# An InceptionTime-Inspired Convolutional Neural Network to Detect Cardiac Abnormalities in Reduced-Lead ECG Data

Harry J Crocker, Aaron W Costall

University of Bath, Claverton Down, United Kingdom

## Abstract

*Cardiovascular disease is the leading cause of death worldwide. The twelve-lead electrocardiogram (ECG) is a common tool for diagnosing cardiac abnormalities, but its interpretation requires a trained cardiologist. Thus there is growing interest in automated ECG diagnosis, especially using fewer leads. Hence the PhysioNet-CinC Challenge 2021: Will two (leads) do? The University of Bath team (UoB_HBC) developed InceptionTime-inspired deep convolutional neural networks, using parallel 1D convolutions of varying length, for twelve-, six-, four-, three-, and two-lead models. The twelve-lead model achieved a Challenge metric score of 0.35 on the test set, placing the University of Bath team 23rd out of 39 entries. Though the twelve-lead model performed best, three-lead performance was lower by only 0.25 %, suggesting potential for reliable reduced-lead diagnoses. Furthermore, the three-lead model performed consistently better than the six-lead, highlighting the importance of selection of type of lead, not just their number.*

## 1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, accounting for 32 % of all global deaths in 2019 [1]. This is not limited to the elderly, with CVDs representing 38 % of premature deaths (under the age of 70) due to non-transmissible disease. CVDs remain a major concern in high income countries, causing 25 % of UK deaths [2], and the British Heart Foundation estimates millions of people in the UK are living with undetected risk factors. Improved diagnosis of heart conditions would therefore substantially benefit world health. The electrocardiogram (ECG) is among the most common tools used to diagnose CVDs. The non-invasive procedure uses electrodes on the patient's chest and limbs to measure the electrical potential difference across the heart with respect to time. Due to the variety of CVDs and the complexity of the heart's conduction system, interpreting ECGs requires highly trained cardiologists. However, lack of clinicians means ECGs are not routinely carried out, and many are conducted in primary care centres and emergency units, where healthcare professionals lack the required specialist knowledge [3]. This issue is most prevalent in low and middle income countries [4]. Automated ECG interpretation would provide many benefits. ECGs would become more accessible, allowing many more people with CVDs, particularly from lower socioeconomic backgrounds, to be diagnosed earlier. Accessibility would also be extended to primary care and emergency units, where faster and more accurate diagnoses would save lives. Furthermore, such a tool would allow trained cardiologists to concentrate on diagnosing unique edge cases.

## 2. Method

InceptionTime is a state-of-the-art convolutional neural network for time-series classification, featuring Inception modules (Figure 1) with parallel 1D convolutions to identify patterns of different lengths. The original architecture [5] was improved with modifications to the classification head and learning rate schedule, and the final model (Figure 2) was achieved by hyperparameter tuning. The model is trained according to the one cycle policy learning rate and momentum schedule, using the AdamW optimizer with constant weight decay.

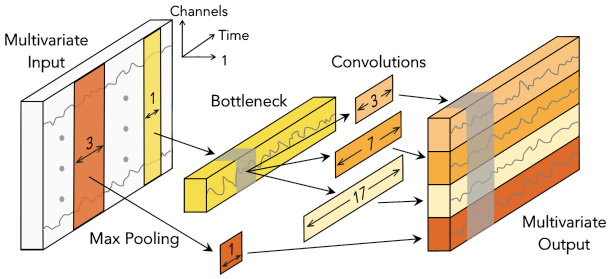


Figure 1. Composition of each Inception module.

The PhysioNet-CinC Challenge 2021 [6] focused on 30 conditions and addressed classification of reduced-lead ECGs as well as full twelve-lead recordings. The reduced-lead sets are the six-lead, four-lead, three-lead, and two-

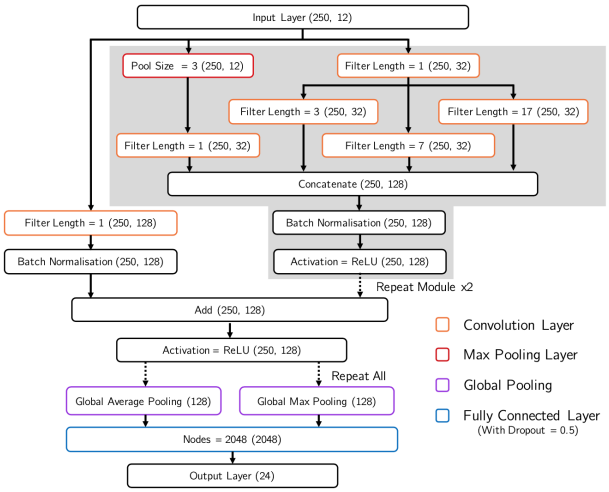


Figure 2. Final model architecture.

Table 1. Lead combinations.

| No. leads | Lead set |
|---|---|
| 12 | I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6 |
| 6 | I, II, III, aVR, aVL, aVF |
| 4 | I, II, III, V2 |
| 3 | I, II, V2 |
| 2 | I, II |

lead. Table 1 defines the various lead combinations used in each case. The dataset provided at the outset of the challenge [7] became the largest collection of twelve-lead ECGs then publicly available, with 43,101 recordings from 34,452 patients. The model was trained on random 2.5 second sub-samples which were down-sampled to 100 Hz using linear interpolation. A randomly selected 10 % of the training data was held back for validation and used to optimize the classification thresholds for maximum Challenge metric. During the Challenge, this dataset was greatly expanded to over 100,000 twelve-lead recordings, comprising a training set of over 88,000, a validation set of 6,630, with 16,630 ECGs retained as hidden test data. However, due to the much enlarged training data it was not possible to achieve cross-validation results on the larger dataset in time. Nevertheless, twelve-, six-, four-, three-, and two-lead models were submitted to the competition. Entries are scored according to the custom Challenge metric [6], which calculates the standard multi-label confusion matrix and normalizes it to give equal weight to each ECG instead of each classification, before taking the dot product with a weighting matrix designed to award partial credit for misclassifications of similar conditions or treatments.

## 3. Results & Discussion

Figure 3 shows the results of the twelve-lead ECG classifier in the form of a multi-class confusion matrix, modified for multi-label problems according to [8]. This awards normalized scores to each class combination such that correct classifications are scored only on the leading diagonal, with misclassifications elsewhere. It also retains the ability to calculate recall from each row and precision from each column. Generally there is classification intensity along the leading diagonal, indicating most ECGs were correctly classified. Off-diagonal, there is also intensity along the Normal Sinus Rhythm (NSR) axes. This is simply due to the large proportion of NSR ECGs in the test set, not classification performance. In the main, there is strong correlation between class-wise performance and number of training examples. Figure 4 shows the F1-Score for each class (except NSR, removed for clarity) against the number of corresponding ECGs in the dataset, demonstrating that more training data generally leads to better performance. A linear trendline is plotted, though in practice more data would see diminishing returns as F1-Score is capped at 1. All classes (NSR included) with more than 2000 recordings performed above average whereas those with below average performance had fewer then 2000 examples.

There are notable outliers, however. Pacing Rhythm (PR), Right Axis Deviation (RAD), and Left Bundle Branch Block (LBBB) show good performance with relatively little data. This suggests other factors, such as similarity to other classes and distinctiveness of features, influence classification performance. Sinus Arrhythmia (SA), however, is consistently misclassified as NSR and vice-versa. This is likely because SA may appear similar to NSR when observed over a short time period. And, despite generally being well classified, there is a noticeable correlation between Complete and Incomplete Right Bundle Branch Block (IRBBB and IRBBB, respectively), again probably due to their similarity as they are variations of the same condition. Another outlier is T-wave Abnormal (TAb), with performance unexpectedly low given the quantity of data, and which is regularly confused with Prolonged QT Interval (LQT) and T-wave Inversion (TInv). Although these conditions are similar, all affecting the T-wave, this response is likely due to the small number of training examples for LQT and TInv. It is expected that more data would not only boost their scores, but also enhance the prediction of TAb and other classes.

Most other misclassifications appear to be without explanation. This is likely due to the large natural variation in ECGs, coupled with relatively limited training data. All classes are generally balanced between precision and accuracy. This is because during development the best model was selected based on AUROC (Area Under Receiver Operating Characteristic) and F1-Score metrics, which penal-

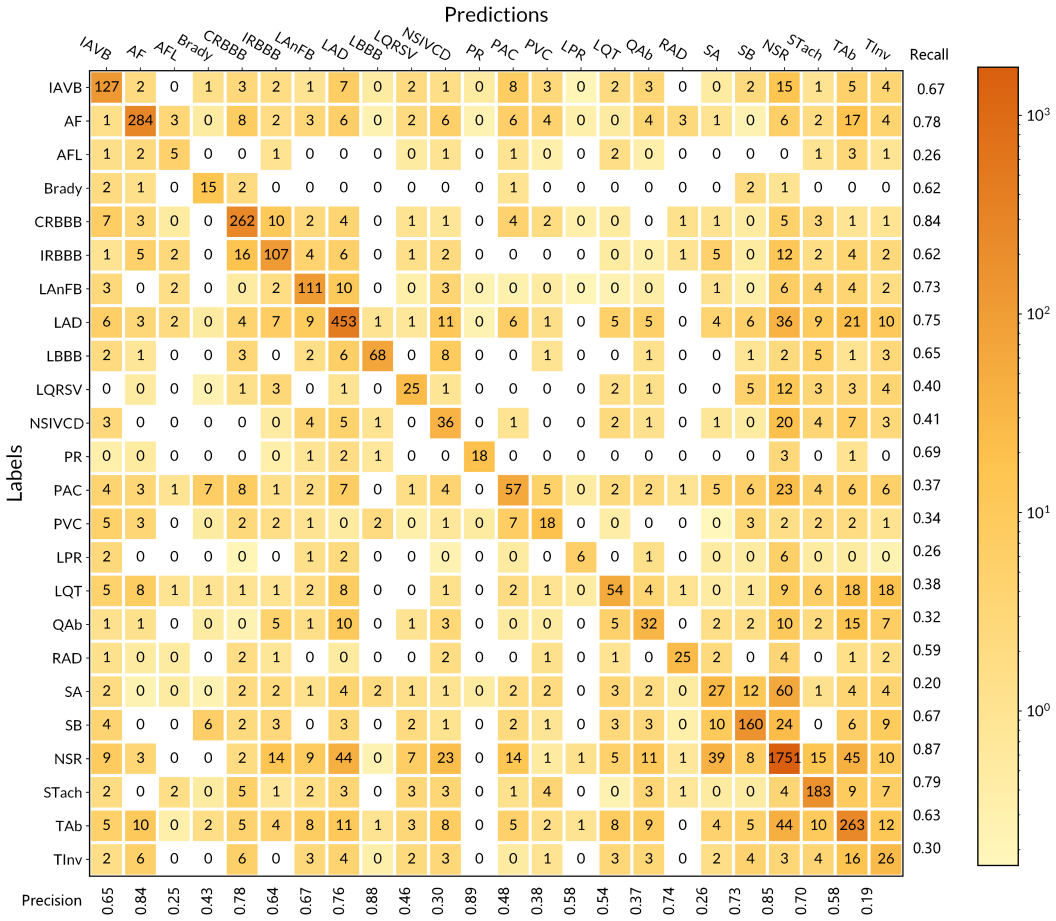


Figure 3. Modified multi-label confusion matrix for 12-lead cross-validation test set results (values are integer rounded).

ize false positives and negatives equally. By requiring high model precision, the model aims to minimize false positives by increasing the thresholds. As a result, 12 % of test set ECGs were unclassified despite only 5 % containing no labels. This is not necessarily an issue as the model may be detecting patterns and conditions it is unable to predict. Furthermore, it is more useful to know when the model is uncertain, rather than it making low confidence predictions, as in these cases a cardiologist can be consulted.

Table 2 compares model performance on the hidden validation data to the cross-validation set, plus the final test set score. In terms of cross-validation performance, the Challenge metric closely correlates with F1-Score, suggesting that optimizing the model and thresholds for the latter translates to good competition performance. Further testing revealed optimizing classification thresholds for maximum Challenge metric score ultimately gives the best competition outcome, as one would expect. Challenge metric scores were consistently 6–8 % lower than cross-validation scores for twelve, three and two leads (up to 15 % for six leads). This suggests the model should generalize well, but there remains a degree of overfitting. This does not necessarily indicate a poor model but highlights the importance of assessing models on unseen, representative data.

Table 2. Comparison of twelve- and reduced-lead model results on cross-validation, hidden validation, and test sets.

| No. | Cross-validation | | Hidden | | Test |
|---|---|---|---|---|---|
| leads | AUROC | F1-Score | Challenge metric | | |
| 12 | 0.95 | 0.57 | 0.54 | 0.402 | 0.35 |
| 6 | 0.93 | 0.52 | 0.47 | 0.385 | - |
| 4 | - | - | - | 0.397 | - |
| 3 | 0.94 | 0.54 | 0.52 | 0.401 | - |
| 2 | 0.93 | 0.50 | 0.45 | 0.391 | - |

Generally, more leads gave better performance in all metrics except for the six-lead set, which had unexpectedly poor results, likely because it doesn't use any chest leads (V1–V6) and so only observes the heart's frontal plane. The three-lead set and the initial two-lead set include chest

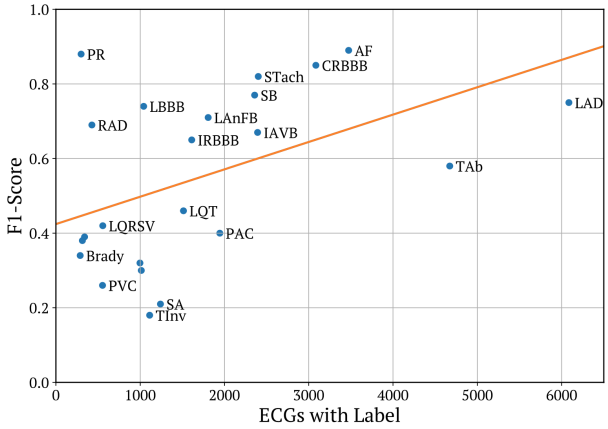


Figure 4. F1-Score versus number of ECGs for each class, illustrating the correlation between performance and quantity of training data (NSR purposely not shown).

leads V2 and V5 respectively, as well as limb leads. This emphasizes that the selection and combination of leads is important, not just their number.

Final performance on the test set achieved a Challenge metric score of 0.35, 13 % down on the hidden validation set score, and placing the University of Bath team 23rd out of 39 entries. This was somewhat expected since, due to computational problems with the larger dataset, it was not possible to generate cross-validation results and make subsequent model adjustments. Nevertheless, there is consistency in terms of the ranking of lead sets and the three-lead remains the best performing reduced-lead model, only 0.25 % short of the twelve-lead result, while the two-lead model was only 2.7 % below twelve-lead performance.

## 4. Conclusions

This paper described a deep convolutional neural network ECG classifier based on InceptionTime architecture. Initial results on twelve-lead ECGs suggested the approach should generalize well but a weakness is performance variation across classes. Classification performance generally correlates with number of training examples. But while more data should improve performance, the models experienced a decrease in Challenge scores on the hidden validation and test sets, despite a doubling of the training data.

The question posed by the 2021 Challenge was: *Will Two Do?* Based on the work here, the answer is: not quite yet. During development, two-lead results were up to 16 % down on twelve-lead. But while results showed that more leads generally produce better results, and that the twelve-lead model always perform best, the worst performing was actually the six-lead, suggesting a sub-optimal combination of leads. On the other hand, the three-lead was surprisingly competitive with twelve-lead performance and outperforms the six-lead consistently across different metrics. This finding may be significant for ECG classifier development as well as cardiology generally. At the very least it suggests excellent potential for reliable diagnoses based on reduced-lead data. It also highlights the importance of appropriate selection of the type of lead, not just their number.

## Acknowledgments

The authors would like to thank PhysioNet for providing the resources and data essential for this project.

## References

[1] World Health Organization. Cardiovascular Diseases (CVDs). Key facts web page, June 2021. URL https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] British Heart Foundation. Statistics Factsheet - UK, July 2021. URL https://www.bhf.org.uk/what-we-do/ our-research/heart-statistics.

[3] Mant J, Fitzmaurice DA, Hobbs FDR, Jowett S, Murray ET, Holder R, Davies M, Lip GYH. Accuracy of Diagnosing Atrial Fibrillation on Electrocardiogram by Primary Care Practitioners and Interpretative Diagnostic Software: Analysis of Data From Screening for Atrial Fibrillation in the Elderly (SAFE) Trial. British Medical Journal June 2007; 335(7616).

[4] Mendis S, Armstrong T, Bettcher D, Branca F, Lauer J, Mace C, et al. Global Status Report on Noncommunicable Diseases 2014. World Health Organization report, 2014.

[5] Ismail Fawaz H, Lucas B, Forestier G, Pelletier C, Schmidt DF, Weber J, et al. InceptionTime: Finding AlexNet for Time Series Classification. Data Mining and Knowledge Discovery September 2020;34(6):1936–1962.

[6] Reyna MA, Sadr N, Perez Alday EA, Gu A, Shah AJ, Robichaux C, Rad AB, Elola A, Seyedi S, Ansari S, Ghanbari H, Li Q, Sharma A, Clifford GD. Will Two Do? Varying Dimensions in Electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021. Computing in Cardiology 2021;48:1–4.

[7] Perez Alday EA, Gu A, Shah AJ, Robichaux C, Wong AI, Liu C, Liu F, Rad AB, Elola A, Seyedi S, Li Q, Sharma A, Clifford GD, Reyna MA. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiol Meas 2020;41(12):124003.

[8] Krstinić D, Braović M, Šerić L, Božić-Štulić D. Multi-label Classifier Performance Evaluation With Confusion Matrix. Proc Computer Science Information Technology Conference Series 2020;10(8):1–14.

Address for correspondence:

A. Costall, University of Bath, Claverton Down BA2 7AY, UK
awc34@bath.ac.uk