Transfer Learning for Predicting Acute Myocardial Infarction Using Electrocardiograms

Axel Nyström, Anders Björkelund, Mattias Ohlsson, Jonas Björk, Ulf Ekelund, and Jakob Lundager Forberg

***Abstract*— At the emergency department it is important to quickly and accurately identify patients at risk of acute myocardial infarction (AMI). One of the main tools for detecting AMI is the electrocardiogram (ECG), which can be difficult to interpret manually. There is a long history of applying machine learning algorithms to ECGs, but such algorithms are notoriously data hungry, and correctly labelled, high-quality ECGs are difficult to obtain. Transfer learning has been a successful strategy for mitigating data requirements in other applications, but the benefits for predicting AMI are under-studied. Here we show that a straight-forward application of transfer learning leads to large improvements also in this domain. We pre-train models to classify sex and age using a collection of 840k ECGs from non-chest pain patients, and fine-tune the resulting models to predict AMI using 47k ECGs from chest pain patients. The results are compared with models trained from scratch (without transfer learning). Our results show a considerable improvement from transfer learning, consistent across multiple state-of-the-art ResNet architectures and data sizes, with the best performing model improving from 0.79 AUC to 0.85 AUC. This suggest that even a simple form of transfer learning from a moderately sized dataset of non-chest pain ECGs can lead to major improvements in predicting AMI. More research is needed to realize the full potential of transfer learning in this context, but it will be increasingly difficult to motivate future studies that do not make use of transfer learning.**

Manuscript received ???. This work was part of the AIR Lund (Artificially Intelligent use of Registers at Lund University) research environment, and received funding from the Swedish Research Council (VR; grant no. 2019-00198). The study also received funding from the Swedish Heart-Lung Foundation (2018-0173) and Sweden's innovation agency (Vinnova; DNR 2018-0192). (*Corresponding author: Axel Nyström.*)

Axel Nyström is with the Department of Laboratory Medicine, Lund University, Lund, Sweden, and also with the Center for Environmental and Climate Science, Lund University, Lund, Sweden (e-mail: axel.nystrom@med.lu.se).

Anders Björkelund is with the Center for Environmental and Climate Science, Lund University, Lund, Sweden (e-mail: anders.bjorkelund@cec.lu.se).

Mattias Ohlsson is with the Center for Environmental and Climate Science, Lund University, Lund, Sweden, and also with the Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Halmstad, Sweden (e-mail: mattias.ohlsson@cec.lu.se).

Jonas Björk is with the Department of Laboratory Medicine, Lund University, Lund, Sweden, and also with Clinical Studies Sweden, Forum South, Skåne University Hospital, Lund, Sweden (e-mail: jonas.bjork@med.lu.se).

Ulf Ekelund is with the department of Internal and Emergency Medicine, Skåne University Hospital, Lund, Sweden, and also with the Department of Clinical Sciences, Lund University, Lund, Sweden.

Jakob Lundager Forberg is with the Department of Clinical Sciences, Lund University, Lund, Sweden, and also with the Department of Emergency Medicine, Helsingborg Hospital, Helsingborg, Sweden (e-mail: jakob.lundager-forberg@skane.se).

***Index Terms*— Machine learning, transfer learning, electrocardiogram, acute myocardial infarction**

I. Introduction

ACUTE myocardial infarction (AMI) is one of the leading causes of death in the developed world, and chest pain is the second most common complaint at the ED [1]. Early detection of AMI is crucial for initiating timely treatment and to reduce mortality and morbidity [2]. Yet fewer than 10% of chest pain patients are diagnosed with AMI, which presents a challenging problem for ED physicians [3]. In order to accurately diagnose or rule out AMI, patient management includes serial electrocardiograms (ECGs), blood tests, and often prolonged and costly observations.

The ECG is one of the cornerstones of diagnosis of AMI, being fast, cheap, non-invasive and widely available across different medical facilities. The interpretation of an ECG requires comprehensive training however, with notable variability in diagnostic skills among physicians, across different medical facilities, and outside regular office hours [4]. This inconsistency extends to expert cardiologists, highlighting a broad variance in detecting various heart conditions [5]. To assist clinicians, automated ECG analysis has been evolving for many years, accelerated by advancements in machine learning (ML) and the digitization of healthcare records [6], [7].

Machine learning (ML) in general and deep learning (DL) in particular almost always profits from larger training datasets [8]. Large models can learn more complex patterns, but require more data to achieve good results compared to smaller models [9]. Acquiring more data is among the most straight-forward and reliable ways of increasing model performance, but doing so can be both difficult and expensive. In the pursuit of better models to predict AMI using ECGs, the standard approach is to train models using ECGs where the outcome (AMI) is known. But ECGs are also routinely collected outside the scope of the ED and for reasons beyond trying to detect short term AMI.

Transfer learning is a collection of techniques through which a downstream, also known as target task is improved by means of first learning an upstream or source task in a step called pre-training. The pre-trained model is then fine-tuned on the target task, with the expectation that some of what was learned from the source task will generalize — transfer — to the target task, thereby achieving higher performance on the target task.

There are multiple variations of transfer learning, but typically the source task utilizes a different dataset and/or a different outcome. Transfer learning has become a staple in several fields where ML is applied, particularly image recognition. Only recently has transfer learning also been successfully applied to ECG classification. A possible reason for the relatively slow adaptation may be the lack of sufficiently large open ECG datasets, which remains an obstacle in the field even now.

Efforts to utilize transfer learning for ECG classification can be broadly divided into two categories. The first uses the raw ECG signals and the second uses image-based methods in which models are pre-trained on large image datasets (usually non-medical images) such as ImageNet and fine-tuned on ECGs converted to images (typically either as spectrograms or time-series plots) [10], [11]. Here we limit our overview to the former approach.

In 2020, Strodthoff *et al.* [12] published one of the first successful results of transfer learning on ECGs. The authors used data from two public ECG databases (PTB-XL [13], containing 21 837 12-lead ECGs, and ICBEB2018 [14], containing 6 877 12-lead ECGs) and found significant improvements in terms of macro AUC for a collection of 71 diverse outcomes.

Also in 2020, Jang *et al.* [15] used 2.6M unlabeled single-lead ECGs to pre-train an autoencoder, which was then fine-tuned on 10k ECGs to predict a collection of 11 different ECG rhythms.

In 2021, Weimann *et al.* [16] showed consistently improved performance on atrial fibrillation (AF) detection on the PhysioNet/CinC 2017 dataset [17] through the use of transfer learning. Their pre-training on the Icentia11k dataset [18] consisted of over 630k hours of single lead ECG data with labeled heart beats. Their ResNet models also showed good performance when evaluated on the PTB-XL dataset.

In 2022, Mehari *et al.* [19] reported success with a self-supervised learning algorithm called Contrastive Predictive Coding (CPC). This can be viewed as a form of inductive transfer learning where the source and target datasets are the same, but the source labels are automatically induced from the data itself without human annotation [20]. The authors found that pre-training improved the downstream performance on PTB-XL in terms of macro AUC, data efficiency and robustness against input perturbations.

Following the thread of self-supervised learning, in 2023, Mehari *et al.* [21] combined a type of model called structured state space models (S4) with the CPC architecture for self-supervised learning, improving state-of-the-art on the PTB-XL dataset. The authors also indicated that the addition of patient metadata (speficially age, sex, height, and weight) alongside the ECG signal further improved downstream performance on PTB-XL.

These results are all promising indications that transfer learning can lead to improvements for a large range of classification tasks and pre-training scenarios. However, most related work either does not focus on AMI, or only includes AMI together with many other outcomes, as is the case in the PTB-XL outcomes.

In this paper, we propose to use age and sex as pre-training tasks, either individually or in combination. Unlike potentially stronger, more clinically relevant labels that might be used for pre-training, age and sex are virtually always available. Using a target and source datasets of 44k and 830k ECGs respectively, we show that pre-training on age and sex offers a substantial improvement in terms of AUC for predicting AMI. This improvement is consistent across several state-of-the-art models described in the recent literature. We further analyze the role of the sizes of source and target datasets and show both that in (a) a small-target data setting where the size of the target training set is reduced to 10% (2.5k ECGs), the gap to training with the full target dataset can be fully bridged by pre-training. Conversely (b), assuming the full target dataset for training (25k ECGs), the AUC on AMI prediction can be improved from 0.79 to 0.85 through transfer learning using 830k ECGs that are unrelated to AMI but are coupled with age and sex information, which must be regarded as a considerable improvement.

II. Methods

A. Data sources

Our data source is the Skåne Emergency Medicine (SEM) cohort, which contains two years of consecutive ED visits from eight different hospitals in Skåne, Sweden [22]. Besides data from the ED visits, each patient also has a five year medical history of diagnoses and ECGs from all health care visits, collated from comprehensive national and regional registers. The ECGs are all 10s, 12-leads, sampled at 500Hz with 16 bits precision.

Of the twelve leads, we only use leads V1-V6, I, and II (in that order), since the remaining four can be constructed as linear combinations of the others. The models were adapted where necessary to comply with this standard, to simplify comparisons.

In the following sections, we describe how the Source and Target datasets were constructed from the SEM cohort. Below we discuss how the data was divided into source and target data and further subdivisions into training, validation, and test sets. The overall process and exclusion criteria are shown in Fig. 1.

1) Target data: The target dataset consists of ECGs from patients arriving at the ED with chest pain as primary complaint, and the label is the binary outcome of AMI within 30 days of arrival at the ED, defined as the AMI diagnosis (ICD-10 code I21) being set according to the Swedish national patient register [23]. We focus on the Area Under the Curve (AUC) of the receiving operator characteristic as our evaluation metric for AMI predictions. We use a single ECG per patient visit, chosen as the first ECG within 3h of arrival, or if there is no such ECG, the last ECG within 2h prior to arrival. The reason for reverting to ECGs from before ED arrival is to catch ECGs recorded by paramedics or at the ED before registration. The target dataset contains 44 370 visits from 37 447 patients, where 3 431 visits from 1 947 patients were excluded due to missing ECGs of sufficient quality and 2 717 visits from 1 276 patients were excluded as a result of the temporal test split (see below). The dataset is partitioned into training, validation and

test sets, where the test set is further divided into a random and a temporal split, allowing us to evaluate the performance both on data with the same underlying distribution as the training data (the random test split) and on data which is separated in time from the training data (the temporal split). The temporal test split consists of the chronologically last visits from the dataset, such that 15% (5 618) of the patients in the whole target dataset are included. All visits from the patients in the temporal test split are then removed from the remainder of the target dataset, for a total of 2 717 visits and 1 276 patients excluded. The random test split as well as the training and validation splits are a random partition on the patient level (i.e. a patient can only be represented in at most one of the splits), with 55% of the patients in the training set, 15% in the validation set and 15% in the random test set. The motivation for this exercise is to avoid ECGs from the same patient ending up in both training and test sets, ensuring that the test set is truly disjoint from the training and development data.

2) Source data: The source dataset contains 836 972 ECGs from 162 903 patients in the SEM cohort. The dataset excludes all ECGs from all patients in the target dataset in order to minimize the interaction between the two datasets. The source dataset is partitioned into a training and validation set using a random split on the patient level, with 5% of the patients in the validation set and the rest in the training set. We did not hold out a test set for the source data since our primary concern is the final performance on the downstream task, rather than the generalizability of the pre-training task itself. For the supervised pre-training task, we use age and sex as labels. We use the Mean Absolute Error (MAE) as primary evaluation metric for age regression, and AUC as primary evaluation metric for sex classification.

The age and sex distributions for each dataset and split is summarized in Table I, which also shows the incidence of AMI for the target dataset splits. The sex distribution is close to 50% for the source data, but the target dataset has more visits from men. Furthermore, the patients in the source dataset is roughly 10 years older than those in the target dataset. These observations are not particularly surprising: men are generally more exposed to cardiovascular disease than women, and we expect general health-care visits to be more strongly correlated with age than chest-pain ED visits. Finally, we note that the incidence of AMI is fairly low, at 5.7–5.8%.

B. Models

In this work we consider four different convolutional models applied to 10s, 12-lead ECGs described in the literature: a simple model from our own prior research that serves as a baseline model [24], together with three variations of residual neural networks (ResNets) of different size and complexity. The selected models (except our baseline model) achieve state-of-the-art performance on their respective tasks. We have not performed any parameter tuning for the models under consideration, with the exception of learning rate and number of training epochs. Unless otherwise noted, the the code implementation of the models were adapted from the respective author's public github repositories (with some cleanup and refactoring).

The full source code for all models, parameters, and experiments, including data processing (but excluding the data, which is sensitive) is available on github [25].

A brief overview of the models and their complexity in terms of number of parameters and convolutional layers is shown in Table II. In the following sections we describe the origins and general structure of each model.

1) CNN-20k: This was the best performing model from [24] using only the raw input ECG to predict major adverse cardiac events (MACE) within 30 days of arrival at the ED. It contains two convolutional layers followed by two fully connected layers. Although small in size compared to the other models, it should be noted that this model was tuned for a very similar task (the majority of MACE outcomes are due to AMI) on a very similar dataset (ESC-TROP, which is a subset of SEM).

2) RN-900k: The xresnet1d50 (RN-900k) model [19] is an adaptation of the original ResNet-50 model from [28], which also incorporates a number of tricks from [29]. The model consists of 51 convolutional layers arranged into four stages, where each stage further contains 3-6 residual block. The residual blocks each contain 3 convolutional layers and a shortcut connection. The RN-900k model was found to outperform other ResNet variations on the PTB-XL dataset for a variety of tasks. In particular, the model has been fine-tuned to perform well on the macro-AUC score of 71 different outcomes from PTB-XL, including AMI.

The model also incorporates a type of data augmentation and ensembling: During training, the 10s input ECG is randomly cropped to a 2.5s long sequence (random sliding windows), whereas during validation and testing, the input ECG is sliced into 10 equidistant, overlapping 2.5s slices, and the predictions for all 10 slices are aggregated to a single prediction. We consider this augmentation/ensembling scheme as part of the model itself. The other models do not use augmentation or ensembling.

3) RN-7M: The RN-7M model [26] is a ResNet style model based on the architecture described by [30] and was developed and tuned using over 2M ECGs to recognize six types of ECG abnormalities. The model consists of 9 convolutional layers arranged into 4 residual blocks. Compared to the RN-900k model, this model is "wide and shallow": each convolution uses more filters, has a larger kernel-size and downsamples the input signal less aggressively, but there are far fewer layers.

The model downsamples the input ECG from 500Hz to 400Hz, and adds zero-padding to make each ECG 4 096 samples along the time-axis. We adjusted it slightly to use only 8 leads instead of the full 12-lead input that was described in the article.

4) RN-33M: The RN-33M model [27] is an extension of the RN-7M model, consisting of 25 convolutional layers arranged into 12 residual blocks. It was developed using 500k ECGs from the ED to distinguish between STEMI, NSTEMI and non-AMI. The main difference from the RN-7M model besides the number of layers is the addition of a Squeeze and Excite block within each residual block, which is intended to help weighting the channel-wise information [31]. Although still only using half the number of convolutions as the RN-900k model, the RN-33M model is by far the largest in terms of

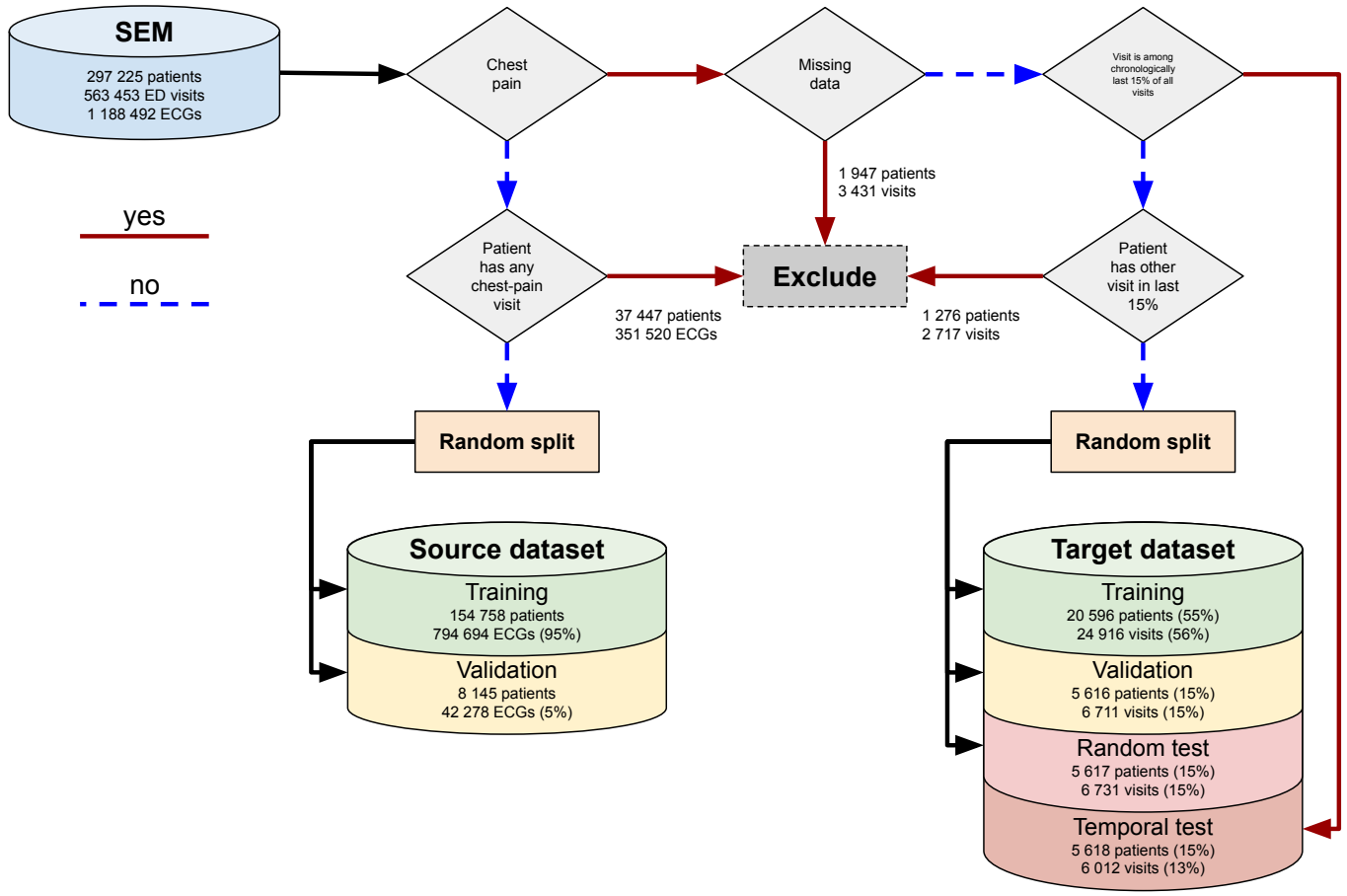

Fig. 1. Overview of exclusion criteria and data splits. SEM refers to the Skåne Emergency Medicine cohort [22], and includes all ED visits from 8 hospitals in Skåne, Sweden for two years between 2017 and 2018. The flow-chart shows which dataset and split each individual ED visit belongs to.

TABLE I
PATIENT CHARACTERISTICS. ECG, ELECTROCARDIOGRAM. IQR, INTER-QUARTILE RANGE. AMI, ACUTE MYOCARDIAL INFARCTION.

	Source dataset		Target dataset			
Patient characteristics	**Train**	**Validation**	**Train**	**Validation**	**Random test**	**Temporal test**
Patients	154 758	8 145	20 596	5 616	5 617	5 618
ECGs	794 694	42 278	24 916	6 711	6 731	6 012
Female sex, %	50.4	49.7	48.0	47.1	47.8	48.3
Age, median (IQR)	71 (57–81)	71 (57–81)	61 (46–75)	61 (46–74)	61 (46–75)	61 (45–74)
AMI, %	N/A	N/A	5.76	5.72	5.81	5.81

TABLE II
LIST OF MODELS USED IN THIS WORK, THE NUMBER OF CONVOLUTIONAL LAYERS, NUMBER OF TRAINABLE PARAMETERS, AND THE ORIGINAL AUTHORS OF EACH ARCHITECTURE.

Model	**Conv layers**	**Parameters**	**Source**
CNN-20k	2	20 479	Nyström *et al.* [24]
RN-900k	51	892 449	Mehari *et al.* [19]
RN-7M	9	6 784 561	Ribeiro *et al.* [26]
RN-33M	25	33 062 569	Gustafsson *et al.* [27]

number of parameters at just over 33M.

Similarly to the RN-7M model, the input ECGs are down-sampled to 400Hz and employs zero-padding to make the signal 4 096 samples wide.

The code for the RN-33M model was not publicly available at the time of writing, but because it is described as an updated version of the RN-7M model, for which code was available, we based our implementation on the RN-7M code and followed the described architectural changes as closely as possible. We did not include the ECG pre-processing step, which in Gustafsson *et al.* [27] consisted of an elliptic high-pass filter.

C. Transfer learning strategy

In order to simplify the analysis, we used the same general transfer learning strategy for each of the four models under consideration, and repeated the process of pre-training and fine-tuning on varying proportions of the available source and target data.

The transfer learning strategy, illustrated in Fig. 2, consisted of three stages: In the first stage, a model was pre-trained on the source dataset to predict age, sex, or both, saving the model weights after each epoch. In the second stage, we loaded the weights from the epoch where the best validation loss was observed. The classification head (i.e. the final layers of the model, everything after the flatten layer) was then replaced by a fully connected layer of size 100, followed by a dropout layer with $p = 0.3$, followed by a fully connected output layer with sigmoid activation. The replacement classification head was randomly initialized, and training proceeded on the target dataset to predict AMI. During the second stage, the weights of the model body were frozen, so that only the weights of the classification head was updated during backpropagation. Finally, in the third stage, fine-tuning continued with all model weights unfrozen.

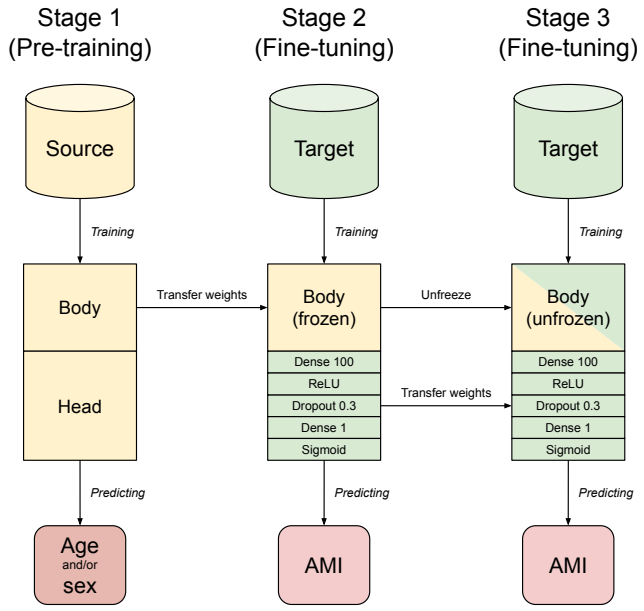

Fig. 2. Visualization of the transfer learning strategy used in this paper. The model is divided into two segments which we call the body and the head. In the second stage, the head of the pre-trained model is replaced by a small neural network with a single hidden layer. The model body is frozen in stage 2, meaning that the parameters are not updated during back-propagation.

For the first two stages we used a one-cycle learning rate schedule consisting of an initial linear warmup, followed by a cosine decay with a factor of 100, so the learning rate at the end is two orders of magnitude smaller than the warmup target. The final stage used a constant learning rate.

For the target task we used binary crossentropy (CE) as the loss function, and performed early stopping where the AUC on the validation set was maximized. For the source task loss function, we used CE for sex prediction and MAE for age prediction. When predicting both age and sex simultaneously, we used a weighted sum of the MAE of age and CE of sex as the loss, where the weight was chosen as 1 for CE and 0.045 for AGE. The weights were chosen empirically from early experiments such that each target would contribute roughly the same to the total loss.

III. Results and Discussion

A. Main results

The overall AUC on the target task for each model is shown in Table III, with the two test-sets shown separately. There is a consistent improvement in AUC for the pre-trained models compared to the models without pre-training. The baseline CNN-20k model displays the smallest improvement overall, ranging from 1 to 4 percent absolute, the smallest when pre-trained on combined age and sex outcome and the largest when pre-trained for sex. The other models display the greatest improvements (up to 7 percent absolute) when pre-trained on age or the combination of age and sex, with sex appearing to be slightly behind. All the models perform worse (about 1 percent absolute) on the temporal test set compared to the random test set. Aside from the baseline CNN-20k model, the improvement from pre-training is similar between the two test sets. The CNN-20k model performed particularly poorly on the temporal test set without pre-training, which leads to an apparently greater improvement from pre-training when compared to the random test set.

Overall, the benefit of pre-training is clear and the trend is similar across both test sets. The best choice of pre-training label is somewhat inconclusive, with a slight edge towards age or a combination of age and sex.

The pre-training results on the validation set are summarized in Table IV. All three ResNet models perform roughly similar on both pre-training tasks, irrespective of whether they are trained towards a single target or both. The MAE for predicting age lies in the range 6.6 – 6.8 years and the accuracy of 88–90% for predicting sex. These numbers are in line with previous studies; for instance, Attia *et al.* achieved a MAE of 6.9 for age regression and an accuracy of 90.4% for sex classification [33]. Strodthoff *et al.* obtained an MAE of 6.8 years for age regression for healthy patients and 7.2 years for non-healthy patients in PTB-XL, as well as an accuracy for sex classification of 81% for healthy and 90% for non-healthy patients [12]. Lima *et al.* achieved an MAE of 8.4 years for predicting age [34].

The baseline CNN-20k model performs reasonably well on the sex classification task, but not so well on the age regression, and struggles especially on the combined task. In contrast, the ResNet models are able to learn both tasks at once with little degradation.

To simplify the presentation, the next section will focus on the results using age as the pre-training task, and consider only the results on the full test set (i.e. the combined random and temporal test sets). Specifically, the test sets are concatenated and metrics are computed on the concatenation, thus resulting in micro-averages.

B. Dependency on dataset size

In order to determine how the size of the target dataset influences the benefit of transfer learning, we fine-tuned each model on progressively smaller subsets of the target dataset. Specifically, we consider subsets of 10%, 20%, 30%, ..., 80%, 90% of the target training data, where each set is a strict subset of the previous one, chosen randomly at the patient level, so

TABLE III
MAIN RESULTS, SHOWING THE AUC ON THE TWO TEST-SETS FOR EACH MODEL, FOR EACH COMBINATION OF PRE-TRAINING TARGETS. 95% CONFIDENCE INTERVAL IN PARENTHESIS IS APPROXIMATED WITH (BASIC) BOOTSTRAPPING [32], WITH $B = 1000$ BOOTSTRAP SAMPLES.

Model	Test set	No pre-training	Age	Sex	Age & Sex
CNN-20k	**Random**	0.785 (0.764 – 0.807)	0.798 (0.777 – 0.820)	0.806 (0.786 – 0.828)	0.794 (0.772 – 0.816)
	Temporal	0.750 (0.724 – 0.775)	0.778 (0.755 – 0.802)	0.792 (0.768 – 0.818)	0.775 (0.751 – 0.799)
RN-900k	**Random**	0.799 (0.777 – 0.823)	0.857 (0.839 – 0.875)	0.845 (0.827 – 0.866)	0.858 (0.840 – 0.876)
	Temporal	0.787 (0.760 – 0.814)	0.846 (0.826 – 0.868)	0.836 (0.815 – 0.857)	0.844 (0.823 – 0.865)
RN-7M	**Random**	0.796 (0.774 – 0.820)	0.852 (0.834 – 0.870)	0.832 (0.811 – 0.852)	0.830 (0.809 – 0.850)
	Temporal	0.771 (0.744 – 0.799)	0.819 (0.799 – 0.842)	0.808 (0.785 – 0.833)	0.817 (0.795 – 0.839)
RN-33M	**Random**	0.729 (0.705 – 0.756)	0.797 (0.777 – 0.818)	0.774 (0.751 – 0.797)	0.796 (0.774 – 0.818)
	Temporal	0.716 (0.687 – 0.743)	0.778 (0.754 – 0.805)	0.770 (0.746 – 0.796)	0.792 (0.769 – 0.816)

TABLE IV
PRE-TRAINING RESULTS. MAE: MEAN ABSOLUTE ERROR, r^2: COEFFICIENT OF DETERMINATION, AUC: AREA UNDER THE CURVE (OF THE RECEIVING OPERATOR CHARACTERISTIC).

Model	Label	MAE	r^2	AUC	Accuracy
CNN-20k	**Age**	10.32	0.50		
	Sex			0.92	0.81
	Age & Sex	15.81	-0.05	0.89	0.80
RN-900k	**Age**	6.69	0.76		
	Sex			0.96	0.88
	Age & Sex	6.80	0.75	0.96	0.88
RN-7M	**Age**	6.64	0.75		
	Sex			0.96	0.89
	Age & Sex	6.81	0.75	0.96	0.90
RN-33M	**Age**	6.60	0.75		
	Sex			0.96	0.89
	Age & Sex	6.70	0.74	0.96	0.89

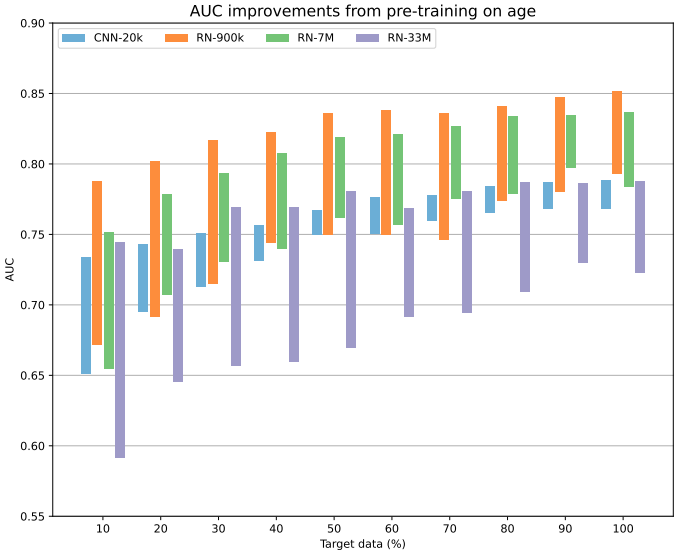

Fig. 3. Bar chart showing the AUC for predicting AMI, micro-averaged over both test-sets, for different sizes of the target training set. The bottom of each bar show the AUC for models without pre-training, and the top of each bar show the AUC for models pre-trained on age, using the full source dataset. The size of each bar thus corresponds to the improvement from transfer learning.

that all examples from a single patient are excluded at once. The validation and test sets remain the same for each subset. The results on the (combined) test set is illustrated in Fig. 3.

Unsurprisingly, more target data leads to overall better performance but with a reduced benefit from transfer learning. Without pre-training, the CNN-20k baseline model performs similarly to the much larger ResNet models, except for the largest model (RN-33M) which struggles with overfitting and performs much worse than the others. The RN-33M model is clearly too large for our small target dataset. When we introduce transfer learning however, the RN-33M model improves enough to catch up with the baseline CNN-20k model, but both are still behind the other two ResNet models. With pre-training, it seems like the baseline is too small and the RN-33M model is too large. The RN-900k and RN-7M models perform roughly the same with a sizable improvement over the other two models.

Next, we investigate how much source data is required for pre-training to be effective. To this end, we divided the training set of the source dataset into smaller subsets, similarly to what we did for the target dataset, again taking care that all ECGs from a single patient is either completely covered or not covered at all by each dataset. We pre-trained each model to predict age on 14 overlapping subsets of sizes 2%, 4%, 6%, 8%, 10%, 20%, 30%, ..., 80%, 90%, 100%, where the full training dataset consists of 794k ECGs. The MAE of predicting age is plotted for all four models in Fig. 4.

The RN-900k network is remarkably good at predicting age even for small datasets, easily beating the other models until about 30% of the training data is available, after which it gets overtaken by the RRN model. The CNN-20k baseline model on the other hand visibly struggles with the task, particularly when trained with fewer than 30% of the data points. Part of this could be a result of poorly optimized hyper-parameters for the model but, even discounting that, it seems clear that the model is not complex enough to properly learn the relevant features for predicting age.

In order to determine how the effectiveness of transfer learning depends on the size of the pre-training dataset, we focused on the best performing model, the RN-900k on the pre-training task of predicting age.

In order to determine how the effectiveness of transfer learning depends on the size of the pre-training dataset, we considered each combination of the 14 subsets of the source dataset and 10 subsets of the target datasets, focusing on the best performing model (RN-900k) and pre-training task (age).

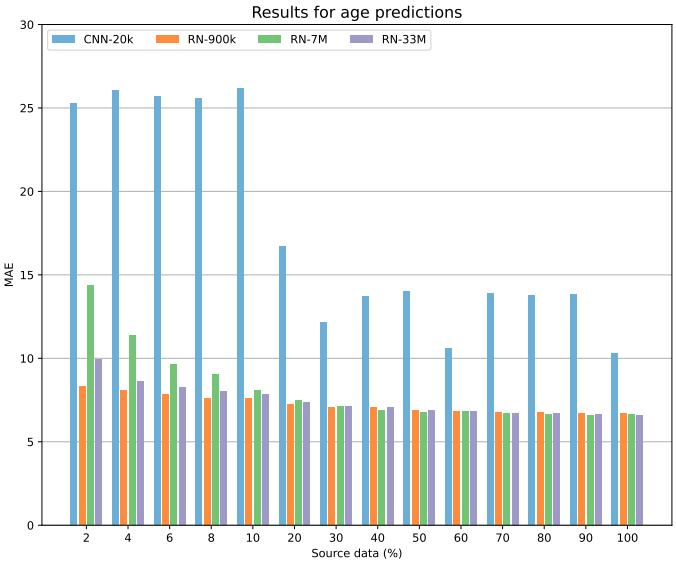

Fig. 4. Mean absolute error for age predictions, for increasing amounts of training data. Lower is better. The x-axis shows percent of the full source dataset, 100% corresponds to 794k ECGs.

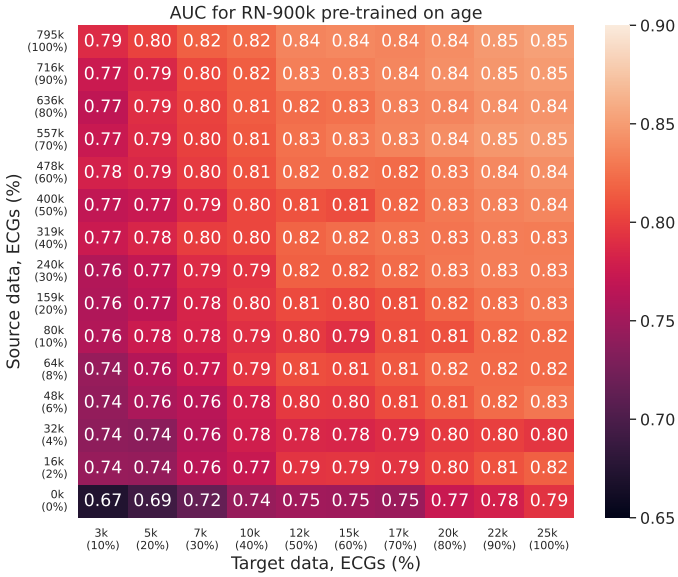

Fig. 5. Heatmap showing the AUC for predicting AMI with the RN-900k model, for different amounts of source and target training data. The x-axis shows the number of ECGs in the target training dataset, and the y-axis shows the number of ECGs in the source training dataset. The bottom row shows results without pre-training.

The AUC on the target task (predicting AMI) for each combination of dataset size is shown in Fig. 5, where the bottom row corresponds to no pre-training. Although there are some exceptions due to the stochastic nature of the optimization procedure, the general trend is clear and expected: The more data the better the final results, regardless of whether that data is in the source or target domain or some combination of the two. In particular, a 90% reduction in the size of the target dataset can be compensated for by pre-training with approximately 800k ECGs. Most of the improvement from pre-training is realized already with relatively small source datasets, although this effect is more pronounced when the target dataset is also small. For instance, the improvement from pre-training on 80k ECGs is between 9 percentage points (for small target sets) and 3 percentage points (for large target sets), but increasing the pre-training dataset by an order of magnitude only yields an additional improvement of roughly 3 percentage points regardless of the size of the target dataset. An intuitive explanation for this is that in the low-data domain, the pre-training task primarily makes up for rudimentary patterns that are not as impactful when the target dataset is larger.

C. Limitations

Perhaps the most glaring limitation of this study is that we have not attempted to adjust the hyper-parameters of the models, except for the learning rate and number of training epochs. Instead, we opted to use a selection of recently published model architectures, complete with all the specified settings, and use them as is, even though they were tuned to different datasets on different tasks. This somewhat naive approach is of course not generally recommended if one wants to maximize the performance of a specific model for a given task. In our case however, the idea was never to maximize the performance per se, but rather to assess the benefit of transfer learning across a range of models on our specific downstream task. We expect that all the models we have used in this project would benefit to varying degrees by further optimizing the hyper-parameters, but in order to get a truly fair comparison between the models, we would have to tune the parameters for each task, model and dataset size separately, which would be prohibitively expensive. The magnitude and consistency of our results across models suggests that the general trend in terms of benefit from transfer learning would remain even if all the relevant hyper-parameters were optimized for each model.

A proper analysis of the relationship between model size and transfer learning is hampered in this study due to fundamental differences in model architecture, primarily between the RN-900k model and the RN-7M and RN-33M models. Although the RN-900k model is much smaller in terms of number of parameters, it outperforms the other models in most situations. However, some of that performance is likely due to clever tricks being employed by the RN-900k model, perhaps most importantly the data augmentation and ensembling, which acts as an effective form of regularization. Incorporating these ideas into the other models would require substantial adjustments contrary to our initial philosophy of using these state-of-the-art models as is, but it would nevertheless be a promising future research direction. We note in passing that the evaluated architechtures and ideas may profit from cross-fertilization – e.g., the ensemble and data augmentation techniques used in the RN-900k network (cf. Section II-B.2) could potentially be employed in the RN-7M and RN-33M architectures. Similarly, Squeeze and Excite-blocks used in the RN-33M architechture (cf. Section II-B.4) could potentially be employed in the RN-900k network.

We have primarily focused on the benefit of a relatively simple transfer learning approach, and although it appears to work well, other techniques might be even better. Further

research would be required to appropriately compare age and sex predictions as transfer learning tasks to more involved methods, such as the self-supervised methods proposed in [21].

IV. Summary and conclusions

In this study, we compared three different recently published state-of-the-art ResNet models and a simple baseline CNN-20k model on the task of predicting AMI using ECGs. We explored the effects of a simple supervised transfer learning approach in which a separate collection of unrelated ECGs (lacking the AMI outcome label) were first pre-trained to predict age and/or sex, and then fine-tuned to the target task of predicting AMI. This simple transfer learning scheme consistently improved our downstream predictions for all models, effectively increasing the maximum viable model size. Although all models were improved by pre-training, the smallest model improved the least, and the largest model, while benefitting the most from transfer learning, was still outperformed in absolute numbers by a substantially smaller model, underlining the important lesson that larger models are not always better.

Our results shows that transfer learning can have a substantial positive impact on classifying AMI using ECGs. Since age and sex are typically recorded in conjunction with the ECG, using them for pre-training serves as a natural and easy to implement first step when considering transfer learning, potentially obviating the need for more complicated unsupervised or self-supervised approaches.

Competing interests

The authors have no competing interests to declare.

Code and data availability

All code developed in this project is freely available on github [25]. Due to the sensitive nature of the dataset (SEM [22]), the data is not publicly available, although we welcome initiatives on international collaborative projects. Anonymized parts of the database, including detailed variable lists, are available for sharing on reasonable requests.

[1] A. Timmis, P. Vardas, N. Townsend, A. Torbica, H. Katus, D. De Smedt, C. P. Gale, A. P. Maggioni, S. E. Petersen, R. Huculeci, D. Kazakiewicz, V. de Benito Rubio, B. Ignatiuk, Z. Raisi-Estabragh, A. Pawlak, E. Karagiannidis, R. Treskes, D. Gaita, J. F. Beltrame, A. McConnachie, I. Bardinet, I. Graham, M. Flather, P. Elliott, E. A. Mossialos, F. Weidinger, and S. Achenbach, "European society of cardiology: Cardiovascular disease statistics 2021," *European Heart Journal*, vol. 43, no. 8, pp. 716–799, 01 2022, https://doi.org/10.1093/eurheartj/ehab892.

[2] R. A. Byrne, X. Rossello, J. J. Coughlan, E. Barbato, C. Berry, A. Chieffo, M. J. Claeys, G.-A. Dan, M. R. Dweck, M. Galbraith, M. Gilard, L. Hinterbuchner, E. A. Jankowska, P. Jüni, T. Kimura, V. Kunadian, M. Leosdottir, R. Lorusso, R. F. E. Pedretti, A. G. Rigopoulos, M. Rubini Gimenez, H. Thiele, P. Vranckx, S. Wassmann, N. K. Wenger, B. Ibanez, and E. S. D. Group, "2023 ESC Guidelines for the management of acute coronary syndromes: Developed by the task force on the management of acute coronary syndromes of the European Society of Cardiology (ESC)," *European Heart Journal*, vol. 44, no. 38, pp. 3720–3826, 08 2023, https://doi.org/10.1093/eurheartj/ehad191.

[3] T. Nilsson, E. Johannesson, J. L. Forberg, A. Mokhtari, and U. Ekelund, "Diagnostic accuracy of the heart pathway and edacs-adp when combined with a 0-hour/1-hour hs-ctnt protocol for assessment of acute chest pain patients," *Emergency Medicine Journal*, vol. 38, no. 11, pp. 808–813, 2021, https://doi.org/10.1136/emermed-2020-210833.

[4] D. A. Cook, S.-Y. Oh, and M. V. Pusic, "Accuracy of Physicians' Electrocardiogram Interpretations: A Systematic Review and Meta-analysis," *JAMA Internal Medicine*, vol. 180, no. 11, pp. 1461–1471, 11 2020, https://doi.org/10.1001/jamainternmed.2020.3989.

[5] J. M. McCabe, E. J. Armstrong, I. Ku, A. Kulkarni, K. S. Hoffmayer, P. D. Bhave, S. W. Waldo, P. Hsue, J. C. Stein, G. M. Marcus, S. Kinlay, and P. Ganz, "Physician accuracy in interpreting potential st‐segment elevation myocardial infarction electrocardiograms," *Journal of the American Heart Association*, vol. 2, no. 5, p. e000268, 2013, https://doi.org/10.1161/JAHA.113.000268.

[6] S. Ansari, N. Farzaneh, M. Duda, K. Horan, H. B. Andersson, Z. D. Goldberger, B. K. Nallamothu, and K. Najarian, "A Review of Automated Methods for Detection of Myocardial Ischemia and Infarction Using Electrocardiogram and Electronic Health Records," *IEEE Reviews in Biomedical Engineering*, vol. 10, pp. 264–298, 2017, https://doi.org/10.1109/RBME.2017.2757953.

[7] X. Liu, H. Wang, Z. Li, and L. Qin, "Deep learning in ECG diagnosis: A review," *Knowledge-Based Systems*, vol. 227, p. 107187, Sep. 2021, https://doi.org/10.1016/j.knosys.2021.107187.

[8] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, p. 78–87, oct 2012, https://doi.org/10.1145/2347736.2347755.

[9] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *Computer Vision – ECCV 2020*. Cham: Springer International Publishing, 2020, pp. 491–507, 10.1007/978-3-030-58558-7_29.

[10] S. Somani, A. J. Russak, F. Richter, S. Zhao, A. Vaid, F. Chaudhry, J. K. De Freitas, N. Naik, R. Miotto, G. N. Nadkarni, J. Narula, E. Argulian, and B. S. Glicksberg, "Deep learning and the electrocardiogram: review of the current state-of-the-art," *EP Europace*, vol. 23, no. 8, pp. 1179–1191, 02 2021, https://doi.org/10.1093/europace/euaa377.

[11] A. Ebbehoj, M. Ø. Thunbo, O. E. Andersen, M. V. Glindtvad, and A. Hulman, "Transfer learning for non-image data in clinical research: A scoping review," *PLOS Digital Health*, vol. 1, no. 2, pp. 1–22, 02 2022, https://doi.org/10.1371/journal.pdig.0000014.

[12] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ecg analysis: Benchmarks and insights from ptb-xl," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1519–1528, 2021, 10.1109/JBHI.2020.3022989.

[13] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "Ptb-xl, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, no. 154, 2020, 10.1038/s41597-020-0495-6.

[14] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, J. Li, and E. N. Yin Kwee, "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, 2018, 10.1166/jmihi.2018.2442.

[15] Y. D. Jang Jong-Hwan, Kim Tae Young, "Effectiveness of transfer learning for deep learning-based electrocardiogram analysis," *Healthc Inform Res*, vol. 27, no. 1, pp. 19–28, 2021, 10.4258/hir.2021.27.1.19.

[16] K. Weimann and T. O. F. Conrad, "Transfer learning for ecg classification," *Scientific Reports*, vol. 11, 2021, 10.1038/s41598-021-84374-8.

[17] G. D. Clifford, C. Liu, B. Moody, L.-w. H. Lehman, I. Silva, Q. Li, A. E. Johnson, and R. G. Mark, "Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017," in *2017 Computing in Cardiology (CinC)*, 2017, pp. 1–4, 10.22489/CinC.2017.065-469.

[18] S. Tan, G. Androz, A. Chamseddine, P. Fecteau, A. Courville, Y. Bengio, and J. P. Cohen, "Icentia11k: An unsupervised representation learning dataset for arrhythmia subtype discovery," *arXiv preprint arXiv:1910.09570*, 2019, https://doi.org/10.48550/arXiv.1910.09570.

[19] T. Mehari and N. Strodthoff, "Self-supervised representation learning from 12-lead ecg data," *Computers in Biology and Medicine*, vol. 141, pp. 105–114, 2022, 10.1016/j.compbiomed.2021.105114.

[20] H. H. Mao, "A survey on self-supervised pre-training for sequential transfer learning in neural networks," *arXiv preprint arXiv:2007.00800*, 2020, 10.48550/arXiv.2007.00800.

[21] T. Mehari and N. Strodthoff, "Towards quantitative precision for ecg analysis: Leveraging state space models, self-supervision and patient metadata," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 11, pp. 5326–5334, 2023, 10.1109/JBHI.2023.3310989.

[22] U. Ekelund, B. Ohlsson, O. Melander, J. Björk, M. Ohlsson, J. Lundager Forberg, P. Olsson de Capretz, A. Nyström, and A. Björkelund, "The Skåne Emergency Medicine (SEM) cohort," *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 2024, https://doi.org/10.1186/s13049-024-01206-0.

[23] Swedish National Board of Health and Welfare, "The national patient register," 2023, https://www.socialstyrelsen.se/en/statistics-and-data/registers/national-patient-register/ [Accessed: (2024-03-08)].
[24] A. Nyström, P. Olsson de Capretz, A. Björkelund, J. Lundager Forberg, M. Ohlsson, J. Björk, and U. Ekelund, "Prior electrocardiograms not useful for machine learning predictions of major adverse cardiac events in emergency department chest pain patients," *Journal of Electrocardiology*, vol. 82, pp. 42–51, 2024, 10.1016/j.jelectrocard.2023.11.002.
[25] A. Nyström, "The mim repository," https://github.com/Tipulidae/mim, 2024.
[26] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira, T. B. Schön, and A. L. P. Ribeiro, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Communications*, vol. 11, no. 1, p. 1760, Dec. 2020, 10.1038/s41467-020-15432-4.
[27] S. Gustafsson, D. Gedon, E. Lampa, A. H. Ribeiro, M. J. Holzmann, T. B. Schön, and J. Sundström, "Development and validation of deep learning ecg-based prediction of myocardial infarction in emergency department patients," *Scientific Reports*, vol. 12, no. 1, p. 19615, Nov 2022, 10.1038/s41598-022-24254-x.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, 10.1109/CVPR.2016.90.
[29] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 558–567, 10.1109/CVPR.2019.00065.
[30] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, pp. 65–69, 2019, 10.1038/s41591-018-0268-3.
[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, https://doi.org/10.48550/arXiv.1709.01507.
[32] R. Wehrens, H. Putter, and L. M. Buydens, "The bootstrap: a tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 54, no. 1, pp. 35–52, 2000, https://doi.org/10.1016/S0169-7439(00)00102-7.
[33] Z. I. Attia, P. A. Friedman, P. A. Noseworthy, F. Lopez-Jimenez, D. J. Ladewig, G. Satam, P. A. Pellikka, T. M. Munger, S. J. Asirvatham, C. G. Scott, R. E. Carter, and S. Kapa, "Age and sex estimation using artificial intelligence from standard 12-lead ecgs," *Circulation: Arrhythmia and Electrophysiology*, vol. 12, no. 9, p. e007284, 2019, 10.1161/CIRCEP.119.007284.
[34] E. M. Lima, A. H. Ribeiro, G. M. M. Paixão, M. H. Ribeiro, M. M. Pinto-Filho, P. R. Gomes, D. M. Oliveira, E. C. Sabino, B. B. Duncan, L. Giatti, S. M. Barreto, W. Meira Jr, T. B. Schön, and A. L. P. Ribeiro, "Deep neural network-estimated electrocardiographic age as a mortality predictor," *Nature Communications*, vol. 12, no. 5117, 2021, 10.1038/s41467-021-25351-7.