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

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

***Abstract*— Electrocardiography (ECG) is a very common, non-invasive diagnostic procedure and its interpretation is increasingly supported by algorithms. The progress in the field of automatic ECG analysis has up to now been hampered by a lack of appropriate datasets for training as well as a lack of well-defined evaluation procedures to ensure comparability of different algorithms. To alleviate these issues, we put forward first benchmarking results for the recently published, freely accessible clinical 12-lead ECG dataset PTB-XL, covering a variety of tasks from different ECG statement prediction tasks to age and sex prediction. Among the investigated deep-learning-based timeseries classification algorithms, we find that convolutional neural networks, in particular resnet- and inception-based architectures, show the strongest performance across all tasks. 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. Our results emphasize the prospects of deep-learning-based algorithms in the field of ECG analysis, not only in terms of quantitative accuracy but also in terms of clinically equally important further quality metrics such as uncertainty quantification and interpretability. 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 practitioners [4], [5] or doctors in the emergency room who have to interpret 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]–[10] or enabled statements that were very difficult to make for cardiologists e.g. to accurately infer age and sex from the ECG [11]. 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 [12] 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 [13] 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 [14], [15] hosted by PhysioNet [16], which represents the to-date largest freely accessible ECG dataset. In addition, many of the freely accessible databases contain only single lead recordings, which makes comprehensive diagnosis and clinical validation difficult. Large and comprehensive databases with 12-lead recordings, however, are rather an exception such as [17] focusing on arrhythmia, which is why the underlying data set is of great importance for the development of algorithmic solutions. 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 comparabil-

This work was supported by the Bundesministerium für 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).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3022989, IEEE Journal of Biomedical and Health Informatics

ity 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 to age and sex prediction. For these tasks, 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 [18] 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 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. To summarize, our main contributions in this paper are the following:

- We propose different benchmarking tasks on the recently published PTB-XL dataset [15] ranging from ECG statement prediction from different subsets of ECG statements and label granularities to age and sex prediction within a structured evaluation methodology.
- We implement and adapt different state-of-the-art deep-learning-based time series classification algorithms and adapt recent image classification algorithms to the ECG context including a new resnet-adaptation (*xresnet1d101*) that turns out to be the best-performing algorithm across all tasks. For full reproducibility, we release the full source code, trained models and the benchmarking infrastructure in a corresponding code repository [19].
- We provide the first reliable assessment of transfer learning in the ECG context demonstrating the promising prospects of transfer learning from PTB-XL to other ECG classification datasets in the small dataset regime.
- We provide evidence for the phenomenon of hidden stratification, a first evaluation of the diagnosis likelihood information provided within the dataset in comparison to model uncertainty and present an outlook to possible applications of interpretability methods in the field.

II. Materials & methods

A. PTB-XL and IBEB2018 datasets

In this section, we briefly introduce the PTB-XL dataset [15] 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 [20] and were assigned to three non-mutually exclusive categories *diag* (short for diagnostic statements such as “anterior myocardial infarction”), *form* (related to notable changes of particular segments within the ECG such as “abnormal QRS complex”) and *rhythm* (related to particular changes of the rhythm such as “atrial fibrillation”). In total, there are 71 different statements, which decompose into 44 diagnostic, 12 rhythm and 19 form statements, 4 of which are also used as diagnostic ECG statements. For diagnostic statements also a hierarchical organization into five coarse superclasses and 24 sub-classes is provided, see Figure 1 for a graphical summary in terms of diagnostic superclasses. For further details on the dataset, the annotation scheme, and other ECG datasets we refer the reader to the original publication [15]. To summarize, PTB-XL does not only stand out by its size as the to-date largest publicly accessible clinical ECG dataset but also through its rich set of ECG annotations and further metadata, which turns the dataset into an ideal resource for the training and evaluation of machine learning algorithms. Throughout this paper we use the recommended train-test splits provided by PTB-XL [15], which consider patient assignments and use input data at a sampling frequency of 100 Hz.

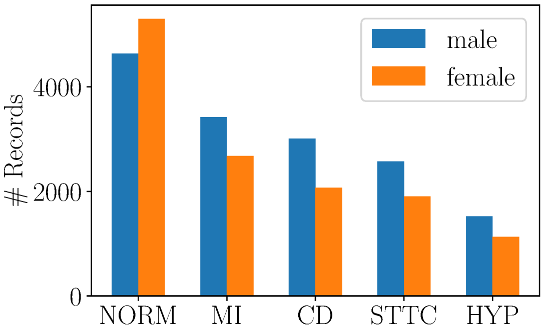


Fig. 1: Summary of the PTB-XL dataset [15] in terms of diagnostic superclasses, where *NORM*: normal ECG, *MI*: myocardial infarction, *CD*: conduction disturbance, *STTC*: ST/T-changes, *HYP*: hypertrophy.

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 [18] 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) [21]. It comprises 6877 12-lead ECGs lasting between 6 and 60 seconds. 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 2). We use the union of labels turning the dataset into a multi-label dataset. As the original test set is not available, we divide the original training sets into 10 folds by stratified sampling preserving the overall label distribution in each fold following [15].

By default for both datasets, 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.

B. Time series classification algorithms

In this work, we address different prediction task based on ECG data,

- inferring diagnostic ECG statements at three different label granularites (multilabel classification),
- inferring ECG statements related to the rhythm of the ECG signal (multilabel classification),
- inferring ECG statements related to the form of the signal (multilabel classification),
- inferring a subject's sex (binary classification),
- inferring a subject's age (regression),

all of which can be broadly characterized as time series classification/regression tasks. For benchmarking different classification algorithms, we focus on algorithms that operate on raw multivariate time series data. Deep learning approaches for time series classification are covered in a variety of recent, excellent reviews [22]–[24].

We evaluate adaptations of a range of different algorithms from the literature that can be broadly categorized as convolutional neural networks and recurrent neural networks, both operating on the raw ECG signal, as well as feature-based approaches. The convolutional neural networks can be further subdivided into standard feed-forward architectures, resnet-based architectures and inception-based architectures. Standard feed-forward architectures include fully convolutional networks [25] (*fcn_wang*) or Deep4Net [26] (*schirrmeister*), which is often used for EEG classification. Resnet-based architectures were proposed for time series classification already a while ago, see e.g. [25] (*resnet1d_wang*), and were successfully applied in different large-scale studies [10], [27]. Here we also propose and evaluate a range of additional one-dimensional adaptations of resnet-based architectures inspired by recent improved resnet-architectures such as xresnets [28] (*xresnet1dxxx*). For comparison, we also adapt standard resnet [25], [29] (*resnet1d_wang*, *resnet1dxxx*) and wide resnet [30] (*wrn1d_22*) architectures. As final convolutional architecture we report on InceptionTime [31] (*inception1d*), an adaptation of the popular inception architecture to the time series domain. 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 [32] 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 slightly enlarge the kernel sizes to 5 as this slightly improved the performance, consistent with observations in the literature [25], [31]. 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. As for recurrent neural networks, we consider unidirectional and bidirectional LSTMs [33] and GRUs [34] (*lstm*,*gru*,*lstm_bidir*,*gru_bidir*) with two layers and 256 hidden units, whose outputs are aggregated using a concat pooling layer [35]. 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 [19]. Finally, in addition to single-model-performance, we also report the performance of an ensemble formed by averaging the predictions of all considered models. The ensemble results are only supposed to serve as rough orientation as the focus of this work is on single-model performance.

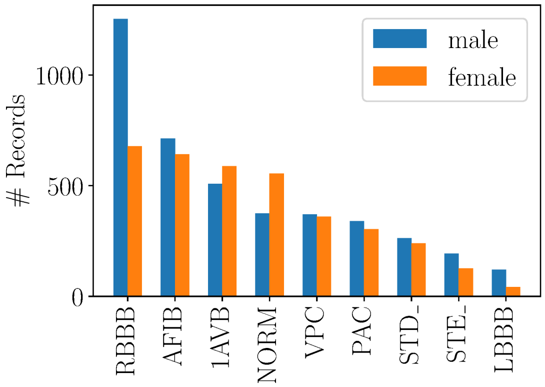


Fig. 2: Summary of the ICBEB2018 dataset [21] in terms of ECG statements, where *RBBB*/*LBBB*: right/left bundle branch block, *AFIB*: atrial fibrillation, *1AVB*: first-degree AV block, *NORM*: normal ECG, *VPC*: ventricular premature complex, *STD_*: non-specific ST depression, *STE_*: non-specific ST elevation.

With the sole exception of sex prediction, where we use mean-squared error as loss functions, we optimize binary cross-entropy, which is appropriate for multi-label classification problems. We use 1-cycle learning rate scheduling during training [36] and the AdamW optimizer [37]. During finetuning a pretrained classifier for transfer learning from PTB-XL to ICBEB2018, we use gradual unfreezing and discriminative learning rates [32], [35] to avoid catastrophic forgetting i.e. overwriting information captured during the initial training phase on PTB-XL. Deep-learning models were implemented using PyTorch [38], fast.ai [32] and Keras [39]. We release our implementations in the accompanying code repository [19].

During training, we follow the sliding window approach that is commonly used in time series classification, see e.g. [24], [26], [40], [41]. 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 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 sex 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.

Feature-based approaches, where a classifier is trained on precomputed statistical features such as Fourier or Wavelet coefficients have been the predominant approach in the ECG analysis literature until fairly recently, see [42], [43] for reviews. Similar to the image domain, there is increasing evidence that deep learning algorithms trained in an end-to-end fashion are able to outperform feature-based approaches [41], [44]. To the best of our knowledge, there is no public implementation of a state-of-the-art feature-based algorithm for multi-channel ECG classification. Nevertheless, we wanted to present a feature-based approach for comparability and loosely followed [45] and trained a classifier on wavelet features. More specifically, we compute a multilevel (5 level) 1d discrete wavelet transform (Daubechies `db6`) for each lead independently leveraging the implementation from [46]. 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 [45], the features were then used to train a shallow neural network with a single hidden layer (*Wavelet+NN*) in order to be able to address multi-label classification problems with a large number of classes in a straightforward manner.

To encourage future benchmarking on this dataset, we release our repository [19] 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.

III. Benchmarking results on PTB-XL and ICBEB2018

A. Tasks and Metrics

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-B, we evaluate classifiers for different selections and granularities of ECG statements, which represents the core of our analysis. It is complemented by Section III-C, 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 sex prediction models, see Section III-D.

As primary performance metric for all classification experiments, we report the macro-averaged area under the receiver operating characteristic curve (henceforth referred to as AUC), which is obtained by averaging class-wise AUCs over all classes. 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. 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. As a final comment, metrics for multi-label classification problems are a wide field, see [47] for a review on multi-label classification metrics and algorithms. On the most fundamental level, one distinguishes sample-centric and label-centric metrics (such as macro AUC), see e.g. [48]. For completeness, we also provide results for sample-centric metrics in the code repository.

For ICBEB2018, which was recently selected as training dataset for the PhysioNet/CinC challenge 2020, we report for reasons of comparability two further performance metrics that were used as evaluation metrics during the first stage of 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 discriminative performance of the classification algorithm with the process of threshold determination.

To assess the uncertainty of the classifiers' scores, we provide 95% confidence intervals via empirical bootstrapping on the test set, in our case with 10,000 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 this case, non-overlapping confidence intervals signify statistically significant differences between the classifiers whereas the converse is not true [49]. To circumvent this issue, we also calculate bootstrap estimates of the difference of the best-performing and all other classifiers. If the corresponding confidence intervals for the difference do not cover zero, the two classifiers are considered statistically significant at a confidence level of 0.05 in this case. 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 [15] point estimates can be 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

TABLE I: Overall discriminative performance of ECG classification algorithms on PTB-XL in terms of macro AUC. For each experiment the best-performing single model is underlined and marked in bold face. All models that do not perform statistically significantly worse than the best-performing model are also marked in bold face. The ensemble score is underlined if the ensemble performs statistically significantly better than the best single model. For all results, we also indicate 95% confidence intervals obtained via bootstrapping on the test set.

Method	all	diag.	sub-diag.	super-diag.	form	rhythm
inception1d	**.925(08)**	**.931(09)**	**.930(10)**	.921(06)	**.899(22)**	**.953(13)**
xresnet1d101	**.925(07)**	**.937(08)**	**.929(14)**	**.928(05)**	**.896(12)**	**.957(19)**
resnet1d_wang	.919(08)	**.936(08)**	**.928(10)**	**.930(05)**	.880(15)	**.946(10)**
fcn_wang	.918(08)	.926(10)	**.927(11)**	.925(06)	.869(12)	.931(08)
lstm	.907(08)	.927(08)	**.928(10)**	.927(05)	.851(15)	**.953(09)**
lstm_bidir	.914(08)	.932(07)	**.923(12)**	.921(06)	.876(15)	**.949(11)**
Wavelet+NN	.849(13)	.855(15)	.859(16)	.874(07)	.757(29)	.890(24)
ensemble	**.929(07)**	**.939(08)**	**.933(11)**	**.934(05)**	**.907(12)**	**.965(07)**

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.

B. 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 [15]). For each experiment, we select only samples with at least one label in the given label selection.

In Table I, we report the results for all six experiments each applied to all models (as introduced in Section II-B). In all six experiments, deep-learning-based methods show a high predictive performance. 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 results can in principle be 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 and significantly outperform other investigated convolutional architectures for selected experiments such as *all*, *diag.* and *form*. These results support well-known findings from the imaging domain in the sense that more modern convolutional architectures involving skip-connections such as resnet- or inception-based architectures allow to train deeper and more performant models as compared to standard feed-forward convolutional architectures. Architectural improvements seem to carry over from the imaging domain, for example xresnets perform significantly better than their standard resnet counterparts. Across all categories, the newly proposed *xresnet1d101*-model either represents the best-performing result or does not perform significantly worse than the best-performing model. In this context, it is worth noting that it performs on par or even outperforms the recently proposed *inception1d*-model that was specifically engineered toward time series classification. Recurrent architectures are consistently slightly less performant than their convolutional counterparts but, at least for sub-diagnostic and rhythm statements, still competitive.

The second general observation is that the performances of both convolutional as well as recurrent deep learning models is significantly better than the performance of the baseline algorithm operating on wavelet features in line with literature results [41], [44]. 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 preprocessing procedure. Note that the classifier from [45] included a number of additional features and preprocessing steps and might therefore lead to an improved score compared to our implementation. We also tested for different classifiers like decision trees or support vector machines, different input features based on Fourier coefficients. However, we did not observe any improvements compared to the presented baseline result. A detailed comparison between deep-learning-based and feature-based approaches is beyond the scope of this manuscript. In fact, the results for feature-based approaches were only included to provide a rough orientation for the reader.

As third observation, it is notable that forming ensemble models leads in many case to slight performance increases. However, only in the super-diagnostic case the ensemble outperforms the best-performing single model in a statistically significant manner.

C. ECG statement prediction on ICBEB2018 and transfer learning

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 II shows the performance of classifiers that were trained using the the same experimental setup as in Section III-B. Nevertheless, both F_β and G_β show a quantitative similarity in terms of ranking compared to our threshold-free AUC metric. Comparing to the quantitative classification performance on PTB-XL as presented in Section III-B, we see a largely consistent

TABLE II: Classification performance on the ICBEB2018 dataset. In addition to macro-AUC, we also report the term-centric $F_{\beta=2}$ and $G_{\beta=2}$ used in the PhysioNet/CinC challenge 2020. The notation follows the one introduced in Table I.

Method	AUC	$F_{\beta=2}$	$G_{\beta=2}$
inception1d	.963(09)	**.802(30)**	**.581(40)**
xresnet1d101	**.974(05)**	**.806(31)**	**.587(37)**
resnet1d_wang	.969(06)	**.794(31)**	**.572(35)**
fcn_wang	.957(08)	.764(30)	.541(38)
lstm	.964(06)	.768(32)	.538(36)
lstm_bidir	.959(11)	.779(31)	.544(35)
Wavelet+NN	.905(14)	.664(33)	.407(35)
naive	.500(00)	.368(06)	.115(00)
ensemble	**.975(05)**	**.820(30)**	**.598(39)**

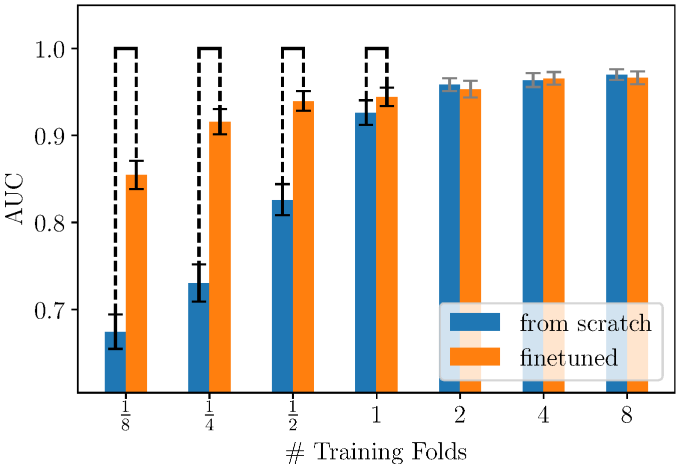


Fig. 3: Effect of transfer learning from PTB-XL to ICBEB2018 upon varying the size of the ICBEB2018 training set. Cases where the difference between pretraining and training from scratch is statistically significant are marked by dashed brackets.

picture on ICBEB2018 in the sense of a similar ranking of the different classifiers. A noticeable observation is the fact that the relative performance of the *xresnet101*-model is even stronger in the sense that it outperforms all other classifiers in terms of macro AUC in a statistically significant manner, whereas it remains compatible with *inception1d* and *resnet1d_wang* in terms of $F_{\beta=2}$ and $G_{\beta=2}$. .

In the next experiment, we leverage PTB-XL by finetuning a classifier 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. The nine ECG statements from ICBEB2018 are a subset of the ECG statements used in PTB-XL, but the latter annotation scheme is much more exact distinguishing complete from incomplete bundle branch blocks and differentiating ST-elevation into non-specific ST-elevations and myocardial infarction or ischemic ECG changes. Here we aim to provide a proof-of-concept for finetuning of ECG classifiers. A detailed investigation of the impact of the differences between the label distributions between the source and the target dataset is therefore beyond the scope of this article.

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 and assess the statistical significance of the corresponding performance differences. Figure 3 summarizes the results of this experiment, and illustrates the fact that pretraining and training from scratch do not deviate significantly for large dataset sizes. However, the performance gap between both approaches widens for smaller training sets and becomes significant (even at a confidence level below 0.0005) for a single training fold or fractions of it. 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 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.

D. Age regression and sex classification

The following experiment is inspired by the recent work from [11] that demonstrated that deep neural networks are capable of accurately inferring age and sex from standard 12-lead ECGs. The motivation for such a prediction task is two-fold: on the one hand, it illustrates the diagnostic potential within ECGs that is difficult to extract for humans but can be uncovered with algorithmic support. On the other hand, in particular the age inference task potentially allows to reach a better understanding of age-induced changes in the ECG.

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-B but with adjusted final layers, where for sex 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 normal subjects and for abnormal subjects. Here, we define the set of normal records as the set of records with *NORM* as the only diagnostic label and the set of abnormal records as its complement.

The results for the age regression experiment are shown in Table III. Overall, testing only on normal subjects yielded better results in each category as compared to testing only on abnormal or all subjects (MAE=6.86 compared to MAE=7.38 and MAE=7.16 respectively). These observations are in line with [11], [50]. In this experiment, the best-performing models are *inception1d* and *resnet1d_wang*, which outperform *xreset1d101* in certain subcategories. Furthermore, the results are competitive in comparison to [11], who reported a value of MAE=6.9 years (R-squared = 0.7) but with thirty times more data (≈20k versus ≈750k samples [11]).

TABLE III: Age regression performance for models trained on all patients and evaluated on all/normal/abnormal subpopulations in terms of mean absolute error (MAE) and R-squared (R2). The notation follows the one introduced in Table I.

	all		healthy		non-healthy	
Method	MAE	R2	MAE	R2	MAE	R2
inception1d	**7.16(19)**	**.728(19)**	**6.89(28)**	**.715(29)**	**7.38(27)**	**.580(40)**
xresnet1d101	7.35(20)	.713(20)	**6.93(29)**	**.711(28)**	7.68(28)	.543(46)
resnet1d_wang	**7.17(19)**	**.728(19)**	**6.86(28)**	**.721(28)**	**7.41(27)**	**.573(43)**
fcn_wang	**7.28(20)**	**.719(18)**	**6.96(29)**	**.712(27)**	**7.54(27)**	.557(42)
lstm	7.54(20)	.703(19)	7.22(30)	.688(29)	7.78(27)	.541(43)
lstm_bidir	**7.42(20)**	**.709(19)**	**7.07(31)**	**.696(30)**	**7.69(27)**	**.550(43)**
ensemble	**7.12(19)**	**.734(17)**	**6.80(29)**	**.724(27)**	**7.37(26)**	**.586(39)**

Table IV shows the corresponding results for sex prediction. As already suggested in [51], [52] 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(normal) patients. This performance level, in particular on the normal subpopulation, is competitive with results from the literature [11] (90.4% accuracy and an AUC of 0.97). In the sex prediction experiment, *xresnet1d101* represents the best-performing model across all subpopulations but with consistently high scores for *inception1d* and even bidirectional recurrent architectures (*lstm_bidir*). As a final word of caution, we want to stress that the results for age and sex 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.

TABLE IV: Sex prediction performance for models trained on all patients and evaluated on all/normal/abnormal subpopulations in terms of accuracy (acc) and area under the receiver operating curve (AUC). The notation follows the one introduced in Table I.

	all		healthy		non-healthy	
Method	ACC	AUC	ACC	AUC	ACC	AUC
inception1d	.834(13)	**.916(09)**	**.896(16)**	**.958(10)**	.785(18)	**.876(16)**
xresnet1d101	**.849(12)**	**.920(09)**	**.900(15)**	**.960(09)**	**.806(18)**	**.881(16)**
resnet1d_wang	**.840(12)**	.909(10)	**.892(16)**	.955(10)	**.797(18)**	.869(17)
fcn_wang	.831(13)	.909(10)	.882(17)	.949(11)	**.796(18)**	**.875(16)**
lstm	.834(13)	.911(10)	.884(16)	.952(10)	**.789(19)**	**.874(16)**
lstm_bidir	**.837(12)**	**.908(10)**	**.893(16)**	**.954(10)**	**.796(19)**	**.868(17)**
ensemble	**.845(12)**	**.928(08)**	**.898(16)**	**.962(09)**	**.805(18)**	**.894(15)**

IV. Deeper insights from classification models

Until now we investigated our experiments quantitatively in order to compare different model architectures. However, a quantitative 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 assignments 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 4, we show the hierarchical decomposition (tree-like structure) 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 [53] for a dedicated review and [54] 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 [55] and describes potential inferior algorithmic performance on certain diagnostic subpopulations that remains hidden if only the superclass performance is reported. In Figure 4, we illustrate how the label AUC of a particular superclass or subclass decomposes into the label AUCs of the corresponding subclasses. One trivial reason for weak classifier performance are ECG statement 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 small sample size. However, there are further ECG statements that stand out from other members of the same subclass,

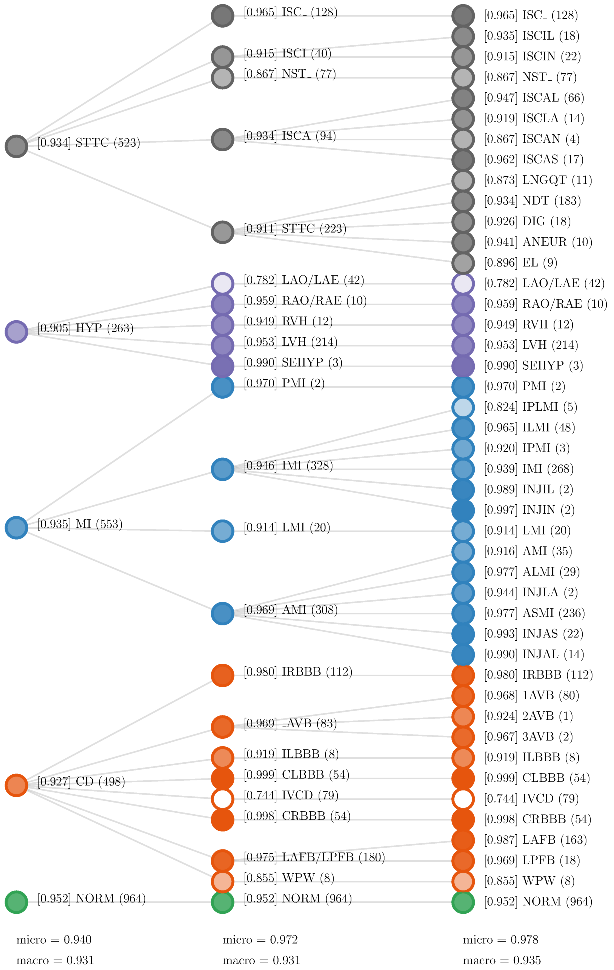


Fig. 4: Hierarchical decomposition of class-specific AUCs onto subclasses and individual diagnostic statements exhibiting hidden stratification, i.e. inferior algorithmic performance on certain diagnostic subpopulations that remains hidden when considering only the superior superclass performance, see 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.

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 subsume rather heterogeneous groups of findings.

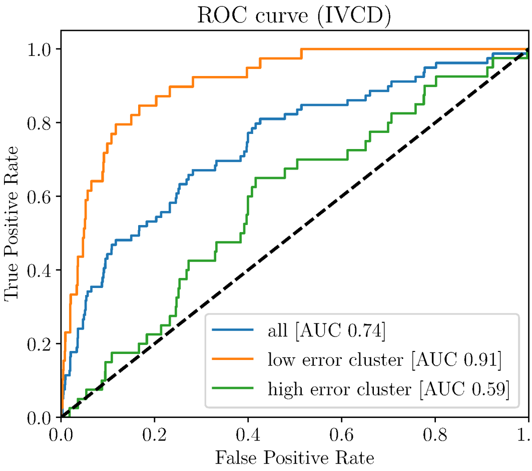


Fig. 5: AUC curves for two subset of samples revealing hidden stratification within the *IVCD* class. While the samples from the low-error cluster are mostly samples without *NORM* as additional label, samples from the high error cluster are mostly with co-occurring *NORM*.

Although identifying hidden stratification is straightforward to identify in hindsight given the hierarchical organization of the diagnostic labels, [55] 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 with k-means clustering revealed two clusters, where one cluster performed much better than the other as can be seen in Figure 5. 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 [55]. Models trained on coarse labels might hide this kind of clinically relevant stratification, because of both subtle discriminative features and low prevalence. 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

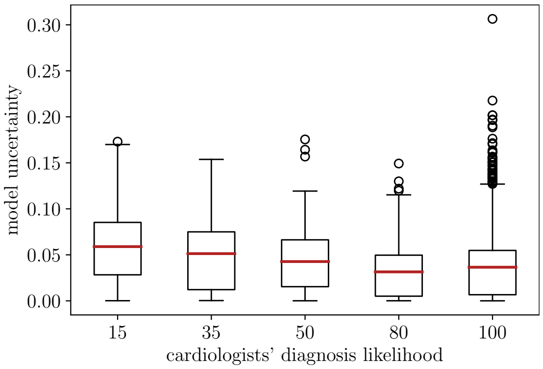


Fig. 6: Relation between model uncertainty (standard deviation of ensemble predictions as in [56]) and diagnosis likelihood as quantified by the annotating cardiologist, see Section IV-C for details.

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 [15]. It is important to stress that this likelihood information represents an individual uncertainty assessment of the annotating cardiologist and is therefore not directly comparable to uncertainty assessments from inter-rater agreement, which is not available for PTB-XL.

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 [56] 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 [56], in this first exploratory analysis. In Figure 6, 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 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. [57]. However, we perceive the overall alignment of diagnosis likelihood with model uncertainty as an interesting observation 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

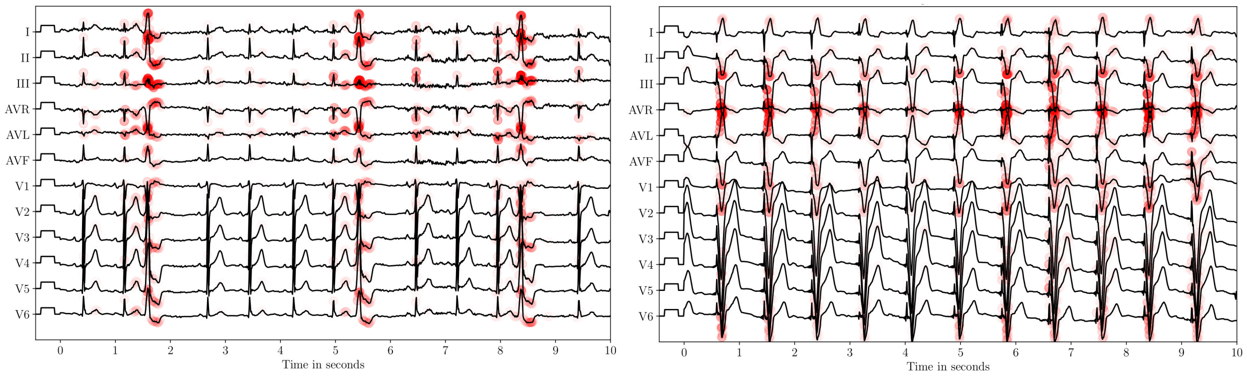
The acceptance of machine learning and in particular deep learning algorithms in the clinical context is often limited by the fact that data-driven algorithms are perceived as black boxes by doctors. In this direction, the recent advances in the field of explainable AI has the prospect to at least partially alleviate this issue by allowing the clinician to align indicative features for the classification decision with medical background knowledge. In particular, we consider post-hoc interpretability that can be applied for a trained model, see e.g. [58]. The general applicability of interpretability methods to multivariate timeseries and in particular ECG data was demonstrated in [41], see also [59], [60] for further 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 7, 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 [61]. 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 along with the relevance distribution onto the different leads 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 procedure. We propose a variety of benchmarking tasks based on the PTB-XL dataset [15] and put forward first baseline results for deep-learning-based time classification algorithms that are supposed to guide future researchers working on this dataset. We find that modern resnet- or inception-based convolutional architectures and in particular a newly proposed resnet-variant *xresnet1d101* 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. We release the training and evaluation code for all ECG statement prediction tasks, trained models as well as the complete model predictions in an accompanying code repository [19].

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.



(a) PVC

(b) PACE

Fig. 7: Two exemplary attribution maps for a resnet model for the classes PVC (left) and PACE (right).

[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] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart *et al.*, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Communications*, vol. 11, no. 1, Apr. 2020.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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*, vol. 7, no. 1, p. 154, 2020.

[16] 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.

[17] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," *Scientific Data*, vol. 7, no. 1, pp. 1–8, 2020.

[18] 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.

[19] P. Wagner and N. Strodthoff, "Code Repository: Deep Learning for ECG Analysis," https://github.com/helme/ecg_ptbxl_benchmarking, 2020.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," *CoRR*, vol. abs/1707.01836, 2017.

[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] 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.

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

[31] 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.

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

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

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

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

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

[41] 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.

[42] 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.

[43] 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.

[44] S. W. Smith, B. Walsh, K. Grauer, K. Wang, J. Rapin, J. Li *et al.*, "A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation," *Journal of Electrocardiology*, vol. 52, pp. 88–95, Jan. 2019.

[45] 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.

[46] 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.

[47] 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.

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

[49] P. C. Austin and J. E. Hux, "A brief note on overlapping confidence intervals," *Journal of Vascular Surgery*, vol. 36, no. 1, pp. 194–195, Jul. 2002.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

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

[60] 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.

[61] 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, p. e0130140, 2015.