medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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]

Corresponding Author:

Andrew Barros

ajb5d@virginia.edu

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

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

ABSTRACT

Objective
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. Several models have been published but no direct comparisons exist.

Approach
We implemented three previously published models and one unpublished model to predict age and sex from a 12-lead ECG and then compared their performance on an open-access data set.

Main results
All models converged and were evaluated on the holdout set. The best preforming age prediction model had a hold-out set mean absolute error of 8.06 years. The best preforming sex prediction model had a hold-out set area under the receiver operating curve of 0.92.

Significance
We compared performance of four models on an open-access dataset.

INTRODUCTION

The electrocardiogram (ECG) entered clinical use in the early 1900's and earned Willem Einthoven a Nobel prize in 1924(Pahlm and Uvelius 2019). 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. Many conditions manifest as alterations of the ECG. In the 1980's automated interpretation was developed and added to many machines(Willems *et al* 1987).

Deep learning has diffused throughout the scientific community(Bianchini *et al* 2020). In cardiology, deep learning has been used automated arrythmia classification and has achieved cardiologist level performance(Chen *et al* 2020, Ribeiro *et al* 2020, Nejedly *et al* 2022, Zhao *et al* 2022). The 2020 Physionet/Computing in Cardiology Challenge yielded 212 entries and the most

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

common technique was deep learning and convolutional neural networks(Perez Alday *et al* 2020).

Beyond features immediately identifiable to human interpreters, deep neural networks have been shown to accurately identify hyperkalemia(Galloway *et al* 2019), reduced left ventricular ejection fraction(Attia *et al* 2019b), risk of incident atrial fibrillation(Raghunath *et al* 2021), and age(Attia *et al* 2019a, Lima *et al* 2021, Baek *et al* 2023). Furthermore, an ECG predicted age that exceeds chronological age has been associated with mortality(Lima *et al* 2021, Baek *et al* 2023, Lorenz *et al* 2023, Ladejobi *et al* 2021). 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(Gow *et al* n.d.) a collection of 800,000 diagnostic electrocardiograms collected at a single hospital and linked to the Medical Information Mart for Intensive Care (MIMIC-IV)(Johnson *et al* n.d., 2023, Baek *et al* 2023) a detailed, deidentified medical record summary of hospitalizations, emergency department visits, and intensive care unit stays. All data is available at PhysioNet(Goldberger *et al* 2000) 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)

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

recommendations(Collins *et al* 2015) and the supplementary recommendations from the editors of Respiratory, Sleep, and Critical Care journals(Leisman *et al* 2020).

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 wide elliptical infinite impulse response[IIR] high-pass filter centered at 0.8hz with 40db of attenuation), a powerline filter(2nd order digital IIR notch filter centered at 60hz), a 40hz low-

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

pass filter (5th order Butterworth IIR filter), down sampled to 240hz, truncated (or zero padded if required) to 2048 samples, and then standardized (to mean zero and variance 1). Records with missing values were dropped. The signal processing was implemented using the SciPy package(Virtanen *et al* 2020). The 40hz low-pass filter was chosen based on the empiric observation that down sampling produced similar or improved results with a substantial improvement in training times. A comparison of results is shown in eFigure 3.

Outcomes

Our primary outcomes were age in years at the time ECG acquisition and the patient's sex recorded in the EHR. In the MIMIC-ECG dataset each subject has a baseline age and an a achor age (The date shifted year in which the baseline age was observed). Each exam has a date shifted timestamp. We calculated the at the age at time of the exam as $[\text{Baseline Age}] + ([\text{Exam Timestamp Year}] - [\text{Anchor Year}])$.

Metrics

For the models predicting age our primary metric was mean absolute error (MAE) with additional metrics of mean squared error (MSE) and coefficient of determination (R2). For the models predicting sex we used the maximal Youden's J-index(Youden 1950) on the validation set to identify the optimal cutoff for classification metrics. The primary evaluation metrics were area under the receiver operating characteristic (AUROC) and accuracy (ACC). Additional metrics were the Brier score(Kruppa *et al* 2014), F1(Buckland and Gey 1994), Sensitivity (Sens)(Altman and Bland 1994), and specificity (Spec)(Altman and Bland 1994). Implementations of all metrics came from the Yardstick package(Kuhn *et al* 2024).

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Models

A schematic representation of all four models is shown in Figure 1. Parameter details, including filter sizes, strides, and max pooling factors, are available in the supplement.

Attia (CNN)

Attia and colleagues developed and tested a convolutional neural network (CNN) for the prediction of age and sex(Attia *et al* 2019a) 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 sex. The details of kernel sizes, filter counts, strides, and max pooling factors are reported in their original paper(Attia *et al* 2019a) 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(He *et al* 2016) style) for the prediction of age and associated mortality(Lima *et al* 2021). 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

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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(Nejedly *et al* 2021) to the 2021 Computing in Cardiology Challenge(Reyna *et al* 2021). 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 sex 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. The rates were chosen by manual tuning and the observation that the ResNet+MHA model had loss explosion at higher rates. Each model was trained twice: once for sex prediction (the output predicting the log-odds of male sex) and once for age prediction (with the output predicting age in years). We used binary cross entropy (BCE) as the loss function for sex prediction and mean squared error (MSE) for age

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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(Paszke *et al* 2019) 2.1.1, and Ignite 0.4.13(Fomin *et al* 2020) 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.

RESULTS

The final datasets included 510,316 records in the training set, 127,044 records in the validation set, and 162,675 records in the testing set. In the training set 49.1% of the records came from women while 47.8% of the records in the testing set and 49.6% of the records in the validation set came from women. The median age in each of the dataset was 66 and a histogram of observed ages is included in eFigure 1.

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(Fgure 3). A plot with observed versus predicted age for a random subset of the data is shown in Figure 4.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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). There was visual evidence of heteroskedasticity in the prediction residuals with a decreasing variance for increasing age.

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 (Figure 4). Only the CNN model attempted sex prediction in the prior work. 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).

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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.

A strength of our work includes use of a common open-access dataset and metrics that would allow for direct comparisons of future models against these baselines. This approach is seen in other machine learning domains such as ImageNet(Deng *et al* 2009), SQuAD(Rajpurkar *et al* 2016), and MS-COCO(Lin *et al* 2015) and is partially responsible for the continued progress seen in deep learning.

A significant limitation of both tasks is the lack of direct clinical applicability. Indeed, age prediction in a vacuum is a parlor trick but coupled with real outcomes has potential applications. The observed errors have two parts: the reducible error due to the model and irreducible error due to natural variation. The total amount of reducible error is unknowable and thus we cannot identify when a model is preforming well -- only better than another model. Long term changes in predicted age have been associated with significant outcomes but, to our knowledge, no work has examined if these predictions change over short time scales (say during a hospitalization) and what outcomes these changes might predict. We propose that open,

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

comparable models can facilitate work that examines downstream applications of prediction tasks.

CONCLUSIONS

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

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.

REFERENCES

Altman D G and Bland J M 1994 Diagnostic tests. 1: Sensitivity and specificity *BMJ* **308** 1552

Attia Z I, Friedman P A, Noseworthy P A, Lopez-Jimenez F, Ladewig D J, Satam G, Pellikka P A, Munger T M, Asirvatham S J, Scott C G, Carter R E and Kapa S 2019a Age and Sex Estimation Using Artificial Intelligence From Standard 12-Lead ECGs *Circ: Arrhythmia and Electrophysiology* **12** e007284

Attia Z I, Kapa S, Lopez-Jimenez F, McKie P M, Ladewig D J, Satam G, Pellikka P A, Enriquez-Sarano M, Noseworthy P A, Munger T M, Asirvatham S J, Scott C G, Carter R E and

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Friedman P A 2019b Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram *Nat Med* **25** 70–4

Baek Y-S, Lee D-H, Jo Y, Lee S-C, Choi W and Kim D-H 2023 Artificial intelligence-estimated biological heart age using a 12-lead electrocardiogram predicts mortality and cardiovascular outcomes *Front Cardiovasc Med* **10** 1137892

Bianchini S, Müller M and Pelletier P 2020 Deep Learning in Science Online: http://arxiv.org/abs/2009.01575

Buckland M and Gey F 1994 The relationship between Recall and Precision *J. Am. Soc. Inf. Sci.* **45** 12–9

Chen T-M, Huang C-H, Shih E S C, Hu Y-F and Hwang M-J 2020 Detection and Classification of Cardiac Arrhythmias by a Challenge-Best Deep Learning Neural Network Model *iScience* **23** 100886

Collins G S, Reitsma J B, Altman D G and Moons K G M 2015 Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) *Ann Intern Med* **162** 735–6

Deng J, Dong W, Socher R, Li L-J, Kai Li, and Li Fei-Fei 2009 ImageNet: A large-scale hierarchical image database *2009 IEEE Conference on Computer Vision and Pattern Recognition* 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops) (Miami, FL: IEEE) pp 248–55 Online: https://ieeexplore.ieee.org/document/5206848/

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

Galloway C D, Valys A V, Shreibati J B, Treiman D L, Petterson F L, Gundotra V P, Albert D E, Attia Z I, Carter R E, Asirvatham S J, Ackerman M J, Noseworthy P A, Dillon J J and Friedman P A 2019 Development and Validation of a Deep-Learning Model to Screen for Hyperkalemia From the Electrocardiogram *JAMA Cardiol* **4** 428

Goldberger A L, Amaral L A, Glass L, Hausdorff J M, Ivanov P C, Mark R G, Mietus J E, Moody G B, Peng C K and Stanley H E 2000 PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals *Circulation* **101** E215-220

Gow B, Pollard T, Nathanson L A, Johnson A, Moody B, Fernandes C, Greenbaum N, Berkowitz S, Moukheiber D, Eslami P, Herbst E, Mark R and Horng S MIMIC-IV-ECG - Diagnostic Electrocardiogram Matched Subset Online: https://physionet.org/content/mimic-iv-ecg/0.1/

He K, Zhang X, Ren S and Sun J 2016 Deep Residual Learning for Image Recognition *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016 IEEE Conference on

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV, USA: IEEE) pp 770–8 Online: http://ieeexplore.ieee.org/document/7780459/

Johnson A, Bulgarelli L, Pollard T, Horng S, Celi L A and Mark R MIMIC-IV Online: https://physionet.org/content/mimiciv/2.2/

Johnson A E W, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard T J, Hao S, Moody B, Gow B, Lehman L H, Celi L A and Mark R G 2023 MIMIC-IV, a freely accessible electronic health record dataset *Sci Data* **10** 1

Kruppa J, Liu Y, Diener H-C, Holste T, Weimar C, König I R and Ziegler A 2014 Probability estimation with machine learning methods for dichotomous and multicategory outcome: applications *Biom J* **56** 564–83

Kuhn M, Vaughan D and Hvitfeldt E 2024 *yardstick: Tidy Characterizations of Model Performance* Online: https://github.com/tidymodels/yardstick

Ladejobi A O, Medina-Inojosa J R, Shelly Cohen M, Attia Z I, Scott C G, LeBrasseur N K, Gersh B J, Noseworthy P A, Friedman P A, Kapa S and Lopez-Jimenez F 2021 The 12-lead electrocardiogram as a biomarker of biological age *European Heart Journal - Digital Health* **2** 379–89

Leisman D E, Harhay M O, Lederer D J, Abramson M, Adjei A A, Bakker J, Ballas Z K, Barreiro E, Bell S C, Bellomo R, Bernstein J A, Branson R D, Brusasco V, Chalmers J D, Chokroverty S, Citerio G, Collop N A, Cooke C R, Crapo J D, Donaldson G, Fitzgerald D A, Grainger E, Hale L, Herth F J, Kochanek P M, Marks G, Moorman J R, Ost D E, Schatz M, Sheikh A, Smyth A R, Stewart I, Stewart P W, Swenson E R, Szymusiak R, Teboul J-L, Vincent J-L, Wedzicha J A and Maslove D M 2020 Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals *Critical Care Medicine* **48** 623–33

Lima E M, Ribeiro A H, Paixão G M M, Ribeiro M H, Pinto-Filho M M, Gomes P R, Oliveira D M, Sabino E C, Duncan B B, Giatti L, Barreto S M, Meira Jr W, Schön T B and Ribeiro A L P 2021 Deep neural network-estimated electrocardiographic age as a mortality predictor *Nat Commun* **12** 5117

Lin T-Y, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick C L and Dollár P 2015 Microsoft COCO: Common Objects in Context Online: http://arxiv.org/abs/1405.0312

Lorenz E C, Zaniletti I, Johnson B K, Petterson T M, Kremers W K, Schinstock C A, Amer H, Cheville A L, LeBrasseur N K, Winkelmayer W C, Navaneethan S D, Baez-Suarez A, Attia Z I, Lopez-Jimenez F, Friedman P A, Kennedy C C and Rule A D 2023 Physiological Age by Artificial Intelligence-Enhanced Electrocardiograms as a Novel Risk Factor of Mortality in Kidney Transplant Candidates *Transplantation* **107** 1365–72

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Nejedly P, Ivora A, Smisek R, Viscor I, Koscova Z, Jurak P and Plesinger F 2021 Classification of ECG Using Ensemble of Residual CNNs with Attention Mechanism *2021 Computing in Cardiology (CinC)* 2021 Computing in Cardiology (CinC) (Brno, Czech Republic: IEEE) pp 1–4 Online: https://ieeexplore.ieee.org/document/9662723/

Nejedly P, Ivora A, Viscor I, Koscova Z, Smisek R, Jurak P and Plesinger F 2022 Classification of ECG using ensemble of residual CNNs with or without attention mechanism *Physiol. Meas.* **43** 044001

Pahlm O and Uvelius B 2019 The winner takes it all: Willem Einthoven, Thomas Lewis, and the Nobel prize 1924 for the discovery of the electrocardiogram *J Electrocardiol* **57** 122–7

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J and Chintala S 2019 PyTorch: An Imperative Style, High-Performance Deep Learning Library Online: https://arxiv.org/abs/1912.01703

Perez Alday E A, Gu A, J Shah A, Robichaux C, Ian Wong A-K, Liu C, Liu F, Bahrami Rad A, Elola A, Seyedi S, Li Q, Sharma A, Clifford G D and Reyna M A 2020 Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020 *Physiol. Meas.* **41** 124003

Raghunath S, Pfeifer J M, Ulloa-Cerna A E, Nemani A, Carbonati T, Jing L, vanMaanen D P, Hartzel D N, Ruhl J A, Lagerman B F, Rocha D B, Stoudt N J, Schneider G, Johnson K W, Zimmerman N, Leader J B, Kirchner H L, Griessenauer C J, Hafez A, Good C W, Fornwalt B K and Haggerty C M 2021 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* **143** 1287–98

Rajpurkar P, Zhang J, Lopyrev K and Liang P 2016 SQuAD: 100,000+ Questions for Machine Comprehension of Text Online: http://arxiv.org/abs/1606.05250

Reyna M A, Sadr N, Alday E A P, Gu A, Shah A J, Robichaux C, Rad A B, Elola A, Seyedi S, Ansari S, Ghanbari H, Li Q, Sharma A and Clifford G D 2021 Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021 *2021 Computing in Cardiology (CinC)* 2021 Computing in Cardiology (CinC) (Brno, Czech Republic: IEEE) pp 1–4 Online: https://ieeexplore.ieee.org/document/9662687/

Ribeiro A H, Ribeiro M H, Paixão G M M, Oliveira D M, Gomes P R, Canazart J A, Ferreira M P S, Andersson C R, Macfarlane P W, Meira W, Schön T B and Ribeiro A L P 2020 Automatic diagnosis of the 12-lead ECG using a deep neural network *Nat Commun* **11** 1760

Virtanen P, Gommers R, Oliphant T E, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt S J, Brett M, Wilson J, Millman K J, Mayorov N, Nelson A R J, Jones E, Kern R, Larson E, Carey C J, Polat İ, Feng Y, Moore E W, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero E A, Harris C R,

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Archibald A M, Ribeiro A H, Pedregosa F, van Mulbregt P, and SciPy 1.0 Contributors 2020 SciPy 1.0: fundamental algorithms for scientific computing in Python *Nat Methods* **17** 261–72

Willems J L, Abreu-Lima C, Arnaud P, van Bemmel J H, Brohet C, Degani R, Denis B, Graham I, van Herpen G and Macfarlane P W 1987 Testing the performance of ECG computer programs: the CSE diagnostic pilot study *J Electrocardiol* **20 Suppl** 73–7

Youden W J 1950 Index for rating diagnostic tests *Cancer* **3** 32–5

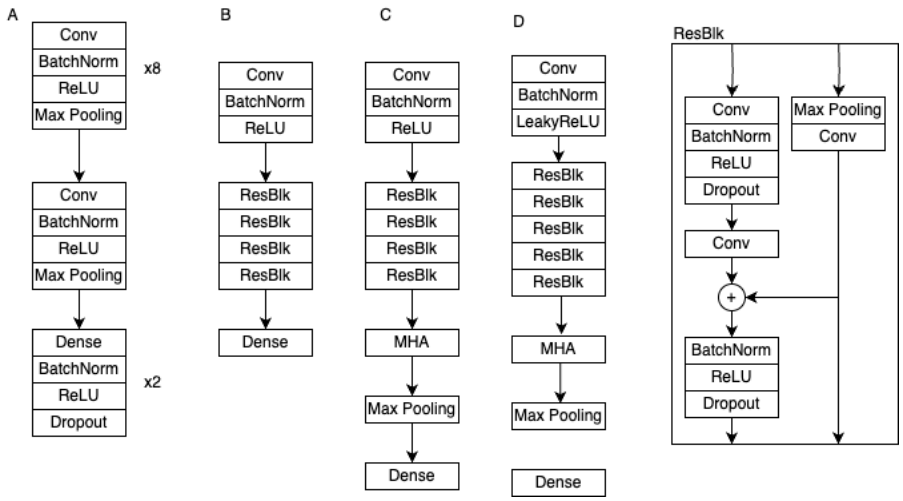
Zhao Z, Murphy D, Gifford H, Williams S, Darlington A, Relton S D, Fang H and Wong D C 2022 Analysis of an adaptive lead weighted ResNet for multiclass classification of 12-lead ECGs *Physiol. Meas.* **43** 034001

DATA AVAILABILITY

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

FIGURES

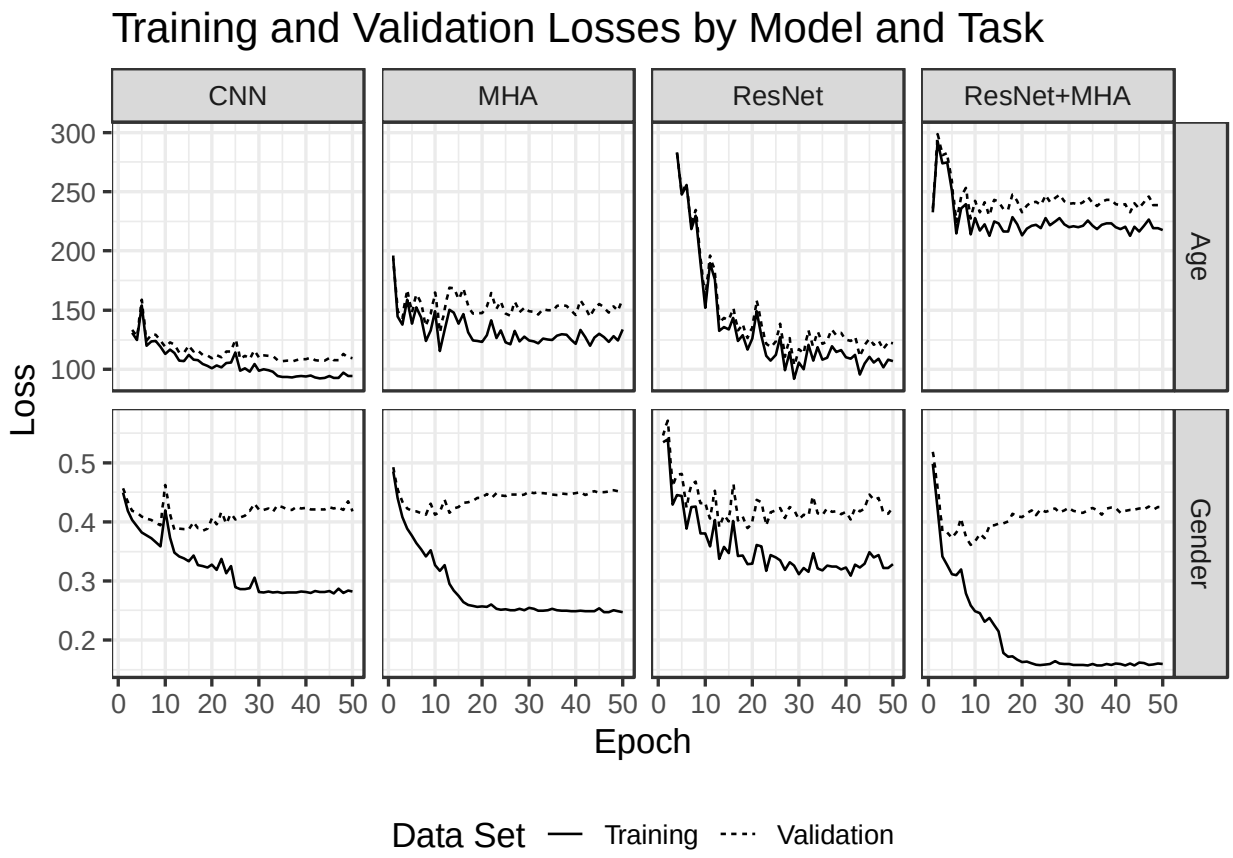
Figure 1



This figure shows the architectures for the a) CNN b) ResNet c) ResNet + MHA and d) MHA models. The ResNet inspired block (ResBlk) is shared by networks B, C, and D.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

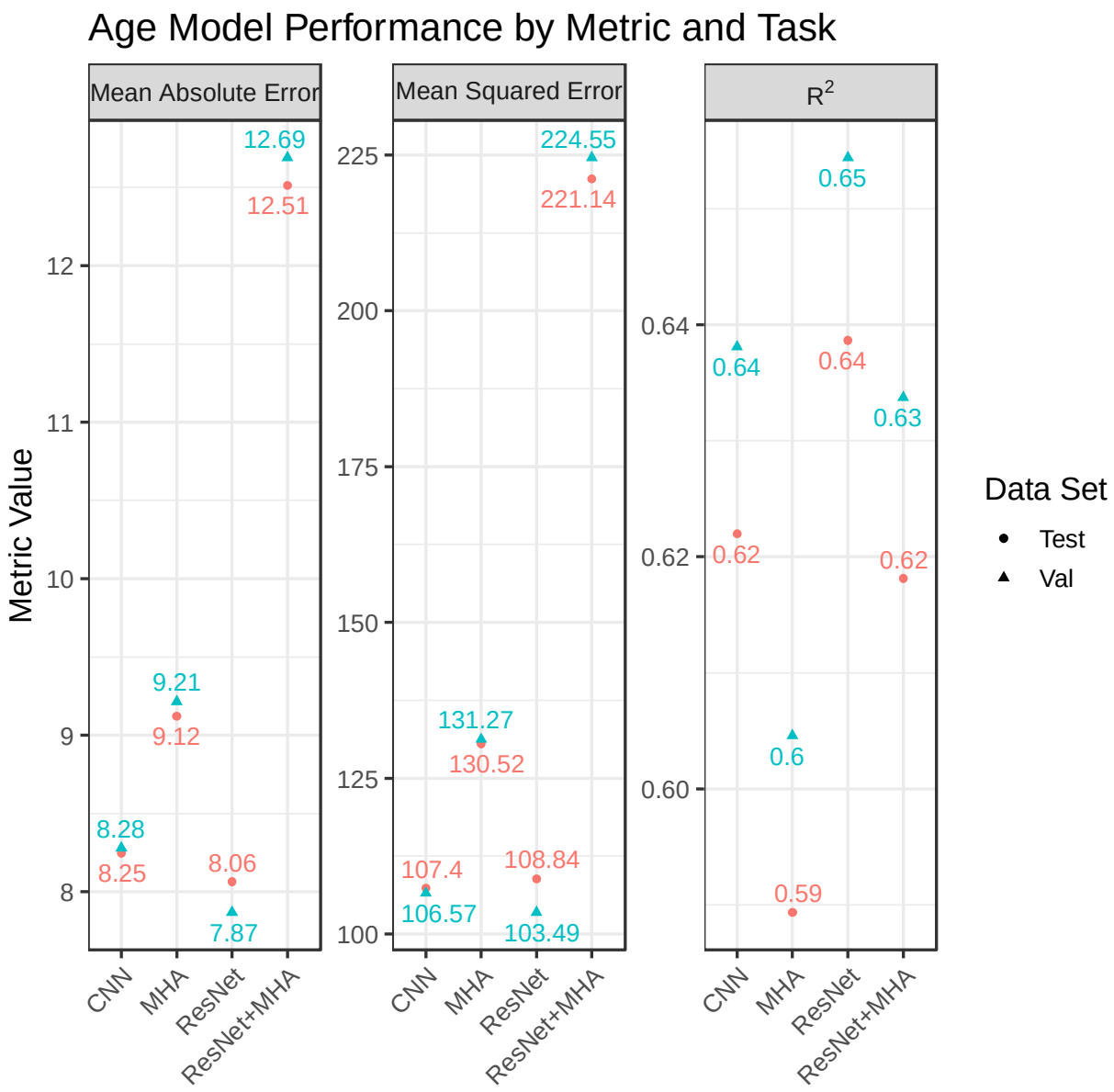
Figure 2



This figure shows the losses by epoch for the age prediction task (top row) and sex prediction task (bottom row). 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.

Figure 3

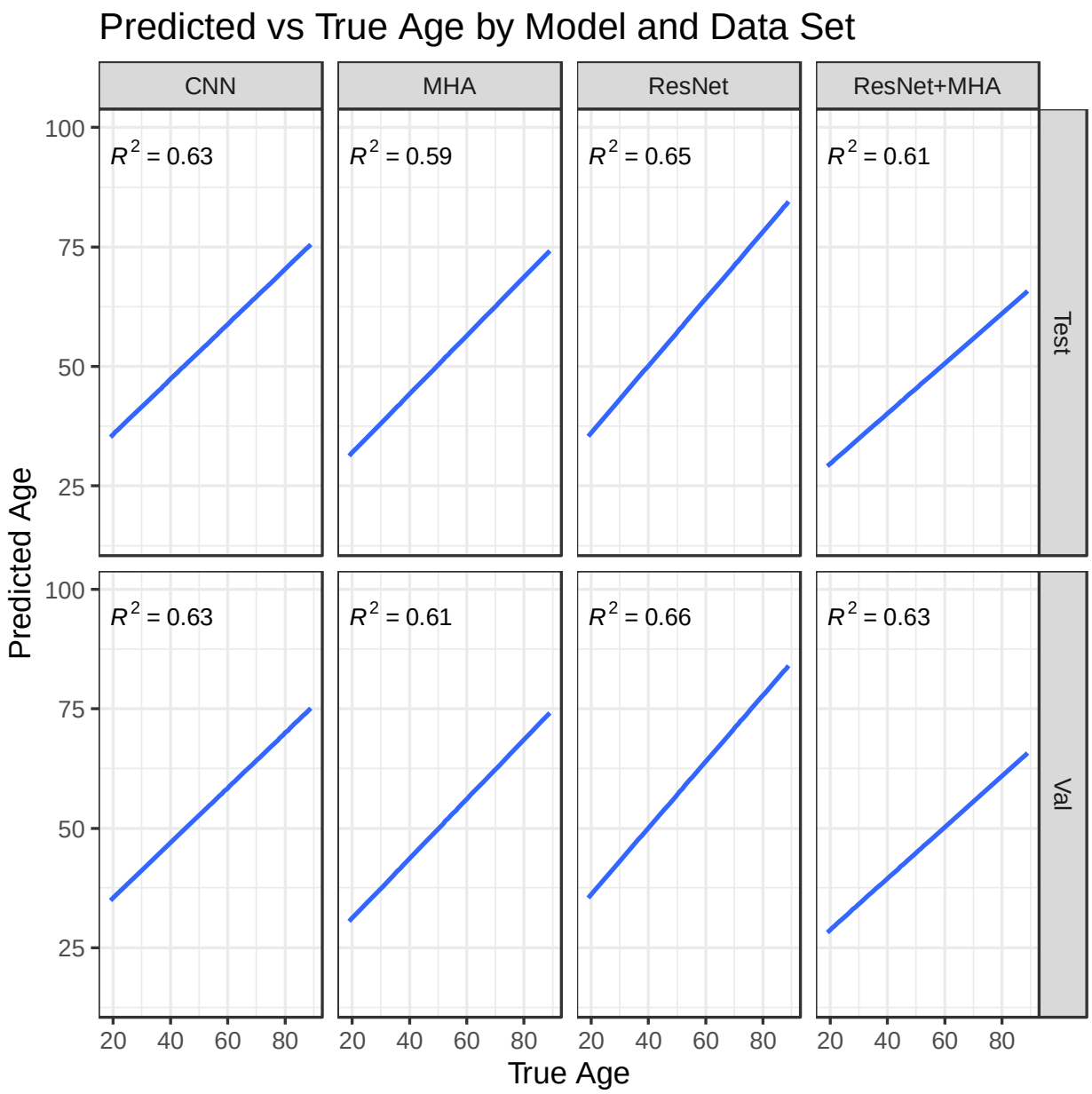
medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



Model performance across the three age prediction metrics. Note that for mean absolute error and mean squared error lower is better while higher is better for R2. The test set is the hold-out set while validation was used for model tuning. Neither set was used for model fitting.

Figure 4

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



Observed versus predicted ages for a random sample of 10,000 patients across the validation and testing sets. The blue line is the linear regression line of best fit.

Figure 5

medRxiv preprint doi: https://doi.org/10.1101/2024.02.02.24302201; this version posted July 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Model performance across the sex prediction metrics: accuracy (ACC), area under the receiver operating curve (AUROC), Brier score, F1, Sensitivity (Sens), and Specificity (Spec). Threshold metrics (ACC, F1, Sens, and Spec) were evaluated at the optimal point selected by Youden's J-index. The test set is the hold-out set while validation was used for model tuning. Neither set was used for model fitting.