ResNet-based ECG Diagnosis of Myocardial Infarction in the Emergency Department

Daniel Gedon[1,*] DANIEL.GEDON@IT.UU.SE
Stefan Gustafsson[1,2,*] STEFAN.GUSTAFSSON@MEDSCI.UU.SE
Erik Lampa[1] ERIK.LAMPA@MEDSCI.UU.SE
Antônio H. Ribeiro[1] ANTONIO.HORTA.RIBEIRO@IT.UU.SE
Martin J. Holzmann[3]
Thomas B. Schön[1] THOMAS.SCHON@IT.UU.SE
Johan Sundström[1,4] JOHAN.SUNDSTROM@MEDSCI.UU.SE
[1] *Uppsala University, Sweden;* [2] *Sence Research AB, Sweden;*
[3] *Karolinska University Hospital and Karolinska Institutet, Sweden;*
[4] *University of New South Wales, Australia.*

Abstract

Myocardial infarctions (MIs) are often missed in the emergency department. In managed settings deep learning models have shown promise in electrocardiogram (ECG) classification. However, in a real-world scenario there is a lack of high performing models for classification of MIs. We developed a ResNet-based deep neural network to classify the ECG between non-ST-elevation MI (NSTEMI), ST-elevation MI (STEMI), and control status in the more challenging real-world setting. In a test set, our model discriminates STEMIs/NSTEMIs with an AUROC of 0.85/0.76 and a Brier score of 0.10/0.18. The model also generalizes well and obtains a similar performance on an additional test set collected in the months following the initial collection and that does not overlap temporally with the set used for developing the model. Our results are on par with human-level performance reported in previous studies for STEMIs and above human-level for NSTEMIs.

1. Introduction

Emergency department care costs are high (Galarraga and Pines, 2016) and rising (Lane et al., 2020) in developed societies. Based on limited data in a chaotic environment, emergency doctors must make quick decisions about patients' probabilities for many diagnoses and risks. Diagnostic error is commonplace (Medford-Davis et al., 2016; Moonen et al., 2017), and there is need for decision support systems (Wright et al., 2019). The emergency department handling of myocardial infarctions (MIs) is especially precarious with 10-50,000 missed cases per year at emergency departments in the United States (Sharp et al., 2021). Less than half of those hospitalized for a suspected MI are diagnosed with it (Caulfield and Stephens, 2018).

The ECG can reveal large ST-elevation MIs (STEMIs), but non-ST-elevation MIs NSTEMIs are often unremarkable to the human eye on the ECG. Physicians at all training levels have deficiencies in ECG interpretation, with an accuracy of 0.69 overall for practicing physicians and 0.75 for cardiologists in controlled test settings (Cook et al., 2020), with similar numbers reported for STEMIs (McCabe et al., 2013; Soares et al., 2019). Diagnosing NSTEMIs with ECGs by physicians is much lower rendering this a particularly difficult problem for humans.

Finding NSTEMIs early and starting treatment improves patient outcomes (Hamm et al., 2011). Since NSTEMIs are impossible to reliably diagnose without blood test, automatically detecting NSTEMIs from an ECG at the emergency department would enable early treatment and potentially prevents missing NSTEMIs.

Deep learning has shown recent promise in ECG classification (Siontis et al., 2021), for common ECG diagnoses (Ribeiro et al., 2020) as well as for traits with unclear ECG diagnostic criteria or those not usually thought of as ECG diagnoses (Cohen-Shelly et al., 2021; Raghunath et al., 2020; Tison et al., 2019). Even ECGs that appear normal to the human eye carry useful information for deep models (Attia et al., 2019; Raghunath et al., 2020). In the diagnosis of MI deep

* These authors contributed equally.

learning is promising (Cho et al., 2020; Liu et al., 2021), but many studies have used limited (Al-Zaiti et al., 2020; Makimoto et al., 2020) or managed (Al-Zaiti et al., 2020; Cho et al., 2020; Liu et al., 2021; Makimoto et al., 2020; Zhao et al., 2020) datasets. Deep learning models using managed datasets and those that sought to discriminate STEMIs often reported super-human-level performance (Cho et al., 2020; Liu et al., 2021). The very few studies that used more real-world-like samples or sought to discriminate also NSTEMIs generally reported human-level or sub-human-level performance (Liu et al., 2021).

We develop and validate a deep learning based model for ECG decision support in diagnosing MIs in the emergency department. We tackle a difficult real-world scenario with immediate benefits for practicing physicians. Our study presents a tentative solution of an unsolved problem using a novel dataset.

2. Methods

Data Sample We utilize a routine 10-second 12-lead ECGs from adult patients attending emergency departments in the Stockholm region between 2007 and 2016, that had such a high risk of an acute coronary syndrome that they were admitted to a coronary care unit, after obtaining an ECG. After applying the sequence of filters (Figure S1), to ensure inclusion of at event before-treatment ECGs as well as confirming the outcome label, 10,583 patients with 12,311 coronary care unit admissions and 16,628 ECGs were available for analysis. For the labels we use of the high-quality SWEDEHEART registry[1]. The labels are the decision of a discharging physician that followed the entire patient journey during hospitalisation (including blood testing for all patients). We can then tie the labels to available electronic health records and national records, e.g. to connect MI patients to increased Troponin levels (Table S1). Details on data sources and exposures, outcomes are in Appendix A.1 and A.2

Training and test datasets The studied patients were divided for training and test in 70%/30% splits with records from the same patient in the same split. Patient characteristics fluctuated with time and dividing the 30% test split into two containing 20% and 10% of the complete data allows us to test the model in different scenarios. The 10% split contains exams with admission date of 2016-01-01 or later which has no temporal overlap with the training dataset. This split can be used to assess the model susceptibility to temporal shifts and trends. We denote this split the *temporal test split*. The 20% split was sampled at random from entries with an admission date before 2016-01-01, the same period as for the training split. We denote this split the *random test split*.

Data pre-processing Data pre-processing steps are described in Figure S2. Next to the ECG tracings as input, we limited ourselves to age and sex, to increase the transportability of the model. The output are the probabilities of the three mutually exclusive outcome classes: NSTEMI/STEMI/control.

Model architecture Our model is an extension of Ribeiro et al. (2020), in which the model classifies six ECG abnormalities (Alkmim et al., 2012). We used a ResNet based architecture for unidimensional signals to process ECG tracings. Age and sex were passed through a fully connected layer and concatenated with the output of the ResNet. The resulting features were used in the linear classification layer. The model architecture with its extensions and the hyperparameters are described in Figure S3 and Appendix A.3.

Generally, ensembles of neural network models improve predictive performance (Hansen and Salamon, 1990) and model calibration. We therefore expanded our model as an ensemble of five model members. The logits were averaged to obtain the final prediction.

Model calibration While discrimination (e.g. AUROC) is important, it contains no information about the reliability of probability estimates. Calibration, i.e. if the model's probability estimates reflect the ground truth empirical class frequencies, is an important model property for clinical use. Since deep models are shown to be poorly calibrated (Guo et al., 2017), one of our concerns regards calibration.

As metrics to evaluate calibration, we focus on the Expected Calibration Error (ECE) and the Brier score estimated on the test sets, see Appendix A.5. We visualize calibration plots of our model in Figure S10.

Model analysis To identify possible patterns in the STEMI/NSTEMI classification, we highlight parts of the ECG that the model focuses on for its prediction using Grad-CAM plots (Selvaraju et al., 2017). Visualization are generated in two steps: In a forward pass we compute the activations of the neural network in an intermediary layer (we use the first convolutional layer). In a backward step we compute the gradients corresponding to these activations. The gradients are

1. ucr.uu.se/swedeheart/dokument-sh/variabellista, accessed: 2021-06-25

averaged to get the proportional importance of each channel, which is then used to compute a proportional mean of the activations. Positive values were plotted as purple disks overlaid on top of the ECG, with size proportional to magnitude. One cardiologist (JS) inspected the Grad-CAM plots from ten cases with highest probability for STEMI/NSTEMI, and selected four representative plots each for illustration.

3. Results

Of the included 12,311 coronary care unit admissions, 3,993 were recorded with NSTEMI, 1,340 with STEMI, and 6,978 without MI. Clinical characteristics of the study sample are listed in Table S3 and stratified for our data splits in Table S2. Patients' age and admission date distributions are shown in Figures S6, S7.

The performance of our model in the two test datasets is listed in Table 1. Note that there exists no direct baseline since we use a novel dataset and there are no openly accessible datasets with NSTEMIs. In the random test set, STEMIs could be discriminated with fair precision, with an AUROC of 0.85 and a Brier score of 0.10. Discrimination of NSTEMIs was poorer, with an AUROC of 0.76 and a Brier score of 0.18. Therefore, our model achieves human-level performance in classifying STEMIs and super-human-level performance for NSTEMI. The performance in the temporal test set, that did not overlap in time with the development set, was similar. Controls were classified with lower accuracy in the temporal test set.

Further results are shown in Figure 1. The left plot shows the development of the model with incorporation of more advanced model and training techniques compared to Ribeiro et al. (2020). The middle and right plots show the receiver-operator and precision-recall curves on the temporal test split.

Inspecting Grad-CAM plots yielded new insights. Figure 2 illustrates STEMIs and an NSTEMI correctly classified with high probability. These illustrations are cropped versions of Figure S4 and S5. For STEMIs in Figure 2 (left) the model focuses on the ST-segment, where a human would look. In Figure 2 (middle), the model focused on the down-sloping part of the T-wave, where a human would not focus for a STEMI diagnosis. For NSTEMIs in Figure 2 (right) the model focuses on the ST-segment. Humans would agree with the model that the ST-segment depressions look potentially ischemic; but would not suspect ischemia based on the ECGs in panels C and D of Figure S5. More results are attached in Appendix B.

Table 1: **Performance of the model in the two test sets.** Given are means and standard deviations over ten different trained models; each of which is an ensemble consisting of five model members. We compute AUROC, Average Precision and Brier score for STEMI *vs* Control and NSTEMI *vs* Control.

		Random	Temporal
Accuracy	Control	0.75 (0.007)	0.44 (0.011)
	NSTEMI	0.57 (0.012)	0.72 (0.010)
	STEMI	0.71 (0.013)	0.72 (0.020)
AUROC	NSTEMI	0.76 (0.003)	0.74 (0.003)
	STEMI	0.85 (0.002)	0.82 (0.003)
Avg. Prec.	NSTEMI	0.69 (0.003)	0.64 (0.005)
	STEMI	0.76 (0.005)	0.64 (0.006)
Brier	NSTEMI	0.19 (0.001)	0.27 (0.002)
	STEMI	0.10 (0.001)	0.13 (0.001)
ECE	Multiclass	0.25 (0.004)	0.11 (0.012)

4. Discussion

This study deals with the relevant population of all-comer patients at emergency departments, representing the real-world experience for doctors with ECGs. Notably, in Sweden pre-hospital ECGs (e.g. in ambulances) are sent to coronary care units for immediate diagnosis. Hence, obvious STEMI cases bypass the emergency department and transfer straight to the coronary intervention lab upon arrival to hospital, rendering the STEMIs in the present study the less obvious cases and the walk-ins. Further, we did not exclude difficult cases, comorbidities, or previous MIs[2].

Other studies have shown similar performance to ours in representative settings (Liu et al., 2021), with reports of very good performance in managed settings (Cho et al., 2020; Liu et al., 2021). Our study differs as it is a multicenter study using data of consecutive patients with very few exclusions, and with labeling by many doctors. As controls, our study uses all hard cases that raised a high suspicion of a MI and were admitted to a coronary care unit, and no easy non-MI cases, which sets it apart from previous studies that included such controls (Liu et al., 2021) or not clearly described the controls (Cho et al., 2020).

2. Except for technical reasons, we removed potentially linked hospitalizations for the same MI, and LBBBs, which cannot *per se* identify an acute MI from a single ECG, but need a prior ECG for comparison.

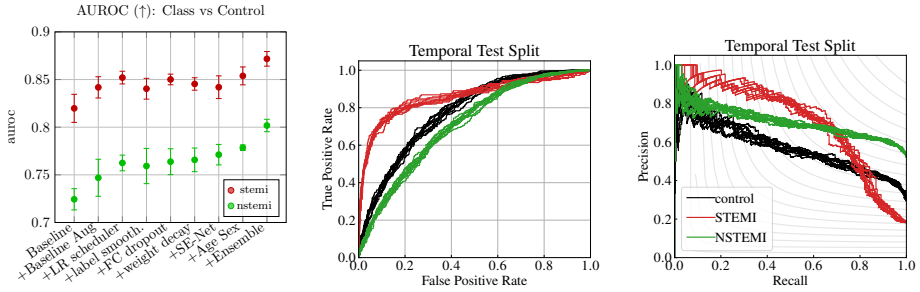

Figure 1: **(Left)** AUROC of model improvements for different training extensions. **(Middle)** ROC curve for temporal test split. All 10 seeds are plotted. Shown is class vs all. **(Right)** Precision-Recall curve for temporal test split. Iso-F_1 curves in the background. Full figures in Appendix A.4 and B.

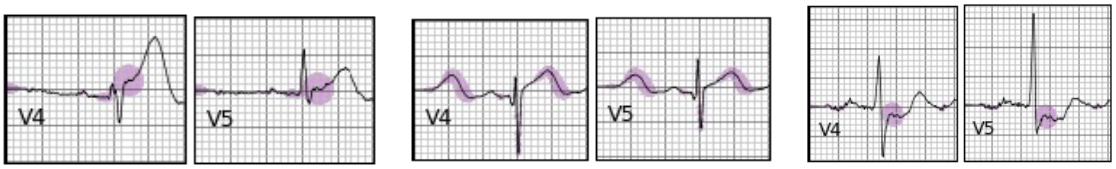

Figure 2: Correctly classified examples with high probability with corresponding Grad-CAM plots. We show leads V4 and V5 of the examples. **(Left)** STEMI with typical ST-segment elevation highlighted. **(Middle)** STEMI with another feature highlighted that is not typical to doctors (establishing the relevance of it would need additional study). **(Right)** NSTEMI with typical but unspecific ST-segment depression highlighted.

The Grad-CAM plots in Figures S4 and S5 provide important insights: In half of the panels the model recognizes the same features that humans would. In the other half the model finds features that are novel or imperceptible to physicians, indicating possible traits. Variants of such model analysis can likely give useful clinical and pathophysiological clues in many medical fields.

Our model's misclassifications as STEMI follows known clinical and machine learning patterns, with myocarditis as an important impostor (Tanguay et al., 2019). The conditions over-represented in those misclassified as NSTEMI were logical to some extent, such as aortic stenosis and pulmonary edema; the late-stage diabetes traits more surprising.

An important limitation is the lack of an external validation sample. We did hold out the 10% most recent episodes for a temporal test set; many circumstances in that set would be similar to those in the training set, but a restructuring of the Stockholm region emergency department logistics during the study period did change the composition of the sample. This likely explains the poorer accuracy of controls in the temporal test set than in the random test set. Furthermore, the label in our study was determined at discharge from the coronary care unit, when the whole care episode could be summarized. The ECGs in the test sets of this study may hence not always be the ones guiding the final diagnosis. We mitigate that to some extent by using multiple ECGs if available within the day before admission in the training set, but not in the test sets. On the other hand, the hindsight allows for more stable labels for the episode as a whole, which is the ultimate goal for the classification.

We present a deep learning model with performance that is comparable to cardiologist performance reported in previous studies (Cook et al., 2020) in classifying STEMIs and above human performance for NSTEMIs. We do so in a real-world sample of emergency department patients with a high suspicion of acute coronary syndrome. Considering the high and rising emergency department costs and the high numbers of missed MIs, our model could be of clinical value for ECG decision support at this stage, with promise of further performance improvement.

Acknowledgments

Author contributions

Conceptualisation, study design, review, editing: All authors. Data acquisition and curation: MJH, SG. Funding acquisition, project administration and supervision: MJH, JS, TBS. Figures, data analysis, data interpretation: SG, DG, EL, AHR. Writing original draft: SG, DG, EL, AHR, JS. All authors had access and could verify the underlying data.

Non-author contributions

We thank David Widmann for the discussions and his inputs towards analyzing and improving the calibration of our model.

Role of the funding source

The study was funded by The Kjell and Märta Beijer Foundation, Anders Wiklöf, The Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation, and Uppsala University. The computations were enabled by resources in project sens2020005 and sens2020598 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973. The funders had no role in study design; collection, analysis, and interpretation of data, writing of the report, or decision to submit the paper for publication.

Declaration of interests

JS reports stock ownership in companies providing services to Itrim, Amgen, Janssen, Novo Nordisk, Eli Lilly, Boehringer, Bayer, Pfizer and AstraZeneca, outside the submitted work.

Data sharing

Neither data nor other related documents in the study can be made available to others, for medicolegal reasons. The code is available upon request.

References

B. Af Ugglas, T. Djärv, P. L. S. Ljungman, and M. J. Holzmann. Association Between Hospital Bed Occupancy and Outcomes in Emergency Care: A Cohort Study in Stockholm Region, Sweden, 2012 to 2016. *Ann Emerg Med*, 76:179–190, 2020.

S. Al-Zaiti, L. Besomi, Z. Bouzid, Z. Faramand, S. Frisch, C. Martin-Gill, R. Gregg, S. Saba, C. Callaway, and E. Sejdić. Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. *Nat Commun*, 11:3966, 2020.

M. B. Alkmim, R. M. Figueira, M. S. Marcolino, C. S. Cardoso, M. Pena de Abreu, L. R. Cunha, D. F. da Cunha, A. P. Antunes, A. G. Resende, E. S. Resende, and A. L. Ribeiro. Improving patient access to specialized health care: the Telehealth Network of Minas Gerais, Brazil. *Bull World Health Organ*, 90:373–378, 2012.

Z. I. Attia, P. A. Noseworthy, F. Lopez-Jimenez, S. J. Asirvatham, A. J. Deshmukh, B. J. Gersh, R. E. Carter, X. Yao, A. A. Rabinstein, B. J. Erickson, S. Kapa, and P. A. Friedman. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet*, 394:861–867, 2019.

I. Bello, W. Fedus, X. Du, E. D. Cubuk, A. Srinivas, T.-Y. Lin, J. Shlens, and B. Zoph. Revisiting resnets: Improved training and scaling strategies. *CoRR*, abs/2103.07579, 2021.

C. A. Caulfield and J. R. Stephens. Things We Do for No Reason: Hospitalization for the Evaluation of Patients with Low-Risk Chest Pain. *J Hosp Med*, 13:277–279, 2018.

Y. Cho, J. M. Kwon, K. H. Kim, J. R. Medina-Inojosa, K. H. Jeon, S. Cho, S. Y. Lee, J. Park, and B. H. Oh. Artificial intelligence algorithm for detecting myocardial infarction using six-lead electrocardiography. *Sci Rep*, 10:20495, 2020.

M. Cohen-Shelly, Z. I. Attia, P. A. Friedman, S. Ito, B. A. Essayagh, W. Y. Ko, D. H. Murphree, H. I. Michelena, M. Enriquez-Sarano, R. E. Carter, P. W. Johnson, P. A. Noseworthy, F. Lopez-Jimenez, and J. K. Oh. Electrocardiogram screening for aortic valve stenosis using artificial intelligence. *Eur Heart J*, 2021. In Press.

D. A. Cook, S. Y. Oh, and M. V. Pusic. Accuracy of Physicians' Electrocardiogram Interpretations: A Systematic Review and Meta-analysis. *JAMA Intern Med*, 180:1461–1471, 2020.

J. E. Galarraga and J. M. Pines. Costs of ED episodes of care in the United States. *Am J Emerg Med*, 34: 357–365, 2016.

C. Guo, G. Pleiss, Sun Y., and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017.

C. W. Hamm, J. P. Bassand, S. Agewall, J. Bax, E. Boersma, H. Bueno, P. Caso, D. Dudek, S. Gielen, K. Huber, M. Ohman, M. C. Petrie, F. Sonntag, M. S. Uva, R. F. Storey, W. Wijns, D. Zahger, J. J. Bax, A. Aurricchio, H. Baumgartner, C. Ceconi, V. Dean, C. Deaton, R. Fagard, C. Funck-Brentano, D. Hasdai, A. Hoes, J. Knuuti, P. Kolh, T. McDonagh, C. Moulin, D. Poldermans, B. A. Popeswcu, Ž. Reiner, U. Sechtem, P. Anton Sirnes, A. Torbicki, A. Vahanian, S. Windecker, S. Achenbach, L. Badimon, M. Bertrand, H. E. Bøtker, J. P. Collet, F. Crea, N. Danchin, E. Falk, J. Goudevenos, D. Gulba, R. Hambrecht, J. Herrmann, A. Kastrati, K. Kjeldsen, S. D. Kristensen, P. Lancellotti, J. Mehilli, B. Merkely, G. Montalescot, F. J. Neumann, L. Neyses, J. Perk, M. Roffi, F. Romeo, M. Ruda, E. Swahn, M. Valgimigli, C. J. Vrints, and P. Widimsky. ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: The Task Force for the management of acute coronary syndromes (ACS) in patients presenting without persistent ST-segment elevation of the European Society of Cardiology (ESC). *Eur Heart J*, 32(23): 2999–3054, Dec 2011.

L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:993–1001, 1990.

T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

T. Jernberg, M. F. Attebring, K. Hambraeus, T. Ivert, S. James, A. Jeppsson, B. Lagerqvist, B. Lindahl, U. Stenestrand, and L. Wallentin. The Swedish Web-system for enhancement and development of evidence-based care in heart disease evaluated according to recommended therapies (SWEDEHEART). *Heart*, 96:1617–1621, 2010.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

M. Kull, M. Perello Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, 2019.

B. H. Lane, P. J. Mallow, M. B. Hooker, and E. Hooker. Trends in United States emergency department visits and associated charges from 2010 to 2016. *Am J Emerg Med*, 38:1576–1581, 2020.

L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations*, 2020.

W. C. Liu, C. S. Lin, C. S. Tsai, T. P. Tsao, C. C. Cheng, J. T. Liou, W. S. Lin, S. M. Cheng, Y. S. Lou, C. C. Lee, and C. Lin. A Deep-Learning Algorithm for Detecting Acute Myocardial Infarction. *EuroIntervention*, 2021. In Press.

I. Loshchilov and F. Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

H. Makimoto, M. Höckmann, T. Lin, D. Glöckner, S. Gerguri, L. Clasen, J. Schmidt, A. Assadi-Schmidt, A. Bejinariu, P. Müller, S. Angendohr, M. Babady, C. Brinkmeyer, A. Makimoto, and M. Kelm. Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction. *Sci Rep*, 10:8445, 2020.

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*, 2:e000268, 2013.

L. Medford-Davis, E. Park, G. Shlamovitz, J. Suliburk, A. N. Meyer, and H. Singh. Diagnostic errors related to acute abdominal pain in the emergency department. *Emerg Med J*, 33:253–259, 2016.

P. J. Moonen, L. Mercelina, W. Boer, and T. Fret. Diagnostic error in the Emergency Department: follow up of patients with minor trauma in the outpatient clinic. *Scand J Trauma Resusc Emerg Med*, 25:13, 2017.

R. Müller, S. Kornblith, and G. E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, 2019.

S. Raghunath, A. E. Ulloa Cerna, L. Jing, D. P. vanMaanen, J. Stough, D. N. Hartzel, J. B. Leader, H. L. Kirchner, M. C. Stumpe, A. Hafez, A. Nemani, T. Carbonati, K. W. Johnson, K. Young, C. W. Good, J. M. Pfeifer, A. A. Patel, B. P. Delisle, A. Alsaid, D. Beer, C. M. Haggerty, and B. K. Fornwalt. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med*, 26:886–891, 2020.

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. *Nat Commun*, 11:1760, 2020.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

A. L. Sharp, A. Baecker, N. Nassery, S. Park, A. Hassoon, M. S. Lee, S. Peterson, S. Pitts, Z. Wang, Y. Zhu, and D. E. Newman-Toker. Missed acute myocardial infarction in the emergency department-standardizing measurement of misdiagnosis-related harms using the SPADE method. *Diagnosis (Berl)*, 8:177–186, 2021.

K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol*, 18:465–478, 2021.

W. E. Soares, L. L. Price, B. Prast, E. Tarbox, T. J. Mader, and R. Blanchard. Accuracy Screening for ST Elevation Myocardial Infarction in a Task-switching Simulation. *West J Emerg Med*, 20:177–184, 2019.

K. Szummer, L. Wallentin, L. Lindhagen, J. Alfredsson, D. Erlinge, C. Held, S. James, T. Kellerth, B. Lindahl, A. Ravn-Fischer, E. Rydberg, T. Yndigegn, and T. Jernberg. Improved outcomes in patients with ST-elevation myocardial infarction during the last 20 years are related to implementation of evidence-based treatments: experiences from the SWEDEHEART registry 1995-2014. *Eur Heart J*, 38:3056–3065, 2017.

K. Szummer, L. Wallentin, L. Lindhagen, J. Alfredsson, D. Erlinge, C. Held, S. James, T. Kellerth, B. Lindahl, A. Ravn-Fischer, E. Rydberg, T. Yndigegn, and T. Jernberg. Relations between implementation of new treatments and improved outcomes in patients with non-ST-elevation myocardial infarction during the last 20 years: experiences from SWEDEHEART registry 1995 to 2014. *Eur Heart J*, 39:3766–3776, 2018.

A. Tanguay, J. Lebon, E. Brassard, D. Hébert, and F. Bégin. Diagnostic accuracy of prehospital electrocardiograms interpreted remotely by emergency physicians in myocardial infarction patients. *Am J Emerg Med*, 37:1242–1247, 2019.

G. H. Tison, J. Zhang, F. N. Delling, and R. C. Deo. Automated and Interpretable Patient ECG Profiles for Disease Detection, Tracking, and Discovery. *Circ Cardiovasc Qual Outcomes*, 12:e005289, 2019.

T. van der Ploeg, P. C. Austin, and E. W. Steyerberg. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol*, 14:137, 2014.

B. Wright, N. Faulkner, P. Bragge, and M. Graber. What interventions could reduce diagnostic error in emergency departments? A review of evidence, practice and consumer perspectives. *Diagnosis (Berl)*, 6:325–334, 2019.

Y. Zhao, J. Xiong, Y. Hou, M. Zhu, Y. Lu, Y. Xu, J. Teliewubai, W. Liu, X. Xu, X. Li, Z. Liu, W. Peng, X. Zhao, Y. Zhang, and Y. Xu. Early detection of ST-segment elevated myocardial infarction by artificial intelligence with 12-lead electrocardiogram. *Int J Cardiol*, 317:223–230, 2020.

Appendix A. Supplementary Methods

A.1. Data sources

Adult patients ($\geq$ 18 years old) with available emergency department data from 6 emergency departments in the Stockholm region, Sweden, between 2003 and 2017 were collected. The sample was linked to national registries (the in-patient, prescribed drug, and death registries), national quality registries (SWEDEHEART [Swedish Web-system for Enhancement and Development of Evidence-based care in Heart disease Evaluated According to Recommended Therapies; a Swedish nation-wide quality register] sub-registries RIKS-HIA [Register of Information and Knowledge About Swedish Heart Intensive Care Admissions] and SCAAR [Swedish Coronary Angiography and Angioplasty Registry]), as well as a regional database of ECGs (Karolinska ECG database) and electronic health records. All data sources covered the time period 2007-2016 or longer. Characteristics have been described for STEMI (Szummer et al., 2017) and NSTEMI (Szummer et al., 2018) patients in SWEDEHEART during the present study period, and the study sample has been partially described previously (Af Ugglas et al., 2020).

The procedure and criteria used to define the study sample are described in Figure S1. In total, 23,244 patients had at least one registered coronary care unit admission at any time and at least one valid ECG recording at any time. We apply the filters in Figure S1 to ensure inclusion of at-even before-treatment ECGS (ECG collected on or one day before the day of the coronary care unit admission without any record of recent intervention) as well as confirming the outcome label (NSTEMI/STEMI/control status available, and ICD10:I21 without a left bundle branch block for the myocardial infarctions) which both reduces the sample size to the studied 10,583 patients.

The outcome label used, NSTEMI/STEMI/control, is the standard INFARCTTYPE variable in SWEDEHEART RIKS-HIA. It captures the view of the whole cycle of care by the attending cardiologist at time of discharge from the coronary care unit, who has access to all relevant patient data, including but not limited to singles ECGs, continuous ECG monitoring, cardiac enzyme series and other lab data, and angiographic and echocardiographic results. Regular monitoring of the SWEDEHEART registry shows a data accuracy of around 96% (Jernberg et al., 2010).

The study was approved by the Swedish Ethical Review Agency, application number 2020-01654.

A.2. Exposures and outcomes

High-quality data on the exposures and outcomes were available for all included patients from discharge records from the emergency departments, from linked hospitalizations, and from the SWEDEHEART registry. Definitions used are listed in Table S3. As exposures, we used digital ECG data, age and sex, as in a previous study (Cho et al., 2020). Standard 10-second 12-lead ECG recordings sampled at 250 to 500Hz were used; 8 leads were used in the present study as 4 of the standard leads are functions of these 8 and are hence redundant. The outcome label was NSTEMI/STEMI/control status, as registered in SWEDEHEART by the attending cardiologist at time of discharge from the coronary care unit. Details are described in the Supplementary Methods. We only included cases with complete data on these few exposures and outcomes.

A.3. Model architecture

Our model is an extension of a previous study (Ribeiro et al., 2020) where the authors performed an extensive hyperparameter search. We performed on top a study about the scalability of the model. We scaled model depth by factors of $\{1, 2, 3, 4\}$ and width by $\{1, 1.5, 2\}$ (only for depth up to 2 because of model size issues) and found that a depth scaling with a factor of 2 from the original study helps performance, which yields the ResNet model structure of Figure S3. We extend the model with SE blocks. We experiment with SE reduction factors of $\{2, 4, 8, 16, 32\}$. For the embedding of the phenotypes age and sex we test with adding {[batch normalization, ReLU], ReLU, [batch normalization, ReLU, Dropout (0.2)]} after the linear layer and find that ReLU activation alone is the best option. The size of the linear layer for the phenotypes embedding is heuristically chosen.

A.4. Model training procedure

The model was trained by minimizing the cross entropy loss using the Adam optimizer with default parameters and learning rate of 10^{-3} for 200 epochs using a batch size of 256. We used a cosine learning rate scheduler (Loshchilov and Hutter, 2017) which reduces the learning rate according to a cosine function from the initial learning rate to zero over the epochs. Initially, we warmed up the learning rate linearly over 15 epochs which helps to improve generalization. In addition to the dropout with dropout probability of 0.5 within the ResNet blocks we used dropout on the linear classifier with value 0.2. Furthermore, we regularized with weight decay of 0.005 and label smoothing (Müller et al., 2019) with value of 0.15. In addition to the training data set we made use of ECG exams of patients who had multiple exams in the same coronary care unit admission for training. We consider this as a form of data augmentation since these exams have the same label but are recorded at different times and therefore with a different state of the patient and possibly placement of the ECG leads. We denoted this additional dataset an augmentation dataset, and included it for training but not for validation. This is common practice in deep learning training in order to get correct validation metrics which resemble the test dataset.

We evaluated the hyperparameters with a 5-fold cross-validation approach where the metrics were averaged over all folds. Our hyperparameter tuning objective was to reduce overfitting while increasing accuracy, AUROC, average precision and model calibration in terms of ECE.

We consider the basic model (Ribeiro et al., 2020) with the depth scaling factors and without SE blocks and iteratively extend the model training procedure. First, we test different training and validation batch sizes $\{32, 64, 128, 256\}$ and find no significant performance differences and therefore choose 256 which yields the fastest training speed. For the optimizer and learning rate we experiment with {SGD, ADAM} (Kingma and Ba, 2015) both using learning rates in $\{0.5, 0.1, 0.01, 0.001\}$ and momentum $\{0.7, 0.8, 0.9\}$ for SGD. We denote the model trained with these settings as the baseline model.

We extend the training procedure iteratively with eight training procedure and architecture options following previous recommendations of modern ResNet tuning to improve model performance (Bello et al., 2021; He et al., 2019). For each of the options we choose the best performing parameters before checking the next extension. We highlight the best performing parameter for an option italic. 1. Training dataset: only training data or with *additional augmentation dataset*. 2. Learning rate scheduler: multistep learning rate scheduler with decrease by factor 10 at epoch 75, 125, 175 or *cosine learning rate scheduler* with linear warmup for $\{0, 5, 10, 15, 20\}$ epochs (Liu et al., 2020). 3. Label smoothing: smoothing value $\{0.0, 0.05, 0.1, 0.15, 0.2\}$. 4. Additional dropout on the final linear layer with value $\{0.2, 0.3, 0.4, 0.5\}$. 5. Re-evaluation of weight decay value with value $\{0.05, 0.01, 0.005, 0.001, 0.0005\}$. 6. Additional SE net with reduction factor $\{32, 16, 8, 4, 2\}$. 7. Additional age and sex embedding with architecture choice as described above. 8. Additional ensemble-based model with heuristically chosen five ensemble members.

The results for the eight options are shown in Figure S8 for accuracy, AUROC, average precision and ECE as examples. The plots show mean and standard deviation over the five cross-validation folds. This indicates that each option increases at least one of the metrics on average and does not reduce the others significantly. Not that ECE should be minimized.

A.5. Model calibration

We use ECE and Brier score as metrics to evaluate model calibration. ECE is the weighted absolute difference between the class membership and the estimated probability for that class averaged over 15 bins, while the Brier score measures the average squared error on the probability scale. Both metrics are applicable to both binary and multiclass problems.

We tried to improve upon the original calibration by temperature scaling (Guo et al., 2017), vector calibration (Guo et al., 2017), and Dirichlet calibration (Kull et al., 2019). None of those methods succeeded in improving the model calibration. Further work is necessary to investigate the possibility to improve model calibration beyond our current setting.

A.6. Model evaluation

We tested the over/underrepresentation of inpatient care diagnoses at the time of the cardiac intensive care among correctly classified versus misclassified patients with a predicted probability > 0.5 in pairwise independent tests.

Appendix B. Supplementary Results

While each of the model and training modifications to the original model architecture (Ribeiro et al., 2020) were effective in some metrics and overall contributed to the performance of our model, Figure S8 shows that two changes were more important: the extension of our original dataset with the augmentation training dataset, and the use of an ensemble-based model.

Additionally to the results in Figure 1 (middle and right) where the AUROC and precision-recall curve for the temporal test split are shown, Figure S9 shows the curves for the random test split.

Characteristics of hospitalizations with misclassified ECGs are described in Table S4. Among the misclassified ECGs, those misclassified as STEMI more often had myocarditis, other severe infections, emaciation, neurological traits, and adverse drug reactions among the diagnoses at the coronary care unit hospitalization; those misclassified as NSTEMI more often had late-stage diabetes, aortic stenosis, and pulmonary edema.

Appendix C. Supplementary Discussion - Limitations

More limitations than the ones mentioned in section 4 are the following. The sample size was limited, albeit on par with other similar studies (Cho et al., 2020). Sample sizes needed for machine learning methods are often tenfold larger than those needed for traditional statistical modeling (van der Ploeg et al., 2014).

While the calibration of our model was better than that of comparable models, there is still room for improvement; this is an underappreciated property in general.

We did not consider transferring learned features, only model architecture, from a previous study (Ribeiro et al., 2020). An exploration of potential improvements in model convergence speed and final performance boost by pretraining on a different ECG classification task with a dataset in a different context may be useful, but may also introduce model biases from the other dataset.

Appendix D. Supplementary Figures

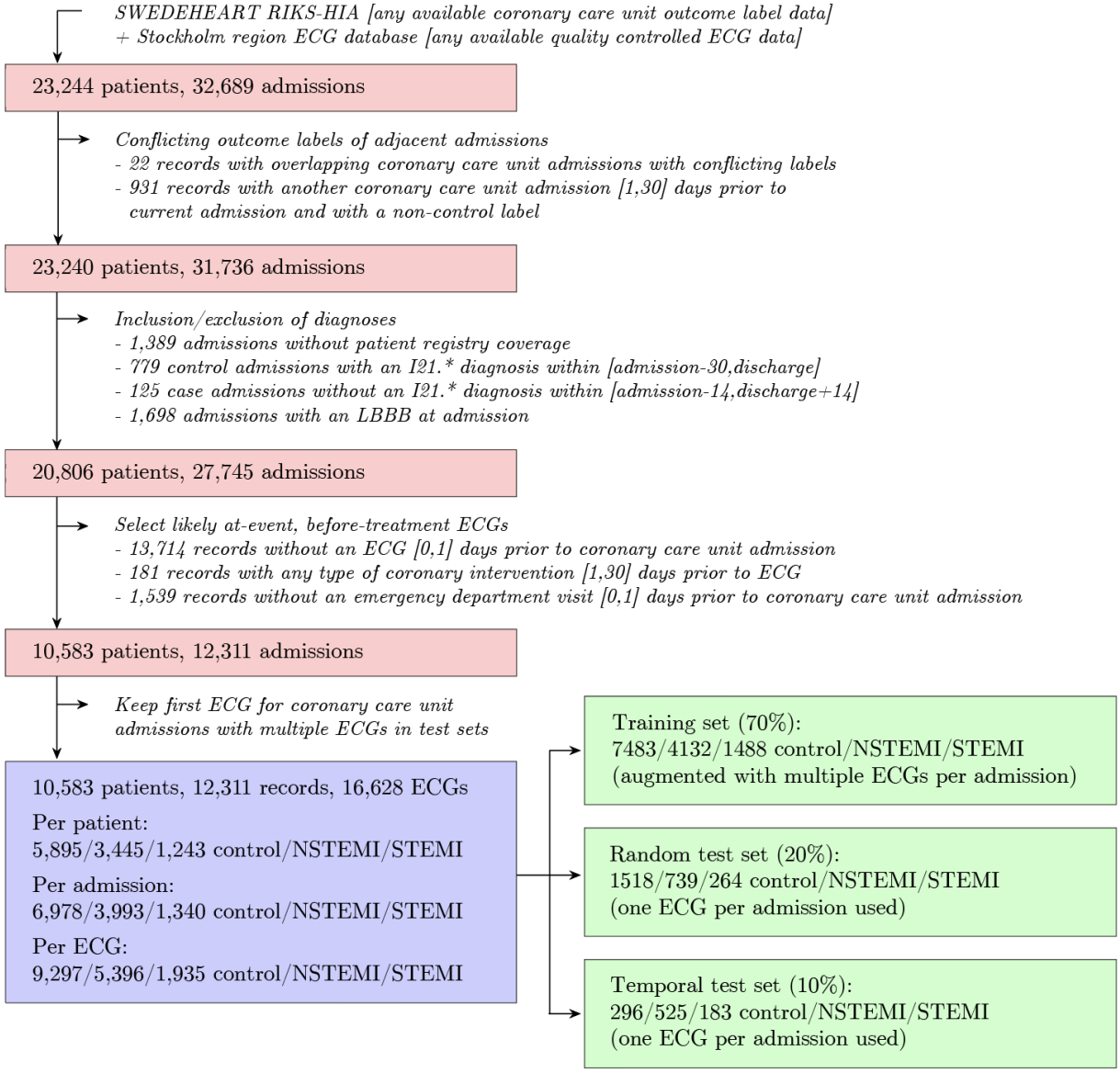

Figure S1: **Derivation of the study sample.** Data linkage of Stockholm region emergency department visits to national and regional registries and electronic health records, together with inclusion/exclusion criteria applied to define the study sample. SWEDEHEART, Swedish Web-system for Enhancement and Development of Evidence-based care in Heart disease Evaluated According to Recommended Therapies; RIKS-HIA, Register of Information and Knowledge About Swedish Heart Intensive Care Admissions.

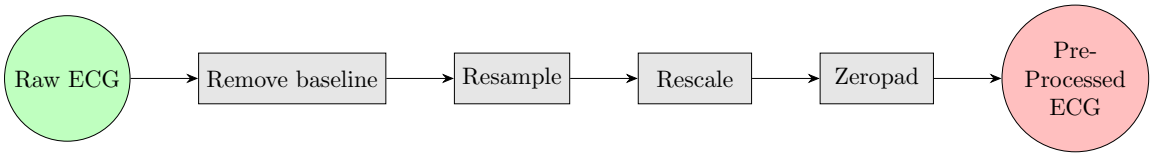

Figure S2: **Data pre-processing.** We re-sampled all ECGs to 400 Hz and zero-padded to a fixed length of 4096 samples, since the convolution-based model requires a fixed input size. For duplicated ECGs with identical data and collection time the first copy was kept and ECGs where one or more required leads were missing or contained all-zero entries were removed. We used one-hot encoding for sex and normalized the age with mean and standard deviation of the training dataset.

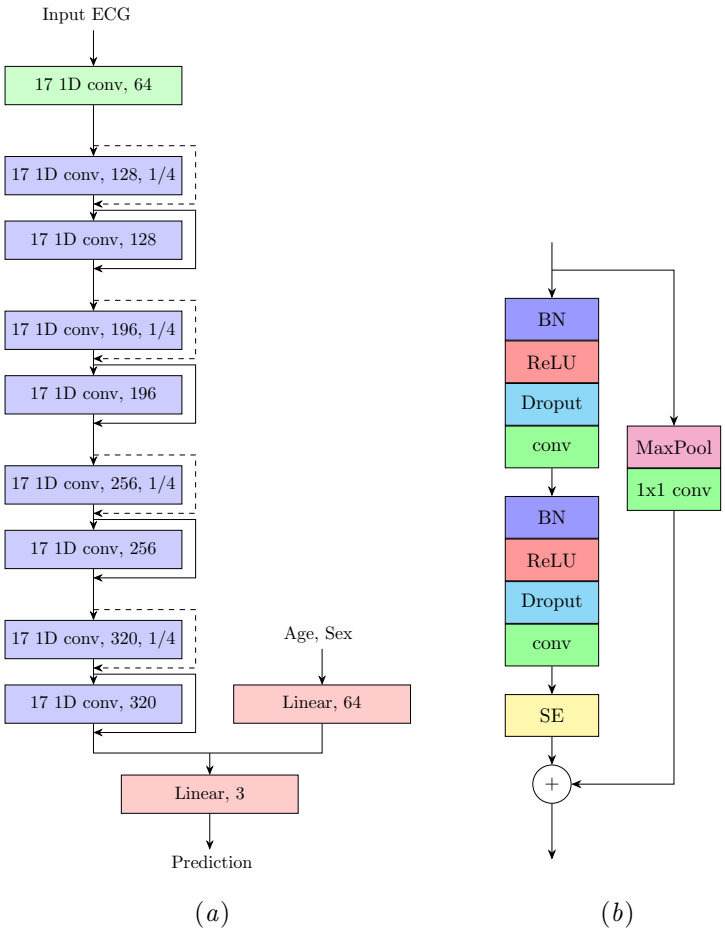

Figure S3: **Deep neural network model architecture.** The left panel is a high-level model for ECG classification consisting of one line to extract features from the ECG exam and one for features from phenotypes age and sex. The light green block contains a convolutional layer followed by a batch normalization for rescaling the output and a ReLU activation function. This layer is followed by four sets of residual blocks in light blue. The name of the block indicates the filter size of the convolutions, the number of filters and the downsampling factor (if applicable). Note that we downsample the signal by a factor of $\frac{1}{4}$ in the beginning of each set of residual blocks. The right panel illustrates the content of each residual block. Dropout is used after each nonlinear activation function as regularization. Only the first residual block does not contain the first batch normalization, ReLU and dropout layer since these layers are already applied after the initial convolution layer. We extended the initial architecture (Ribeiro et al., 2020) with Squeeze and Excite (SE) blocks with reduction factor of 16. This operation helps to weight the channel wise information. Downsampling residual blocks consist in the residual skip connection (dashed lines) of a MaxPooling operation followed by a convolutional layer with filter length 1 to match the dimensions with the main branch for the summation. The remaining skip connections (full lines) do not contain any operations since input and output dimensions of the residual block are equal.

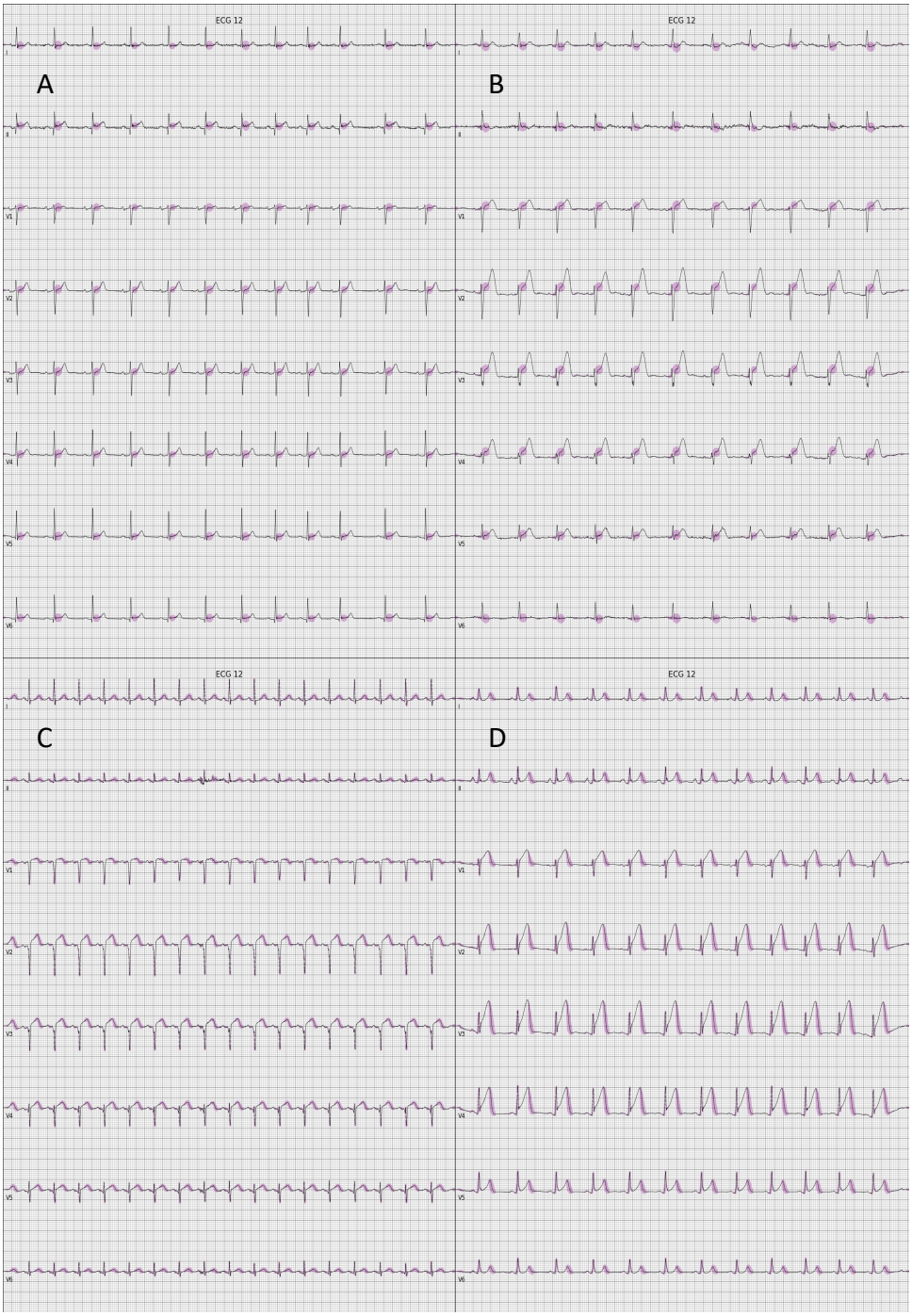

Figure S4: **Four representative STEMIs correctly classified with high probability.** Four Grad-CAM plots of STEMIs correctly classified with high probability, highlighting the parts in the ECG that the model focuses on for its prediction. Gradients corresponding to the activations in the first convolutional layer of the neural network are averaged to get the proportional importance of each channel, which is then used to compute a proportional mean of the activations. Positive values obtained were plotted as blue disks overlaid on top of the ECG, with size proportional to its magnitude.

Figure S5: **Four representative NSTEMIs correctly classified with high probability.** Four Grad-CAM plots of NSTEMIs correctly classified with high probability, highlighting the parts in the ECG that the model focuses on for its prediction. Gradients corresponding to the activations in the first convolutional layer of the neural network are averaged to get the proportional importance of each channel, which is then used to compute a proportional mean of the activations. Positive values obtained were plotted as blue disks overlaid on top of the ECG, with size proportional to its magnitude.

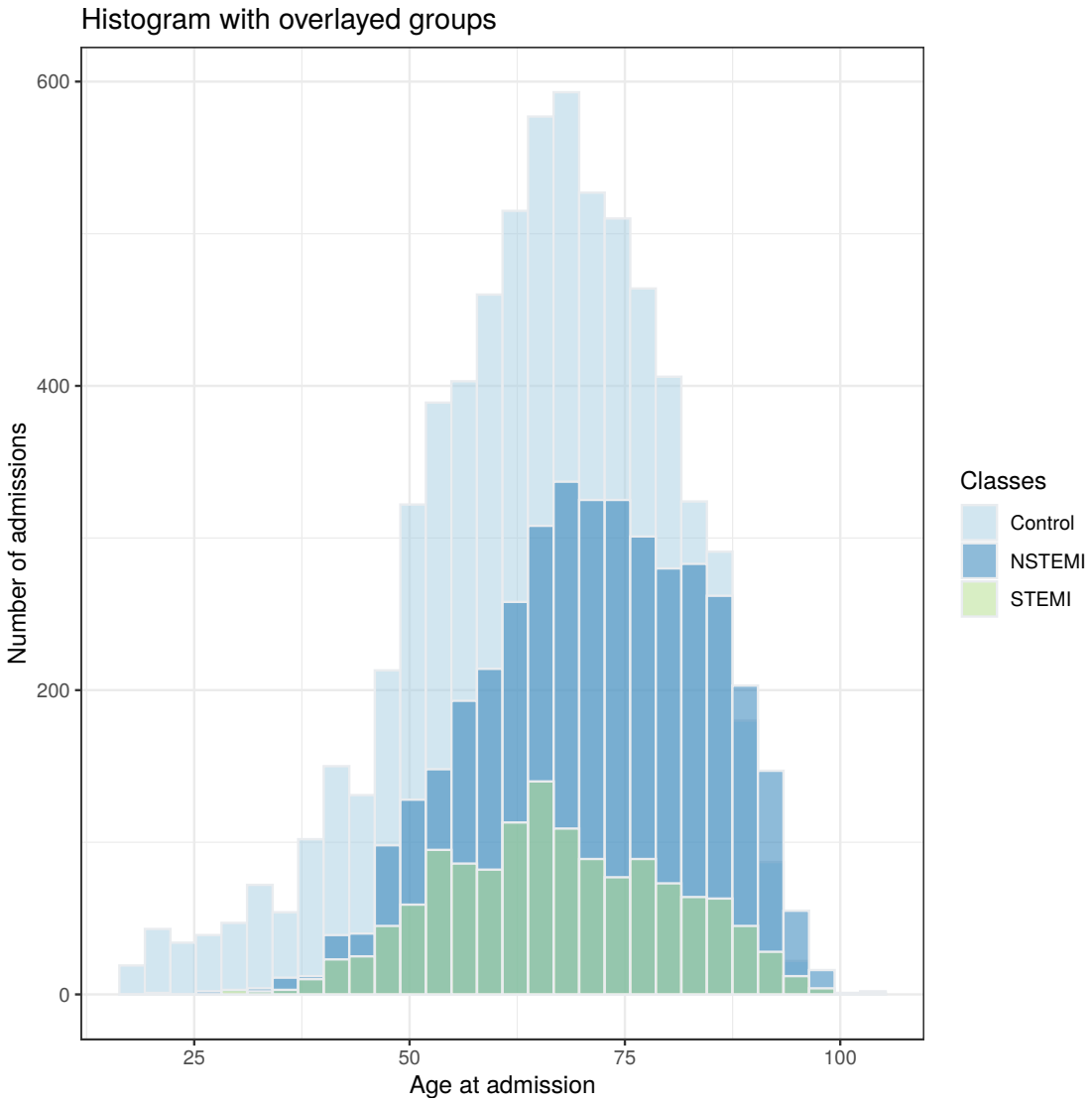

Figure S6: **Distribution of admission age for cardiac intensive care admissions in overlayed histograms, by control/NSTEMI/STEMI.**

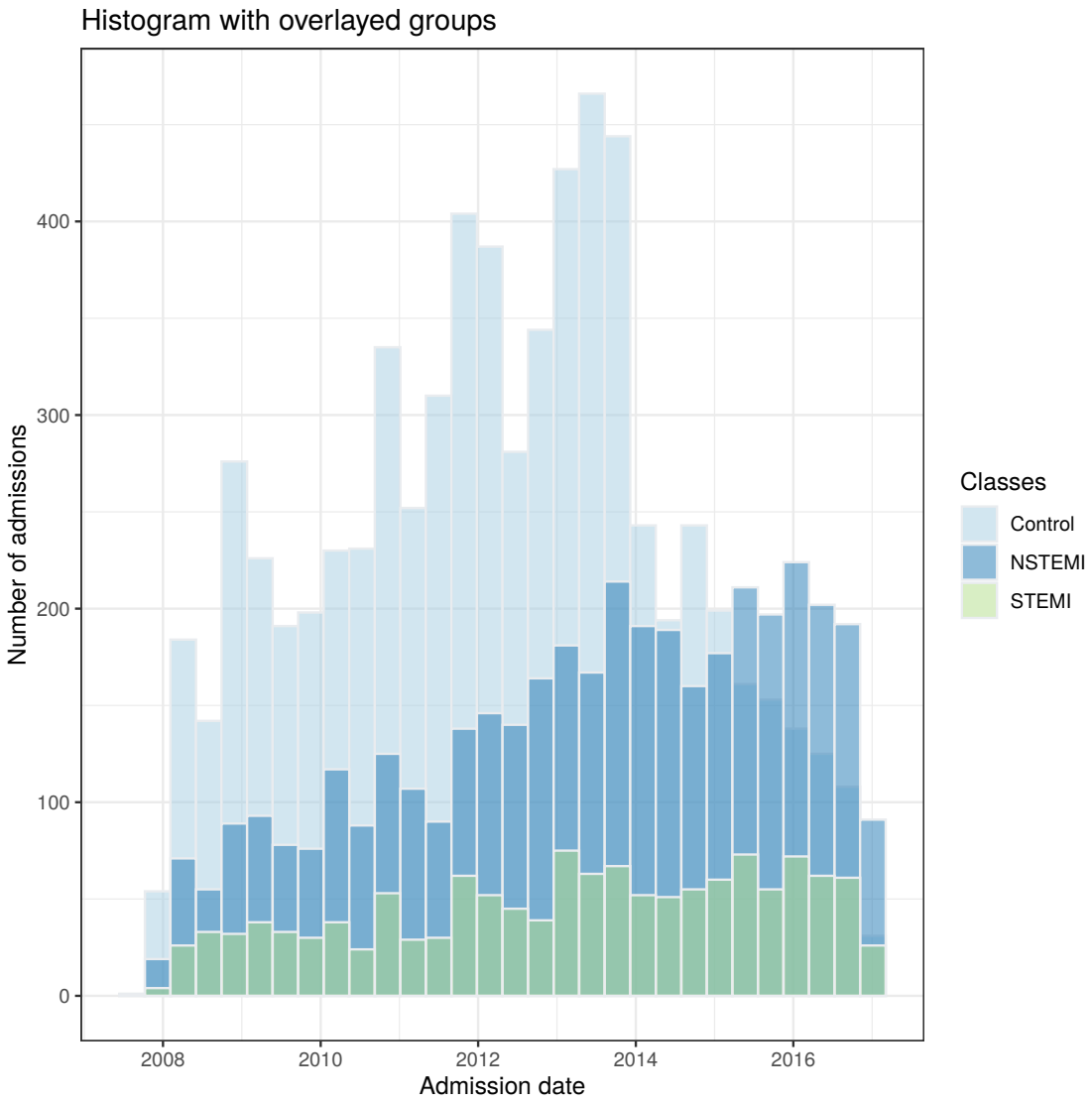

Figure S7: **Distribution of admission date in overlayed histograms, by control/NSTEMI/STEMI.** A decrease in number or controls admitted in 2014 due to administrative changes in the Stockholm healthcare system, with a limitation of walk-ins at the Karolinska emergency department from 2014.

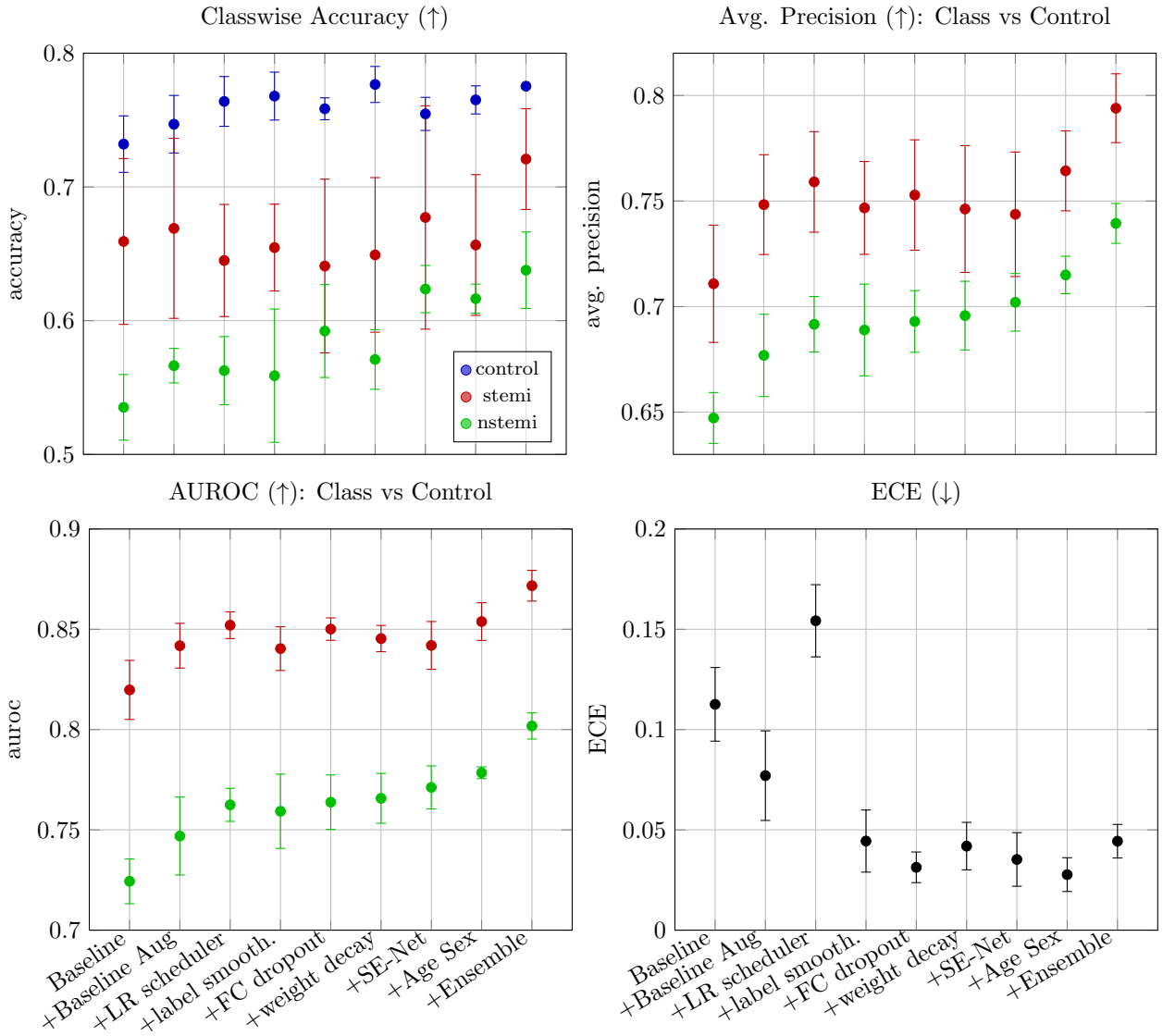

Figure S8: **Model improvements by tuning of hyperparameters.** Evaluated metrics for hyperparameter tuning. Classwise accuracy, average precision, Cstatistics and expected calibration error (ECE) were evaluated. Shown is mean and standard deviation over the validation sets of the 5-fold cross-validation. The final model contains all model and training extensions. Arrows in the title indicate the optimization objective. Aug, augmentation; LR, learning rate; label smooth, label smoothing; FC, fully connected layer; SE-Net, squeeze and excite network.

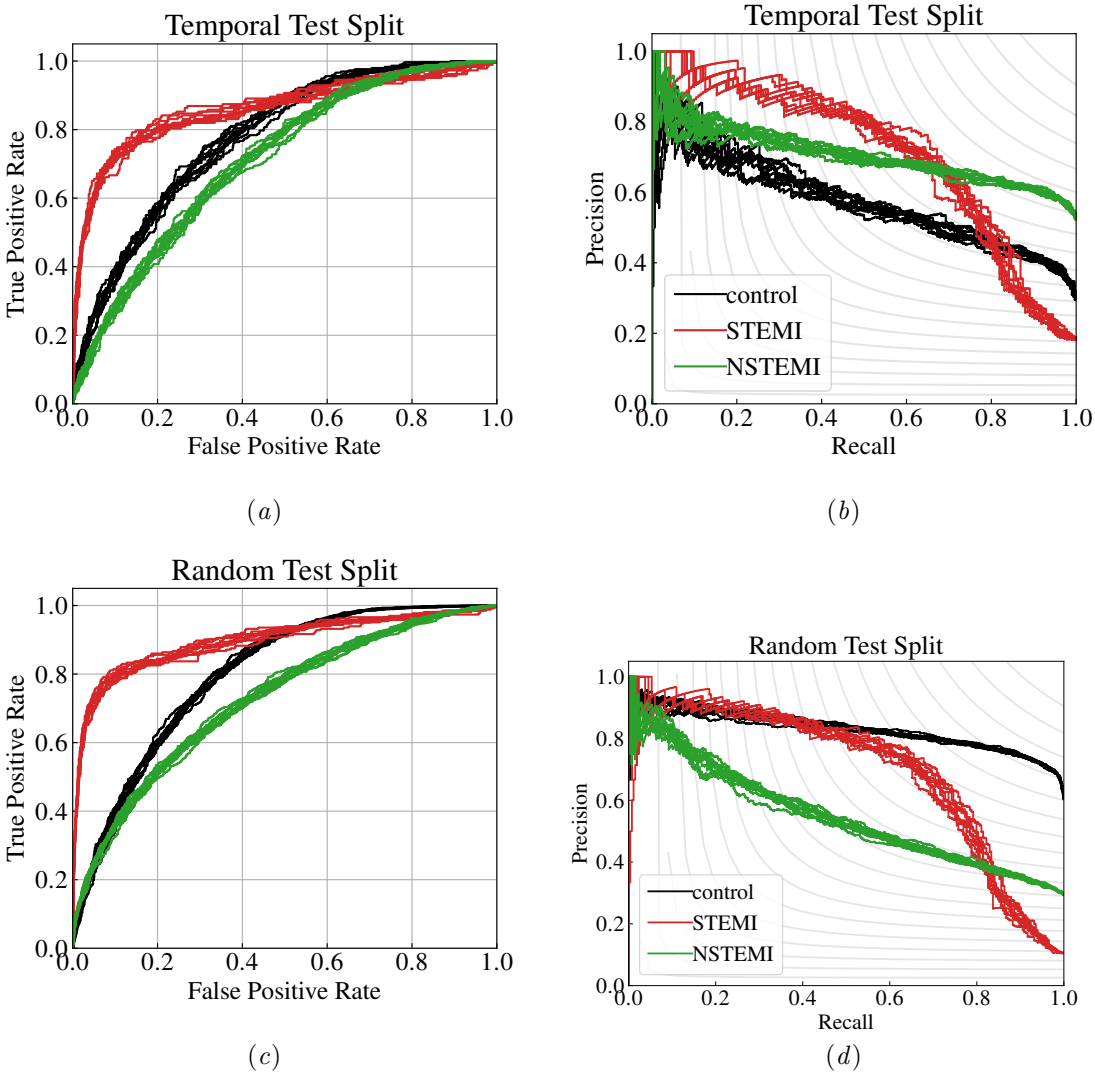

Figure S9: **ROC and Precision Recall curves.** Top row: Temporal test split, repetition from figure 1 (middle and right). Bottom row: Random test split. Each curve for the 10 seeds of our ensemble based model are shown. Each curve is a class vs all curve, not a class vs control as in Table 1.

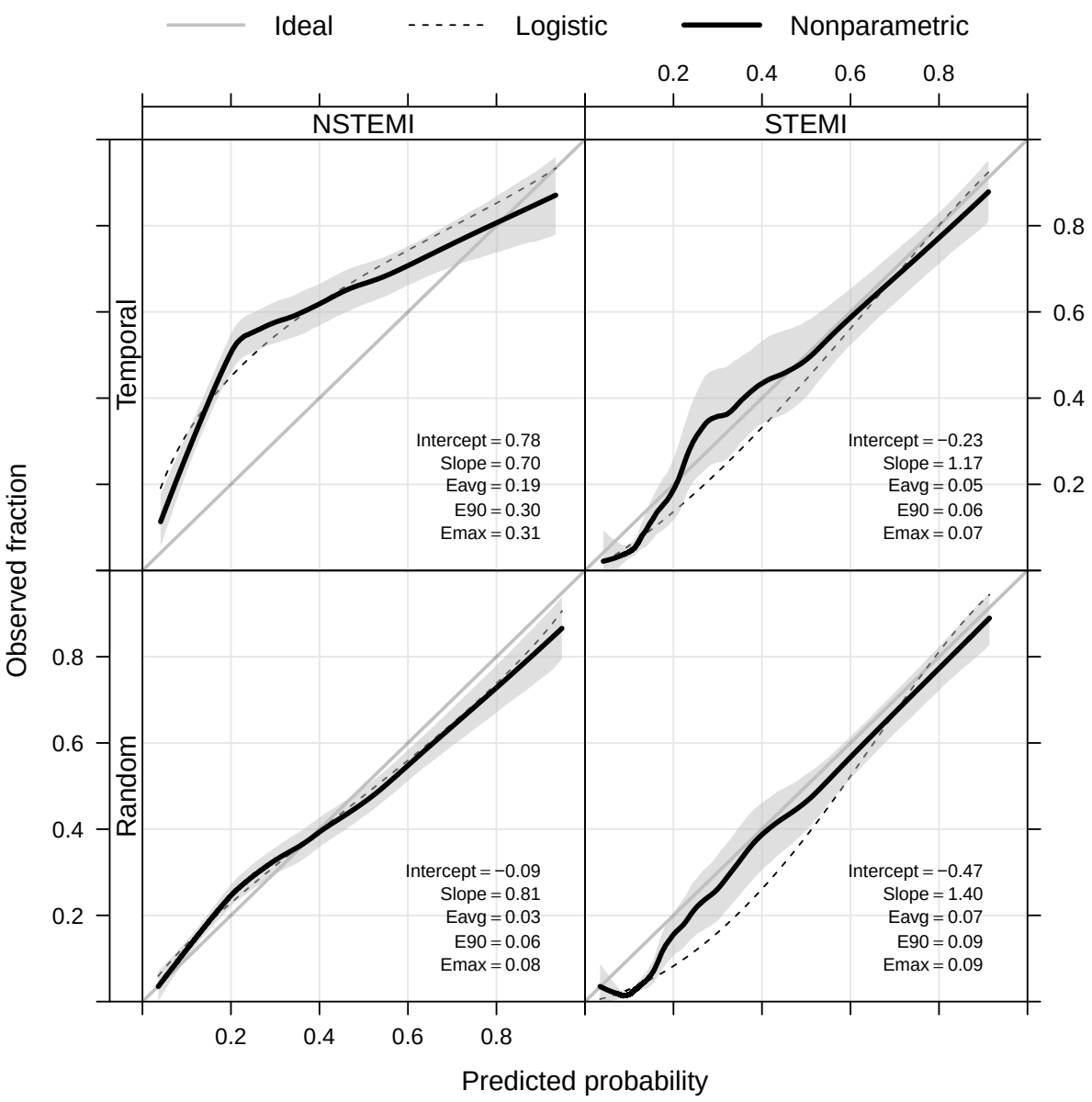

Figure S10: **Calibration of final model.** Calibration plot (also called reliability diagram) for NSTEMI vs all and STEMI vs all for both of our test sets. The grey solid lines indicate perfect calibration. Black solid lines are LOESS fits and shaded areas are 95% bootstrap confidence intervals. Black broken lines are fits resulting from logistic regressions using the estimated logits as the sole predictors. Intercept denotes the intercept from the logistic models and should be as close to zero as possible. Slope denotes the slope of the logistic regression fit and should be as close to 1 as possible. A slope > 1 indicates underfitting and a slope < 1 indicates overfitting with the degree of under-/overfitting directly proportional to the absolute size of the slope. Eavg, E90 and Emax correspond to the average absolute error, the 90th percentile of the absolute error and the maximum absolute error between the predicted probabilities and the LOESS fit.

Appendix E. Supplementary Tables

Table S1: **Clinical characteristics of the study sample** Patient characteristics of coronary care unit admissions included in the study, by control/NSTEMI/STEMI outcome. Data are medians (quartiles) or percent.

	Control	**NSTEMI**	**STEMI**
Number of patients	6,978	3,993	1,340
Clinical characteristics at ED visit			
Age	66.0 (56.0,76.0)	71.0 (62.0,81.0)	66.0 (57.0,77.0)
Male	61.6	65.9	73.7
Year	2012 (2010,2013)	2013 (2011,2015)	2013 (2011,2015)
Walked into ED	54.9	45.6	49.0
Presenting complaint:			
Chest pain	69.7	72.2	72.0
Difficulty breathing	10.4	12.0	5.3
Abnormal rhythm	3.9	1.4	0.7
Malaise	1.2	1.5	2.2
Circulatory arrest	0.8	1.0	4.5
Cardiovascular diagnoses prior to ED visit*			
Myocardial infarction	23.6	25.9	15.9
Unstable angina	15.0	11.4	5.1
Ischemic heart disease	45.6	42.7	23.1
Stroke	8.9	10.4	6.4
Peripheral artery disease	9.8	12.1	6.6
Heart failure	22.1	20.1	8.1
Atrial fibrillation	20.4	15.4	7.9
Cardiovascular disease	72.2	70.2	50.9
Drugs with >= 1 dispensation within one year prior to ED visit			
Renin-angiotensin system inhibitors	49.3	51.2	34.8
Calcium channel blockers	26.4	28.3	21.0
Beta-receptor blockers	54.8	50.0	32.3
Mineralocorticoid receptor antagonists	8.7	6.1	2.6
Diuretics	33.3	34.3	18.1
Anti-arrhythmic drugs	1.9	0.5	0.4
Statins	46.6	41.4	23.8
Anticoagulants	13.7	8.9	4.6
Antiplatelets	48.6	48.6	27.2
Cardiac enzymes within ED visit or coronary care unit hospitalization**			
Troponin I measured	10.3	8.4	12.3
Max troponin I (ng/L)	29.7 (29.7,60.0)	2300.0 (540.0,8000.0)	18100.0 (3300.0,50500.0)
Troponin T measured	80.1	84.6	85.5

Continued on next page

Table S1 – *Continued from previous page*

	Control	NSTEMI	STEMI
Max troponin T (ng/L)	13.0 (9.0,39.0)	249.0 (84.8,778.2)	1900.0 (550.5,4817.5)
NTproBNP measured	20.9	18.8	15.1
Max NTproBNP (ng/L)	1190.0 (257.5,3972.5)	2635.0 (678.5,8067.5)	2740.0 (649.0,8160.0)
Main cause of coronary care unit hospitalization***			
Myocardial infarction	0.0	96.1	97.6
Unstable angina	8.5	4.1	1.3
Ischemic heart disease	21.1	96.5	97.6
Stroke	0.3	0.8	0.7
Peripheral artery disease	0.3	0.1	0.1
Heart failure	7.0	1.6	0.8
Atrial fibrillation	6.6	0.5	0.3
Cardiovascular disease	47.6	98.0	99.1
Mortality after coronary care unit admission			
30-day all-cause death	1.4	6.1	11.0
In-hospital all-cause death	1.1	5.1	9.7

*Prevalent disease based on any diagnosis position, inpatient and outpatient specialist care combined.
**Combining all troponin T/I laboratory measurements from the Stockholm region as well as high-sensitive troponin T/I from SWEDEHEART. Maximum of all available measurements within the time window reported.
***Primary diagnosis from inpatient specialist care. ED, emergency department; NTproBNP, N-terminal pro-B-type natriuretic peptide.

Table S2: **Clinical characteristics of the study sample, stratified by control/NSTEMI/STEMI and training/test set.** Data presented as median (interquartile range) or percentages. Abbreviations: cont. is control, NST. is NSTEMI, ST. is STEMI.

	Test Random			Test Temporal			Training		
	con.	**NST.**	**ST.**	**con.**	**NST.**	**ST.**	**cont.**	**NST.**	**ST.**
Baseline characteristics (at ED/HIA admission)									
Age	66.0 (56.2, 75.0)	72.0 (62.0, 82.0)	65.5 (57.8, 76.2)	66.0 (55.0, 75.0)	73.0 (63.0, 81.0)	67.0 (56.0, 76.0)	66.0 (56.0, 76.0)	71.0 (62.0, 81.0)	67.0 (57.0, 78.0)
Year	2012 (2010, 2013)	2013 (2010, 2014)	2012 (2010, 2013)	2016 (2016, 2016)	2016 (2016, 2016)	2016 (2016, 2016)	2012 (2010, 2013)	2013 (2011, 2014)	2013 (2010, 2014)
Walked into ED	57,5	49,3	53,7	68,0	53,6	53,4	54,3	44,5	47,5
Presenting complaint									
– Chest pain	69,1	72,7	72,0	72,8	69,6	78,8	69,5	73,0	71,2
– Difficulty	11,9	9,8	3,0	11,2	14,4	4,2	10,2	11,7	5,8

Continued on next page

Table S2 – *Continued from previous page*

	Test Random			Test Temporal			Training		
	con.	**NST.**	**ST.**	**con.**	**NST.**	**ST.**	**cont.**	**NST.**	**ST.**
breathing									
– Abnormal rhythm	3,3	1,5	1,2	3,0	2,6	0,8	4,1	1,2	0,8
– Malaise	1,3	1,5	1,2	1,2	1,3	0,8	1,2	1,4	2,5
– Circulatory arrest	0,9	0,6	2,4	2,4	1,6	6,8	0,5	0,7	4,5
Cardiovascular diagnoses prior to ED visit*									
AMI	25,0	28,2	13,4	11,8	23,2	11,0	23,9	26,7	18,0
Unstable angina	15,8	12,9	1,2	8,9	11,1	3,4	15,2	11,1	6,0
CHD	48,2	45,3	18,3	36,7	38,6	18,6	46,0	43,5	25,9
All stroke	10,7	9,8	5,5	9,5	10,1	4,2	8,6	10,4	7,1
PAD	9,5	11,7	1,8	5,9	10,1	7,6	10,2	12,2	6,8
HF	23,2	17,3	4,9	11,8	19,6	4,2	22,5	20,4	9,6
Afib	18,6	15,0	4,3	18,3	17,0	5,1	20,9	15,4	8,8
Chapter I	75,3	69,7	43,9	69,8	69,6	50,8	72,4	70,1	52,6
Drugs with 1+ dispensation 1 year prior ECG									
RAAS	51,1	47,2	28,7	52,7	59,5	33,9	49,4	51,2	35,8
CCB	27,3	28,2	17,1	27,8	30,7	20,3	26,2	28,0	21,2
Beta block	57,3	51,6	28,0	47,3	46,7	23,7	55,2	50,9	34,5
MRA	7,4	5,8	1,8	4,7	4,9	2,5	9,4	6,5	2,9
All diuretics	33,2	35,9	15,9	19,5	32,7	11,0	34,3	34,8	20,3
Antiarrhytmia	2,5	1,3	0,0	1,8	0,0	0,0	1,8	0,4	0,3
Statins	48,5	40,7	22,0	48,5	36,3	26,3	46,5	42,2	24,3
Anti-coagulation	14,0	8,1	2,4	16,0	11,8	4,2	13,6	8,9	4,8
Antiplatelet	52,8	50,5	23,8	45,0	41,2	18,6	48,4	49,9	29,3
Cardiac enzymes within ECG date and HIA**									
Has troponin I	12,0	10,9	14,6	0,0	0,3	0,0	10,5	9,5	14,2
Max troponin I (ng/L)	12.0 (9.0, 32.0)	210.0 (67.0, 630.0)	1410.0 (432.2, 5452.5)	21.0 (12.0, 63.5)	195.0 (80.5, 614.5)	2705.0 (889.8, 5972.5)	12.0 (9.0, 36.0)	250.0 (82.0, 779.0)	1945.0 (519.0, 4687.5)
Has NT-proBNP	19,9	13,6	11,6	28,4	16,7	20,3	20,4	18,7	14,1
max NT-proBNP (ng/L)	1190 (274.0, 4350.0)	1540 (592.0, 4440.0)	5890 (940.0,1 2050.0)	930 (219.0,2 640.0)	2190 (455.0,5 800.0)	1310 (585.0, 5175.0)	1250 (271.8, 3942.5)	2925 (829.2, 8082.5)	2975 (779.8, 8995.0)
Main cause of hospitalisation with ECG date and HIA discharge date***									
AMI	0,0	97,1	97,0	0,0	97,4	96,6	0,0	95,9	97,9
Unstable angina	9,0	3,1	1,2	25,4	7,5	0,0	7,9	3,8	1,6
CHD	22,2	97,3	97,0	37,9	97,4	96,6	20,6	96,4	97,9
All stroke	0,5	1,0	0,0	0,0	1,0	0,8	0,3	0,7	1,0

Continued on next page

Table S2 – *Continued from previous page*

	Test Random			Test Temporal			Training		
	con.	**NST.**	**ST.**	**con.**	**NST.**	**ST.**	**cont.**	**NST.**	**ST.**
PAD	0,1	0,0	0,0	0,6	0,0	0,0	0,3	0,1	0,1
HF	7,9	1,3	1,2	5,3	1,0	0,0	7,0	1,6	0,8
Afib	5,8	0,6	0,6	1,8	0,3	0,0	6,9	0,5	0,3
Chapter I	47,0	98,5	98,2	65,7	98,7	99,2	46,5	97,8	99,1
30-day mortality after HIA admission									
All-cause	1,4	5,6	10,4	1,2	6,2	11,0	1,3	5,9	11,2
All-cause, inhospital	1,0	4,8	9,1	1,2	4,9	10,2	1,0	4,9	9,9
All-cause, inhospital, AMI	0,2	2,9	8,5	0,6	2,3	9,3	0,0	2,6	8,7
All-cause, inhospital, Chapter I	0,7	3,8	8,5	1,2	3,6	10,2	0,5	3,6	9,3

*Prevalent disease based on any diagnosis position, inpatient and outpatient specialist care combined.
**Combining all troponin T/I laboratory measurements from Karolinska as well as high-sensitive troponin T/I from SWEDEHEART. Maximum of all available measurements within the time window reported.
***Primary diagnosis from inpatient specialist care.

Table S3: **Definitions used.**

Diagnosis/intervention/treatment	**Codes**
Diagnosis (ICD10)	
Acute myocardial infarction	I21
Unstable angina	I20.0
Coronary heart disease	I20-I25
All stroke	I60-I64,G45
Peripheral artery disease	I70-I74,I77.(3|6|8),I79
Heart failure	I50
Atrial fibrillation	I48
Chapter I	I
Surgical codes (KVÅ)	
PCI/CABG	FNG(00|02|05|10|96),FNC,FND,FNE
Treatment (ATC)	
RAASi	C09
Calcium channel blocker	C08
Beta blocker	C07
MRA	C03DA
All diuretics	C03
Anti-arrhytmia	C01B
Statins	C10AA
Anti-coagulation	B01A(A|E|F)
Anti-platelet	B01AC

Table S4: **Over-/underrepresented diagnoses among misclassified cases.** Over-/underrepresented diagnoses when comparing correct classifications with misclassifications for a given class label in a general independence test. All diagnoses within the coronary care unit admission are included (all diagnosis positions). All results with a false discovery rate (Benjamini-Hochberg) < 0.01 are reported.

ICD10 code	**Odds ratio**	**Class label**	**Misclassification**
M90	Inf	control	NSTEMI
Z89	Inf	control	NSTEMI
A04	Inf	control	STEMI
I22	Inf	control	STEMI
E34	Inf	control	STEMI
E42	Inf	control	STEMI
A41	Inf	NSTEMI	STEMI
G90-G99	141,8	control	STEMI
I97	68,4	control	STEMI
J09	68,4	control	STEMI
J10	68,4	control	STEMI
T88	34,5	control	STEMI
I24	23,0	control	STEMI
I46	9,8	control	STEMI
I42	9,4	control	STEMI
J81	9,4	control	NSTEMI
I40	6,8	control	STEMI
I35	3,8	control	NSTEMI