Improving Machine Learning Education During the COVID-Pandemic Using Past Computing in Cardiology Challenges

Maurice Rohr[1], Filip Plešinger[2], Veronika Bulkova[3], Christoph Hoog Antink[1]

[1]KIS*MED – AI Systems in Medicine,
Technische Universität Darmstadt, Darmstadt, Germany,
[2]Institute of Scientific Instruments of the Czech Academy of Sciences, Brno, Czechia,
[3]Medical Data Transfer, s.r.o., Brno, Czechia

Abstract

During the COVID-Pandemic and the lockdown of universities, the need for stimulating, novel teaching methods was high, as most students were confined to their homes. For over 20 years, the annual PhysioNet / CinC Challenges not only lead to technological advances for specific problems, they have also proven repeatedly to be of immense value from an educational point of view. In this paper, we report results from the class "Artificial Intelligence in Medicine Challenge", which was implemented as an online project seminar at TU Darmstadt and which was heavily inspired by the PhysioNet / CinC Challenge 2017 "AF Classification from a Short Single Lead ECG Recording". In particular, we show numeric results of the developed approaches on several datasets, highlight themes commonly observed among participants, and report the results from student evaluation. Several teams were able to implement approaches based on state-of-the-art algorithms achieving F_1 scores above / close to 90 % on a hidden test-set of Holter recordings. Moreover, the self-assessment of the students reported a notable increase in machine learning knowledge.

1. Introduction

In the wake of the COVID-Pandemic and the lockdown of universities, novel teaching concepts combining online teaching, experimenting, and self-learning with a motivational environment are needed. Challenge-based gamification aspects such as clear tasks, leaderboards, instant feedback, and points showed promising results in improving statistics and engineering education in recent studies [1,2]. In particular, leaderboards offer a system of self-feedback and goal-setting to students [3]. It comes to no surprise that the annual PhysioNet / CinC Challenges not only lead to technological advances for specific problems, they also repeatedly proved to be of immense educational value for participants.

Thus, when faced with the task of designing the project seminar "Artificial Intelligence in Medicine Challenge" as part of the electrical / biomedical engineering curriculum at TU Darmstadt, the CinC Challenge 2017 "AF Classification from a Short Single Lead ECG Recording" served as prototype. Since research [4] has shown that gamification alone may not be the "holy grail of education", we tried to counteract potential negative effects. In particular, we offered weekly sessions to discuss ideas, checked problems with the teams, and had a voluntary mid-semester presentation about intermediate results. Most importantly, we emphasized from the start that a good score and ranking was neither required nor sufficient for a good grade but that the originality of approaches and a good analysis were paramount to us. Thus, the award for the winning team consisted of a certificate and a small price but was not tied to the grade.

2. Methods

As in the original CinC challenge [5], the goal of the course was to detect atrial fibrillation in one-lead ECGs. The ECG recordings provided could have various labels and so the secondary goal was to solve the multi-label problem consisting of the classes normal sinus rhythm ('N'), atrial fibrillation ('A'), other rhythm ('O') or noisy recording ('~'). In total, three datasets were used:

A From the official CinC 2017 training set [5], 6000 randomly selected samples were handed out as training set and 2528 samples were held back as test set. Four classes ['N','A','O','~'] are available. The recordings are short single-channel ECGs.

B A "quasi hidden" test set was sampled from an openly-available ECG-database containing 3652 examples. Three classes ['N','A','O'] are available.

C A "true hidden" set was provided by Medical Data Transfer, s.r.o. containing 1,000 Holter recordings with

 ISSN: 2325-887X DOI: 10.22489/CinC.2021.055

two classes ['N','A'].

Because the datasets differed in signal amplitudes (mV v.s. raw ADCs outputs), we additionally provided two examples that resembled datasets B & C by re-scaling and debiasing a sinus rhythm example from dataset B.

3. Set-Up of the Class

In a kick-off video meeting, the students were instructed about the problem. We provided a simple example for detecting atrial fibrillation in the form of a jupyter-notebook[1]. The examples exploited that Afib is often characterized by irregular beat-to-beat intervals (BBI)[6]. Therefore our model simply computes the BBIs from detected QRS-complexes and classifies the training data based on a threshold on the standard deviation of BBIs. Students were encouraged to use the model (KIS*MED Model)[2] as a baseline and as an easy starting point to explore more sophisticated methods. The students were asked to form groups of 2-3 members or alternatively were grouped by us. We offered a weekly video-meeting where teams could discuss their main problems and ask questions. After roughly 2 months, all teams presented their general ideas and the difficulties they were facing. Tricks used for achieving good scores were mostly kept secret. We provided example python-files for training and inference from the model, as well as standard functionality to compute the score, load in the data and save predictions. In contrast to the fully automated CinC challenges, teams were allowed to give additional installation instructions (conda, compilation, etc.)

The time-frame from kick-off to final code submission was 3 months. We limited the number of submissions during that time-frame to 5, only counting successful runs, plus one final submission. Only the final submission was used for the final evaluation and ranking. The main reason for limiting the number of submissions was to counteract overfitting on the test sets, a minor reason was the human supervision required for the submission system. For each submission, pretrained models were used at inference time and for scoring. Only in the final submission each team's training code was used on our training set to compare the resulting model performance to the pre-trained model.

4. Evaluation and Ranking

For the evaluation, similar metrics as in the 2017 CinC challenge were used.

[1]Google Colab: https://colab.research.google.com/drive/1AoloKP-ZfZ7rRJu6-aq1cG1PkJHS7KJS (in German)

[2]KIS*MED on GitHub: https://github.com/KISMED-TUDa/18-ha-2010-pj

4.1. Scoring Metrics

We used two metrics to score the submissions of the participants. Both scores where provided and visualized in a table to generate ongoing insights about the performance of each team during the 3 months. The ranking based on F_1 was visible to participants only.

$$F_1 = \frac{\text{TP}}{\text{FP} + \frac{1}{2}(\text{FP} + \text{FN})} \tag{1}$$

Where TP is the number of recordings correctly labeled 'A' , FP is the number of recordings that are labeled as 'A' for which the ground truth is 'N', FN is the number of recordings labeled as 'N' whereby ground truth is 'A'.

$$\text{Multilabel Score} = \frac{1}{N} \cdot \sum_i F_{1,i} \tag{2}$$

Where $F_{1,i}$ is the F_1 score for assigning the recordings to class i or not.

All unlabeled recordings were scored as if they were labeled as 'N'. For the F1-Score, only recordings with ground-truth 'N' and 'A' were evaluated and predicted labels ['O','~'] were relabeled as 'N'. Besides the ranking, each team was notified about their result by eMail and additional remarks about code execution and implemented warnings were shared (provided they did not reveal anything about the test set).

4.2. System and Setup

We used a system with two NVIDIA Quadro RTX5000 GPUs and two Intel Xeon @ 3.8GHz and 256 GB RAM. For each submission we copied the provided model and code (git) to a team folder, installed the required packages in a virtual environment, and executed additional instructions. We overwrote the scoring and prediction scripts and executed these scripts. The scores for each dataset, alongside with date, dataset name and team name were stored in one CSV-file per team.

The students were encouraged to use their own PCs or Google Colab for free gpu computing and were given access to the TU Darmstadt Lichtenberg high performance computer (HPC), which provides high performance parallel computing capabilities. A short introduction to the usage of the HPC was given but to our knowledge only one team made significant use of the HPC.

5. Results

All teams achieved competitive results in the binary classification setting for almost all datasets (Table 1), including set C where data was definitively not available. The superb results for dataset A for some teams stem from

overfitting on the test set. As expected, one team found out that the class was based on the CinC Challenge 2017 (an information we had kept secret at the start) and shared this knowledge with the other groups.

The F1-Scores for the test set B were very good on average with low standard deviation between teams, which can be explained by set B being less noisy. Teams that overfit on test set A, teams that used pretraining on different openly available data, and those that only used the training data provided by us performed well on dataset B, which is an indicator for relatively good generalization of most models. All but two teams tried to optimize the multilabel score but did not put the same effort to the task, as can be seen when looking at the difference between teams that have good scores on Multi Set A as opposed to Multi Set B.

5.1. Common Themes

Apparently, most teams used *pre-training* as a technique to train large scale models, often using the Icentia11k dataset [7]. All teams used at least a *separate validation set (7/7)* generated from their data, even though one team trained their final submission on all data at hand. Three teams used some kind of *cross-validation (3/7)* to compare their models or check for overfitting. Almost all teams used *CNN* for some part of their models, *ResNet (4/7)* architectures were used by four. Four teams used *hand-crafted features (4/7)*, of which two teams relied *solely on handcrafted features (2/7)* for their final model. Four teams applied *ensemble methods (4/7)*, either training two different models and *averaging / voting (3/7)* the end results, or partly *trained different models together (1/7)*. Three teams used a *spectrogram (3/7)* of the ECG as input for their Classification network. Four teams applied *data augmentation methods (4/7)* and all found that this improved the performance of their models significantly. Interestingly, some teams saw improvements from using ensemble models while others did not. Four teams tried, but saw *no improvement from additional datasets (4/7)* in F_1 after pretraining the models on these while training and validating on the CinC 2017 dataset.

Another interesting yet expected outcome can be seen in Figure 1: As with the original challenge, submissions clustered at the end of the semester.

5.2. Student Self-Assessment

We designed a short survey which asked questions about the perceived impact of the course on the methodological knowledge of the students and the impact of the leaderboard on specific motivation (n=11). The survey was taken anonymously after the final grade was assigned. As can be seen in Figure 2, the most common answer in terms

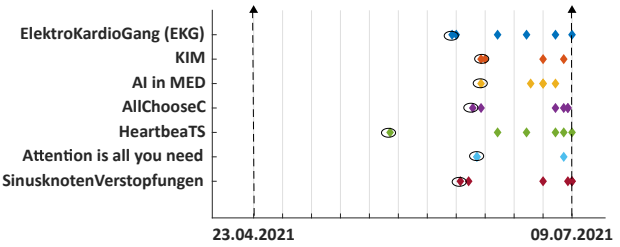

Figure 1. Submissions per Team plotted from kick-off to final submission. Vertical lines indicate weeks.

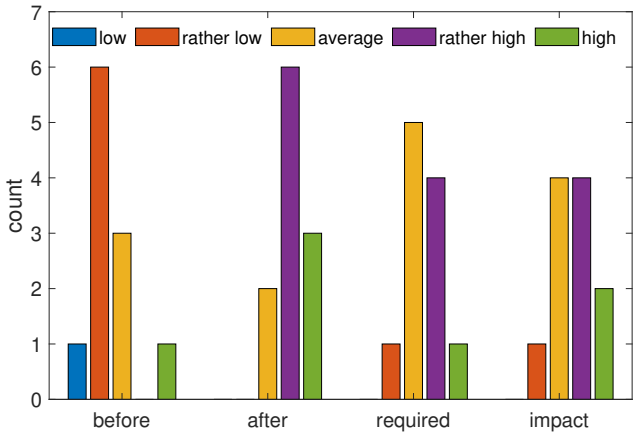

Figure 2. The plot shows the self-assessed knowledge in machine learning and ECG-Analysis of the students before and after the course as well as the perceived required knowledge for participating and the impact of the course on knowledge-gain.

of knowledge before and after the course changed from "rather low" to "rather high". Interestingly, both the requirements on prior knowledge for the class as well as the impact of the class on knowledge gain was rated "average" and "rather high" by the majority of participants. Figure 3 reveals that the leaderboard probably motivated the students far more to try new methods than it did to perform parameter tuning. The main reason for participating in our class was interest in AI and Machine Learning, which was the selected answer of 9 participants. The good time/credit point ratio and interest in teamwork were selected once each, while interest in medicine was only a minor reason. Particular noteworthy from the free-text responses is that students liked the fact that there were only few restrictions regarding code requirements and that working on the same task led to seeing multiple solutions. Interestingly, students wished, among other things, to be actually more restricted by introducing mandatory submissions during the course and a mandatory halftime presentation. Environmental aspects of machine learning were also addressed by proposing limits on computation-time and dataset-size.

Final Ranking

Pos.	Team Name	F1 Set A	F1 Set B	**F1 Set C**	Multi Set A	Multi Set B
1	SinusknotenVerstopfungen	0.986	0.977	0.911	0.887	0.576
2	Attention is all you need	0.939	0.963	0.906	0.831	0.566
3	HeartbeaTS	1.000	0.949	0.881	1.000	0.459
4	AllChooseC	0.935	0.914	0.867	0.878	0.598
5	AI in MED	0.779	0.938	0.803	0.330	0.465
6	KIM	0.894	0.989	0.725	0.896	0.393
7	ElektroKardioGang (EKG)	0.993	0.935	0.554	0.367	0.464
mean		0.932	0.952	0.806	0.741	0.503
sd		0.072	0.024	0.119	0.253	0.071

Table 1. Final ranking of the AI in Med. Challenge sorted by F1 Set C which is also the final score.

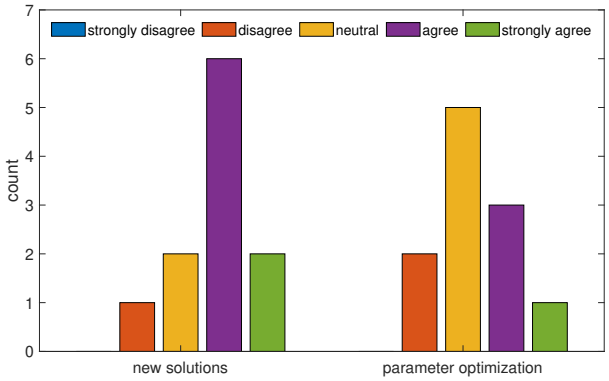

Figure 3. "The leaderboard of the challenge motivated students to focus on..."

6. Conclusion and Lessons Learned

Even though we had emphasized originality and good analysis was more important for grading than a good final score, all teams were fueled by the competitive nature of the seminar. The high scores and individual statements by participants on the voluntary but high workload demonstrate known aspects of gamification. Overall, several similarities to the CinC challenges could be observed, such as the well-documented clustering of submissions towards the end as well as the (subjectively perceived) growing interest from an AI rather than a medical perspective. Interestingly, the suggested mandatory mid-term submission would mimic the unofficial / official phase of the challenge. Also, the semi-automatic analysis method intended to lower the threshold for beginners compared to a fully automated analysis system resulted in a significant overhead, even for this small cohort. In the future, we plan to switch to a Jupyter-Notebook-based evaluation system to reduce overhead on both students and tutors.

Acknowledgements

We would like to thank all students for participating in the evaluation, making useful suggestions as well as excelling in the competition. Additional compute power was provided by Lichtenberg high performance computer of the TU Darmstadt.

References

[1] Legaki NZ, Xi N, Hamari J, Karpouzis K, Assimakopoulos V. The Effect of Challenge-based Gamification on Learning: An Experiment in the Context of Statistics Education. International Journal of Human Computer Studies 2020; 144:102496.

[2] Colombari R, D'Amico E, Paolucci E. Can Challenge-based Learning be Effective Online? A Case Study using Experiential Learning Theory. CERN IdeaSquare Journal of Experimental Innovation 2021;5(1):40–48.

[3] Nah FFH, Zeng Q, Telaprolu VR, Ayyappa AP, Eschenbrenner B. Gamification of Education: A Review of Literature. In International Conference on HCI in Business. Springer, 2014; 401–409.

[4] Toda AM, Valle PH, Isotani S. The Dark Side of Gamification: An Overview of Negative Effects of Gamification in Education. In Researcher Links Workshop: Higher Education for all. Springer, 2017; 143–156.

[5] Clifford G, Liu C, Moody B, Li-Wei L, Silva I, Li Q, Johnson A, Mark R. AF Classification from a Short Single Lead ECG Recording: The PhysioNet/Computing in Cardiology Challenge 2017. In 2017 Computing in Cardiology (CinC). IEEE, 2017; 1–4.

[6] Couceiro R, Carvalho P, Henriques J, Antunes M, Harris M, Habetha J. Detection of Atrial Fibrillation using Model-based ECG Analysis. In 2008 19th International Conference on Pattern Recognition. IEEE, 2008; 1–5.

[7] Tan S, Androz G, Chamseddine A, Fecteau P, Courville A, Bengio Y, Cohen JP. Icentia11k: An Unsupervised Representation Learning Dataset for Arrhythmia Subtype Discovery. arXiv preprint arXiv191009570 2019;.

Address for correspondence:

Maurice Rohr
KIS*MED
TU Darmstadt
Magdalenenstr. 2a, 64289 Darmstadt, Germany
rohr@kismed.tu-darmstadt.de