Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL

Nils Strodthoff*, Patrick Wagner*, Tobias Schaeffter and Wojciech Samek, *Member, IEEE*

***Abstract*— *Objective:* We put forward first benchmarking results for the recently published, freely accessible PTB-XL clinical 12-lead ECG dataset to address the lack of relieable benchmarking results that hampers the progress in the field of automatic ECG analysis. *Methods:* We define a variety of predictive tasks ranging from different ECG statement prediction tasks over age and gender prediction to signal quality assessment and compare the performance of different state-of-the-art deep learning and feature-based time series classifications algorithms. *Results:* We find that convolutional neural networks, in particular resnet- and inception-based architectures, show the strongest performance across all tasks outperforming feature-based algorithms by a large margin. We find consistent results on the ICBEB2018 challenge ECG dataset and discuss prospects of transfer learning using classifiers pretrained on PTB-XL. These benchmarking results are complemented by deeper insights into the classification algorithm in terms of hidden stratification, model uncertainty and an exploratory interpretability analysis, which provide connecting points for future research on the dataset. *Conclusion:* Our results emphasize the tremendous prospects of deep-learning-based algorithms in the field of automatic ECG interpretation not only in terms of quantitative accuracy but also in terms of further quality metrics such as uncertainty quantification and interpretability, which are of utmost for clinical applications. *Significance:* With this resource, we aim to establish the PTB-XL dataset as a resource for structured benchmarking of ECG analysis algorithms and encourage other researchers in the field to join these efforts.**

***Index Terms*— Decision support systems, Electrocardiography, Machine learning algorithms**

I. Introduction

CARDIOVASCULAR diseases (CVDs) rank among diseases of highest mortality [1] and were in this respect only recently surpassed by cancer in high-income countries [2]. Electrocardiography (ECG) is a non-invasive tool to assess the general cardiac condition of a patient and is therefore as first-in-line examination for diagnosis of CVD. In the US, during about 5% of the office visits an ECG was ordered or provided [3]. In spite of these numbers, ECG interpretation remains a difficult task even for cardiologists [4] but even more so for residents, general practioners [4], [5] or doctors in the emergency room who have to interprete ECGs urgently. A second major application area that will even grow in importance in the future is the telemedicine, in particular the monitoring of Holter ECGs. In both of these exemplary cases medical personnel could profit from significant reliefs if they were supported by advanced decision support systems relying on automatic ECG interpretation algorithms.

During recent years, we have witnessed remarkable advances in automatic ECG interpretation algorithms. In particular, deep-learning-based approaches have reached or even surpassed cardiologist-level performance for selected subtasks [6]–[9] or enabled statements that were very difficult to make for cardiologists e.g. to accurately infer age and gender from the ECG [10]. Due to the apparent simplicity and reduced dimensionality compared to imaging data, also the broader machine learning community has gained a lot of interest in ECG classification as documented by numerous research papers each year, see [11] for a recent review.

We see deep learning algorithms in the domain of computer vision as a role model for the deep learning algorithms in the field of ECG analysis. The tremendous advances for example in the field of image recognition relied crucially on the availability of large datasets and the competitive environment of classification challenges with clear evaluation procedures. In reverse, we see these two aspects as two major issues that hamper the progress in algorithmic ECG analysis: First, open ECG datasets are typically very small [12] and existing large datasets remain inaccessible for the general public. This issue has been at least partially resolved by the publication of the PTB-XL dataset [13], [14] hosted by PhysioNet [15], which provides a freely accessible ECG dataset of unprecedented size with predefined train-test splits based on stratified sampling. Second, the existing datasets typically provide only the raw data, but there exist no clearly defined benchmarking tasks with corresponding evaluation procedures. This severely restricts the comparability of different algorithms, as experimental details such as sample selection, train-test splits, evaluation metrics and score estimation can largely impact the final result. To address this second issue, we propose a range of different tasks showcasing the variability of the dataset ranging from the prediction of ECG statements over age and gender prediction to the assessment of signal quality. For these tasks,

This work was supported by the Bundesministerium fr Bildung und Forschung through the BIFOLD - Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref 01IS18037A) and by the EMPIR project 18HLT07 MedalCare.

Nils Strodthoff, Patrick Wagner and Wojciech Samek are with Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany (email: firstname.lastname@hhi.fraunhofer.de). Tobias Schaeffter is with Physikalisch-Technische Bundesanstalt, 10587 Berlin, Germany, Technical University Berlin, 10587 Berlin, Germany, and King's College London, London WC2R 2LS, UK (email: tobias.schaeffter@ptb.de).

*Both authors contributed equally to this work (Corresponding authors: Nils Strodthoff, Wojciech Samek).

we present first benchmarking results for deep-learning-based time series classification algorithms. We use the ICBEB2018 dataset to illustrate the promising prospects of transfer learning especially in the small dataset regime establishing PTB-XL as a pretraining resource for generic ECG classifiers, very much like ImageNet [16] in the computer vision domain.

Finally, assessing the quantitative accuracy is an important but by far not the only important aspect for decision support systems in the medical domain. To develop algorithms that create actual clinical impact, the topics of interpretability, robustness in a general sense and model uncertainty deserve particular attention. Such deeper insights, which go beyond benchmarking results, are discussed in the second part of the results section highlighting various promising directions for future research. In particular, we present a first evaluation of the diagnosis likelihood information provided within the dataset in comparison to model uncertainty as well as an outlook to possible applications of interpretability methods in the field.

II. Materials & methods

A. PTB-XL dataset

In this section, we briefly introduce the PTB-XL dataset [14] that underlies most experiments presented below. The PTB-XL dataset comprises 21837 clinical 12-lead ECG records of 10 seconds length from 18885 patients, where 52 % were male and 48 % were female. The ECG statements used for annotation are conform to the SCP-ECG standard [17] and were assigned to three non-mutually exclusive categories *diag.* (short for diagnostic), *form* and *rhythm*. In total, there are 71 different statements, which decompose into 44 diagnostic, 12 rhythm and 19 form statements. Note that there are 4 form statements that are also assigned to the set of diagnostic ECG statements. For diagnostic statements also a hierarchical organization into five coarse superclasses (*NORM*: normal ECG, *CD*: conduction disturbance, *MI*: myocardial infarction, *HYP*: hypertrophy and *STTC*: ST/T changes) and 24 sub-classes is provided, see Figure 1. For further details on the dataset and the annotation scheme, we refer the reader to the original publication [14]. To illustrate the versatility of tasks that can be addressed within the dataset, we also incorporate the further metadata provided, namely demographic information such as age and gender or signal quality as assessed by a technical expert.

B. Time series classification algorithms

For benchmarking different classification algorithms, we focus on algorithms that operate on raw multivariate time series data. An alternative class of algorithms operates on derived or transformed features such as Fourier or Wavelet coefficients or handcrafted features extract from single beats after beat segmentation, see [18], [19] for review in the context of ECG classification and [20] for (mostly univariate) time series classification in general. Deep learning approaches for time series classification are covered in a variety of recent, excellent reviews [21]–[23].

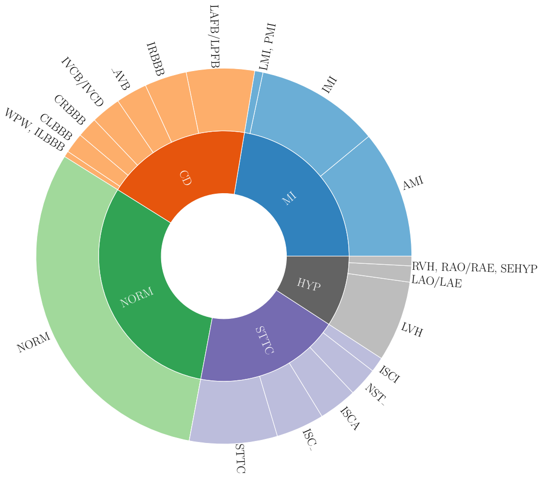


Fig. 1: Summary of the PTB-XL dataset in terms of diagnostic super and subclasses where the size of area represents the fraction of samples (figure reprinted from [14]).

We evaluate adaptations of a range of different algorithms from the literature that can be broadly categorized as follows, see Appendix I for experimental details:

- **convolutional neural networks**:
  - **standard**: fully convolutional [24], Deep4Net [25]
  - **resnet-based**: one-dimensional adaptations of standard resnets [24], [26], wide resnets [27] and xresnets [28]
  - **inception-based**: InceptionTime [29]
- **recurrent neural networks**: LSTM [30], GRU [31]
- **baseline classifiers**:
  - **feature-based:** Wavelet + shallow NN inspired by [32]
  - **naive**: predicting the frequency of each term in the training set

For reasons of clarity, we only report the performance for selected representatives including the best-performing method for each group. Typically the differences within the different groups are rather small. For completeness, the full results including all architectures are available in the accompanying code repository.

To encourage future benchmarking on this dataset, we release our repository[1] used to produce the results presented below along with instructions on how to evaluate the performance of custom classifiers in this framework. Finally, we would like to stress that the deep learning models were trained on the original time series data without any further preprocessing such as removing baseline wander and/or filtering, which are commonly used in literature approaches but introduce further hyperparameters into the approach.

C. Multi-label classification metrics

In this subsection, we review metrics for multi-label classification problems, see [33] for a review on multi-label classification metrics and algorithms. Multi-label classification

[1] https://github.com/nstrodt/PTBXL_benchmarking

metrics can be categorized broadly as *sample-centric* and *label-centric* metrics. The main difference between metrics from both categories is the question whether to first aggregate the scores across labels and then across samples or vice versa. To obtain a comprehensive view of the classification performance, we pick one exemplary metric from each category as proposed on theoretical grounds by [34]. Here, we focus on metrics that can be evaluated based on soft classifier outputs, where no thresholding has been applied yet, as this allows to get a more complete picture of the discriminative power of a given classification algorithm. In addition, it disentangles the selection of an appropriate classifier from the issue of threshold optimization, that will anyway have to be adjusted to match the clinical requirements rather than to optimize a certain global target metric.

a) Term-centric metrics: In general label-centric metrics are based on averages across class-specific metrics, which can further subdivided into micro- and macro-averages. In our setting, macro-averaging is preferred, since we expect class imbalance and do not want the score to be dominated by a few large classes. In addition, the distribution of pathologies in the dataset does not follow the natural distribution in the population but rather reflects the data collection process. Averaging class-wise AUCs over all classes yields the term-centric macro AUC (henceforth abbreviated as AUC), which we will use as primary evaluation metric.

b) Sample-centric metrics: Sample-centric evaluation metrics measure how accurately classification algorithms assign labels to a given sample, which is an information-retrieval point of view. For the selection of sample-centric metrics, we follow the evaluation procedures over the course of to-date three CAFA classification challenges [35]. The CAFA challenges address protein function prediction, which is also an inherent multi-label problem and shows strong structurally similarities to the task of ECG classification. For a given prediction $P_i(\tau)$ for given threshold $\tau \in [0, 1]$ and corresponding ground-truth annotations T_i, we can define sample-centric precision $\mathrm{pr}(\tau)$, recall/sensitivity $\mathrm{rc}(\tau)$

$$\mathrm{pr}(\tau) = \frac{1}{N_\tau} \sum_{i \in \mathcal{N}_\tau} \frac{\sum_f \mathbb{1}(f \in P_i(\tau) \wedge f \in T_i)}{\sum_f \mathbb{1}(f \in P_i(\tau))}\,,$$

$$\mathrm{rc}(\tau) = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_f \mathbb{1}(f \in P_i(\tau) \wedge f \in T_i)}{\sum_f \mathbb{1}(f \in T_i)}\,, \qquad (1)$$

where $\mathcal{N}_\tau = \{i \in 1, \ldots, N | \sum_f \mathbb{1}(f \in P_i(\tau)) > 0\}$ and $N_\tau = |\mathcal{N}_\tau|$. Here, we handle a possibly vanishing denominator when calculating the average precision in the same way as it is done in the CAFA challenges [36] by restricting the mean to the subset of samples with at least one prediction at the given threshold. Note that this procedure assumes a single threshold rather than class-dependent thresholds. We focus on Fmax as secondary performance metric, which was considered as main metric in the CAFA challenge. To this end, one defines a threshold-dependent F_1-score as the harmonic mean of precision and recall, i.e.

$$F_1(\tau) = \frac{2\,\mathrm{pr}(\tau)\mathrm{rc}(\tau)}{\mathrm{pr}(\tau) + \mathrm{rc}(\tau)}\,. \qquad (2)$$

TABLE I: Number of ECG statments per sample for a given level.

Level	# classes	0	1	2	3	≥4
diag.	44	407	15019	4242	1515	654
sub diag.	24	407	16272	4079	920	159
super-diag.	5	407	15239	4171	1439	581
form	19	12849	6693	1672	524	99
rhythm	12	771	20923	142	1	0
all	71	0	705	11247	5114	4771

To summarize $F_1(\tau)$ by a single number, the threshold is varied and the maximum score, from now on referred to as Fmax, is reported. As in the CAFA challenge, the threshold is optimized on the respective test set for each classification task and classifier under consideration. This procedure allow for a black-box evaluation just based on soft classifier outputs.

III. Benchmarking results on PTB-XL and ICBEB2018

PTB-XL comes with a variety of labels and further metadata. The presented experiments in this section serve two purposes: On the one hand, we provide first benchmarking results for future reference and, on the other hand, they illustrate the versatility of analyses that can be carried out based on the PTB-XL dataset. In Section III-A, we evaluate classifiers for different selections and granularities of ECG statements, which represents the core of analysis. It is complemented by Section III-B, where we validate our findings on the ICBEB2018 dataset and investigate aspects of transfer learning using PTB-XL for pretraining. Finally, we illustrate ways of leveraging further metadata within PTB-XL to construct age and gender prediction models, see Section III-C, and to build signal quality assessment models based on the provided signal quality annotations, see Section III-D.

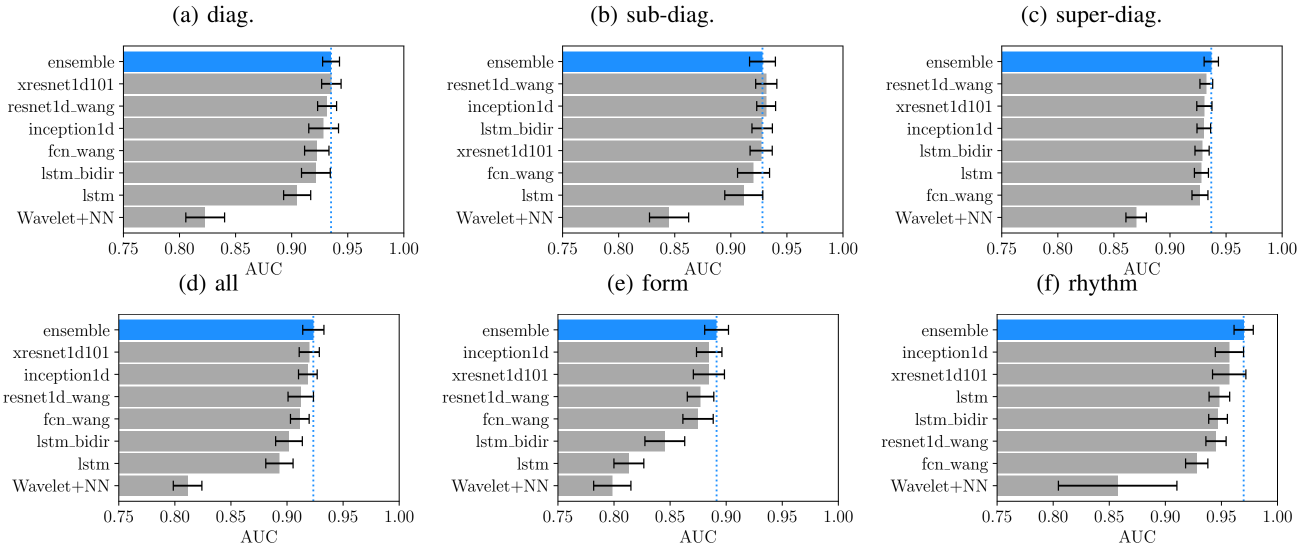
A. ECG statement prediction on PTB-XL

We start by introducing, performing and evaluating all experiments that are directly related to ECG-statements, where we cover the three different major categories diagnostic *diag.*, *form* and *rhythm* and level (*sub-diag.* and *super-diag.* as proposed in [14]) resulting in different number of labels per experiment and per sample as can seen in Table I. In the next step, we select only samples with at least one label in the given label selection. Our proposed evaluation as described in Section II-C is applied the same way for each experiment, where we report the term-centric macro-averaged AUC and the sample-centric Fmax-score.

In Table II, we report the results for all six experiments each applied to all models (as introduced in Section II-B), Figure 2 shows the result for all six experiments using barplots with associated bootstrap confidence intervals, see Appendix I for details. In all six experiments, deep-learning-based methods show a high predictive performance. Interestingly, even though all models are optimized based on binary cross-entropy loss rather than on the target metrics directly, the ranking according to both sample-based and term-based metrics largely coincides across all algorithms, which is why we focus on macro AUC

TABLE II: Overall discriminative performance of ECG classification algorithms on PTB-XL. For each experiment and each metric the best mean performing model is highlighted in bold font. For all experiments, 95% confidence intervals were calculating via bootstrapping on the test set, see Appendix I for notation and further details.

	all		diag.		sub-diag.		super-diag.		form		rhythm	
Method	AUC	Fmax	AUC	Fmax	AUC	Fmax	AUC	Fmax	AUC	Fmax	AUC	Fmax
lstm_bidir	.902(11)	.749(10)	.922(12)	.729(14)	.928(09)	.756(12)	.929(06)	.817(12)	.845(17)	.605(22)	.947(10)	.908(09)
lstm	.893(12)	.745(08)	.905(12)	.724(13)	.912(16)	.753(10)	.928(06)	.819(11)	.813(17)	.596(25)	.948(09)	.907(10)
fcn_wang	.911(10)	.754(08)	.922(10)	.731(14)	.920(14)	.752(11)	.927(07)	.815(12)	.875(18)	.625(23)	.928(10)	.899(11)
resnet1d_wang	.912(11)	.764(08)	.932(08)	.741(15)	**.932(09)**	.760(12)	**.932(06)**	**.825(12)**	.877(14)	.620(23)	.945(09)	.908(09)
xresnet1d101	**.920(08)**	**.765(08)**	**.935(08)**	**.743(13)**	.927(09)	.759(10)	.931(06)	.819(11)	**.885(13)**	**.629(20)**	**.957(20)**	.915(08)
Wavelet+NN	.811(14)	.678(10)	.823(19)	.627(15)	.845(17)	.654(14)	.870(10)	.731(13)	.798(21)	.526(22)	.857(52)	.866(13)
inception1d	.919(08)	**.765(07)**	.929(13)	.737(12)	**.932(08)**	**.763(10)**	.930(06)	.819(11)	**.885(14)**	.627(20)	**.957(14)**	**.917(09)**
ensemble	**.923(09)**	**.767(08)**	**.935(07)**	.740(12)	.928(11)	**.764(11)**	**.937(06)**	**.827(12)**	**.891(12)**	**.638(23)**	**.970(08)**	.916(08)
naive	*.500(00)*	*.557(11)*	*.500(00)*	*.440(18)*	*.500(00)*	*.440(18)*	*.500(00)*	*.448(09)*	*.500(00)*	*.365(19)*	*.500(00)*	*.797(13)*



(a) diag. (b) sub-diag. (c) super-diag. (d) all (e) form (f) rhythm

Fig. 2: Graphical summary of experiments described in Section III-A. For comparability, the algorithms are ranked according to prediction performance in each category.

in the following. The best-performing resnet or inception-based models reach macro AUCs ranging from 0.89 in the *form* category, over around 0.93 in the *diagnostic* categories to 0.96 in the *rhythm* category. These performance metrics can in principle used for a rudimentary assessment of the difficulty of the different prediction tasks. However, one has to keep in mind that for example the *form* prediction task has a considerably smaller training set compared to the other experiments due to approximately 12k ECGs without any *form* annotations.

As first general observation upon investigating the different model performances in more detail, we find that resnet-architectures and inception-based architectures perform best across all experiments, but all convolutional architectures show a comparable performance level. In fact, the results of all convolutional models, up to very few exceptions, remain compatible within error bars. Recurrent architectures are consistently slightly less performant than their convolutional counterparts but, at least for diagnostic and rhythm statements, still competitive. The second general observation is that the performances of both convolutional as well as recurrent deep learning models turn out to be considerably stronger than the performance of the baseline algorithm operating on wavelet features. However, this statement has to be taken with caution, as the performance of feature-based classifiers is typically rather sensitive to details of feature selection choice of derived and details of the proprocessing procedure.

In addition to single-model-performance, we also report the performance of an ensemble formed by averaging the predictions of all considered models (except the naive model). As can be seen in Table II, ensembling leads in many case to slight performance increases, but the best-performing single resnet or inception models always remain compatible with the ensemble result within error bars. The largest performance improvement of the ensemble model compared to single model performance is observed in the *rhythm* category, where the ensemble model outperforms all convolutional models except for xresnet1d101 and inception1d (as can be seen in Figure 2f). The ensemble results are only supposed to serve as rough orientation as the focus of this work is on single-model performance.

As a final remark, throughout this paper we use the recommended train-test splits provided by PTB-XL [14], which consider patient assignments and use input data at a sampling frequency of 100 Hz. Deviations from this setup are investigated in Appendix II.

TABLE III: Classification performance on the ICBEB2018 dataset. In addition to sample-centric Fmax and term-centric macro-AUC, we also report the term-centric $F_{\beta=2}$ and $G_{\beta=2}$ to be used in the PhysioNet/CinC challenge 2020.

Method	AUC	Fmax	$F_{\beta=2}$	$G_{\beta=2}$
lstm	0.953(07)	0.804(20)	0.770(31)	0.536(37)
lstm_bidir	0.954(13)	0.828(17)	0.781(30)	0.556(31)
xresnet1d101	**0.970(06)**	**0.862(16)**	**0.821(33)**	**0.607(39)**
resnet1d_wang	0.968(07)	0.847(19)	0.790(31)	0.578(36)
fcn_wang	0.959(07)	0.824(21)	0.771(31)	0.552(36)
inception1d	0.962(08)	0.835(18)	0.798(34)	0.575(35)
Wavelet+NN	0.905(14)	0.691(25)	0.665(26)	0.414(29)
naive	0.500(00)	0.289(23)	0.376(05)	0.119(01)
ensemble	***0.972(06)***	*0.852(18)*	*0.809(26)*	*0.591(33)*

B. ECG statement prediction on ICBEB2018 and transfer learning

Beyond analyses on the PTB-XL dataset itself, we see further application of it as generic pretraining resource for ECG classification task, in a similar way as ImageNet [16] is commonly used for pretraining image classification algorithms. One freely accessible dataset from the literature that is large enough to reliably quantify the effects of transfer learning is the ICBEB2018 dataset, which is based on data released for the 1st China Physiological Signal Challenge 2018 held during the 7th International Conference on Biomedical Engineering and Biotechnology (ICBEB 2018) [37]. It comprises 6877 12-lead ECGs lasting between 6 s and 60 s. Each ECG record is annotated by up to three statements by up to three reviewers taken from a set of nine classes (one normal and eight abnormal classes, see Figure 3). We use the union of labels turning the dataset into a multi-label dataset. As the original test set is not available, we define 10 cross-validation folds by stratified sampling preserving the overall label distribution in each fold following [14].

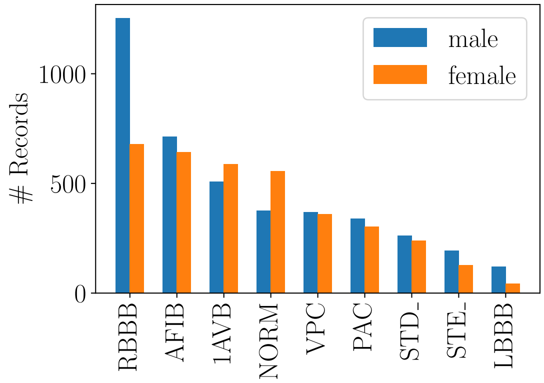


Fig. 3: Summary of the ICBEB2018 dataset [37] in terms of ECG statements.

We start by analyzing the classification performance of classifiers trained on ICBEB2018 from scratch as an independent validation of the results obtained on PTB-XL. Table III shows the performance of classifiers that were trained using the the same experimental setup as in Section III-A. In all cases, we train a classifier from scratch by training on the first eight folds using the ninth and tenth fold as validation and test sets, respectively. Interestingly, the ICBEB2018 dataset was recently selected as training dataset for the PhysioNet/CinC challenge 2020 [2]. For this reason we also report two further label-based performance metrics that will supposedly serve as evaluation metrics in the challenge, namely a macro-averaged F_β-score ($\beta = 2$) and a macro-averaged G_β-score with $\beta = 2$, where $G_\beta = TP/(TP + FP + \beta \cdot FN)$, in both cases with sample weights chosen inversely proportional to the number of labels. Values of $\beta > 1$ allow to assign more weight to recall than precision, which might be a desirable property. However, applying this equally to the *NORM*-class seems questionable since high precision is required in this case. In addition, the corresponding scores are sensitive to the chosen classification threshold, which we determine by maximizing the F_β/G_β-score on the training set, which is an undesirable aspect as it entangles the discriminitive performance of the classification algorithm with the process of threshold determination. Nevertheless, both F_β and G_β show a quantitative similarity in terms of ranking between our threshold-free metrics. Comparing to the quantitative classification performance on PTB-XL as presented in Section III-A, we see a largely consistent picture on ICBEB2018 in the sense of a rather uniform performance level among the convolutional architectures, all of which remain consistent within error bars, a slightly weaker performance of the recurrent architectures and a considerable performance gap to the feature-based baseline classifier.

In the next experiment, we leverage PTB-XL by finetuning a classifer trained on PTB-XL on ICBEB2018 data. To this end, we take a classifier trained on PTB (using *all* ECG statements) and replace the top layer of the fully connected classification head to account for the different number ECG statements in ICBEB2018. This classifier is then finetuned on ICBEB2018 data. To systematically investigate the transition into the small dataset regime, we do not only present results for finetuning on the full dataset (8 training folds) but for the full range of one eighth to eight training folds i.e. from 85 to 5500 training samples. For each training size and fixed model architecture (xresnet1d101), we compare models trained from scratch to models that pretrained on PTB-XL and then finetuned on ICBEB2018. Figure 4 summarizes the results of this experiment, and illustrates the fact that for large dataset sizes pretraining on PTB-XL does not improve the performance compared to training from scratch but even potentially slightly deteriorates it, even though the two results remain compatible within error bars. However, for smaller dataset sizes of a single training fold or fractions of it, we see a clear advantage from pretraining. Most notably, the performance of the finetuned model remains much more stable upon decreasing the size of the training set and consequently outperforms the model trained from scratch by a large margin in the the case of small training sizes. In the most extreme case of one eighth of the original training fold corresponds to 85 samples, where the performance of the finetuned classifier only drops by about 10% in terms of AUC compared to a classifier trained on a training set that is 64 times larger. Since the

[2] https://physionetchallenges.github.io/2020/

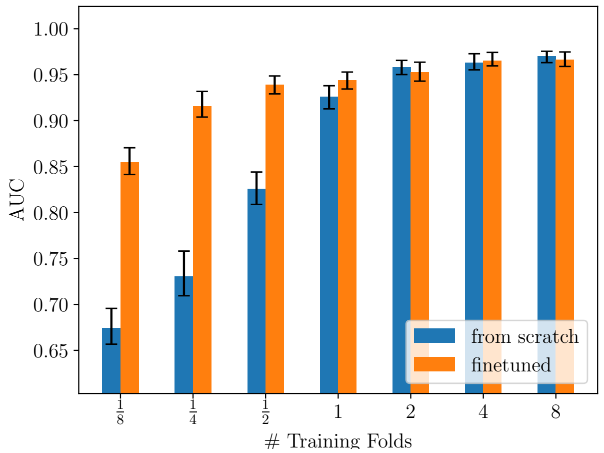


Fig. 4: Effect of transfer learning from PTB-XL to ICBEB2018 upon varying the size of the ICBEB2018 training set.

TABLE IV: Age regression performance for models trained on all patients and evaluated on all/healthy/non-healthy subpopulation in terms of mean absolute error (MAE) and R-squared (R2).

	all		healthy		non-healthy	
Method	MAE	R2	MAE	R2	MAE	R2
lstm	7.54(22)	.703(17)	7.22(32)	.688(28)	7.78(28)	.541(42)
lstm_bidir	7.42(20)	.709(22)	7.07(31)	.696(24)	7.69(26)	.550(39)
xresnet1d101	7.35(22)	.713(19)	6.93(26)	.711(25)	7.68(27)	.543(44)
resnet1d_wang	7.17(18)	**.728(20)**	**6.86(30)**	**.721(24)**	7.41(25)	.573(37)
fcn_wang	7.28(23)	.719(21)	6.96(27)	.712(23)	7.54(24)	.557(43)
inception1d	**7.16(18)**	**.728(21)**	6.89(30)	.715(25)	**7.38(21)**	**.580(38)**
ensemble	**7.12(20)**	**.734(19)**	**6.80(28)**	**.724(24)**	**7.37(26)**	**.586(33)**

small dataset regime is the most natural application domain for pretraining on a generic ECG dataset, we see this as a very encouraging sign for future applications of PTB-XL as a pretraining resource for relatively small datasets.

C. Age regression and gender classification

The following experiment is inspired by the recent work from [10] that demonstrated that deep neural networks are capable of accurately inferring age and gender from standard 12-lead ECGs. Here, we look into both tasks again based on PTB-XL. The experiment is supposed to illustrate the possibility of leveraging demographic metadata in the PTB-XL dataset. We applied the same model architectures from Section III-A but with adjusted final layers, where for gender prediction a binary and for age prediction a linear output neuron was trained and optimized such that the binary cross-entropy or mean squared error is minimized respectively. Both networks were trained separately but with the same train-test-splits and identical hyperparameters as in previous experiments, except that for final output prediction where we computed the mean of all windows instead of the maximum (as used above). In order to study the effect of pathologies on performance for this task, in addition to all subjects we also evaluated the models only for healthy subjects and for non-healthy subjects. Here, we define the set of healthy records as the set of records with *NORM* as the only diagnostic label and non-healthy as its complement.

TABLE V: Gender prediction performance for models trained on all patients and evaluated on all/healthy/non-healthy subpopulations in terms of accuracy (acc) and area under the receiver operating curve (AUC).

	all		healthy		non-healthy	
Method	ACC	AUC	ACC	AUC	ACC	AUC
lstm	.833(14)	.911(11)	.886(18)	.952(11)	.785(20)	.874(17)
lstm_bidir	.838(14)	.908(12)	.893(19)	.954(14)	.796(23)	.868(17)
xresnet1d101	**.849(14)**	**.920(10)**	**.898(19)**	**.960(10)**	**.806(17)**	**.881(17)**
resnet1d_wang	.840(14)	.909(11)	.895(21)	.955(12)	.799(19)	.869(18)
fcn_wang	.832(13)	.909(11)	.882(22)	.949(12)	.796(22)	.875(18)
inception1d	.836(15)	.916(09)	.896(20)	.958(12)	.787(18)	.876(15)
ensemble	.847(15)	**.928(09)**	.896(22)	**.962(11)**	.801(17)	**.894(15)**

The results for the age regression experiment are shown in Table IV. Overall, testing only on healthy subjects yielded better results in each category as compared to testing only on non-healthy or all subjects (MAE=6.86 compared to MAE=7.38 and MAE=7.16 respectively). These observations are in line with [10], [38]. Furthermore, these results are competitive to [10], who reported a value of MAE=6.9 years (R-squared = 0.7) but with thirty times more data (≈20k versus ≈750k samples [10]). Table V shows the corresponding results for gender prediction. As already suggested in [39], [40] the differences between male and female are also present in ECG, which is also confirmed by our model yielding a accuracy of 84.9%(89.8%) and an AUC of 0.92(0.96) on all(healthy) patients. This performance level, in particular on the healthy subpopulation, is competitive with results from the literature [10] (90.4% accuracy and an AUC of 0.97). As a final word of caution, we want to stress that the results for age and gender prediction algorithms are not directly comparable across different datasets due to different dataset distributions not only in terms of the labels themselves but also in terms of co-occurring diseases. This is apparent from the performance differences of our classifier for both subtasks when evaluated on the full dataset and on the two different subpopulations.

D. Signal quality assessment

As part of a technical validation of the database each sample underwent a second iteration by a technical expert to annotate the data with respect to signal artifacts, see [14] for a detailed description. The annotations were given without any regular syntax, for this reason the annotations were coded as a binary targets, where targets are set to one if any annotation is given for *NOISE* (either globally present static noise (*static_noise*) or local bursts of high voltage induced by external sources (*burst_noise*)), *DRIFT* (baseline wandering). In total this binary target is set for ≈ 22% (i.e. ≈ 78% of signals contain no artifacts). Using these annotations coded as binary targets might help to develop a signal quality classifier for creating validation data to test for robustness with respect to artifacts. For this purpose we conducted experiments along the lines of Section III-A i.e. again using the same models, hyperparameters and train-test-splits as above. Overall, our models reach AUC scores around 0.81, which seems to indicate a slightly weaker predictive performance compared to ECG statement prediction models discussed in Section III-A, even though performance measures for different tasks are obviously not directly comparable. According to a first analysis, a significant portion of this performance deficiency can be attributed label noise (i.e. missing annotation in case of artifacts or misleading annotation in case of normal signals). However, a more thorough analysis should attempt to incorporate the full report strings instead of just binary labels. In any case, models trained on this task can still be used as a prescreening procedure for ECG quality assessment.

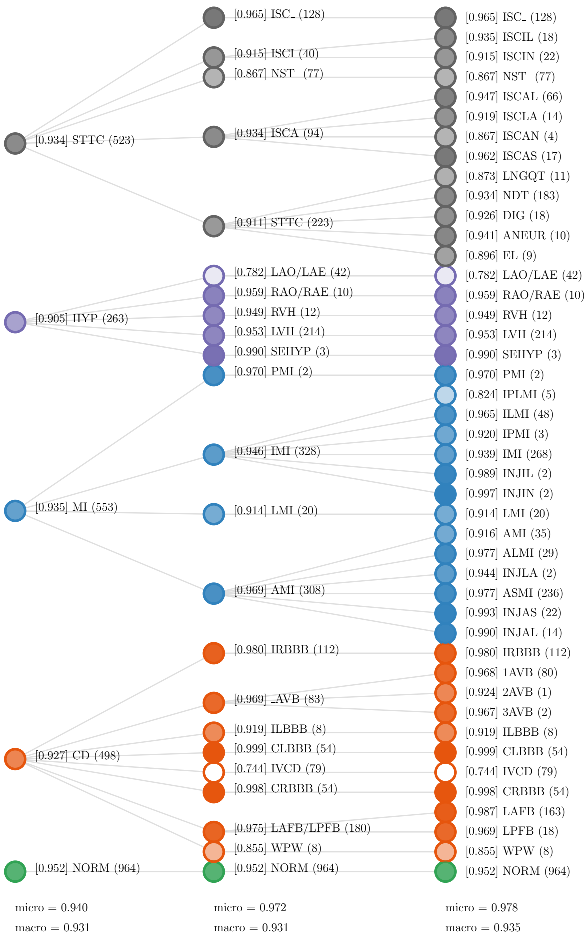


Fig. 5: Hierarchical decomposition of class-specific AUCs onto subclasses and individual diagnostic statements exhibiting hidden stratification, i.e. inferior algorithmic performance on certain diagnostic subpopulations that remains hidden when considering only the superior superclass performance, see the description in Section IV-B for details. AUC is given in square brackets and the number of label occurrences in the test set in parentheses. The transparency of each colored node is relative to the minimum and maximum AUC in the last layer.

IV. Deeper insights from classification models

Until now we investigated our experiments quantitatively in order to compare different model architectures. However, a quantitatively evaluation focusing on overall predictive performance, as presented in the previous section, might not take important qualitative aspects into account, such as the predictive performance for single, potentially sparsely populated ECG statements. Here, we focus our analysis on a single xresnet1d101 model, but we verified that the results presented below are largely consistent across different model architectures.

A. Hierarchical organization of diagnostic labels

As first analysis, we cover the hierarchical organization of diagnostic labels and its impact on predictive performance. The PTB-XL dataset provides proposed assignements to one of five superclasses and one of 23 subclasses for each diagnostic ECG statement, which represents one possible ontology that can be used to organize ECG statements. In Figure 5, we show the hierarchical decomposition for the diagnostic labels in sub- and superclasses, where we propagated predictions from experiment *diag.* upwards the hierarchy over *sub-diag.* to *super-diag.* by summing up prediction probabilities of the corresponding child nodes and limiting the output probabilities to one. We experimented with other aggregation strategies such as using the maximum or the mean of the predictions of the child nodes but observed only minor impact on the results. The same holds for models trained on the specific level, where no propagation is needed. The training of hierarchical classifiers is a topic with a rich history in the machine learning literature, see for example [41] for a dedicated review and [42] for a recent deep learning approach to the topic. Extensive experiments on this topic are beyond the scope of this manuscript, but our first experiments on this topic indicate that the performance of a model trained on a coarser granularity is largely compatible or in some cases even slightly inferior to a model trained on the finest label granularity and propagating prediction scores upwards the label hierarchy.

B. Hidden stratification and co-occurring pathologies

The hierarchical organization of the diagnostic labels allows for deeper insights and potential pitfalls of model evaluation that are crucial for clinical applications. In particular, we focus on the issue of *hidden stratification* that was put forward in [43] and describes potential inferior algorithmic performance on certain diagnostic subpopulations that remains hidden from the outside if only the superclass performance is reported. We analyze this effect in a top-down fashion using the results obtained by propagating the finest granularity scores upwards the label hierarchy as described above. In Figure 5, we illustrate how the label AUC of a particular superclass or subclass decomposes into the label AUCs of the corresponding subclasses. One reason for weak classifier performance are ECG statements classes that are too scarcely populated to allow training a discriminative classifier on them and for which also the score estimate on the test set is unreliable due to the

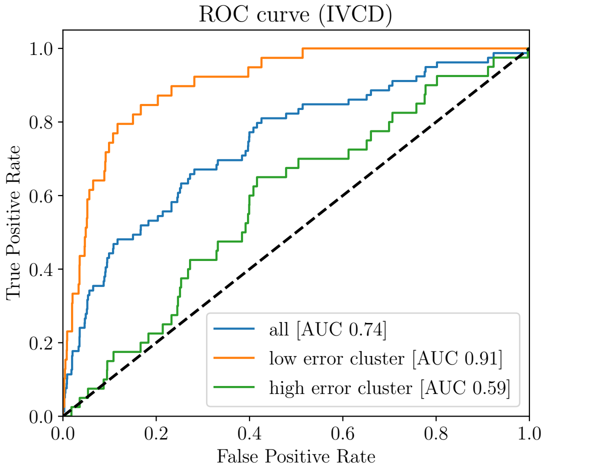


Fig. 6: AUC curves for two subset of samples revealing hidden stratification within the *IVCD* class.

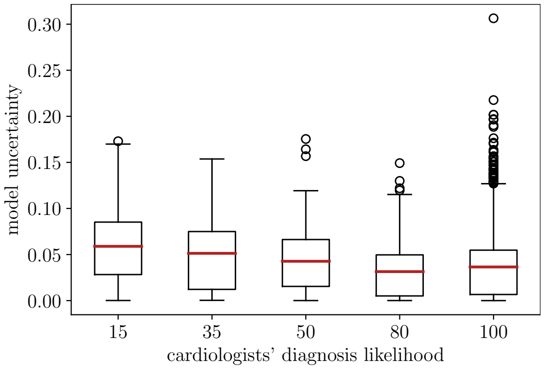


Fig. 7: Relation between model uncertainty (standard deviation of ensemble predictions as in [44]) and diagnosis likelihood as quantified by the annotating cardiologist, see Section IV-C for details.

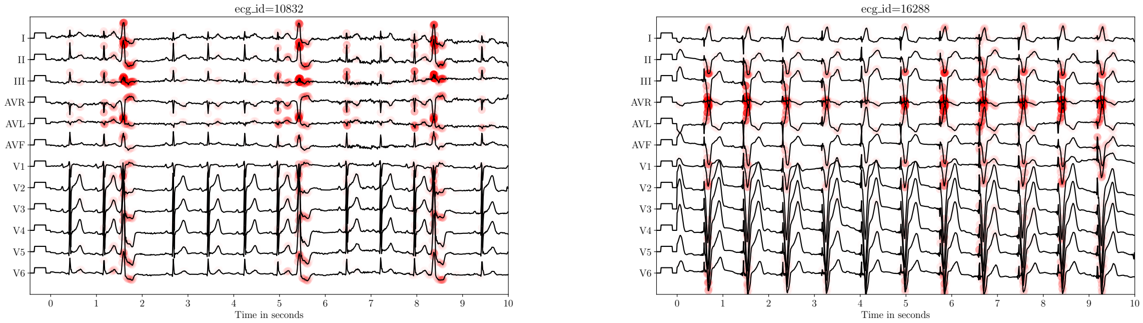
small sample size. However, there are further ECG statements that stand out from other members of the same subclass, where the performance deficiency cannot only be attributed to effects of small sample sizes. For example, consider the classes *NST_* (non-specific ST changes), *LAO/LAE* (left atrial overload/enlargement) and *IVCD* (non-specific intraventricular conduction disturbance (block)) in the bottom layer of the hierarchy, where the classifier shows a weak performance, which is in fact hidden when reporting only the corresponding superclass or subclass performance measures. At least for *NST_* and *IVCD*, these findings can be explained by the fact that both statements are by definition non-specific ECG statements and potentially subsum rather heterogenous groups of findings.

Although identifying hidden stratification is straightforward to identify in hindsight given the hierarchical organization of the diagnostic labels, [43] also demonstrated how to identify groups of samples exhibiting hidden stratification for a given class label under consideration using an unsupervised clustering approach. For demonstration, we carried out such a comparable analysis for *IVCD* in order to understand the comparably weak classification performance on the particular statement compared to other conduction disturbances. Indeed, clustering the model's output probabilities revealed two clusters, where one subset performed much better than the other as can be seen in Figure 6. Interestingly, it turned out that the two clusters largely align with the presence/absence of *NORM* as additional ECG statement. The blue line (all) represents the performance as is (AUC 0.74), the green line is the performance for samples out of one cluster (AUC 0.59, for which most of the sample were also associated with *NORM*), the orange line for the second cluster (AUC 0.91, predominantly samples without *NORM*). As can be seen clearly, samples with *IVCD* in combination with *NORM* are much harder to classify.

These kinds of investigations are very important for the identification of hidden stratification in the model which are induced by data and their respective labels [43]. Models trained on coarse labels might hide this kind of clinically relevant stratification, because of both subtle discriminative features and low prevalence. Further studies might investigate hidden stratification below our deepest level of labels. At this point, it remains to stress that the PTB-XL dataset does not provide any clinical ground truth on the considered samples but only provides cardiologists' annotations based on the ECG signal itself, which could compromise the analysis. However, we still see an in-depth study towards the identification subgroups with certain combinations of co-occurring ECG statements/pathologies, along the lines of the example of *IVCD* presented above, as a promising direction for future research in the sense that it can potentially provide pointers for future clinical investigations.

C. Model uncertainty and diagnosis likelihoods

Besides this hierarchical organization of diagnostic labels, PTB-XL comes along with associated likelihoods for each diagnostic label ranging from 15 to 100, where 15 indicates less and 100 strong confidence for one label. These likelihoods were extracted from the original ECG report string for all diagnostic statements based on certain keywords [14]. As an initial experiment to assess the quality of this likelihood information, we compare the likelihoods to model uncertainty estimates for a model trained on diagnostic statements. To quantify the model uncertainty, we follow the simple yet very powerful approach put forward in [44] that defines model uncertainty via the variance of an ensemble of identical models for different random initializations. Here, we use an ensemble of 10 models and for simplicity even omit the optional stabilizing adversarial training step, which was reported to lead to slightly improved uncertainty estimates [44], in this first exploratory analysis. In Figure 7, we plot model uncertainty versus diagnosis likelihood and observe the expected monotonic behavior. Only the likelihood 100 stands out from this trend and shows a large number of outliers. One possible explanation for this observation is an overconfidence of human annotators when it comes to seemingly very obvious statements that goes in with the human inability to precisely quantify uncertainties, which is a well-known phenomenon in cognitive psychology, see e.g. [45]. However, we perceive the overall alignment of diagnosis likelihood with model uncertainty as an interesting observation



(a) PVC

(b) PACE

Fig. 8: Two exemplary attribution maps for a resnet model for the classes PVC (left) and PACE (right).

as it correlates perceived human uncertainty with algorithmic uncertainty, a statement that is normally impossible for clinical datasets due to the unavailability of appropriate labels.

D. Prospects of interpretability methods

The acceptance of machine learning and in particular deep learning algorithms in the clinical context is often limited by the fact that data-driven algorithms are perceived as black boxes by doctors. In this direction, the recent advances in the field of explainable AI has the prospect to at least partially alleviate this issue. In particular, we consider post-hoc interpretability that can be applied for a trained model, see e.g. [46]. The general applicability of interpretability methods to multivariate timeseries and in particular ECG data was demonstrated in [47], see also [48], [49] for futher accounts on interpretability methods for ECG data. Here, we focus on exemplary for the form statement “premature ventricular complex” (*PVC*) and the rhythm statement *PACE* indicating an active pacemaker. The main reason for choosing these particular classes is the easy verifiable also for non-cardiologists. In Figure 8, we show two exemplary but representative attribution maps obtained via the ϵ-rule with $\epsilon = 0.1$ within the framework of layer-wise relevance propagation [50]. For *PVC* the relevance is located at the extra systole across all leads. For *PACE*, the relevance is scattered across the whole signal aligning nicely with the characteristic pacemaker spikes (just before each QRS complex) in each beat. It is a non-trivial finding that the relevance patterns for the two ECG statements from above align with medical knowledge. A more extensive, statistical analysis of the attribution maps both within patients across different beats and across different ECGs with common pathologies is a promising direction for future work.

V. Summary and conclusions

Electrocardiography is among the most common diagnostic procedures carried out in hospitals and doctor’s offices. We envision a lot potential for automatic ECG interpretation algorithms in different medical application domains, but we see the current progress in the field hampered by the lack of appropriate benchmarking datasets and well-defined evaluation procedures. We propose a variety of benchmarking tasks based on the PTB-XL dataset [14] and put forward first baseline results for deep-learning-based time classification algorithms that are supposed to guide future reasearchers working on this dataset. We find that convolutional, in particular resnet- and inception-based, architectures show the best performance but recurrent architectures are also competitive for selected prediction tasks. Furthermore, we demonstrate the prospects of transfer learning by finetuning a classifier pretrained on PTB-XL on a different target dataset, which turns out to be particularly effective in the small dataset regime. Finally, we provide different directions for further in-depth studies on the dataset ranging from the analysis of co-occurring pathologies, over the correlation of human-provided diagnosis likelihoods with model uncertainties to the application of interpretability methods.

References

[1] E. Wilkins, L. Wilson, K. Wickramasinghe, P. Bhatnagar, J. Leal, R. Luengo-Fernandez *et al.*, *European Cardiovascular Disease Statistics 2017*. Belgium: European Heart Network, 2 2017.

[2] G. R. Dagenais, D. P. Leong, S. Rangarajan, F. Lanas, P. Lopez-Jaramillo, R. Gupta *et al.*, “Variations in common diseases, hospital admissions, and deaths in middle-aged adults in 21 countries from five continents (PURE): a prospective cohort study,” *The Lancet*, Sep. 2019.

[3] CDC, “National Ambulatory Medical Care Survey: 2016 National Summary Tables,” Centers for Disease Control and Prevention, Tech. Rep., 2019.

[4] S. M. Salerno, P. C. Alguire, and H. S. Waxman, “Competency in interpretation of 12-lead electrocardiograms: A summary and appraisal of published evidence,” *Annals of Internal Medicine*, vol. 138, no. 9, p. 751, May 2003.

[5] G. Fent, J. Gosai, and M. Purva, “Teaching the interpretation of electrocardiograms: Which method is best?” *Journal of Electrocardiology*, vol. 48, no. 2, pp. 190–193, Mar. 2015.

[6] Z. I. Attia, C. V. DeSimone, J. J. Dillon, Y. Sapir, V. K. Somers, J. L. Dugan *et al.*, “Novel bloodless potassium determination using a signal-processed single-lead ECG,” *Journal of the American Heart Association*, vol. 5, no. 1, Jan. 2016.

[7] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia *et al.*, “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” *Nature Medicine*, vol. 25, no. 1, pp. 65–69, 2019.

[8] Z. I. Attia, P. A. Noseworthy, F. Lopez-Jimenez, S. J. Asirvatham, A. J. Deshmukh, B. J. Gersh *et al.*, “An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction,” *The Lancet*, vol. 394, no. 10201, pp. 861–867, Sep. 2019.

[9] Z. I. Attia, S. Kapa, F. Lopez-Jimenez, P. M. McKie, D. J. Ladewig, G. Satam *et al.*, "Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram," *Nature Medicine*, vol. 25, no. 1, pp. 70–74, Jan. 2019.

[10] Z. I. Attia, P. A. Friedman, P. A. Noseworthy, F. Lopez-Jimenez, D. J. Ladewig, G. Satam *et al.*, "Age and sex estimation using artificial intelligence from standard 12-lead ECGs," *Circulation: Arrhythmia and Electrophysiology*, vol. 12, no. 9, Sep. 2019.

[11] S. Hong, Y. Zhou, J. Shang, C. Xiao, and J. Sun, "Opportunities and challenges in deep learning methods on electrocardiogram data: A systematic review," *arXiv preprint arXiv:2001.01550*, 2020.

[12] J. Schläpfer and H. J. Wellens, "Computer-Interpreted Electrocardiograms," *Journal of the American College of Cardiology*, vol. 70, no. 9, pp. 1183–1192, Aug. 2017.

[13] P. Wagner, N. Strodthoff, R.-D. Bousseljot, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," *PhysioNet*, https://doi.org/10.13026/6sec-a640, 2020.

[14] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek *et al.*, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, 2020, in press.

[15] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[17] ISO Central Secretary, "Health informatics – Standard communication protocol – Part 91064: Computer-assisted electrocardiography," International Organization for Standardization, Geneva, CH, Standard ISO 11073-91064:2009, 2009.

[18] N. Maglaveras, T. Stamkopoulos, K. Diamantaras, C. Pappas, and M. Strintzis, "ECG pattern recognition and classification using non-linear transformations and neural networks: A review," *International Journal of Medical Informatics*, vol. 52, no. 1-3, pp. 191–208, Oct. 1998.

[19] E. H. Houssein, M. Kilany, and A. E. Hassanien, "ECG signals classification: a review," *International Journal of Intelligent Engineering Informatics*, vol. 5, no. 4, p. 376, 2017.

[20] A. Bagnall, A. Bostrom, J. Large, and J. Lines, "The great time series classification bake off: An experimental evaluation of recently proposed algorithms. extended version," *arXiv preprint arXiv:1602.01711*, 2016.

[21] J. C. B. Gamboa, "Deep learning for time-series analysis," *arXiv preprint arXiv:1701.01887*, 2017.

[22] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data Mining and Knowledge Discovery*, pp. 1–47, 2019.

[23] R. Yannick, B. Hubert, A. Isabela, G. Alexandre, F. Jocelyn *et al.*, "Deep learning-based electroencephalography analysis: a systematic review," *arXiv preprint arXiv:1901.05498*, 2019.

[24] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 International joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 1578–1585.

[25] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann *et al.*, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, aug 2017.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[28] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 558–567.

[29] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber *et al.*, "Inceptiontime: Finding alexnet for time series classification," *arXiv preprint arXiv:1909.04939*, 2019.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734.

[32] L. D. Sharma and R. K. Sunkaria, "Inferior myocardial infarction detection using stationary wavelet transform and machine learning approach," *Signal, Image and Video Processing*, vol. 12, no. 2, pp. 199–206, Jul. 2017.

[33] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[34] X.-Z. Wu and Z.-H. Zhou, "A unified view of multi-label performance measures," in *ICML*. JMLR, 2017, pp. 3780–3788.

[35] N. Zhou, Y. Jiang, T. R. Bergquist, A. J. Lee, B. Z. Kacsoh, A. W. Crocker *et al.*, "The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens," *Genome Biology*, vol. 20, no. 1, Nov. 2019.

[36] Y. Jiang, T. R. Oron, W. T. Clark, A. R. Bankapur, D. D'Andrea, R. Lepore *et al.*, "An expanded evaluation of protein function prediction methods shows an improvement in accuracy," *Genome Biology*, vol. 17, no. 1, Sep. 2016.

[37] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu *et al.*, "An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection," *Journal of Medical Imaging and Health Informatics*, vol. 8, no. 7, pp. 1368–1373, Sep. 2018.

[38] R. L. Ball, A. H. Feiveson, T. T. Schlegel, V. Starc, and A. R. Dabney, "Predicting heart age using electrocardiography," *Journal of personalized medicine*, vol. 4, no. 1, pp. 65–78, 2014.

[39] M. Malik, K. Hnatkova, D. Kowalski, J. J. Keirns, and E. M. van Gelderen, "Qt/rr curvatures in healthy subjects: sex differences and covariates," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 305, no. 12, pp. H1798–H1806, 2013.

[40] G. Salama and G. C. Bett, "Sex differences in the mechanisms underlying long qt syndrome," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 307, no. 5, pp. H640–H648, 2014.

[41] C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining and Knowledge Discovery*, vol. 22, no. 1-2, pp. 31–72, Apr. 2010.

[42] J. Wehrmann, R. Cerri, and R. Barros, "Hierarchical multi-label classification networks," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmssan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 5075–5084.

[43] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. R, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in *Machine Learning for Health (ML4H) at NeurIPS 2019 - Extended Abstract*, 2019.

[44] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in neural information processing systems*, 2017, pp. 6402–6413.

[45] P. D. Windschitl and G. L. Wells, "Measuring psychological uncertainty: Verbal versus numeric methods." *Journal of Experimental Psychology: Applied*, vol. 2, no. 4, p. 343, 1996.

[46] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer International Publishing, 2019.

[47] N. Strodthoff and C. Strodthoff, "Detecting and interpreting myocardial infarction using fully convolutional neural networks," *Physiological Measurement*, vol. 40, no. 1, p. 015001, jan 2019.

[48] J. van der Westhuizen and J. Lasenby, "Techniques for visualizing lstms applied to electrocardiograms," in *ICML Workshop on Human Interpretability in Machine Learning*, 2018.

[49] S. Vijayarangan, B. Murugesan, V. R, P. SP, J. Joseph, and M. Sivaprakasam, "Interpreting deep neural networks for single-lead ecg arrhythmia classification," *arXiv preprint 2004.05399*, 2020.

[50] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, 2015.

[51] J. Howard *et al.*, "fast.ai," http://fast.ai, 2018.

[52] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 328–339.

[53] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay," *arXiv preprint arXiv:1803.09820*, 2018.

[54] I. Loshchilov and F. Hutter, "Fixing Weight Decay Regularization in Adam," *International Conference on Learning Representations (ICLR)*, 2019.

[55] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[56] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[57] Z. Cui, W. Chen, and Y. Chen, "Multi-scale convolutional neural networks for time series classification," *arXiv preprint arXiv:1603.06995*, 2016.

[58] G. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt, and A. O'Leary, "PyWavelets: A python package for wavelet analysis," *Journal of Open Source Software*, vol. 4, no. 36, p. 1237, Apr. 2019.

[59] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet," *Biomedizinische Technik/Biomedical Engineering*, pp. 317–318, 1995.

[60] O. Kwon, J. Jeong, H. B. Kim, I. H. Kwon, S. Y. Park, J. E. Kim *et al.*, "Electrocardiogram sampling frequency range acceptable for heart rate variability analysis," *Healthcare informatics research*, vol. 24, no. 3, pp. 198–206, 2018.

Appendix I
Experimental details

In general, our implementations follow the implementations of the architectures described in the original publications and reference implementations as closely as possible. The most significant modification in our implementations is the use of a concat-pooling layer [51] as pooling layer, which aggregates the result of a global average pooling layer and a max pooling layer along the feature dimension. For resnets we enlarge the kernel sizes as this slightly improved the performance, consistent with observations in the literature [24], [29]. All convolutional models then use the same fully connected classification head with a single hidden layer with 128 hidden units, batch normalization and dropout of 0.25 and 0.5 at the first/second fully connected layer, respectively. For recurrent neural networks we use concat pooling as in [52]. For reference, we typically report the performance of both unidirectional and bidirectional recurrent models, in our case LSTMs/GRUs with two layers and 256 hidden units. As we are dealing with a multi-label classification problem, we optimize binary cross-entropy. We use 1-cycle learning rate scheduling during training [53] and the AdamW optimizer [54]. During finetuning a pretrained classifier for transfer learning from PTB-XL to ICBEB2018, we use gradual unfreezing and discriminative learning rates [51], [52] to avoid catastrophic forgetting i.e. overwriting information captured during the initial training phase on PTB-XL. Deep-learning models were implemented using PyTorch [55], fast.ai [51] and Keras [56]. We release our implementations in the accompanying code repository.

During training, we follow the sliding window approach that is commonly used in time series classification, see e.g. [23], [25], [47], [57]. Here, the classifier is trained on random segments of fixed length taken from the full record. This allows to easily incorporate records of different length (as it is the case for ICBEB2018) and effectively serves as data augmentation. During test time, we use test time augmentation. This means we divide the record into segments of the given window size that overlap by half of the window size and obtain model predictions for each of the segments. These predictions are then aggregated using the element-wise maximum (or mean in case of age and gender prediction) in order to produce a single prediction for the whole sample. This procedure considerably increases the overall performance compared to the performance on random sliding windows without any aggregation. If not mentioned otherwise, we use a fixed window size of 2.5 seconds.

Besides our end-to-end trainable models we also compare to classic machine learning models, where a classifier is trained on precomputed statistical features such as wavelets. Here we loosely follow [32] and train a classifier on wavelet features. More specifically, we compute a multilevel 1d discrete wavelet transform (Daubechies `db4`) for each lead independently leveraging the implementation from [58]. From the resulting coefficients we compute a variety of statistical features such as entropy, 5%, 25%, 75% and 95% percentiles, median, mean, standard deviation, variance, root of squared means, number of zero and mean crossings. Different from [32], the features were then used to train a shallow neural network with a single hidden layer in order to be able to address multi-label classification problems with a large number of classes in a straightforward manner. Note that the classifier from [32] included a number of additional features and preprocessing steps and might therefore lead to an improved score compared to our implementation.

Again following the example of the CAFA challenge, we provide 95% confidence intervals via empirical bootstrapping on the test set, in our case with 1,000 iterations. More specifically, we report the point estimate from evaluating on the whole test set and estimate lower and upper confidence intervals using the bootstrap examples. In summary tables, we typically report only the point estimate and the maximal absolute deviation between point estimate and lower and upper bound, where for example $0.743(09)$ is supposed to be understood as 0.743 ± 0.009. We deliberately decided not to exclude sparsely populated classes from the evaluation. Due to the stratified sampling procedure underlying the fold assignments in [14] point estimates can evaluated for all metrics. However, during the bootstrap process it is not guaranteed that at least one positive sample for each class is contained in each bootstrap sample. In such a case, metrics such as the term-centric macro-AUC cannot be evaluated. To circumvent this issue, we discard such bootstrap samples and redraw until we find at least one positive sample for each class. For metrics such as sample-centric Fmax that can be evaluated without any constraints on the bootstrap samples, we verified empirically that this procedure only marginally impacts the corresponding confidence intervals. For later reference, we store the selection of bootstrap samples, evaluate the confidence intervals for all algorithms on this fixed set of samples and provide this as part of our code repository for later reference.

Appendix II
Train-test splits and sampling frequency

In this section, we investigate the impact of two crucial experimental parameters on the classification performance, namely the effect of using random splits disregarding patient assignments and using 500 Hz compared to 100 Hz data as input.

TABLE VI: Investigating impact of random train-test splits (disregarding patient assignments) and increasing the temporal resolution (sampling frequency of 500 Hz) compared to the setup used throught this article (train-test splits considering patient assignments and a sampling frequency of 100 Hz).

	strat. & 100Hz		strat. 500Hz		rnd. 100Hz	
Method	AUC	Fmax	AUC	Fmax	AUC	Fmax
xresnet1d101	.931(06)	.819(11)	**.933(06)**	.821(11)	.938(05)	.827(11)
Wavelet+NN	.870(10)	.731(13)	.852(08)	.709(12)	–	–
lstm	.928(06)	.819(11)	.922(06)	.809(11)	–	–
inception1d	.930(06)	.819(11)	.931(05)	**.824(10)**	.936(05)	**.829(10)**
fcn_wang	.927(07)	.815(12)	.919(06)	.806(10)	–	–
lstm_bidir	.929(06)	.817(12)	.919(06)	.807(13)	–	–
resnet1d_wang	**.932(06)**	**.825(12)**	.922(06)	.809(09)	**.939(05)**	.828(12)
ensemble	**.937(06)**	**.827(12)**	.931(05)	.818(11)	**.942(04)**	**.834(12)**
naive	*.500(00)*	*.448(09)*	*.500(00)*	*.448(12)*	*.500(00)*	*.455(13)*

Concerning the first aspect, we noticed that many literature approaches in the field ECG analysis perform train-test splits using random splits on individual ECGs or even individual beats rather than patients. This leads to a systematic overestimation of generalization performance, since during prediction on the test set the model can exploit training data for patients with multiple ECGs having both samples in train and test split. We substantiate this claim by comparing the results from Section III-A with models trained on random splits for the experiment *super-diag.*. By random we refer to random splits based on ECGs rather than patients but with the same splitting procedure i.e. stratified sampling in order to ensure balanced label distributions. And even more importantly, we also maintain two *clean* folds for validation and testing i.e. only samples where *validated_by_human* is set to true. This point is important since splitting disregarding clean folds yields even better results, which might be attributed to slight mismatches in the distribution of the clean folds and the remaining training folds. Table VI shows the results for this experiment, where the overestimation becomes apparent when comparing the ensemble model for example where in terms of both metrics the performance is increased only by this effect. This is a strong indication for our claim that most of results reported in previous literature overestimates generalization performance due to a data leakage arising from having one patient in both training and test set. The effect is even more pronounced for datasets where the fraction of test set samples of patients with records already contained in the training set is larger, as observed by [32] for the original PTB diagnostic ECG dataset [59]. We tested only those samples where the leakage was most severe (i.e. patients having most ECGs in training data) and observed a significant gain in performance and gap in loss confirming our claim.

Finally, we investigate the prospects of using input data at a higher sampling frequency of 500 Hz compared to 100 Hz that was used throught the rest of the manuscript. To investigate this claim, we applied the same pipeline as described in Section III-A to input data sampled at 500 Hz. Also our inputs to the networks were adjusted such that each window consists of five seconds, this results in 2500 timesteps (as compared to 500 timesteps for 100 Hz). In order to compensate for the higher sampling frequency we also increased the filter size of convolutional filters by a factor of five, while the rest of hyperparameters (number of filter, layers etc.) were left unchanged compared to the case of 100 Hz. In our experiments, we found no compelling evidence for this apprehension, i.e. no significant gain in performance in all our metrics for diagnostic tasks, as can be seen in Table VI. This observation is in accordance with [60], where the authors also came to the conclusion that 100 Hz is still sufficient for models operating in time domain but not for models in frequency domain. For this reason it is might not even be a fair comparison to compare our baseline Wavelet model for 100 Hz as it was done in Section III-A. Although this seems plausible, additional considerations might be necessary to compare this issue in a more reliable way. Nevertheless we believe that the gain in performance will be negligible for our models, since there are more obvious issues affecting performance, e.g. dealing with all sorts of artifacts and label noise, preprocessing and more resilient and effective training procedures.