Exploring Best Practices for ECG Pre-Processing in Machine Learning

Amir Salimi ASALIMI@UALBERTA.CA
Abram Hindle ABRAM.HINDLE@UALBERTA.CA
Osmar Zaiane ZAIANE@UALBERTA.CA
others CVC@UALBERTA.CA
University of Alberta

Editors: List of editors' names

Abstract

We are looking for universal best practices in pre-processing of Electrocardiogram (ECG) signals in order to train better models for the diagnosis of heart disease. State of the art machine learning algorithms have achieved remarkable results in classification of heart disease using ECG data, yet there appears to be no consensus on pre-processing best practices. Is this lack of consensus due to different diseases and model architectures requiring different processing steps for optimal performance? Is it possible that state-of-the-art deep-learning models have rendered pre-processing unnecessary? In this work we apply downsampling, normalization, and filtering functions to 3 different multi-label ECG datasets and measure their effects on 3 different SOTA time-series classifiers. We find that sampling rates as low as 50Hz can yield comparable if not better results than the commonly used 500Hz sampling rate. This is significant as smaller sampling rates will result in smaller models, which in turn require less time and resources to train. We found min-max normalization to be slightly detrimental overall, and band-passing to make no measurable difference.

Keywords: electrocardiogram, machine learning, signal processing

1. Introduction

We focus on the Electrocardiogram (ECG) pre-processing decisions faced by researchers working towards automatic classification of cardiovascular conditions. The ever-growing number of tools and approaches available for machine learning often requires a number of a priori guesses to be made by those who process the training data and design the architectures. The task of training models which predict heart disease is no exception.

Routinely utilized by clinicians for diagnoses of cardiovascular abnormalities, an ECG is a recording of the electrical activity of the cardiovascular system Uwaechia and Ramli (2021); Hong et al. (2022); Ismail Fawaz et al. (2019); Strodthoff et al. (2020); Chen et al. (2020). In recent years, machine learning models have achieved remarkable results in automatic diagnosis of some heart conditions when trained with enough labeled ECG data Reyna et al. (2021a,b). However, training such models requires a large amount of data Reyna et al. (2021a,b); Natarajan et al. (2020); Ribeiro et al. (2020), where decisions regarding how the data is pre-processed (e.g., filtering, scaling, augmentation) can be critical for both the model's performance as well as the amount of time and hardware required.

In order to automatically classify cardio-vascular conditions using ECG data, we are faced with many choices for the pre-processing functions. Pre-processing functions are transformation functions applied to signals such as ECGs in order to reduce noise and simplify the learning task. Previous works utilize a diverse range of pre-processing functions, datasets, and architectures, often with great results Hong et al. (2022). As we will discuss in Section 2 there is a large variety of implementations for such functions in past works, and finding a trend between pre-processing functions used and performance results is not feasible. We believe that in order to make claims about the viability of ECG pre-processing methods (that is, whether they result in better outcomes, and their effect on time and hardware requirements) it is best to consider multiple datasets, architectures, and heart conditions.

In this work we aim to simplify this decision space for other researchers by using our code to test the effect of *downsampling*, *band-passing*, and *normalizing* on the outcome of 3 different multi-label SOTA classifiers for time-series when trained on 3 different ECG datasets. We show that downsampling from the standard recording rate of 500Hz can significantly reduce training times and hardware requirements, without being detrimental to model performance. We found min-max normalization to be slightly detrimental overall, and band-passing to make no measurable difference. In Section 4 we discuss the experiment setup and performance measures, and the discuss the results of downsampling, bandpassing, and normalization in Sections 5, 6,and 7.

2. Previous Work

needs:
works which used the pre-processing methods in question

The state of the art for the automatic classification of cardio-vascular disease continually improves as more data sets become available and computation costs decrease. However, no consensus seems to exist on the best pre-processing steps for such datasets. Here, we define pre-processing as "any function which aims to simplify or transform the data in-order to simplify the learning process". Hong et al. (2022) highlight 4 different popular pre-processing methods for ECG signals, based on their analysis of the entries to the Physionet2020 Challange Goldberger et al. (2000); Alday et al. (2020):

- Resizing: Signals are time-series, and time-series can vary in length. Resizing refers to having a fixed length for all signals. In this work, we do not explore various approaches to resizing, rather, we pad samples with zeros on the right when lower than our desired fixed length.

- Resampling (downsampling): ECGs are typically sampled at the rate of 500Hz Luo and Johnston (2010); Uwaechia and Ramli (2021), meaning that each second of an ECG recording contains 500 samples. The act of downsampling an ECG is akin to resolution reduction of an image. This can drastically lower training times and hardware requirements, yet it is possible that important information could be lost. In this works we explore the effect of different downsampling (or scaling) rates on the performance of ML models. We discuss our findings in Section 5.

- Band-Passing: Many sources can intorduce noise to ECG signals. Powerlines are a common source Uwaechia and Ramli (2021), and depending on the region they can introduce noise at various

Table 1: Top 5 teams in Physionet2020 and pre-processing functions used.

Rank	Group	Scaling	Normalize	BandPass
1	Wide and Deep Transformer Natarajan et al. (2020)		✓	✓
2	Adaptive ResNet Zhao et al. (2020)	✓	✓	
3	SE-ResNet Zhu et al. (2020)			
4	Scatter Transform and DNNs Oppelt et al. (2020)		✓	
5	Adversarial Domain Generalization Hasani et al. (2020)			✓

frequencies. Some ECG recording hardware have band rejection (or notch filters) built in to deal with such issues. However, there are other unaccounted sources of noise which need to be dealt with at the software level. A band-pass filter is used for the removal of frequencies outside of its two cut-offs. All frequencies below the high-pass cut-off and above the low-pass cut-off are removed or reduced. This can improve performance as noise can interfere with the learning process. In this work we compare the use of different cut-offs as well as the case when no filtering was applied to the dataset. The results are shown in Section 6. *notch filters*

- Normalizing: Normalization of input values has been shown to produce better or equal value results in image and time-series classification Bhanja and Das (2018)cite more/better. Particularly in the case of ECGs, there can be differences in recorded amplitudes depending on the equipment used Uwaechia and Ramli (2021), making normalization a logical step in pre-processing. Here we compare 3 different normalization algorithms proposed by previous works as well as the case when no normalization was applied. We show the results in Section 7

There is also great variety in the architectures used to learn from ECG datasets, or time-series data more generally. Here, we use three models with proven results in state of the art time-series classification tasks: the Inception-Time Network, which is the 1-dimensional application of the Inception-Network Szegedy et al. (2017); Ismail Fawaz et al. (2020) and MiniRocket, a quick and mostly deterministic feature extractor for time-series Dempster et al. (2021). A recent survey of time-series classification methods by Ruiz et al. (2021) highlights both of these models as excellent performers in various multi-variable time-series classification benchmarks. In the context of ECG classification, Inception-Time has achieved state of the art performance ECG classification tasks Strodthoff et al. (2021). The third architecture tested here is xresnet1d101, a resnet model for time-series which was found to be the overall best performer in a benchmark study of best architectures for classification of ECGs Strodthoff et al. (2020) .

3. Datasets

We conduct our experiments on three datasets, the **CPSC** dataset Liu et al. (2018), **Chapman-Shaoxing** dataset Zheng et al. (2020), and **PTB-XL** Wagner et al. (2020). The datasets are all multi-label, and recorded at the sampling rate of 500Hz. All have been released as part of the 8 datasets of labeled 12-lead ECGs provided by the Physionet2021 challenge Reyna et al. (2021a,b). To simplify our datasets and experiments, we only use the labels which appear in more than 5% of the ECGs in each dataset. Here we only use 10 seconds of

Table 2: Modified CPSC Dataset

Label	Count
right bundle branch block	1857
ventricular ectopics	700
atrial fibrillation	1221
1st degree av block	722
premature atrial contraction	616
sinus rhythm	918
st depression	869
Total	**6903**

Table 3: Modified Chapman Dataset

Label	Count
left ventricular high voltage	1295
atrial fibrillation	1780
t wave abnormal	1876
sinus bradycardia	3889
supraventricular tachycardia	587
sinus rhythm	1826
sinus tachycardia	1568
nonspecific st t abnormality	1158
Total	**9910**

data from each ECG. ECGs shorter than this length are padded with zeros (on the left side), such that we can represent each ECG as a matrix with dimensions of 12x5000 (before downsampling). The breakdown of these datasets and their label counts after our modifications are given in Tables 2, 3, and 4.

- **CPSC** is an open access dataset released in 2018 as part a multi-label ECG classification competition Liu et al. (2018) and used in SOTA benchmarks Strodthoff et al. (2020). ECG signal duration in this dataset is between 6 and 60 seconds, with an average duration of 15.79 seconds. The the 12 leads were recorded at the frequency of 500 Hz, and each ECG can have up to 9 labels.

Table 4: Modified PTB-XL

Label	Count
left axis deviation	5146
myocardial ischemia	2175
myocardial infarction	5261
left ventricular hypertrophy	2359
ventricular ectopics	1154
atrial fibrillation	1514
t wave abnormal	2345
abnormal QRS	3389
sinus rhythm	18092
left anterior fascicular block	1626
incomplete rbbb	1118
Total	**21311**

- **Chapman-Shaoxing** is an open access dataset which has not been subject to many benchmarks. This dataset contains 12-lead ECGs of 10,646 patients with a 500 Hz sampling rate and 54 labels Zheng et al. (2020).

- **PTB-XL** is an open access dataset, used in recent benchmarks Strodthoff et al. (2020). This dataset contains 12-lead ECGs of 21,837 ECGs with a 500 Hz sampling rate Zheng et al. (2020). The original dataset has 54 labels, however after dropping uncommon labels, we are left with 11.

4. Experiment Methodology

In Section 2 we discussed common preprocessing functions, and how some SOTA classifiers have used one or more of these functions, while others have omitted them entirely as seen in Table 1 as well as works by Ribeiro et al. (2020)(should put in more works here). We also discussed three SOTA

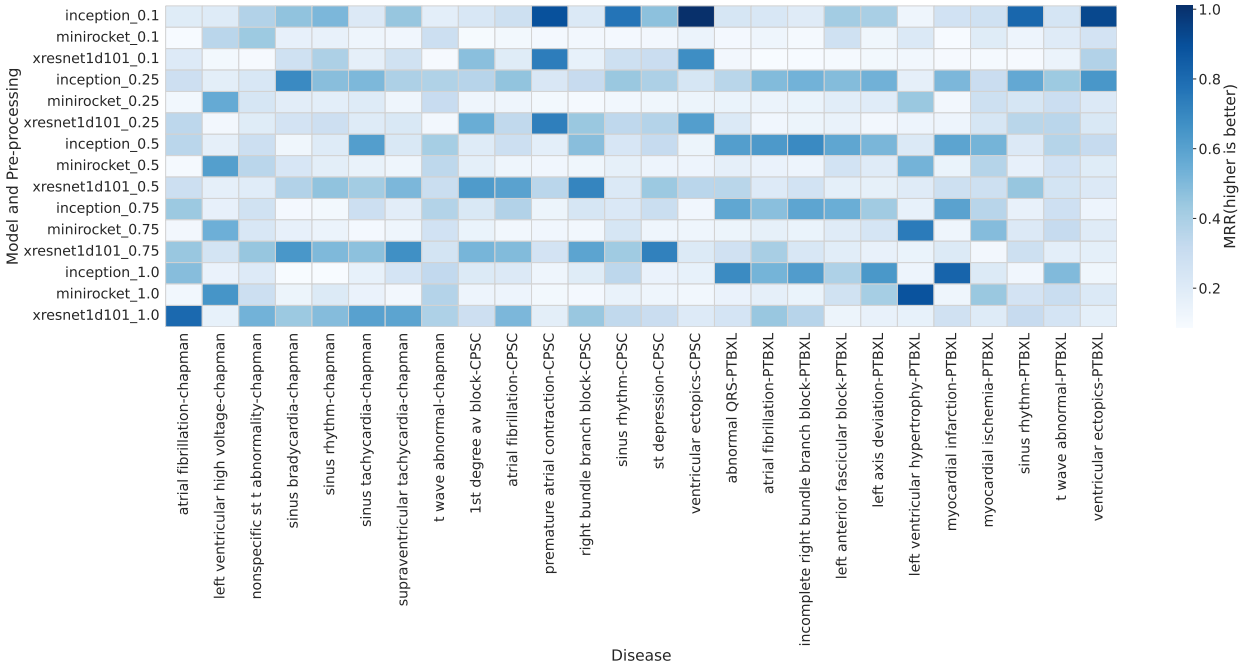

Figure 1: Performance of the models for each dataset and disease at varying scaling rates. The 15 models are ranked against each other, the close the MRR is to 1, the better the model performed relative to others.

architectures for time-series classification, and in Section 3, we described the three datasets of multi-labeled ECGs that the models can learn from. *Best practices for ECG pre-processing can be established if we see consistent patterns of improvement in performance scores when applying a pre-processing function to different datasets, and comparing the performances of the models.*

To compare pre-processing methods, we apply the methods we want to compare to each dataset, and train models on the transformed datasets. We use the performance of the model to give a score to the pre-processing function. Given enough of these performance scores for the various models, datasets, and diseases, we use statistical tests such as wilcoxon or Kruskal-Wallis to determine whether the pre-processing method had an effect on the performance of the models. For example, we would have the CPSC dataset at various sampling rates: CPSC_500Hz, CPSC_250Hz, etc. For each version of the dataset (i.e, the dataset pre-processed using a different function), we train and test each model 20 times. For 20 rounds, the data is randomly split in 80/10/10 training/validation/testing sets. Each model is trained on the training set, and the best model (according to the F1 score on the validation set) is saved to disk. When no improvements in the F1 measure is seen after 30 epochs, the training stops, and the best model's F1 score on the test set is

saved as the performance score of the experiment. This means that for each dataset/preprocessing/model combination, we have 20 performance scores.

In this case, we can use the Wilcoxon signed rank test to measure whether each model out-performs every other model. To rank and compare the models, we use the Wilcox-Holm post-hoc analysis used for creating critical-difference diagrams by Ismail Fawaz et al. (2019). This analysis uses a pairwise Wilcoxon test with Holm-Bonferroni p-value correction, and shows whether the differences between each model pair is significant. While Ismail Fawaz et al. (2019) uses the average rank to show relative performance, we use the "Mean Reciprocal Rank" (MRR) for each model. Here, $rank_i$ refers to the model's rank relative to the other models when trained/validated/tested on the same subsets of the data, and $Q = 20$ since for each model/dataset combination we have 20 F1 scores:

$$\text{MRR} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}$$

Another method for comparison of results is the Kruskal-Wallis test Kruskal and Wallis (1952); Ostertagova et al. (2014). This test measures whether the median of several sets are *different*. We use this test to determine whether the pre-processing method had any effect on the performance of models for each disease. Since we are comparing 3 different pre-processing methods for 20 different heart conditions, we use the Holm-Bonferroni correction Holm (1979); Abdi (2010), and divide our initial alpha of 0.05 by 60, which gives us the new alpha $\alpha^* = 8.33e^{-4}$. We reject the null hypothesis (i.e, the pre-processing method has an effect on the outcome) if $p < \alpha^*$. We've provided the raw p-value results in Table 5, such that less conservative means of determining pre-processing significance can be applied.

5. Scaling/Downsampling

The ECG signals in our dataset are recorded at 500Hz. We apply the scaling rates of 0.1, 0.25, 0.5, 0.75, and 1, which downsamples the signals to sampling rates of 50Hz, 125hz, 250Hz, 375Hz, and the unmodified 500Hz. The downsampling method used here is Pytorch's "nearest-exact" interpolation function Paszke et al. (2019). Effectively, each dataset has 5 different sampling rates, and each of the 3 models learns from these datasets by randomly selecting 80% of the dataset for training, 10% for validation, and 10% for testing. In total, for each dataset, we run 15 experiments (3 architectures, and 5 different sampling rates) and for each experiment, we measure 20 performance scores by running the training/validation/testing step and measuring the macro-average F1 score on the test set. Show how high the performance cost for this is, in time and VRAM required. We use the Wilcox-Holm post-hoc analysis described in the methodology (Section 4) to get the rank and relative performance for each model.

5.1. Downsampling Results

The MRR for each model when predicting each label is shown in Figure 1. Here, no consistent pattern can be seen across diseases and scaling rates, however, some trends are observable. For example, for Atrial fibrillation which is common label in all 3 datasets, we see that DL models with higher sampling rates performed the best. Another interesting label is Sinus Rythm, also common in all 3 datasets, where the inception network with lowest possible sampling rate clearly out-performs. Kruskal-Wallis p-values in Table 5 show that sampling rate is an important factor in 10 out of 20 diseases. From these observations we can conclude that sampling rate is an important factor for better diagnosis of *some* heart conditions.

Table 5: Kruskal-Wallis P-Value for importance of factor in diagnosis of each disease

disease	Sampling Rate	Bandpassing	Normalizing
1st degree av block	3.50e-02	2.128e-01	2.74e-01
abnormal QRS	**3.56e-08**	7.514e-01	**2.41e-08**
atrial fibrillation	**6.53e-04**	9.828e-02	4.14e-03
incomplete right bundle branch block	**2.13e-11**	8.064e-01	3.30e-02
left anterior fascicular block	2.06e-01	1.707e-01	**2.00e-04**
left axis deviation	**2.33e-04**	3.918e-02	**7.56e-06**
left ventricular high voltage	2.84e-01	9.227e-02	**1.44e-14**
left ventricular hypertrophy	**6.34e-05**	6.495e-01	**2.09e-12**
myocardial infarction	**2.86e-10**	3.122e-02	1.62e-02
myocardial ischemia	**2.82e-04**	4.939e-01	**2.80e-07**
nonspecific st t abnormality	2.51e-02	5.428e-01	1.49e-02
premature atrial contraction	**5.97e-08**	2.842e-01	3.28e-01
right bundle branch block	1.89e-02	7.902e-01	2.38e-02
sinus bradycardia	**1.43e-05**	7.902e-01	9.28e-01
sinus rhythm	3.23e-01	9.566e-01	3.23e-01
sinus tachycardia	2.54e-03	2.685e-02	4.93e-01
st depression	6.40e-03	9.764e-01	**4.43e-05**
supraventricular tachycardia	9.73e-01	2.441e-01	5.42e-01
t wave abnormal	1.29e-01	7.262e-01	1.53e-01
ventricular ectopics	**4.02e-21**	2.357e-01	3.56e-01

We also show the Spearman correlation between sampling rate and performance in Table 6. There is a per-model breakdown of the correlation values, as well as the values when all datasets and all models are considered. Here, correlation measures the mapping between two continuous variables: sampling rates between 50-500Hz and F1 performance. Overall, when considering all models and all datasets, we see a negligible correlation of 0.039. This suggests that sampling-rate is **not** an important factor in the overall results of our training.

Main Takeaway: Based on these results, sampling rate does appear to be an important factor in diagnosis of some labels, however, it does not standout as a particularly important factor when training a multi-label model, unless some labels in the models are weighted differently than others. This is a significant find as sampling-rate has a near 1-to-1 effect on the speed and hardware required for training SOTA models on large datasets of ECG signals, and these results suggest that it is possible to reduce sampling rates from the standard 500Hz to 250Hz and lower without major loss in over-all performance.

Table 6: Spearman correlation between F1 measures and sampling rate (50Hz-500Hz), and F1 measures and normalization (0 → raw, 1 → minmax). Overall, we see weak correlations for all scenarios.

	Rate	**Norm**
Inception	-5.23e-02	-1.17e-01
MiniRocket	2.70e-02	-1.21e-02
xresnet	5.17e-02	-9.50e-02
All Models	8.00e-3	-7.41e-02

6. Band-Pass Filtering

A band-pass filter removes all frequencies outside its range, which can be an effective noise removal technique. The band-pass filter range applied to ECG data varies greatly in past works. Typically, the high-pass filter takes on a value of 0.05-1Hz, and the low-pass filter is applied somewhere between 30-150Hz Uwaechia and Ramli (2021); Luo and Johnston (2010). there are a lot more works to cite here. Is there an "optimal" range for a band-pass filter in this context? Is band-pass filtering even necessary when using large datasets and large DL models?

In Section 5.1, we discussed how downsampling does not appear to have a major effect on the over-all model performance. Here, in order to speed up training, we downsample each dataset to 250Hz. We then apply 3 different band-pass filters to the data before training the 3 models. The 3 ranges are 1-30Hz, 1-50Hz, 1-100Hz. We compare these 3 pre-processing methods to see if the change in the low-pass filter frequency makes a difference in the final outcome, we then make a comparison to the case when no filtering was applied to the dataset.

6.1. Band-Pass Filtering Results

The Kruskal-Wallis results in Table 5 do not indicate a difference between the performance of the 4 model groups in question (models trained using the 3 different band-passing functions, and models trained on the raw data). This is an interesting result which may explain the lack of consistency in the application of signal filtering in SOTA ECG classification works.

Main Takeaway: When training on multiple labels from large, carefully curated datasets of ECGs, band-passing is not a necessary step. Band-pass filtering does not hurt the final outcomes, and can be applied at the GPU level with low overhead.

Therefore in most situations, particularly when dealing with potentially noisy data, it is advisable to use filtering. We cannot make any concrete statements about best cut-off frequencies, as it likely varies based on the recording hardware used and the cardiovascular conditions being diagnosed.

7. Normalization

What we refer to as "normalization" is the transformation of an input signal—which in this case is an ECG—to a distribution of values in the range of [0,1] Hong et al. (2022). Here we use the "Min-Max" normalization method, a commonly used pre-processing technique for ECG signals Uwaechia and Ramli (2021); Li and Narayanan (2010); Li and Cui (2019); Fang and Chan (2009). This technique uses the following formula where $Y[n]$ is the new amplitude for time-step n, $X[n]$ is the old amplitude, min is the minimum value for $X[n]$, and max is the maximum value for $X[n]$ Uwaechia and Ramli (2021).

$$Y[n] = \frac{x[n] - min}{max - min}$$

By default, all our experiments have used this technique in the pre-preocessing chain. Here we ran our 3 models on the 3 datasets with no pre-processing, and compare the results to the models with only normalization applied (these would be the models from Section 5 with the unmodified sampling rate of 500Hz). The Kruskal-Wallis results are shown in Table 5

7.1. Normalization Results

The Kruskal-Wallis results in Table 5 show normalization to be an important factor in the outcomes of 7 out of 20 diseases. To calculate the correlation values in Table 6, we added a new label to the results by mapping raw, none-normalized ECGs to 0, and min-max normalized ECGs to 1. We then measure the correlation between F1 values and the normalization type. The correlation values have a slight negative bias, which indicates that in the cases where normalization type was important, more often than not models training on raw ECGs performed better.

Main Takeaway Surprisingly, Min-Max normalization has a slight negative impact on the performance relative to raw ECGs. However, min-max normalization one of multiple common normalization methods used for time-series classification Uwaechia and Ramli (2021), therefore we cannot discourage the use of all normalization methods. However, these results show that although normalization is a staple pre-processing method in some fields of deep-learning, we cannot presume its effectiveness for the task of ECG classification.

8. Discussion

We found no consistency in the pre-processing methods used in recent research on the detection of cardiovascular conditions using ECG signals. To address this gap in knowledge, we trained 3 different time-series classifiers on 3 open datasets of ECGs in order to examine the effect of 3 popular pre-processing methods. Our findings suggest that the performance of different pre-processing and architecture combinations varies depending on the condition of interest. This would explain the lack of consistency in previous works, which may be looking at different datasets, using different label weights, or looking at different performance measures.

Many recent works have highlighted the effectiveness of an ensemble method using various architectures relative to any single model Uwaechia and Ramli (2021); Hong et al. (2022); Ismail Fawaz et al. (2019); Strodthoff

et al. (2020); Chen et al. (2020). Our findings here suggest that use of various architectures *as well* as different pre-processing methods could yield slight improvements in SOTA benchmarks. However, we would be brute-forcing the problem, rather than solving it.

References

Hervé Abdi. Holm's sequential bonferroni procedure. *Encyclopedia of research design*, 1(8):1–8, 2010.

Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, et al. Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. *Physiological measurement*, 41(12):124003, 2020.

Samit Bhanja and Abhishek Das. Impact of data normalization on deep neural network for time series forecasting. *arXiv preprint arXiv:1812.05519*, 2018.

Tsai-Min Chen, Chih-Han Huang, Edward SC Shih, Yu-Feng Hu, and Ming-Jing Hwang. Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model. *Iscience*, 23(3):100886, 2020.

Angus Dempster, Daniel F Schmidt, and Geoffrey I Webb. Minirocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 248–257, 2021.

Shih-Chin Fang and Hsiao-Lung Chan. Human identification by quantifying similarity and dissimilarity in electrocardiogram phase space. *Pattern Recognition*, 42(9): 1824–1831, 2009.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Hosein Hasani, Adeleh Bitarafan, and Mahdieh Soleymani Baghshah. Classification of 12-lead ecg signals with adversarial multi-source domain generalization. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

Shenda Hong, Wenrui Zhang, Chenxi Sun, Yuxi Zhou, and Hongyan Li. Practical lessons on 12-lead ecg classification: Meta-analysis of methods from physionet/computing in cardiology challenge 2020. *Frontiers in Physiology*, page 2505, 2022.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4): 917–963, 2019.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

Ming Li and Shrikanth Narayanan. Robust ecg biometrics by fusing temporal and cepstral information. In *2010 20th International Conference on Pattern Recognition*, pages 1326–1329. IEEE, 2010.

Yaoguang Li and Wei Cui. Identifying the mislabeled training samples of ecg signals using machine learning. *Biomedical signal processing and control*, 47:168–176, 2019.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, 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(7):1368–1373, 2018.

Shen Luo and Paul Johnston. A review of electrocardiogram filtering. *Journal of electrocardiology*, 43(6):486–496, 2010.

Annamalai Natarajan, Yale Chang, Sara Mariani, Asif Rahman, Gregory Boverman, Shruti Vij, and Jonathan Rubin. A wide and deep transformer neural network for 12-lead ecg classification. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020.

Maximilian P Oppelt, Maximilian Riehl, Felix P Kemeth, and Jan Steffan. Combining scatter transform and deep neural networks for multilabel electrocardiogram signal classification. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020.

Eva Ostertagova, Oskar Ostertag, and Jozef Kováč. Methodology and application of the kruskal-wallis test. In *Applied Mechanics and Materials*, volume 611, pages 115–120. Trans Tech Publ, 2014.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Matthew A Reyna, Nadi Sadr, Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, et al. Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021. In *2021 Computing in Cardiology (CinC)*, volume 48, pages 1–4. IEEE, 2021a.

Matthew A Reyna, Nadi Sadr, Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, et al. Issues in the automated classification of multilead ecgs using heterogeneous labels and populations. *personnel*, 4:5, 2021b.

Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1–9, 2020.

Alejandro Pasos Ruiz, Michael Flynn, James Large, Matthew Middlehurst, and Anthony Bagnall. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 35(2):401–449, 2021.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE Journal of Biomedical and Health Informatics*, 25(5): 1519–1528, 2020.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE Journal of Biomedical and Health Informatics*, 25(5): 1519–1528, 2021. doi: 10.1109/JBHI.2020.3022989.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Anthony Ngozichukwuka Uwaechia and Dzati Athiar Ramli. A comprehensive survey on ecg signals as new biometric modality for human authentication: Recent advances and future challenges. *IEEE Access*, 2021.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

Zhibin Zhao, Hui Fang, Samuel D Relton, Ruqiang Yan, Yuhong Liu, Zhijing Li, Jing Qin, and David C Wong. Adaptive lead weighted resnet trained with different duration signals for classifying 12-lead ecgs. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1):1–8, 2020.

Zhaowei Zhu, Han Wang, Tingting Zhao, Yangming Guo, Zhuoyang Xu, Zhuo Liu, Siqi Liu, Xiang Lan, Xingzhi Sun, and Mengling Feng. Classification of cardiac abnormalities from ecg signals using se-resnet. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020.

Appendix A. First Appendix

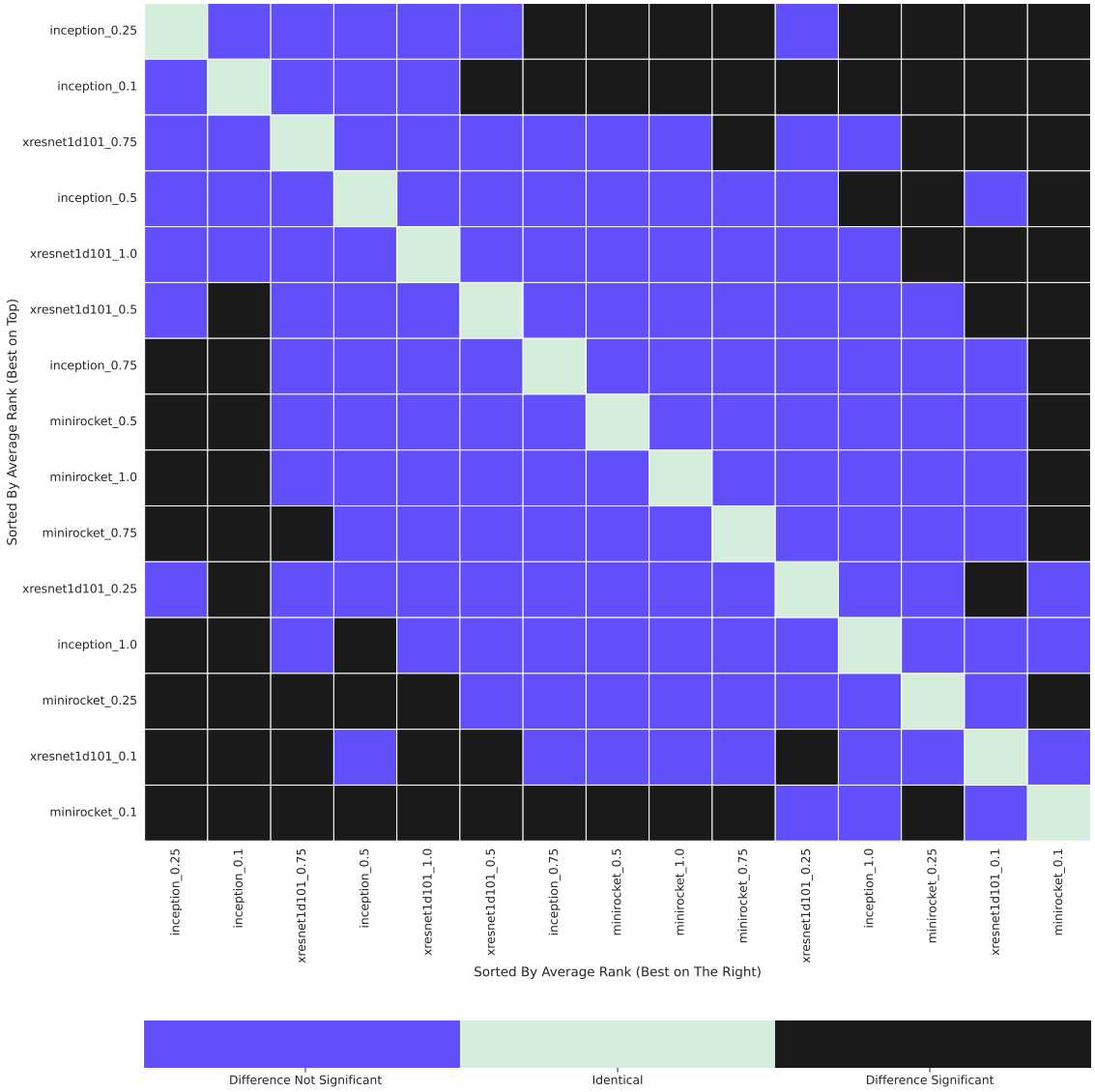

Figure 2: Model performance comparison for all datasets. Best models (those with the lowest average ranking) are higher on the y-axis.

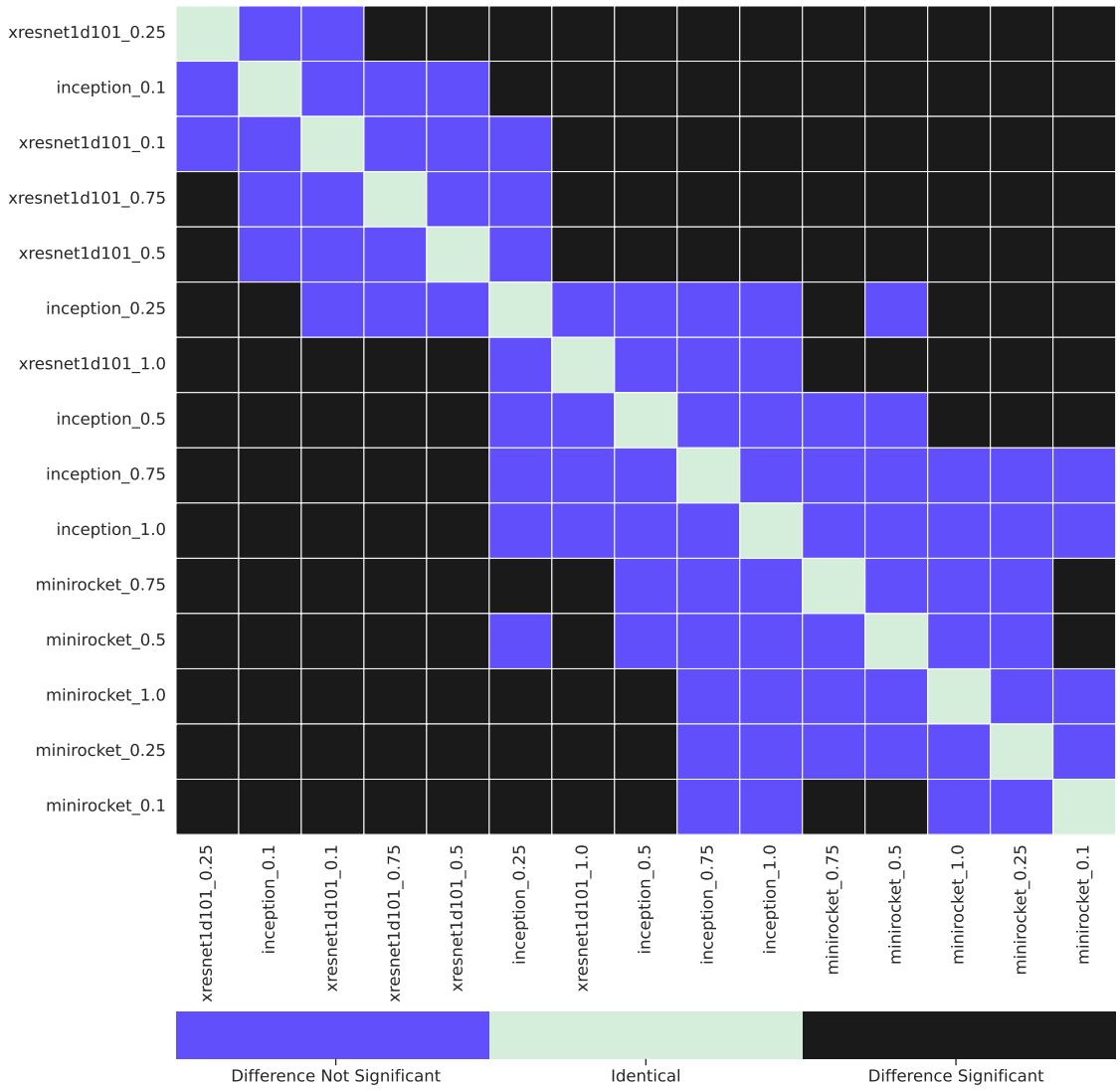

Figure 3: Model performance comparison for CPSC

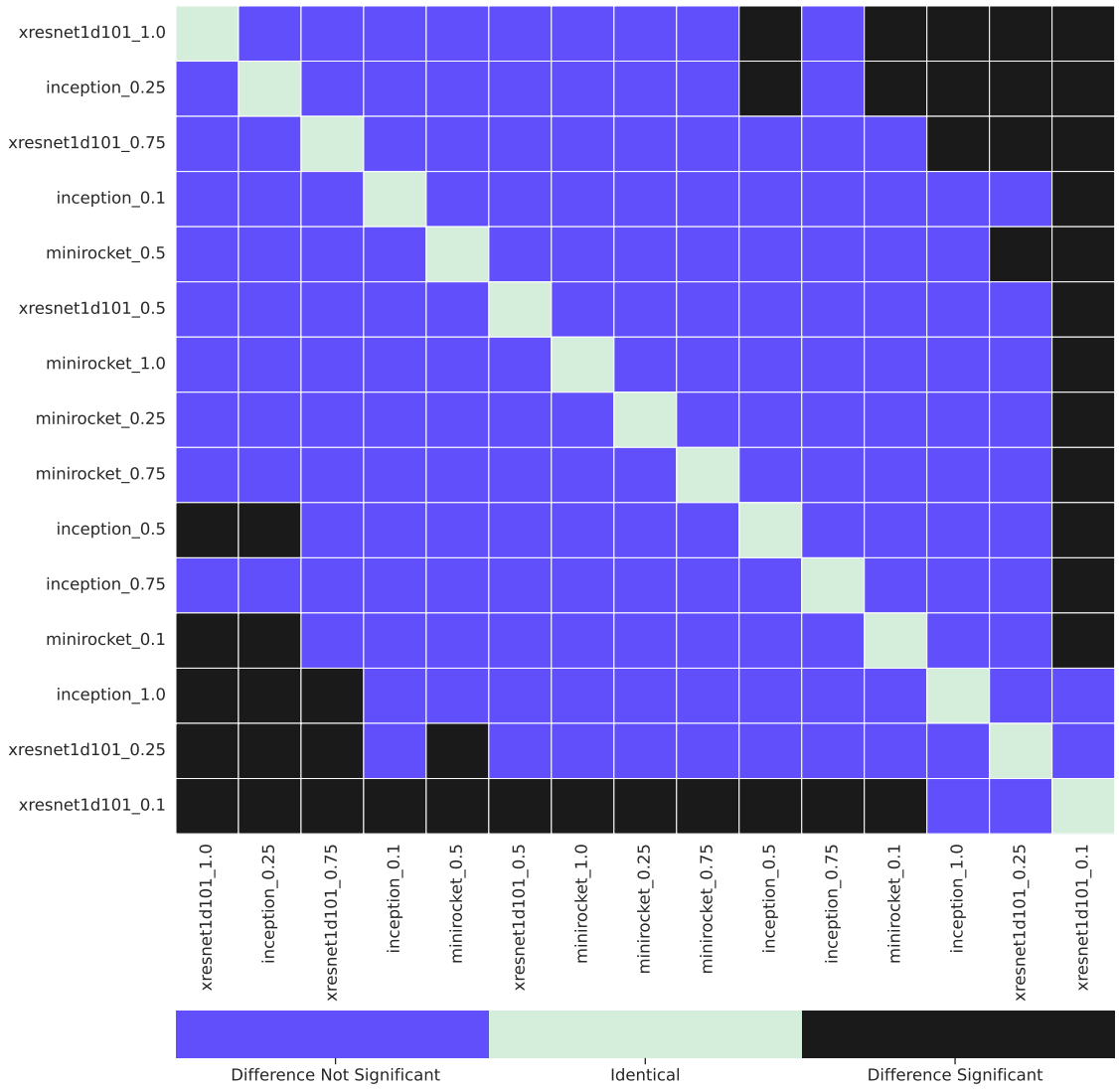

Figure 4: Model performance comparison for Chapman

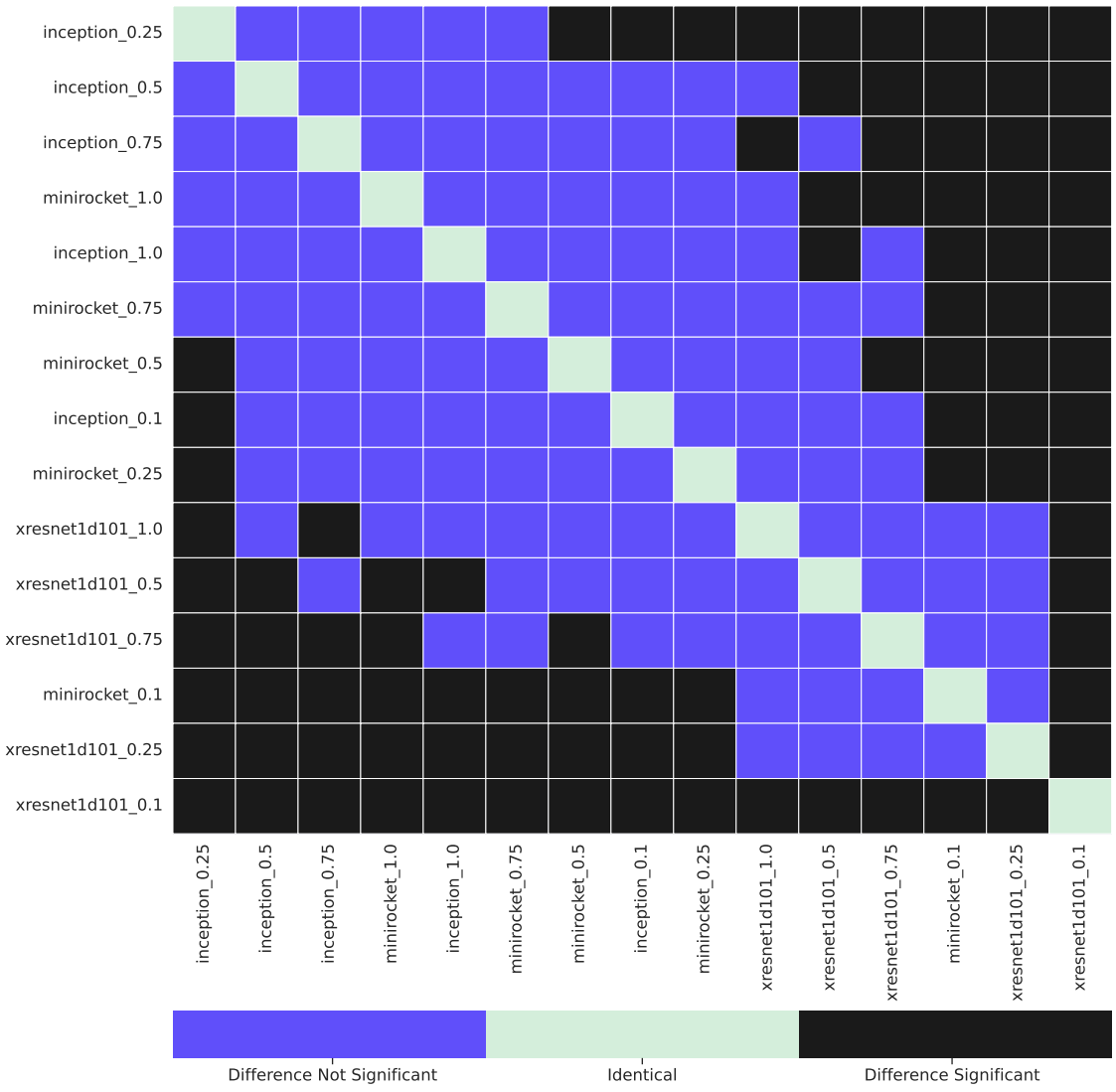

Figure 5: Model performance comparison for PTBXL