	LQT
236	LQT	LQT	LQT
273	LQT	LQT	LQT
300	LQT	LQT	LQT
360	SR	LQT	LQT
395	SR	LQT	LQT

In the Kardiovize database there were 6 subjects automatically diagnosed with atrial fibrillation by Mortara system. As is shown in Table 2, neither one of them was identified as AF by our deep neuronal model. Interestingly, non of the subject in the cohort was also newly diagnosed with AF which does not correspond to commonly reported prevalence. There were also three patients newly diagnosed with LQT and not captured by Mortara system. In Table 3, we present overall results of the ResNet model and Mortara system on the Kardiovize dataset. The ResNet model outperform currently used system in both main categories reaching Sp of 1.0 for atrial fibrillation, and F_1 of 0.875 for long QT syndrome. Values marked by a dash symbol were could not be calculated due to missing diagnostic classes in the dataset.

Table 3: Classification results of the ResNet model and the Mortara system on Kardiovize database. AF – atrial fibrillation; PAC – premature atrial contraction; PVC – premature ventricular contraction; LQT – long QT syndrome; SR – sinus rhythm.

Lbl.	Classifier	Dataset	F_1	Se	Sp
SR	Ours	PhCi2020$_H$	.633	n/a	n/a
	Ours	Kardiovize	**.999**	**.997**	–
	ELI^{TM} 350	Kardiovize	.992	.985	–
AF	Ours	PhCi2020$_H$	.871	n/a	n/a
	Ours	Kardiovize	–	–	**1.00**
	ELI^{TM} 350	Kardiovize	.000	–	.985
LQT	Ours	PhCi2020$_H$	.227	n/a	n/a
	Ours	Kardiovize	**.875**	**.875**	.997
	ELI^{TM} 350	Kardiovize	.769	.625	**1.00**
PAC	Ours	PhCi2020$_H$	.627	n/a	n/a
	Ours	Kardiovize	.727	**1.00**	.952
	ELI^{TM} 350	Kardiovize	**.909**	.833	**1.00**
PVC	Ours	PhCi2020$_H$	.238	n/a	n/a
	Ours	Kardiovize	.417	.455	.979
	ELI^{TM} 350	Kardiovize	**.667**	**.636**	**.992**

4. Discussion

Presented pilot group has almost 400 ECGs, being the largest study to knowledge of the authors in context of "healthy" Central European population. On the other hand, still small from computational perspective. The approach using ResNet mode is reaching relatively high Sp and Se, in case of groups of AF and SR 100 % and 99,7 % respectively. Nevertheless as there are no AF in the probed group, only one of the parameters can be calculated each time. These would be surpassing other groups Sp and Se, however it is impossible to compare results from differing datasets. One should note that Asgari et al. [11] used stationary wavelet transform (SWT) combined with Support Vector Machine (SVM) model to detect AF in short-term ECG signals. The Se and Sp of the method reached 97.0 % and 97.1 %, respectively. Faust et al. in [12] reported a six-layer bidirectional long-short term memory (LSTM) network to classify AF/non-AF within ECG signals of 100 beats. The accuracy, Se and Sp of the model, reached 98.51 %, 98.32 %, and 98.67 %, respectively. Dang et al. [13] added a bidirectional LSTM network to form a nine-layer deep neural network, and classified AF for 100 consecutive R peak sample points. The accuracy, Se and Sp of the model were 96.59 %, 99.93 %, and 97.03 %. The accuracy of the training set and validation set were 99.94 % and 98.63 %. Short 1s ECG fragments may be analyzed nevertheless accuracy was reaching lower values 81.07 % and 84.85 %, respectively [14] This short overview shows that

ML algorithms could be used for detection and recognition of either shorter and longer ECG segments, reaching reasonable experimental results in multiple classes. ResNet model had similar lower Sp/Se only in case of PVC, due to misclassification to PAC.

In summary, our study is comparable to other ML methods, applied to unique set of data. The ML algorithms are robust and more suitable for the automatic detection of AF, than manufacturer provided automatic detection, which may be despite updates unreliable in current clinical settings.

4.1. Limitations

It should be noted that diagnostic labels provided by Mortara/Veritas system are not validated by a cardiologist. The results may be biased by possible faulty ECG interpretation made by another automated system. This is not the case of subject diagnosed with atrial fibrillation which was further evaluated by a clinical professional. Other databases were not yet probed nor integrated in addition to our initial training set. Using a single database is known limitation, despite evaluation indicators of the training set, validation set, and test set performed excellently. Obvious limitation is absence of AF in target group, preventing assessment of Sp of sinus and Se of AF. We aim to combined with other databases available and conduct further study to achieve better generalization performance of the model.

5. Conclusion

Selected ECG recordings Kardiovize study was screened with neural model ResNet, pretrained on external ECG database. ResNet identified 6 otherwise faulty automatically assorted AF diagnoses, which would be time consuming and in the whole Kardiovize study impossible manual task. All ECGs were manually confirmed and overall agreement on classified parameters (SR. AF, LQT, PAC and PVC) was 0.874.

6. Acknowledgements

Supported by the European Union's Horizon 2020 FET research and innovation program no.: 732170 CResPace, by the European Regional Development Fund - Project ENOCH CZ.02.1.01/0.0/0.0/1, and by ICRC IGC 1901 project: ECG machine reading for large patient cohorts.

References

[1] Khurshid S, Trinquart L, Weng LC, et al. Atrial fibrillation risk and discrimination of cardioembolic from noncardioembolic stroke. Stroke 2020;51(5):1396–1403.

[2] Zhou M, Wang H, Zeng X, et al. Mortality, morbidity, and risk factors in china and its provinces 1990–2017: a systematic analysis for the global burden of disease study 2017. The Lancet September 2019;394(10204):1145–1158.

[3] Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 2000;101(23):215–220.

[4] Alday EAP, Gu A, Shah AJ, et al. Classification of 12-lead ECGs: the PhysioNet/computing in cardiology challenge 2020. Physiological Measurement January 2021; 41(12):124003.

[5] Movsisyan NK, Vinciguerra M, Lopez-Jimenez F, et al. Kardiovize brno 2030, a prospective cardiovascular health study in central europe: Methods, baseline findings and future directions. European Journal of Preventive Cardiology August 2017;25(1):54–64.

[6] Vicar T, Hejc J, Novotna P, et al. Ecg abnormalities recognition using convolutional network with global skip connections and custom loss function. In 2020 CinC. 2020; 1–4.

[7] He K, Zhang X, Ren S, at al. Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; 770–778.

[8] Zhang ML, Zhou ZH. A Review on Multi-Label Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering 2013;26(8):1819–1837.

[9] Hansen LK, Salamon P. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 1990;12(10):993–1001.

[10] Izmailov P, Podoprikhin D, Garipov Taa. Averaging Weights Leads to Wider Optima and Better Generalization. arXiv preprint arXiv180305407 2018;.

[11] Asgari S, Mehrnia A, Moussavi M. Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine. Comput Biol Med May 2015; 60:132–142.

[12] Faust O, Shenfield A, Kareem M, et al. Automated detection of atrial fibrillation using long short-term memory network with RR interval signals. Computers in Biology and Medicine November 2018;102:327–335.

[13] Dang H, Sun M, Zhang G, et al. A novel deep arrhythmia-diagnosis network for atrial fibrillation classification using electrocardiogram signals. IEEE Access 2019;7:75577–75590.

[14] Xu X, Wei S, Ma C, et al. Atrial fibrillation beat identification using the combination of modified frequency slice wavelet transform and convolutional neural networks. J Healthc Eng July 2018;2018:1–8.

Address for correspondence:

Martin Pesl, ICRC, St. Anne's University Hospital, Pekarska 53, 656 91 Brno, Czech Republic, martin.pesl@fnusa.cz