Age Prediction From 12-lead Electrocardiograms Using Deep Learning: A Comparison of Four Models on a Contemporary, Freely Available Dataset

Andrew Barros, MD, MS[1,2]

Ian German-Mesner, MS[1]

N. Rich Nguyen, PhD[1,3]

J. Randall Moorman, MD[1,4]

1. Center for Advanced Medical Analytics (CAMA), School of Medicine, University of Virginia, Charlottesville, VA
2. Division of Pulmonary and Critical Care Medicine, Department of Medicine, School of Medicine, University of Virginia, Charlottesville, VA
3. Department of Computer Science, University of Virginia, Charlottesville, VA
4. Division of Cardiovascular Medicine, Department of Medicine, School of Medicine, University of Virginia, Charlottesville, VA

ABSTRACT

The 12-lead electrocardiogram (ECG) is routine in clinical use and deep learning approaches have been shown to have the identify features not immediately apparent to human interpreters including age and sex. ECG predicted age has been identified as a predictor of long-term mortality. Here we compare four models for age and sex prediction on a contemporary, freely available dataset.

INTRODUCTION

The electrocardiogram (ECG) entered clinical use in the early 1900's and earned Willem Einthoven a Nobel prize in 1924[1]. For much of the last 100 years, the ECG has been acquired by a machine, printed in a canonical paper form, and interpreted by physicians and many conditions manifest as alterations of the ECG. In the 1980's automated interpretation was developed and added to many machines[2]. Recently, deep learning approaches have been used to achieve cardiologist level interpretation performance[3].

Beyond features immediately identifiable to human interpreters, deep neural networks have been shown to accurately identify hyperkalemia[4], reduced left ventricular ejection fraction[5], risk of incident atrial fibrillation[6], and age[7–9]. Furthermore, an ECG predicted age that exceeds chronological age has been associated with mortality[8–11]. However, most models are trained and reported on local data preventing between model comparisons. We sought to compare the performance of models on a new, publicly available data set.

METHODS

Data Source, Ethical Declarations, and Reporting Standards

Our data comes from the MIMIC-IV-ECG module[12] a collection of 800,000 diagnostic electrocardiograms collected at a single hospital and linked to the Medical Information Mart for Intensive Care (MIMIC-IV)[9,13,14] a detailed, deidentified medical record summary of hospitalizations, emergency department visits, and intensive care unit stays. All data is available at PhysioNet[15] and was determined to be not human subjects research. This work adheres to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) recommendations[16] and the supplementary recommendations from the editors of Respiratory, Sleep, and Critical Care journals[17].

Patient Inclusion

MIMIC uses time shifting for patient privacy where the first encounter for each patient is randomly shifted into the future and all dates for that patient are shifted by the same amount. Additionally, patient who are older than 89 at any time during their observation period are further shifted so that no age can be identified, and only relative differences are available. We included all ECGs for all patients with an identifiable (i.e. who were younger than 90 at all their time points) and who were older than 18 at the time of their ECG collection.

Data Allocation

We allocated the available data by patient with 20% of the patients reserved for final model testing. Of the remaining 80%, we allocated 80% (64% of the total cohort) for model training

and 20% (16% of the total cohort) for validation. All ECGs from a subject were included in a single data set. The training set was used for model fitting while the validation set was used for model selection and parameter tuning. The test set was only used for final model evaluation.

Preprocessing

Individual records were preprocessed by applying a wandering baseline filter (0.2hz band stop filter centered at 0.8hz with 40db of attenuation), a 40hz low-pass filter, down sampled to 240hz, zero padded (if necessary) to 2048 samples, and then standardized (to mean zero and variance 1). Records with missing values were dropped.

Outcomes

Our primary outcomes were age in years at the time ECG acquisition and the patient's gender recorded in the EHR. We used the maximal Youden's J-index on the validation set to identify the optimal cutoff for classification metrics.

Models

A schematic representation of all four models is shown in Figure 1.

<u>Attia (CNN)</u>

Attia and colleagues developed and tested a convolutional neural network (CNN) for the prediction of age and gender[7] using 774,783 standard ECGs collected at the Mayo Clinic between January 1994 and February 2017. Their model used eight temporal layers, followed by

one spatial layer, and then either a classification or regression head. The temporal and spatial layers consisted of 1 dimensional convolutions (either along the time or space axis), batch normalization, ReLU activation, and max pooling. The classification or regression heads consisted of two repeated blocks of fully connected, batch normalization, ReLU activation, and dropout followed by a single output for age or two softmax outputs for gender. The details of kernel sizes, filter counts, strides, and max pooling factors are reported in their original paper[7] but none of the source code, training data, or model weights are publicly available.

Lima (ResNet)

Lima and colleagues developed and tested a CNN with residual connections (i.e. ResNet[18] style) for the prediction of age and associated mortality[8]. Their model uses an initial 1-D convolutional block, followed by four ResNet-styled blocks, and finally a dense classification or regression head. The ResNet-styled blocks combine a 1-D convolutional block with a "skip pathway" where input data is added to the convolutional output and then passed forward. The source code, model weights, and most of the training data is publicly available.

Nejedly (MHA)

Nejedly et. al. submitted the winning entry[19] to the 2021 Computing in Cardiology Challenge[20]. They used a CNN with ResNet styled blocks and multi-head attention to deliver the best performance on rhythm classification. We have adapted this model for age and gender prediction. The original source code is publicly available, and we have provided our source code, model weights, and input pipeline for the MIMIC-ECG data.

Our Model (ResNet+MHA)

Finally, we adapted the Lima model by adding a multi-head attention block after the ResNet styled blocks. We have provided our source code, model weights, and input pipeline for the MIMIC-ECG data.

Training

All models were trained for 50 epochs with the Adam optimizer and a learning rate of 1e-4 except for the ResNet+MHA model which was trained at 3e-5. We used binary cross entropy (BCE) as the loss function for sex prediction and mean squared error (MSE) for age prediction. The model with the best validation set results was saved and used for the final evaluation.

Software and Source Code Availability

We used Python 3.10.12, PyTorch[21] 2.1.1, and Ignite 0.4.13[22] on the University of Virginia Rivanna high performance computing cluster to develop and evaluate all models. The source code to support this analysis is available at https://github.com/UVA-CAMA/EcgAgePrediction/

RESULTS

All models converged. There was evidence of overfitting in all the sex prediction models with a significant gap between training and validation set performance (Figure 2). In contrast, the gap between training and validation performance was much smaller in the age prediction models.

For the task of age prediction, the ResNet model had the best test set MAE and validation set MAE, MSE, and R^2 while the CNN model had the lowest test set MSE and R^2(Table 1). Overall, both models had very similar results. The observed test set performance for the ResNet model was similar to the original published performance (MAE 8.06 years versus 8.38 years) but the CNN performance was worse that published (MAE 8.25 years versus 6.9 years, R^2 0.6 versus 0.7).

For the task of sex prediction, the ResNet+MHA model had the best accuracy, AUROC, Brier score, F1 score, and specificity on both the testing and validation set (Table 2). The AUROC for the CNN model was lower than previously published (0.902 versus 0.968) as was the accuracy (0.821 vs 0.904).

DISCUSSION

We trained and compared four models on a contemporary, freely available dataset. Although age and sex prediction are not directly clinically applicable this framework allows for comparison of architectures across tasks on a single dataset.

We acknowledge that only the CNN model was initially designed for sex prediction and that the MHA model was designed for neither sex prediction nor age prediction. However, given that Attia uses the same architecture for both tasks we thought it was reasonable to compare all models on all tasks. An ideal architecture would be foundational and allow for task specific fine-

tuning. We found that the best architecture for a sex prediction was not the best for age prediction (and vice versa).

One weakness of our analysis is that we re-implemented parts of all the reference models. The source code for the CNN model was not available and we needed to make some assumptions to develop our implementation. We were able to directly use most of the ResNet implementation but did need to make minor changes due to changes in library versions and data formats. While our implementations are not identical to the original authors, they are a good faith attempt to implement and evaluate the models on a common reference.

CONCLUSIONS

Age can be predicted with reasonable accuracy from a standard 12-lead ECG with a performance that varies by network architecture. Future models should consider using a publicly available dataset so that between model comparisons can be made.

REFERENCES

1. Pahlm O, Uvelius B. The winner takes it all: Willem Einthoven, Thomas Lewis, and the Nobel prize 1924 for the discovery of the electrocardiogram. *J Electrocardiol.* 2019;57:122-127. doi:10.1016/j.jelectrocard.2019.09.012

2. Willems JL, Abreu-Lima C, Arnaud P, et al. Testing the performance of ECG computer programs: the CSE diagnostic pilot study. *J Electrocardiol.* 1987;20 Suppl:73-77.

3. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun*. 2020;11(1):1760. doi:10.1038/s41467-020-15432-4

4. Galloway CD, Valys AV, Shreibati JB, et al. Development and Validation of a Deep-Learning Model to Screen for Hyperkalemia From the Electrocardiogram. *JAMA Cardiol*. 2019;4(5):428. doi:10.1001/jamacardio.2019.0640

5. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med*. 2019;25(1):70-74. doi:10.1038/s41591-018-0240-2

6. Raghunath S, Pfeifer JM, Ulloa-Cerna AE, et al. Deep Neural Networks Can Predict New-Onset Atrial Fibrillation From the 12-Lead ECG and Help Identify Those at Risk of Atrial Fibrillation–Related Stroke. *Circulation*. 2021;143(13):1287-1298. doi:10.1161/CIRCULATIONAHA.120.047829

7. Attia ZI, Friedman PA, Noseworthy PA, et al. Age and Sex Estimation Using Artificial Intelligence From Standard 12-Lead ECGs. *Circ: Arrhythmia and Electrophysiology*. 2019;12(9):e007284. doi:10.1161/CIRCEP.119.007284

8. Lima EM, Ribeiro AH, Paixão GMM, et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nat Commun*. 2021;12(1):5117. doi:10.1038/s41467-021-25351-7

9. Baek YS, Lee DH, Jo Y, Lee SC, Choi W, Kim DH. Artificial intelligence-estimated biological heart age using a 12-lead electrocardiogram predicts mortality and cardiovascular outcomes. *Front Cardiovasc Med*. 2023;10:1137892. doi:10.3389/fcvm.2023.1137892

10. Lorenz EC, Zaniletti I, Johnson BK, et al. Physiological Age by Artificial Intelligence-Enhanced Electrocardiograms as a Novel Risk Factor of Mortality in Kidney Transplant Candidates. *Transplantation*. 2023;107(6):1365-1372. doi:10.1097/TP.0000000000004504

11. Ladejobi AO, Medina-Inojosa JR, Shelly Cohen M, et al. The 12-lead electrocardiogram as a biomarker of biological age. *European Heart Journal - Digital Health*. 2021;2(3):379-389. doi:10.1093/ehjdh/ztab043

12. Gow B, Pollard T, Nathanson LA, et al. MIMIC-IV-ECG - Diagnostic Electrocardiogram Matched Subset. doi:10.13026/TW25-EC93

13. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV. doi:10.13026/6MM1-EK67

14. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10(1):1. doi:10.1038/s41597-022-01899-x

15. Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. 2000;101(23):E215-220. doi:10.1161/01.cir.101.23.e215

16. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Ann Intern Med*. 2015;162(10):735-736. doi:10.7326/L15-5093-2

17. Leisman DE, Harhay MO, Lederer DJ, et al. Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals. *Critical Care Medicine*. 2020;48(5):623-633. doi:10.1097/CCM.0000000000004246

18. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2016:770-778. doi:10.1109/CVPR.2016.90

19. Nejedly P, Ivora A, Smisek R, et al. Classification of ECG Using Ensemble of Residual CNNs with Attention Mechanism. In: *2021 Computing in Cardiology (CinC)*. IEEE; 2021:1-4. doi:10.23919/CinC53138.2021.9662723

20. Reyna MA, Sadr N, Alday EAP, et al. Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. In: *2021 Computing in Cardiology (CinC)*. IEEE; 2021:1-4. doi:10.23919/CinC53138.2021.9662687

21. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Published online 2019. doi:10.48550/ARXIV.1912.01703

22. Fomin V, Anmol J, Desroziers S, Kriss J, Tejani A. High-level library to help with training neural networks in PyTorch. *GitHub repository*. Published online 2020. https://github.com/pytorch/ignite

DATA AVAILABILITY

The source code for this analysis is available at GitHub. The data is available from PhysioNet.

FUNDING

The work of AB was conducted with the support of the iTHRIV Scholars Program. The iTHRIV Scholars Program is supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Numbers UL1TR003015 and KL2TR003016 as well as by The University of Virginia. This content is solely the responsibility of the authors and does not necessarily represent the official views of NIH or UVA.

TABLES

Table 1

Model	MAE		MSE		RSQ	
	Test	Val	Test	Val	Test	Val
ResNet + MHA	12.51	12.69	221.14	224.55	0.17	0.17
ResNet	8.06	7.87	108.84	103.49	0.59	0.62
MHA	9.12	9.21	130.52	131.27	0.51	0.52
CNN	8.25	8.28	107.40	106.57	0.60	0.61

This table shows the validation and test set performance for the age prediction task with the best values for each metric are highlighted in blue. Note that similar performance of the ResNet and CNN models. MAE: Mean absolute error, MSE: Mean squared error, RSQ: R^2 correlation.

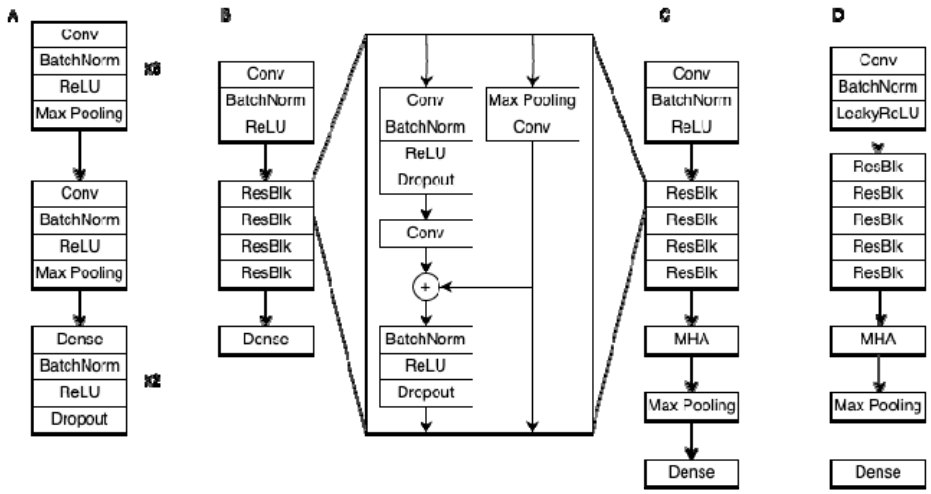
Table 2

Model	ACC		AUROC		Brier		F1		Sens		Spec	
	Test	**Val**	**Test**	**Val**	**Test**	**Val**	**Test**	**Val**	**Test**	**Val**	**Test**	**Val**
ResNet+MHA	0.847	0.852	0.925	0.928	0.112	0.109	0.850	0.851	0.841	0.847	0.854	0.857
ResNet	0.840	0.844	0.919	0.922	0.119	0.116	0.845	0.845	0.842	0.848	0.837	0.840
CNN	0.821	0.829	0.902	0.909	0.126	0.122	0.828	0.831	0.835	0.842	0.805	0.816
MHA	0.813	0.818	0.895	0.898	0.132	0.131	0.819	0.819	0.819	0.823	0.806	0.813

This table shows the validation and test set performance for the sex prediction task with the best values for each metric are highlighted in blue. Note the ResNet+MHA model has the best (or near best) performance across all metrics.

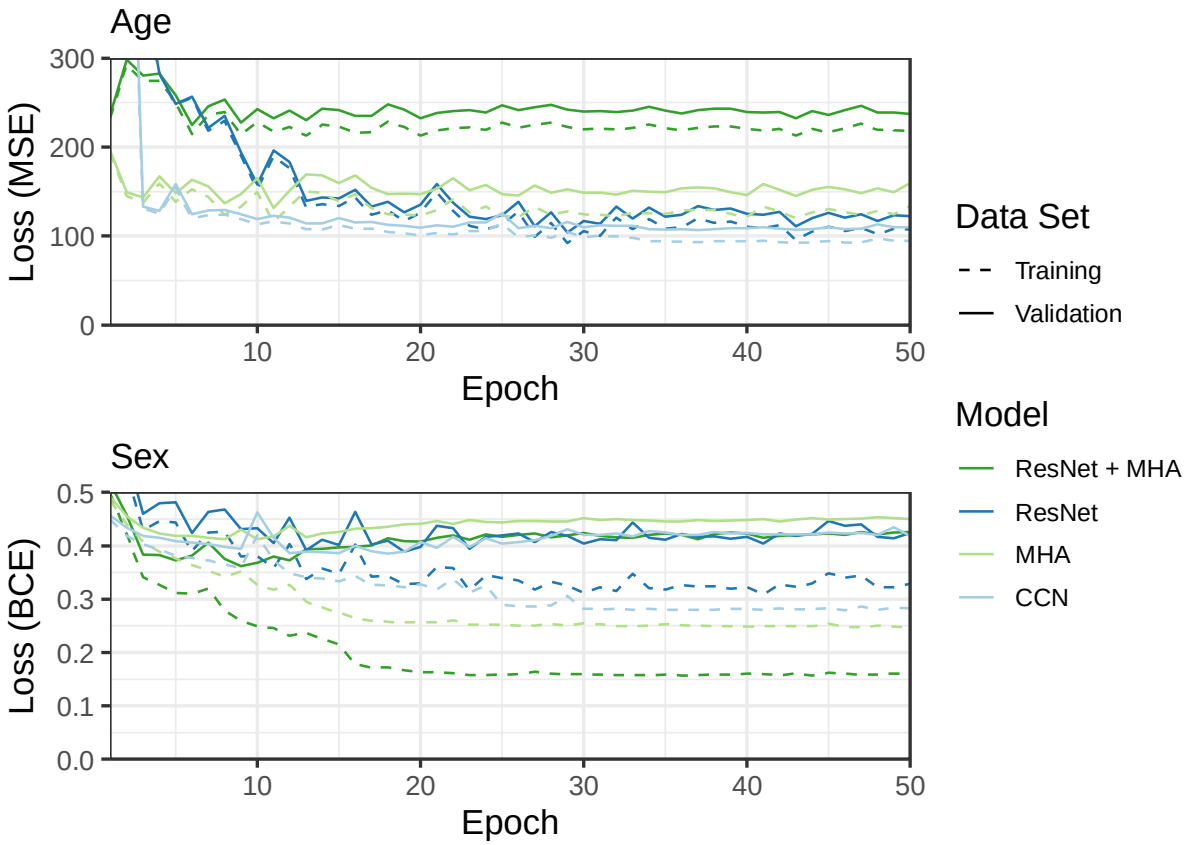
FIGURES

Figure 1



This figure shows the architectures for the a) CNN b) ResNet c) ResNet + MHA and d) MHA models.

Figure 2



This figure shows the losses by epoch for the age prediction task (top panel) and sex prediction task (bottom panel). Note the gap between training and validation set performance on the sex prediction task with essentially similar validation set performance across all four models. In contrast, the gap on the age prediction task is generally less than the between model gap.