Multi-label Classification of Cardiac Abnormalities for Multi-lead ECG Recordings Based on Auto-encoder Features and a Neural Network Classifier

Onno Linschmann[1], Maurice Rohr[2], Klaus Steffen Leonhardt[1], Christoph Hoog Antink[2]

[1] Medical Information Technology, Helmholtz-Institute, RWTH Aachen University, Aachen, Germany

[2] KIS*MED, AI Systems in Medicine, Technische Universität Darmstadt, Darmstadt, Germany

Abstract

A large number of annual deaths are related to cardiac abnormalities. Typically for diagnosis, a 12-lead Electrocardiogram (ECG) is recorded and analysed. However, the electrode placement for all twelve leads is not always correct and twelve-lead ECG systems are not always available. Therefore, the aim of the Computing in Cardiology Challenge 2021 is to correctly classify and label two, three, four, six and twelve-lead ECG recordings regarding 26 classes of cardiac abnormalities.

A deep neural network consisting of three parts is developed. First, a convolutional neural network structure for extending the feature space. Second, a parallel Long-short term memory (LSTM) and linear network structure for feature extraction, and third, a linear layer with Sigmoid activation for multi-label classification. For training, a custom loss function was used which can be assumed to be a weighted, generalised Softmax function with quadratic differences. The network labels 6 random segments of 8 seconds of the lead I and lead II of the ECG recordings and then combines the labels according to a simple rule.

For hyperparameter tuning, an 80/20 split (training data/ test data) was used. The network performance was finally tested using a 5-fold cross-validation and validated on the Challenge System (team = AADAConglomerate). The network scored 0.428 while parameter tuning and 0.183, 0.223, 0.206, 0.246 and 0.223 for the 5-fold cross-validation for all leads. Finally, a score of 0.32 was achieved on the Challenge system during the official phase for all leads.

1. Introduction

Cardiovascular diseases are some of the most prominent causes for death [1]. Therefore, an early diagnosis is crucial for achieving timely treatment. Usually, a twelve-lead ECG recording is used for diagnosis. However, the electrode placement as well as diagnosis is a cumbersome and error-prone process [2]. Algorithms for automated detection as well as the use of ECG recordings with less than 12 leads could highly simplify the process while increasing the accuracy of the diagnosis.

The aim of the Computing in Cardiology Challenge 2021 [3, 4] was to correctly classify twelve-lead, six-lead, four-lead, three-lead and two-lead ECG recordings regarding 26 diagnoses. Similar to last year's challenge, no limitation with respect to the classification algorithm were made.

In last year's challenge [5], where the objective was to correctly classify twelve-lead ECG recordings, a prominent number of high-ranking algorithms used deep neural networks [6–8]. Therefore, a deep neural network for multi-label classification was developed. In the following sections, the methodology and results are presented. First, the preprocessing of the ECG recordings is described followed by the deep neural network structure. Third, hyperparameter tuning and network training are demonstrated, and finally, the results are presented and discussed.

2. Method and Materials

During the course of the Challenge, the neural network structure as well as the training procedure was altered several times. For the unofficial phase, a simple multi-layer perceptron with three linear layers and Sigmoid activations was trained by hand-crafted features. Formerly, an auto-encoder was to be used for the creation of additional features, but this idea was abandoned and a signal deep neural network was used for both feature extraction and classification. The latter is described in the following.

2.1. Preprocessing

For robustness, the beginning and end of each recording were discarded since they often exhibited artifacts. Then, the raw signal was filtered by a fifth order butterworth high-pass filter with cut-off frequency at 0.5 Hz followed by a 50 Hz notch filter to further suppress power line noise

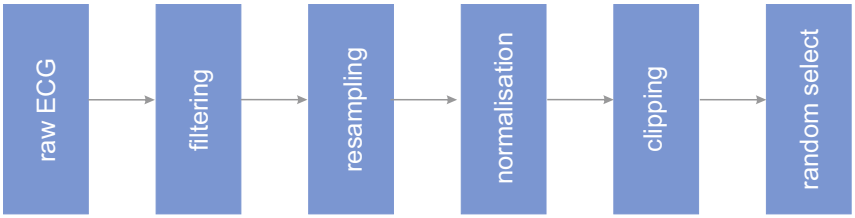

Figure 1. Toolchain of preprocessing. In the 'clipping' block the removal of artifacts at the beginning and end and by the threshold are summarised.

[9]. After resampling the signal to 250 Hz, the signal was normalised and noisy parts were clipped if the absolute value of the signal amplitude exceeded a threshold, i.e. the amplitude was set to zero. Finally, n random segments of t seconds were selected and forwarded to the neural network. Only Einthoven lead I and lead II were used. The preprocessing toolchain is depicted in Fig. 1.

2.2. Deep neural network architecture

The deep neural network consists of three parts (see Fig. 2). First, a convolutional neural network, second a parallel structure of a Long-short term memory (LSTM) network and a linear network, and third, a linear layer with Sigmoid activation for multi-label classification. The first part of the network is used for extending the feature space. The convolutional network's key element is the combination of max pooling, convolutional layer (conv), rectified linear unit (ReLU) and batch normalisation which increase the feature space while binning features using max pooling. Each convolution layer increases the dimensionality by two, except the last which reduces it by two, such that it is first increased up to 256 and finally decreased to 128. The kernel size for all convolutional layers is five and a zero-padding of two is applied. The stride is selected as two for all convolutional layers except the second and second last layers where it is one. The max pooling layers have a kernel size of three, a stride of two, and zero-padding of one is applied. The input size is a segment of around 8 seconds (2048 samples at 250 Hz) of lead I and II in one vector (2048x2). Over-fitting is prevented by dropout layers and max pooling.

The second part of the network is for decreasing the feature space and construct features which are evolving over time (LSTM) and rigid features (Linear network). Also, information from both leads are combined. Finally, the third part of the network is for classification. The network classifies n random segments. Each output which can be assumed to be a probability for the class is transformed to a label by rounding, thus generating n times 26 labels. The labels are then combined by taking the maximum over all segments except for the normal class (sinus rhythm). For the normal class, the minimum operator is applied preventing the network from being influenced by the imbalanced dataset and to prevent illogical classifications.

2.3. Hyperparameter tuning and training

A crucial part for multi-label classification is the loss function which is usually chosen to be the binary cross-entropy with a Sigmoid activation [10]. Since the loss function has an important influence on the training, a custom loss-function was used. The loss-function can be assumed to be a weighted, generalised Softmax with quadratic differences given by

$$\mathrm{loss} = \frac{1}{\mathrm{n}} \sum_{\mathrm{j}=1}^{\mathrm{n}} [\mathrm{W}(\mathrm{y} - \frac{\tilde{\mathrm{y}} \cdot \sum_{\mathrm{i}} \mathrm{y_i}}{\sum_{\mathrm{i}} \tilde{\mathrm{y}}_{\mathrm{i}}})] \tag{1}$$

Here, y is the vector containing the reference labels while $\tilde{y}$ is the vector of predicted labels for n segments. The weighting gives more emphasis to the sinus rhythm to balance the data. For optimisation, the ADAM optimiser is used. Additionally, a learning rate scheduling is used which decreases the learning rate after 20 epochs by 75%. For hyperparameter tuning, the whole dataset is split such that 80% is used for training and 20% for testing. After hyperparameter tuning the network is validated by a 5-fold cross-validation. The final hyperparameter values can be found in Tab. 1.

Table 1. Values of the hyperparameter after tuning.

Hyperparameter	Value
Learning rate	$1.5 \cdot 10^{-5}$
Batch size	128
Dropout probability	0.1
Epochs	40
Number segments n	6

3. Results

The achieved scores for the unofficial phase and the official phase are shown in Tab. 2. This includes 5-fold cross-validation, parameter tuning and the score achieved on the challenge system. First, it can be seen that the

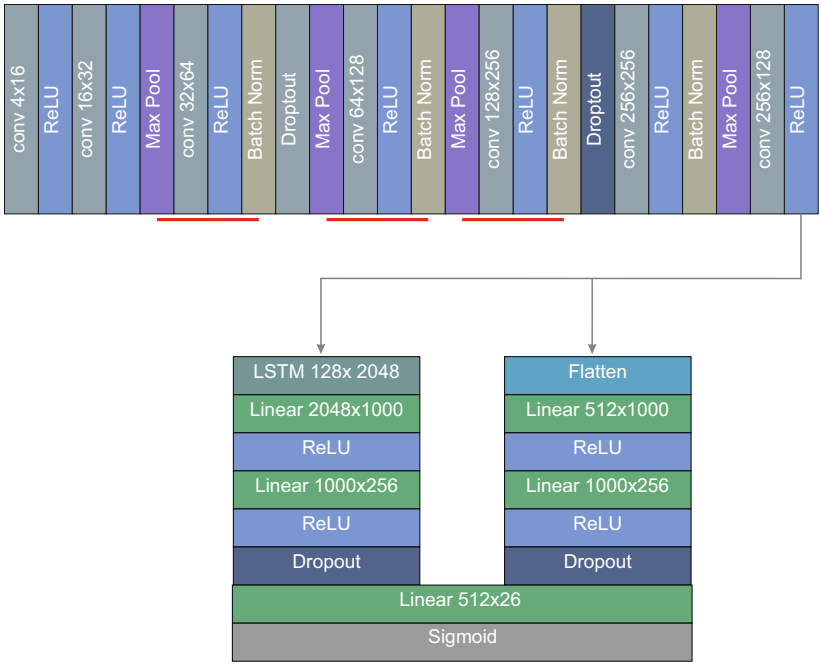

Figure 2. Architecture of the deep neural network. The red underlined layers are the key elements of the convolutional neural network.

performance in the official phase was improved, mainly by using hyperparameter tuning and a convolutional neural network. Hand-crafted features were discarded. Second, it can be seen that over-fitting occurred while tuning the hyperparameters since the validation scores are smaller than the training score. In Fig. 4 and Fig. 3 the confusion matrix for one fold of the 5-fold cross-validation (fold 4) and the 80/20 split are depicted. It can be seen that the neural network over-fits especially to sinus bradycardia (426177001). Additionally, the random dataset for 5-fold cross-validation is rather imbalanced since some classes do not occur at all. The dataset for the 80/20 split used for parameter tuning was constructed more thoroughly. From the confusion matrix of the 80/20 split, it can further be concluded that similar pathologies are confused more often, such as atrial flutter (164890007) and atrial fibrillation (164889003) or bundle branch block (6374002) and complete left/right bundle branch block (733534002 and 713427006). It can also be seen that waveform-related pathologies are detected with less accuracy ($< 10\%$), such as low QRS voltages. Finally, pathologies with respect to the rhythmic changes such as sinus tachycardia (427084000) are detected with rather high accuracies ($> 80\%$). The AUROC value for the 5-fold cross-validation was only around 0.5 while it was at approximately 0.856 for the 80/20 split.

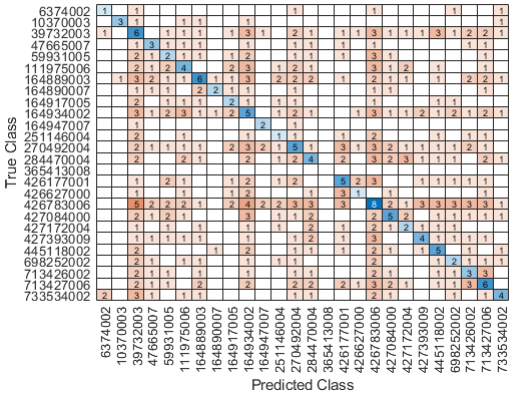

Figure 3. Confusion matrix of the 80/20 split. The colour map was enhanced by taking the rounded logarithm of confusion matrix plus 1, e.g. 6 corresponds to approximately 312. Information about the classes can be found in [3].

4. Discussion

The overall performance of the neural network must be improved to be accurate enough for clinical application. The performance especially for waveform-related pathologies could be improved by using additional hand-crafted features on the raw ECG data since information about the voltages is lost through normalisation. Another problem is over-fitting. Therefore, dropout layers as well as fewer

Table 2. Challenge scores for hyperparameter tuning, testing and validation. CS: Challenge System. The value for the unofficial phase are averaged over all leads.

	80/20	CS	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Unofficial phase	-	-0.3315	-0.441	-0.441	-0.441	-0.441	-0.441
Official phase	0.428	0.32	0.183	0.223	0.206	0.246	0.223

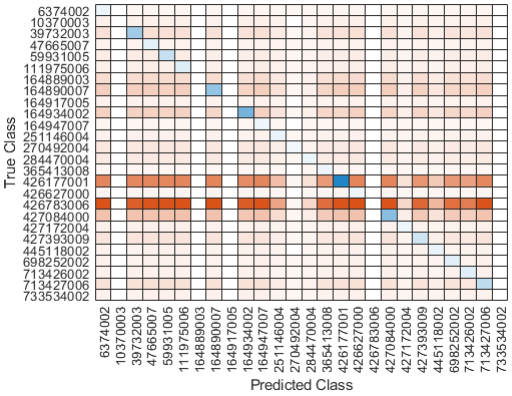

Figure 4. Confusion matrix of the third fold. Information about the classes can be found in [3].

stages of convolutional layers could be helpful. Since only two leads of the ECG recordings were fed into the neural network, it would also be conceivable to use more leads. However, investigation in the unofficial phase as well as the results of the other teams showed that the number of leads does not necessarily have a high influence on the classification results.

5. Conclusion

For the Computing in Cardiology Challenge 2021, a deep neural network for classifying ECG recordings with two, three, four, six and twelve leads was developed and tested. The network achieved scores of 0.428 (test dataset) and 0.32 (Challenge system). Improvements are necessary for usage in a clinical environment.

References

[1] Roth GA, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. The Lancet 2018; 392(10159):1736–1788. ISSN 0140-6736.

[2] Bickerton M, Pooler A. Misplaced ECG electrodes and the need for continuing training. British Journal of Cardiac Nursing 2019;14(3):123–132.

[3] Reyna M, Sadr N, Gu A, Perez Alday EA, Liu C, Seyedi S, Shah A, Clifford G. Will two do? Varying dimensions in electrocardiography: the physionet - computing in cardiology challenge 2021 (version 1.02). PhysioNet, 2021; URL https://doi.org/10.13026/jz9p-0m02.

[4] Goldberger A, Amaral L, Glass L, Hausdorff J. Ivanov P, R. M, Mietus J, Moody G, Peng C, Stanley HE. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation Online ;101(3):pp. e215–e220.

[5] Perez Alday EA, Gu A, Shah A, Liu C, Sharma A, Seyedi S, Bahrami Rad A, Reyna M, Clifford G. Classification of 12-lead ecgs: the physionet - computing in cardiology challenge 2020, Nov. 2020. URL http://doi.org/10.1088/1361-6579/abc960.

[6] Natarajan A, Chang Y, Mariani S, Rahman A, Boverman G, Vij S, Rubin J. A wide and deep transformer neural network for 12-lead ECG classification. In 2020 Computing in Cardiology Conference (CinC). Computing in Cardiology, December 2020; .

[7] Zhu Z, Wang H, Zhao T, Guo Y, Xu Z, Liu Z, Liu S, Lan X, Sun X, Feng M. Classification of cardiac abnormalities from ECG signals using SE-ResNet. In 2020 Computing in Cardiology Conference (CinC). Computing in Cardiology, December 2020; .

[8] Zhao Z, Fang H, Relton S, Yan R, Liu Y, Li Z, Qin J, Wong D. Adaptive lead weighted ResNet trained with different duration signals for classifying 12-lead ECGs. In 2020 Computing in Cardiology Conference (CinC). Computing in Cardiology, December 2020; .

[9] Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, Schölzel C, Chen SHA. Neurokit2: A python toolbox for neurophysiological signal processing. Behavior Research Methods Feb 2021;ISSN 1554-3528.

[10] Bishop CM. Pattern Recognition and Machine Learning (Information Science and Statistics). Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738.

Address for correspondence:

Onno Linschmann

Lehrstuhl für Medizinische Informationstechnik, Helmholtz-Institut, RWTH Aachen, Pauwelsstr. 20, D-52074, Aachen, Germany

linschmann@hia.rwth-aachen.de