Multi-label Classification of multi-lead ECGs using Convolutional Neural Networks with double soft F1-loss

Bjørn-Jostein Singstad [1], Eraraya Morenzo Muten [2] & Pål Haugar Brekke [1]

[1] ProCardio Center for Innovation, Dept of Cardiology, Oslo University Hospital, Rikshospitalet, Oslo, Norway.
[2] Institut Teknologi Bandung, Bandung, Indonesia

E-mail: b.j.singstad@fys.uio.no

January 2022

Abstract. The electrocardiogram (ECG) is an almost universally accessible diagnostic tool for heart disease. An ECG is measured by an electrocardiograph, and today's electrocardiographs use built-in software to interpret the ECGs automatically after recording. However, these algorithms show limited performance, in terms of accuracy, and the clinicians often have to read over the ECG. Manual interpretation of the ECG can be time-consuming and require certain skills. There is clearly a need for better interpretation algorithms. On the other side, algorithms based on artificial intelligence have shown promising performance in many fields, including ECG interpretation, over the last years.

This study is a part of a challenge organized by PhysioNet and Computing in Cardiology, and here we developed and trained a convolutional neural network on the raw 12-lead ECG signals from a development set comprising of more than 80000 patients. 30 different cardiac diseases were used as ground truth for the supervised data-driven model. Furthermore, we compared the classification performance using different subsets of the 12 standard ECG leads and also compared the performance using different sampling rates on the input ECG signals. Finally, we deployed the best model using a Docker container and classified 6630 and 36272 ECGs in a hidden validation and test set.

The best performing model was an Inception model using double soft F1-loss as loss function and ECGs downsampled to 75 Hz. The model achieved a PhysioNet Challenge score of 0.548 ± 0.012 using 10 fold stratified cross-validation on the development set, ... on the hidden validation set and ... on the hidden test set. These findings contribute to our understanding of how we can train generalizable ECG classifiers which also perform well on unseen data.

Introduction

Cardiovascular diseases (CVD) are one of the leading causes of death in the world. Numbers from World Health Organization estimates that 17.9 million people died from

CVD in 2016 which represented 31% of all global deaths that year [1]. Early detection of patients with a risk of CVD could potentially reduce the severeness of the disease and also decrease the number of persons who die from CVD. Electrocardiography is a method already in use to detect cardiac-related pathology and this method has the potential to detect even more pathologies [2]. The electrocardiograph is non-invasive and relatively easy to use, compared to methods like echocardiogram and MRI, which makes it a convenient diagnostic tool. As an example of how widely the electrocardiograph is used, National Ambulatory Medical Care reported that 40 million electrocardiograms (ECG) were recorded in the USA in 2015 [3].

ECG is the result of a recording performed by an electrocardiograph. It measures the electrical activity of the heart by recording the voltage potential from electrodes placed on the patient's skin. This technique has enabled clinicians to interpret, diagnose and prognosticate heart disease since the beginning of the 20th century [4]. While the setup of an ECG recording is quite easy, one of the challenges is that the ECG can be difficult to interpret correctly. Correct interpretation can be time-consuming and require a high degree of expertise [5].

In the 1950s it became possible to convert the analog ECG signals to digital format and this lead to the development of digital interpretation algorithms in the 1960s [6]. Today, most of the modern and clinically used electrocardiographs are equipped with built-in interpretation software. The software interprets the ECG and prints interpretive texts that may indicate different pathologies. Studies show that there are several limitations to the automatic interpretation algorithms [2, 6]. The errors, caused by the automatic interpretation algorithms, imply that doctors or cardiologists have to read over the ECGs to ensure the diagnose is correct. This is time-consuming for the doctors and also leads to high interpretation variability. Thus, there is a need for developing a better ECG interpretation algorithm, since this may lead to less time-consuming interpretation for the doctors and less variability in the interpretation.

In the past decades, several new important trends have converged and may potentially be ushering in a new age with significance to ECG interpretation. Firstly, ECGs are now increasingly stored in digital format, allowing computerized analysis of massive data sets. Secondly, personal sensors such as training monitors and smartwatches (e.g., Apple Watch, Samsung Galaxy Watch) now include simple ECG recording abilities, further expanding access to ECGs and the range of people studied. Finally, artificial intelligence (AI) or more specifically deep learning (DL) has shown remarkable abilities in detecting subtle patterns in large sets of signals data that are not apparent to the human interpreter.

A considerable amount of literature has been published on heartbeat classification using ML and DL [7], single lead sequences [8] and 2-lead sequence classification [9] over the last ten years. In most recent years there has been an increasing focus on 12-lead ECG classification and some recent studies have shown that machine learning has great potential in this field [10, 11, 12, 13, 14]. On the other hand, many of the datasets used have either been small and homogeneous [15] or not accessible to everyone.

Despite the high popularity of AI in the last decade, it has also received a lot of criticism for its lack of explainable decisions. This applies in particular to DL which is, by many [16], considered to be a black box. Explainable decisions are especially important in medicine where lives are at stake and the doctors are responsible for the decision. The need for transparent and explainable decisions has led to a new and emerging field in AI and is called Explainable AI [17], where interpretable model-agnostic models and game-theoretic ideas such as LIME and Shapley values play an important role [18, 19].

This paper is a result of a research challenge organized by PhysioNet and Computing in Cardiology. The objective is to develop models to classify a large variety of cardiovascular diagnoses from ECGs.

Previous work

This paper summarizes our contribution to the PhysioNet/Computing in Cardiology Challenge (now called George Moody Challenge) 2020 and 2021. A lot of the improvements that are presented in the final model are a result of trial and error, but also learning from other participants at the Computing in Cardiology conference in 2020 and 2021.

In our first paper [20], describing our contribution to PhysioNet Challenge 2020, we used the 3 best convolutional neural networks from Fawas HI et al 2019 [21] as a starting point for our classifier. The objective of the PhysioNet Challenge 2020 was to classify 27 cardiac abnormalities based on 12-lead ECGs. The organizers of the challenge provided a training dataset comprising 43101 annotated ECGs.

In our second paper [22] we described our contribution to PhysioNet Challenge 2021 where we developed an algorithm using convolutional neural networks in classifier chains. Furthermore, we compared the performance of the classifier using 2, 3, 4, 6 and 12-lead ECGs as input. In the 2021 challenge, the organizers expanded the data set from the previous year with approximately 45000 new 12-lead ECGs, giving a data set with a total of 88253 ECGs [23].

Methods and materials

Data

A total of 131155 12-lead ECGs from 9 different databases were used in this study. 88253 ECGs were used as the training set (hereby called development set), 6630 ECGs were used as validation set (hereby called hidden validation set) and 36272 ECGs were used as test set (hereby called hidden test set) [23, 24, 25, 26, 27, 28, 29]. The hidden validation and test set were withheld by the organizers of the PhysioNet Challenge, while the development set was made publicly available. The development data contained ECGs from 7 out of the total 9 sources. This means that the model was not only tested and validated on unseen data, but also on 2 new cohorts. Testing on new cohorts can potentially reveal bias acquired by a data-driven model.

Each ECG was stored in a .mat file and had a corresponding .hea file containing metadata such as the ECG recording length, sample frequency, the patient's age, gender and diagnosis. In the development set, there was a total of 133 different diagnoses, but the objective of this challenge was to classify 30 of them. Each patient could have more than one of the 30 diagnoses and there were over 3000 different combinations of diagnoses present in the development set.

In the development set, we excluded all patients with an ECG not equal to a recording length of 10 seconds. Figure 1 shows how we excluded 6926 out of 88253 patients and ended up using 81327 ECGs.

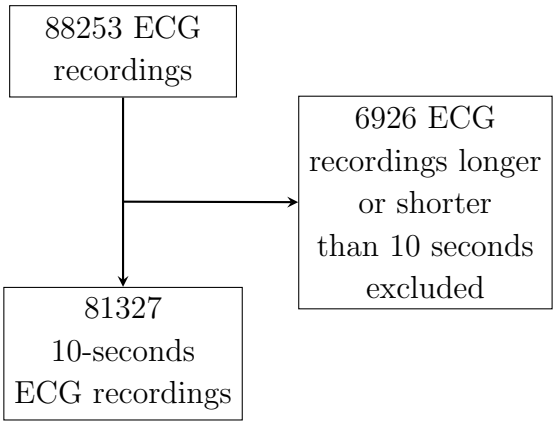


Figure 1: Patients with an ECG recording shorter or longer than 10 seconds were excluded from the development set. 6926 ECGs were excluded and the remaining 81327 ECGs were used to develop and train the model.

After excluding the ECG recordings as shown in Figure 1, the prevalence of the remaining scored diagnoses in the data set is shown in Table 1.

Diagnoses	Prevalence	Diagnoses	Prevalence
1st Degree AV block	3291	Premature Atrial Contraction	2827
Atrial Fibrillation	5062	Premature Ventricular Contractions	1259
Atrial Flutter	8509	Prolonged PR Interval	393
Bradycardia	267	Prolonged QT Interval	2152
Bundle Branch Block	511	Q Wave Abnormal	2261
Complete Left Bundle Branch Block	218	Right Axis Deviation	1482
Complete Right Bundle Branch Block	2015	Right Bundle Branch Block	2331
Incomplete Right Bundle Branch Block	2306	Sinus Arrhythmia	4176
Left Anterior Fascicular Block	2665	Sinus Bradycardia	19303
Left Axis Deviation	7952	Sinus Rhythm	30426
Left Bundle Branch Block	1260	Sinus Tachycardia	10261
Low QRS Voltage	1765	Supraventricular Premature Beats	267
Nonspecific Intraventricular Conduction Disorder	2101	T Wave Abnormal	12673
Pacing Rhythm	1694	T Wave Inversion	4340
Poor R Wave Progression	656	Ventricular Premature Beats	731

Table 1: The prevalence of the 30 scored diagnoses in the development set after excluding patients with an ECG recording shorter or longer than 10 seconds.

Preprocessing development data

To feed the diagnoses as labels to our data-driven model during training, we one-hot encoded the diagnoses, such that each ECG recording had a corresponding 30-bit long array of ones or zeros. A binary one means that the patient has the given diagnose and zero means that the patient does not have the diagnose.

More than 85% of all ECGs in the development set were originally sampled at 500 Hz. In this study, we show how downsampling the ECG may affect the performance in terms of selected scoring metrics.

CNN architectures

In this study, we employed an Inception model architecture as shown in Figure 2. The input to this model was an array, representing the raw ECG. The array containing the ECG signal can be denoted as:

$$\text{number of leads} \times (\text{ ECG length} \cdot \text{sampling frequency})$$

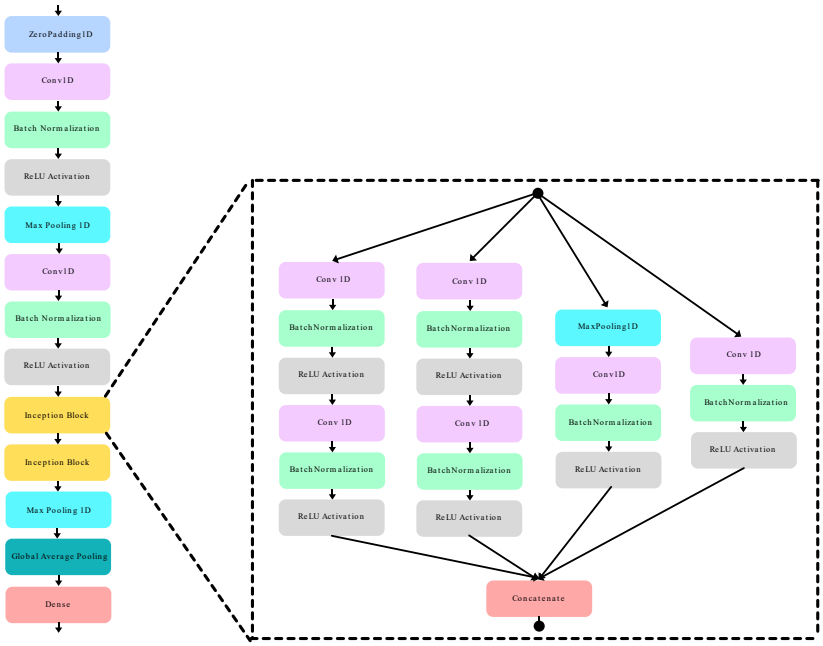


Figure 2: Inception model architecture. Each block represents a mathematical operation in the Convolutional Neural Network. The blocks inside the dashed lines represent an Inception block.

The output layer of the model corresponds to the 30 scored diagnoses in the challenge. A Sigmoid activation was used in the final layer, giving a continuous number between 0 and 1 for each of the 30 diagnoses. A number equal or close to 1 means that the model predict the diagnose to be true for the given patient. A number equal or close to 0 means that the model predicts the diagnose to not be true for the given patient.

Loss function

To deal with the imbalance in our data set and the fact that we did not know the distribution of diagnoses in the hidden validation and test set, we implemented an F1-score inspired loss function, called double soft F1-loss ($\mathcal{L}_{F_1}$), to the Inception model [30, 31]. The purpose of this loss function is to treat each diagnosis with equal importance even though some diagnoses are more represented than others. Equation 1 shows how double soft F1-loss is calculated. The small number ($+10^{-16}$) is added to the denominator to prevent the function to divide by zero.

$$\tilde{tp} = \sum_{i=1}^{n} \hat{y}_i \cdot y_i \quad , \quad \tilde{fp} = \sum_{i=1}^{n} \hat{y}_i \cdot (1 - y_i)$$

$$\tilde{fn} = \sum_{i=1}^{n} (1 - \hat{y}_i) \cdot (1 - y_i) \quad , \quad \tilde{tn} = \sum_{i=1}^{n} (1 - \hat{y}_i) \cdot (1 - y_i)$$

$$\mathcal{L}_{F_1} = 1 - \frac{2 \cdot \tilde{tp}}{2 \cdot \tilde{tp} + \tilde{fp} + \tilde{fn} + 10^{-16}} \tag{1}$$

The double soft F1-loss was used to compute the error of the prediction and an Adam optimizer was used to compute the gradients used to backpropagate the error and update the weights in the Inception model [32, 33]. The learning rate, used by the Adam optimizer, was initially set equal to 0.001 and from previous studies, we found batch size = 30 to be close to optimal.

Model selection

The best model and model parameters were selected based on internal training and validation on the development set using 10-fold stratified cross-validation. The development data was stratified based on the prevalence of the diagnoses and this resulted in a similar distribution of diagnosis in both the train and validation fold. To ensure reproducible splits of the development data we set random state equal to 42 [34].

The learning rate, which was initially set to 0.001, was decreased by a factor of 10 during training. In the model selection phase, during 10-folded stratified cross-validation on the development set, double soft F1-loss on the validation split was used to control the decrease of learning rate. Each time the double soft F1-loss did not improve, the learning rate was reduced. The only exception was if the learning rate just got reduced, it had to run two more epochs before the learning rate could be reduced again (cool down factor = 2).

The best model setting was finally decided based on the performance measured in terms of the PhysioNet Challenge score [24, 23]. In addition, we also scored the model using F1-score, the area under the curve (AUC) of the receiver operating characteristic (ROC) curve and average accuracy across all classes (hereby just referred to as accuracy).

Equation 2 shows how accuracy is calculated by comparing the true label (y) and the predicted label ($\hat{y}$) for each sample, n_s and then finding the average accuracy for each class c and finally taking the average across all classes, n_c.

$$accuracy(y, \hat{y}) = \frac{1}{n_c} \sum_{i=0}^{n_c-1} \frac{1}{n_s} \sum_{j=0}^{n_s-1} 1 \cdot (\hat{y}_{ij} = y_{ij}) \tag{2}$$

Model deployment

To obtain a score on the hidden validation and test set we submitted the model training code as a Docker file to the organizers of the PhysioNet Challenge. The model settings that gave the best performance in the model selection stage were used in the submitted training code. The learning rate schedule in the submitted code was programmed to imitate the best learning rate schedule found during model selection and the model was trained on the whole development set, except for the ECGs with a recording length unequal to 10 seconds. Furthermore, the model was applied to the hidden validation and test set.

The recording length of the ECG signals in the hidden validation and test set were not known by the participants of the PhysioNet Challenge and to be able to classify longer or shorter signal than 10 seconds we zero-padded signals shorter than 10 seconds. If time t represents the recording length in seconds and the sample frequency is represented by f_s, then zero-padding was done by adding a tail of $(f_s \cdot 10) - (f_s \cdot t)$ zeros to any recording of length $< (f_s \cdot 10)$.

Explainability

To highlight the feature importance in the ECG, determined by the Inception model, we employed local interpretive model-agnostic explanation (LIME) models. These interpretive models were applied to the input and output of the Inception model, fitted to the 12-lead ECGs, after training on the first out of 10 cross-validation folds. For each of the 30 diagnoses classified by the Inception model, one LIME model was trained to reveal the feature importance of the input ECGs. 2000 ECGs were used as training data for the LIME model. 1000 of these ECGs were labeled with the diagnose to explain and the other 1000 were not labeled with the diagnose to explain. All of these 2000 ECGs were randomly selected from the ECGs in the training split. Furthermore, the LIME models were set to explain the activation of all samples in all leads from the ECG. Only ECGs from the validation splits were explained to ensure that the ECG were previously unseen by both the Inception and LIME model.

Results

The sampling frequency of the ECG had a major impact on the classification result. In Figure 3 we show the PhysioNet Challenge score achieved on the validation split during 10-folded stratified cross-validation on the development set, using 12-lead ECG sampled from 25 Hz to 500 Hz.

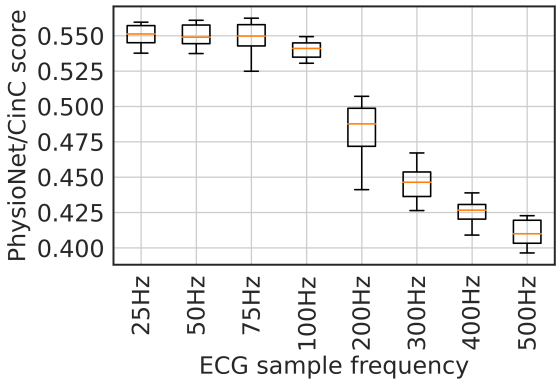


Figure 3: Classification results performed on the validation set during 10-folded cross-validation by the Inception model with F1-loss. The y-axis shows the PhysioNet Challenge score and the x-axis shows the various sampling frequencies of the input 12-lead ECG.

From Figure 3 we see a clear increase in the PhysioNet Challenge score when the sampling frequency is lowered from 500 Hz. The increasing performance seems to saturate around 75 Hz and because of that, we choose to use ECGs downsampled to 75 Hz in our prediction on the hidden validation and test set. The performance of the Inception model on the development set, the hidden validation and test set measured in terms of PhysioNet Challenge score, accuracy, AUROC and F1-score are summarized in Table 2.

Metric	Development set	Hidden validation set	Hidden test set
PhysioNet Challenge score	0.548 ± 0.012	-	-
Accuracy	0.954 ± 0.002	-	-
AUROC	0.832 ± 0.019	-	-
F1-score	0.420 ± 0.017	-	-

Table 2: PhysioNet Challenge score, accuracy, AUROC and F1-score achieved by the Inception model on the development set (10-fold cross-validation), the hidden validation set and the hidden test set, using 12-lead ECGs as input.

By the objective of PhysioNet Challenge 2021, we trained the Inception model on different subsets of the 12 lead ECG. Besides the 12-lead model, we trained a 6-lead model using lead I, II, III, aVR, aVF and aVL, a 4-lead model using lead I, II, III and

V2, a 3-lead model using lead I, II and V2, and a 2-lead model using lead I and II. Figure 4 compares the PhysioNet Challenge score from doing cross-validation on the development set using the five different lead combinations.

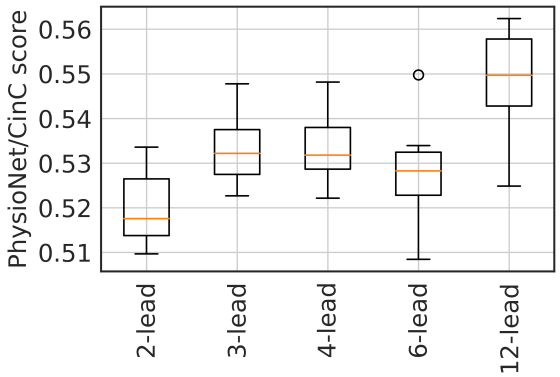


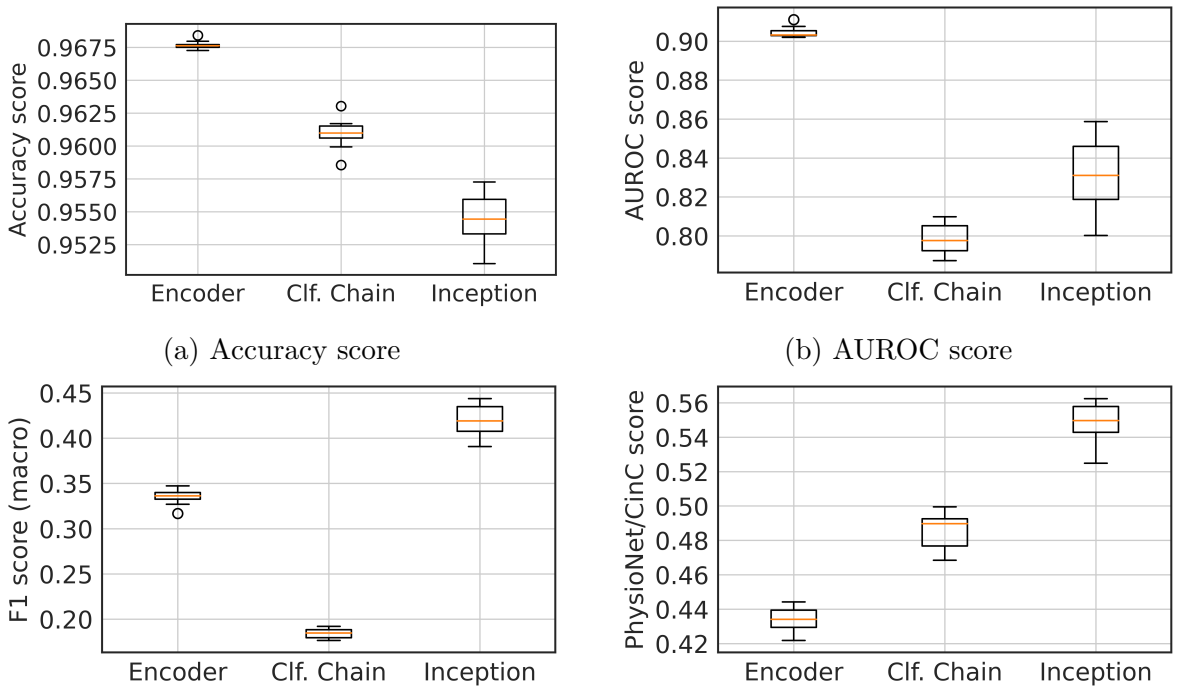
Figure 4: PhysioNet Challenge score achieved by Inception models, using 2, 3, 4, 6 and 12 leads, during 10-folded stratified cross-validation on the development set.

In Table 3 we compare the PhysioNet Challenge score achieved by the Inception model on the development set, the hidden validation set and the hidden test set, using 12, 6, 4, 3 and 2-lead ECGs as input. The scores on the development set are achieved by doing 10-fold stratified cross-validation.

Leads	Development set	Hidden validation set	Hidden test set
12-lead	0.548 ± 0.012	-	-
6-lead	0.528 ± 0.011	-	-
4-lead	0.533 ± 0.009	-	-
3-lead	0.533 ± 0.008	-	-
2-lead	0.520 ± 0.008	-	-

Table 3: PhysioNet Challenge score achieved by the Inception model using 12, 6, 4, 3 and 2 leads, on the development set (10-fold cross-validation), the hidden validation set and the hidden test set.

In Figure 5 we compare the Inception model, proposed in this study, with the Encoder model and the classifier chain model developed in PhysioNet Challenge 2020 and 2021. In Figures 5a and 5b, we observe that the Encoder model achieved a better score than the classifier chain and the Inception model in terms of accuracy and AUROC. On the other hand, we see from Figures 5c and 5d that the Inception model performs better than the classifier chain and the Encoder model in terms of F1-score and PhysioNet Challenge score.



(a) Accuracy score

(b) AUROC score

(c) F1-score

(d) PhysioNet Challenge score

Figure 5: Accuracy, AUROC, F1 and PhysioNet Challenge score achieved on the development set by the Encoder model, classifier chain and the Inception model using 12-lead ECGs. Each box represents the classification performance achieved using 10-fold cross-validation on the development set.

Explainability

All ECGs in the first validation split during 10-fold stratified cross-validation were classified by the Inception model and all true diagnoses were explained by the LIME model. Figure 6 show the aVL lead from a 12-lead ECG recorded from a patient with atrial fibrillation. The vertical red lines show the parts of the ECG that contributed most towards the correct prediction by the Inception model. The aVL lead was selected because it got the highest activation compared with the other leads in this specific case. Examples of activation maps for all 30 scored diagnoses are added to the appendix.

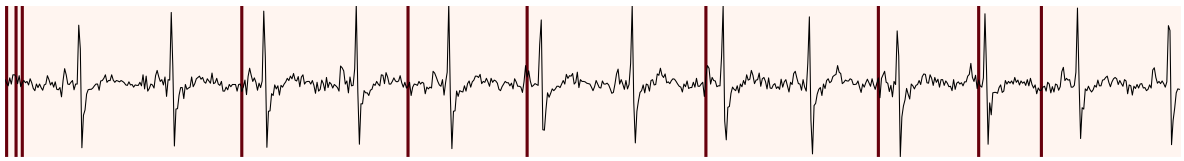


Figure 6: A patient correctly classified with atrial fibrillation with a probability of 0.998 seen from lead aVL. The red vertical lines show the features in the ECG that contributed most towards the prediction, according to the local interpretable model-agnostic explanation (LIME) model.

Discussion

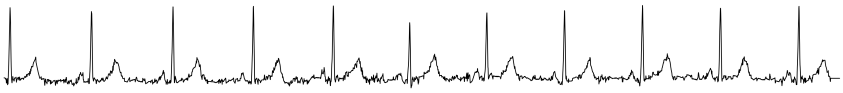
The results of this study show that the proposed Inception model performed better, in terms of F1- and PhysioNet Challenge score, than the Encoder and classifier chain, developed in the official phase of PhysioNet Challenge 2020 and 2021. On the other hand, the Encoder model performed better in terms of accuracy and AUROC. This may be caused by the imbalance in the data set. Its well known that F1-score weights false negatives and false positives more than the accuracy score. Therefore the results from Figure 5a may indicate that the Encoder model are generally better in classifying true positives and true negatives than the Inception model. On the other hand, the results in Figure 5c may indicate that the Inception model are better to limit the amount of false negatives and false positives.

The most surprising finding was the drastic improvement in classification performance when downsampling the input ECG from 500 Hz towards 25 Hz, as seen in Figure 3. At sample frequencies below 100 Hz, the improvement in classification performance seems to reach a plateau and no significant differences were seen while using ECGs sampled at 25, 50 and 75 Hz as input to the Inception model. The increase in classification performance using downsampled signals could be a bit counter-intuitive since one would expect the ECG to lose a lot of subtle, but important features. A possible explanation for this is that there is an ideal ratio between convolution kernel size and the features in the ECGs, such as P-waves T-waves and QRS complex. However, we also did some experiments by increasing the kernel size, but this did not give the same improvement in classification performance as lowering the ECG sampling frequency. However, this needs more research to conclude.

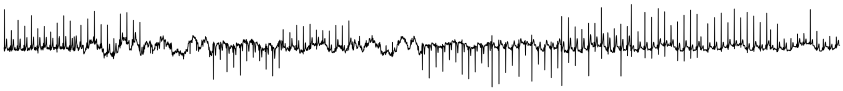
As expected, the Inception trained on all 12-leads performed significantly better than the models trained on 6 leads or less. On the other hand, and contrary to expectations, the 6-lead model performed slightly worse than the 3- and 4-lead model (but not significantly). A possible explanation for this result might be that the 3- and 4-lead model use a combination of standard leads and precordial leads, The 6-lead model on the other hand, only use the standard leads (I, II and III) and the augmented leads (aVR, aVL, aVF), which is mathematically derived from the standard leads and therefore might not provide any new information for an ML model.

An issue we detected in the algorithms developed and used in the submission to PhysioNet challenge 2020 and 2021 was that did an unintentional modification of the ECG signal using NumPy reshape [35]. In the submission related to the current paper, we have corrected this and found NumPy move_axis to work. Figure 7a shows Lead-I from the original ECG, Figure 7b shows the ECG pre-processed using NumPy reshape method and Figure 7c shows Lead-I from the ECG pre-processed using NumPy move_axis.

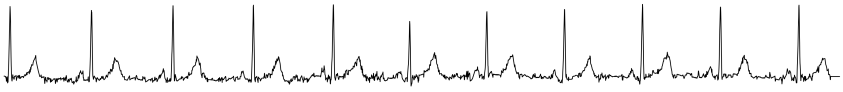
To our surprise, we found that a model using the NumPy reshape method (Figure 7b) performed better than using the NumPy move_axis method (Figure 7c). A closer look at the result of the NumPy reshape method showed us that the ECGs were



(a) Original ECG



(b) ECG processed using NumPy reshape method



(c) ECG processed using NumPy move_axis

Figure 7: Different pre-processing of the ECG

downsampled in such a way that the signals from all 12 leads fitted into the original sample length of one lead and also copied such that all 12 leads of the resulting array were the same. In this case, the ECG was downsampled from 500Hz to 41.7Hz. When downsampling the original ECG and then using NumPy move_axis we saw an increasing performance compared to using 500Hz and also exceeded the performance of the signal processed using NumPy reshape.

Augmentation has shown promising performance in various image classification tasks [36]. In this study, we hypothesized that augmentation techniques might improve the performance of ECG sequence classification as well. The techniques we tested were to add random noise to the signal and simulation of baseline wander.

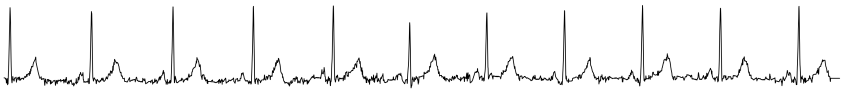
The random noise was induced by adding a random number (N) between $\pm$ the standard deviation (σ) of all values in an ECG recording, shown in Equation 3.

$$y_{i_n} = y_i + N(-\sigma, \sigma) \tag{3}$$

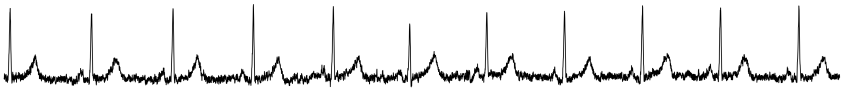
Baseline wander was induced to the signal by adding a cosine wave from 0 to 2π and shifting the cosine wave randomly between 0 and 2π. The amplitude of the signal was randomly set by multiplying a random number (N) between $\pm$ the standard deviation (σ) of all values in an ECG recording, shown in Equation 4.

$$y_{i_{bw}} = cos(2\pi \frac{y_i}{n} + N(0, 2\pi)) \cdot N(-\sigma, \sigma),\ i = 0, 1, 2, 3...(n-1) \tag{4}$$

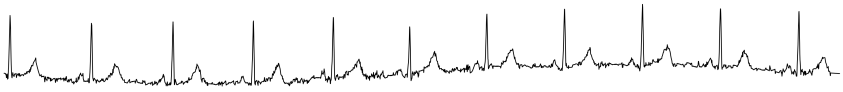
Figure 8a show an example of an unprocessed ECG and Figure 8b shows the same ECG with added random noise using the method described in Equation 3. Figure 8c shows the ECG in Figure 8a with simulated baseline wander as described in Equation 4.



(a) Unprocessed ECG



(b) ECG with added random noise



(c) ECG with simulated baseline wander

Figure 8: Comparing an unprocessed ECG (a) with the same ECG with added noise (b) and simulated baseline wander (c)

Conclusion

The primary aim of the current study was to train an ECG classifier to achieve the highest possible PhysioNet Challenge score on the hidden validation and test set. This study has identified that double soft F1-loss and reducing sample frequency of the ECG to 75Hz increased the performance in terms of F1-score and PhysioNet Challenge score.

The second aim of this study was to investigate the effects of using different subsets of the 12 standard leads as input to the model. We found that the 12-lead model performed significantly better than the other models, but on the other hand, the 3 and 4 lead model also showed promising performance. The combination of at least one precordial and one standard lead seems to give yield better performance than only using standard leads.

Code availability

The complete source code described in this paper is openly available at GitHub (`https://github.com/CardiOUS/PhysioNetChallenge2020-2021`) under a free software license (CC-BY 4.0).

Acknowledgment

This study is entirely based on open source ECG data and we will therefore acknowledge the publisher of the ECG databases we have used, and PhysioNet for facilitating a storage and sharing platform of such data. Finally, we also want to thank Computing in Cardiology and the organizers of the PhysioNet Challenge for organizing a great challenge.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Norwegian Research Council (grant number: 309762 - ProCardio).

References

[1] Cardiovascular diseases (CVDs). URL `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`.

[2] Schläpfer, J. & Wellens, H. J. Computer-Interpreted Electrocardiograms. *Journal of the American College of Cardiology* **70**, 1183–1192 (2017).

[3] of Health and Human Services, U. D. National Ambulatory Medical Care Survey: 2015 State and National Summary Tables (2015). URL `https://www.cdc.gov/nchs/data/ahcd/namcs_summary/2015_namcs_web_tables.pdf`.

[4] Barold, S. S. Willem Einthoven and the birth of clinical electrocardiography a hundred years ago. *Cardiac Electrophysiology Review* **7**, 99–104 (2003).

[5] Bickerton, M. & Pooler, A. Misplaced ECG Electrodes and the Need for Continuing Training. *British Journal of Cardiac Nursing* **14**, 123–132 (2019).

[6] Smulyan, H. The Computerized ECG: Friend and Foe. *The American Journal of Medicine* **132**, 153–160 (2019).

[7] Annam, J. R., Kalyanapu, S., Ch., S., Somala, J. & Raju, S. B. Classification of ECG Heartbeat Arrhythmia: A Review. *Procedia Computer Science* **171**, 679–688 (2020). URL `http://www.sciencedirect.com/science/article/pii/S1877050920310425`.

[8] Mathews, S. M., Kambhamettu, C. & Barner, K. E. A novel application of deep learning for single-lead ECG classification. *Computers in Biology and Medicine* **99**, 53–62 (2018).

[9] Liu, S.-H., Cheng, D.-C. & Lin, C.-M. Arrhythmia Identification with Two-Lead Electrocardiograms Using Artificial Neural Networks and Support Vector Machines for a Portable ECG Monitor System. *Sensors (Basel, Switzerland)* **13**, 813–828 (2013). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3574706/`.

[10] Ribeiro, A. H. *et al.* Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications* **11**, 1760 (2020). URL `https://www.nature.com/articles/s41467-020-15432-4`. Number: 1 Publisher: Nature Publishing Group.

[11] Yao, Q., Wang, R., Fan, X., Liu, J. & Li, Y. Multi-class Arrhythmia detection from 12-lead varied-length ECG using Attention-based Time-Incremental Convolutional Neural Network. *Information Fusion* **53**, 174–182 (2020). URL `http://www.sciencedirect.com/science/article/pii/S1566253518307632`.

[12] Li, D., Wu, H., Zhao, J., Tao, Y. & Fu, J. Automatic Classification System of Arrhythmias Using 12-Lead ECGs with a Deep Neural Network Based on an Attention Mechanism. *Symmetry* **12**, 1827 (2020). URL `https://www.mdpi.com/2073-8994/12/11/1827`. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[13] Chen, T.-M., Huang, C.-H., Shih, E. S., Hu, Y.-F. & Hwang, M.-J. Detection and Classification of Cardiac Arrhythmias by a Challenge-Best Deep Learning Neural Network

Model. *iScience* **23**, 100886 (2020). URL https://linkinghub.elsevier.com/retrieve/pii/S2589004220300705.

[14] Attia, Z. I., Harmon, D. M., Behr, E. R. & Friedman, P. A. Application of artificial intelligence to the electrocardiogram. *European Heart Journal* **42**, 4717–4730 (2021). URL https://doi.org/10.1093/eurheartj/ehab649.

[15] Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. URL https://physionetchallenges.github.io/2020/.

[16] Castelvecchi, D. Can we open the black box of AI? *Nature News* **538**, 20 (2016). URL http://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731. Section: News Feature.

[17] Rai, A. Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science* **48**, 137–141 (2020). URL https://doi.org/10.1007/s11747-019-00710-5.

[18] Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]* (2017). URL http://arxiv.org/abs/1705.07874. ArXiv: 1705.07874.

[19] Ribeiro, M. T., Singh, S. & Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]* (2016). URL http://arxiv.org/abs/1602.04938. ArXiv: 1602.04938.

[20] Singstad, B.-J. & Tronstad, C. Convolutional Neural Network and Rule-Based Algorithms for Classifying 12-lead ECGs. 1–4 (2020).

[21] Fawaz, H. I. & et al. Deep Learning for Time Series Classification: A Review. *Data Mining and Knowledge Discovery* **33**, 917–963 (2019).

[22] Singstad, B.-J. & Brekke, P. H. Multi-label ECG classification using Convolutional Neural Networks in a Classifier Chain. *Computing in Cardiology* (2021). In Review.

[23] Reyna, M. A. & et al. Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2020 **48**, 4.

[24] Alday, E. A. P. *et al.* Classification of 12-lead ECGs: the PhysioNet/ Computing in Cardiology Challenge 2020. *Physiol. Meas. 2020 Nov 11* (2020).

[25] Liu, F. & et al. An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. *Journal of Medical Imaging and Health Informatics* **8**, 1368–1373 (2018).

[26] Tihonenko, V. & et al. St.-Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database (2007).

[27] Wagner, P. & et al. PTB-XL, a Large Publicly Available Electrocardiography Dataset. *Scientific Data* **7**, 154 (2020).

[28] Bousseljot, R., Kreiseler, D. & Schnabel, A. Nutzung der EKG-Signaldatenbank Cardiodat der PTB über das Internet. *Biomedizinische Technik/Biomedical Engineering* 317–318 (2009).

[29] Zheng, J. & et al. Optimal Multi-Stage Arrhythmia Classification Approach. *Scientific Reports* **10**, 2898 (2020).

[30] Bénédict, G., Koops, V., Odijk, D. & de Rijke, M. sigmoidF1: A Smooth F1 Score Surrogate Loss for Multilabel Classification. *arXiv:2108.10566 [cs, stat]* (2021). URL http://arxiv.org/abs/2108.10566. ArXiv: 2108.10566.

[31] van Lohuizen, Q. *Training Deep Neural Networks with Soft Loss for Strong Gravitational Lens Detection*. Master's thesis, University of Groningen.

[32] Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (2017). URL http://arxiv.org/abs/1412.6980. ArXiv: 1412.6980.

[33] Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986). URL https://www.nature.com/articles/323533a0. Number: 6088 Publisher: Nature Publishing Group.

[34] Pedregosa, F. & et al. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[35] Van Der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering* **13**, 22 (2011). Publisher: IEEE Computer Society.

[36] Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* **6**, 60 (2019). URL `https://doi.org/10.1186/s40537-019-0197-0`.