Towards Invariant Soft Biometrics from Electrocardiograms

Arian Ranjbar[1], Bjørn-Jostein Singstad[1,2], Jesper Ravn[1], Henrik Schirmer[1,2]

[1] Akershus University Hospital, Medical Technology and E-health, Lørenskog, Norway
[2] University of Oslo, Institute of Clinical Medicine, Oslo, Norway

Abstract

Medical data such as the electrocardiogram (ECG) has received an increased interest within biometric settings. One of the main benefits is the difficulty in counterfeiting the information due to its hidden nature. However, medical information may be exposed to intra-subject variability. This is evident in extraction of soft biometric traits from ECG, where results can vary widely depending on cardiac condition and status. This work investigates methods of lowering the variability, by employing multi-task learning on a shared feature extractor. Three different architectures are suggested and benchmarked. Specifically, experiments are carried out on age and gender estimation showing state of the art results on two public ECG datasets, while lowering estimation variance among instances.

1. Introduction

The last decade has seen a large increase in biometric research, mostly focusing on authentication. Biometric authentication systems have the benefit of relying on intrinsic characteristics inherent to the person, thereby requiring physical presence. In addition, fraudulent authentication attempts are harder to execute, compared to systems relying on extrinsic information which can be stolen. Several types of sensors have been suggested for the purpose, particularly fingerprints and face recognition. The latter has also been proposed as a method to retrieve soft biometric traits such as age and gender, [1]. Such traits have proven useful to improve accuracy or detecting fraud attempts, but may also have value in other applications like monitoring, [2], continuous authentication, surveillance, [3]; or to provide human interpretable information.

Lately there has been an increased interest in using biomedical sensors for biometrics. Biomedical sensors provide measurements even harder to counterfeit due to the hidden nature of the information, compared to e.g. face and fingerprint data which are more easily spoofed, circumvented or scraped. Of biomedical sensors, the Electrocardiogram (ECG) has proven to be particularly promising in this regard, [4]. Benefits of ECG, in addition to the safety aspects, include being computationally efficient compared to sensor data of higher dimension, and having simple and standardized acquisition. However, ECG tend to suffer from high intra-subject variability, which may severely affect the potential of biometric applications. For example, age prediction using CNNs have shown to correlate with physiological age rather than the typical association of chronological age, [5]. Physiological age may be desirable in diagnostic scenarios or in analysis at an aggregated population level, but could be adverse in e.g. authentication settings.

The aim of this work is to lower variability in extracting soft biometrics from ECG. Particularly performing age and gender prediction, invariant to cardial condition and status. Such information is valuable in medical settings, both for patient identification, [6], and monitoring diagnostic performance of sub-populations, [7]. We propose using MTL together with advancements in time-series modeling, to reduce variability in parallel with the main classification task. In particular, three different variants of MTL will be tested in experiments of age and gender prediction.

2. Related work

Current state of the art soft biometric extraction from ECG mostly rely on CNNs, [4]. In [5] a CNN was suggested for age and gender prediction, trained separately. The model used 8 convolutional blocks on 12 separate channels in a 12-lead ECG, together with a final convolutional block to fuse the channels, attaining state-of-the-art performance on both age and gender prediction, although with separate models. The study also found a strong correlation between higher prediction errors and comorbidities, indicating the prediction of chronological age getting overestimated in presence of cardiovascular disease.

The challenge of intra-subject variability has been studied in other domains of biometrics. In [8], a multitask CNN is suggested to combat pose variation in face recognition, incorporating pose estimation as a network branch. Similarly in [9], a multitask cascaded CNN is developed to jointly perform face detection and alignment.

Lately, there have also been significant developments to

network architectures for time series data. Rather than using regular CNNs, or even deeper models such as ResNet; the use of inception modules, have been proven to provide increased accuracy. Such configurations allows longer filters while using the same amount of parameters, in addition to applying several different filters simultaneously as illustrated in the InceptionTime architecture, [10].

3. Method

Consider the ECG as a time series $X^D = \{X_t^D : t \in T\}$, for an ordered index set T and where D denotes the number of channels (often with $D = 12$). The aim is to find a model, $\phi(X;\, \theta) \in Y$, for the map between ECG and a soft biometric property Y, such that the result is invariant, $\phi(X;\, \theta) = \phi(TX;\, \theta)$, to certain perturbations T. However, the challenge arises since the invariance relation cannot be used as a constraint, due to T not being known beforehand. To circumvent this, MTL is used as proxy, where the soft biometric trait extraction and variance inducing properties are learned in parallel.

3.1. Multi-task learning

The main goal of MTL is to simultaneously learn several tasks, often from the same feature extractor. Apart from benefiting from the increased supervision data, which effectively works as data augmentation, mitigating data noise; it has also been shown to help with both regularization and attention focusing, where additional tasks can be used to induce bias, [11]. The latter will be exploited to enforce the network to learn properties affecting the intra-subject variability. In other words, each property will have its own classifier f_w from the feature extractor, $(\phi \circ f_w)(X; \theta,\, w)$. Learning is then done via multi-task learning over all tasks, replacing a single loss with a combined one:

$$\text{argmin}_{\theta,\, w,\, v} \sum_i \lambda_i^{\text{bio}} L_i^{\text{bio}}(\theta,\, w_i) + \sum_j \lambda_j^{\text{var}} L_j^{\text{var}}(\theta,\, v_j),$$

where L_i^{bio} and L_j^{var} is the task loss for the soft biometric properties and variables affecting variance respectively, and λ_i weights the importance of each task.

3.2. Feature extractor and architecture

The common feature extractor among the tasks, ϕ, is built using inception modules in the InceptionTime configuration. Three network architectures are proposed for testing the framework, using different methods of learning the tasks, illustrated in Fig. 1. The first use an individual fully connected layer attached to the output of the feature extractor for each task. The idea is to rely on the bias induction of learning the variable property in conjunction with the soft biometric properties.

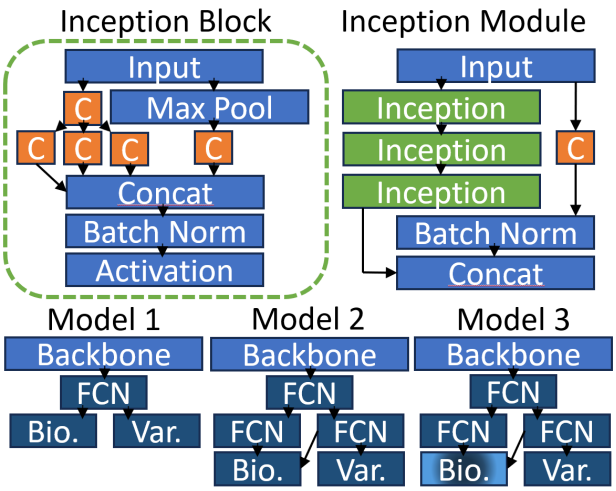

Figure 1. The backbone is built from inception modules, which in turn consists of inception blocks (marked in green). Orange indicate 1D convolutions. The three suggested architectures are mainly made of fully connected layers (dark blue), in various configurations.

The second model first split into two branches of fully connected layers, where the variable property is learned on one of the branches. After, the two branches are connected again into a feature fusion, where the biometric variable is estimated. In other words, the model still use hard parameter sharing in the feature extractor, but soft parameter sharing in the classifier, which may put an even heavier weight on the information causing the variance.

Finally, the second model may be implemented using Bayesian learning on the last layers through variational inference, [12]. This allows for confidence estimations explicitly dependent on the variance induction, [13].

4. Experiments

Experiments are carried out estimating age and gender from ECG, with the aim of lowering variance due to cardial condition. In this scenario, CVD labels will be used to define a patient as either healthy or unhealthy (with at least one positive CVD label). The task weights are set to one, $\lambda = 1$, using binary cross-entropy loss for learning cardial condition and gender, and mean squared logarithmic loss for age prediction.

4.1. Data

PTB-XL is a public ECG dataset collected between 1989 and 1996, containing 21,799 clinical 12-lead ECGs from 18,869 patients, [14]. Each ECG consists of a 10

second time series at 500 Hz, downsampled to 100 Hz. Annotations of the ECGs have been done by cardiologists, and also include age and gender.

CODE15 is another public dataset containing 345,779 12-lead ECGs from 233,770 patients, collected between 2010 and 2016, [15]. CODE15 also provides diagnostic information labeled by cardiologists, albeit at a lower level of detail. The ECG time series are recorded for 10 or 7 seconds at 400 Hz. The 7 second samples are zero-padded to match the larger sample size of the 10 second ECGs. For CODE15 a downsampling to 100 HZ is also done.

Running experiments on both datasets may illustrate where there is an advantage of applying multitask learning. PTB-XL represents the more likely scenario in a medical setting, where the learned trait, is highly skewed within the data; combined with having few samples of the variance inducing traits at the tails. CODE15 have a more uniform age distribution, and enough data that parts of the variability may be captured in a regular model, see Fig. 3.

4.2. Setup

Benchmarking is done for all three model configurations on both PTB-XL and CODE15. The same backbone setup is used, consisting of six inception blocks, where each module use three convolutional filters with length 40, 20 and 10. Experiments are implemented in Tensorflow using Adam optimizer, learning rate 0.001 and batch size of 20. In the Bayesian scenario, Tensorflow Probability is used for variational inference using dense flipout layers. Model performance is validated on 20% of the training data during training. For PTB-XL the standard pre-defined train and test split is used, corresponding to 90% and 10% of the dataset respectively. For CODE15 a 75/25% split is used, with 20% of the training data used for validation.

The first experiment uses a fully connected layer for each task separately with 64 units connected to a single final unit. Similarly, the second experiment uses 64 units in the fully connected layers, with a single unit in the final classification layers. Age and gender are connected to the cardial condition prediction branch separately, with no link between them. Finally, the third experiment uses the same structure, with Bayesian learning on the final layers.

In addition to the three proposed models, the backbone is trained separately using a fully connected layer for age prediction, as comparison. The model proposed in [5] is also implemented for benchmarking purposes.

4.3. Results

Mean average error of all estimations can be found in Table 1. Results are reported for the full test population and stratified on cardial health. Although the multi-task model, especially model 2, shows state-of-the-art performance; the main advantage is lowering variance in the unhealthy population. This is further evident in Fig. 2 showing the distribution of age predictions in different age groups. A sample of predictions from the Bayesian model can be found in Fig. 3, illustrating the potential benefit of the confidence estimations, even if the model performs worse on average.

	Age MAE			Gender AUC		
PTB-XL	All	Heal.	Unhe.	All	Heal.	Unhe.
MTL 1	7.75	7.54	7.95	0.915	0.947	0.879
MTL 2	7.53	7.54	7.51	0.907	0.944	0.872
MTL 3	9.16	9.79	8.77	0.895	0.937	0.861
STL Inc	11.3	8.00	13.8	-	-	-
STL [5]	13.7	10.4	16.1	-	-	-
CODE15						
MTL 1	8.05	8.00	8.45	0.942	0.947	0.901
MTL 2	7.93	7.89	8.14	0.950	0.953	0.921
MTL 3	9.80	9.78	10.1	0.925	0.930	0.896
STL Inc	8.23	8.17	8.72	-	-	-
STL [5]	8.52	8.38	9.01	-	-	-

Table 1. The table shows results in terms of mean average error of age and area under the curve for gender prediction.

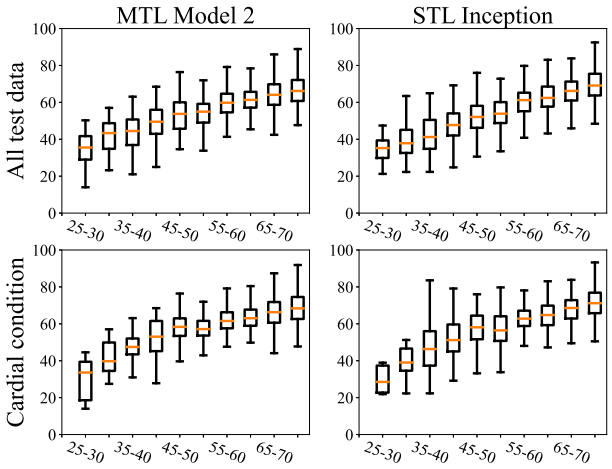

Figure 2. Distribution of age predictions in different age groups on PTB-XL. Higher variance is particularly prevalent in the STL model on cases with cardial conditions.

5. Discussion

The proposed framework produce better results than the single-task model for gender and age prediction, especially at the stratified test group. This is particularly evident on PTB-XL which has less data and more skewed distributions of the involved variables. The overall lower mean average error in PTB-XL for the multi-task models may be due to the age prediction operating on a lower range than in

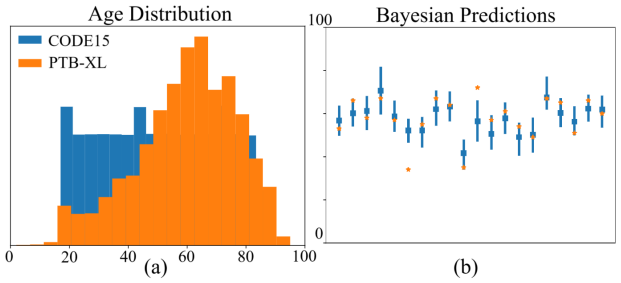

Figure 3. (a) shows the age distribution in the two datasets. (b) shows 20 randomly selected bayesian predictions in the PTB-XL test set. Orange markers indicates the true label, blue marker indicates prediction with confidence interval.

CODE15. For example, PTB-XL has very few adolescents in the dataset while CODE15 is more evenly distributed. Similarly, the single-task models perform much better in CODE15, since parts of this particular variability (CVD annotated by humans) is possible to learn from the data, given the sample size. The Bayesian model has worse performance metrics on average, but could still be useful since it provides confidence on the estimations. To improve the age and gender estimations further, additional tasks affecting variability could be incorporated into the model.

5.1. MTL improvements

One of the main challenges of MTL is accurate loss weighting for λ. Rather than setting all weights to one, as in the experiments above, a more optimal distribution can be derived through e.g. brute force search on the validation set, [16]. In face recognition similar designs have been optimized by dynamically changing the weights during training or incorporating them into the optimization task, [8].

Another challenge is the choice of architecture for the task inference from the feature extractor. Three different methods are explored, however, many other variations are available. As the number of additional tasks and variables are included the interconnection and dependence between them get gradually more complex. Instead of hand-designing the multi-task connections they can also be dynamically changed within the learning task, [17].

6. Conclusion

This work looks into intra-subject variability of ECG and its effect in biometric applications. In particular, MTL is proposed as a method of mitigating variance in soft biometric estimations. Three different multi-task setups were benchmarked showing state-of-the-art performance in age and gender prediction, while lowering estimation variance among instances. The effects are particularly prevalent on the smaller dataset with a skewed variable distribution.

References

[1] Angulu R, Tapamo JR, Adewumi AO. Age estimation via face images: a survey. EURASIP Journal on Image and Video Processing 2018;2018(1):1–35.

[2] Ranjbar A, et al. Enabling clinical trials of artificial intelligence: Infrastructure for heart failure predictions. Studies in health technology and informatics 2023;302:177–181.

[3] Reid DA, et al. Soft biometrics for surveillance: an overview. Handbook of statistics 2013;31:327–352.

[4] Pinto JR, Cardoso JS, Lourenço A. Evolution, current challenges, and future possibilities in ecg biometrics. IEEE Access 2018;6:34746–34776.

[5] Attia ZI, et al. Age and sex estimation using artificial intelligence from standard 12-lead ecgs. Circulation Arrhythmia and Electrophysiology 2019;12(9):e007284.

[6] Agrafioti F, Bui FM, Hatzinakos D. Secure telemedicine: Biometrics for remote and continuous patient verification. Journal of Computer Networks and Communications ;2012.

[7] Ranjbar A, Skolt K, Vik KTA, Øistad BS, Mork EW, Ravn J. Fairness in artificial intelligence: Regulatory sanbox evaluation of bias prevention for ecg classification. Studies in health technology and informatics 2023;302:488–489.

[8] Yin X, Liu X. Multi-task convolutional neural network for pose-invariant face recognition. IEEE Transactions on Image Processing 2017;27(2):964–975.

[9] Zhang K, Zhang Z, Li Z, Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters 2016;23(10).

[10] Fawaz HI, et al. Inceptiontime: Finding alexnet for time series classification. Data Mining and Knowledge Discovery 2020;34(6).

[11] Caruana R. Multitask learning. Springer, 1998.

[12] Graves A. Practical variational inference for neural networks. Advances in neural information processing systems 2011;24.

[13] Kristiadi A, Hein M, Hennig P. Being bayesian, even just a bit, fixes overconfidence in relu networks. In International conference on machine learning. PMLR, 2020; 5436–5446.

[14] Wagner P, et al. Ptb-xl, a large publicly available electrocardiography dataset. Scientific data 2020;7(1):154.

[15] Ribeiro AH, et al. Code-15%: a large scale annotated dataset of 12-lead ecgs. `https://doi.org/10.5281/zenodo.4916206`, 2021.

[16] Tian Y, Luo P, Wang X, Tang X. Pedestrian detection aided by deep learning semantic tasks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2015; 5079–5087.

[17] Lu Y, et al. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In Proceedings of the IEEE CVPR. 2017; 5334–5343.

Address for correspondence:

Arian Ranjbar
Sykehusveien 25, 1478 Nordbyhagen, Norway
arian.ranjbar@ahus.no