



## UvA-DARE (Digital Academic Repository)

### Artificial intelligence in the prognostication and classification of cardiovascular diseases

Ramos, L.A.

**Publication date**

2021

**Document Version**

Final published version

[Link to publication](#)

**Citation for published version (APA):**

Ramos, L. A. (2021). *Artificial intelligence in the prognostication and classification of cardiovascular diseases*.

**General rights**

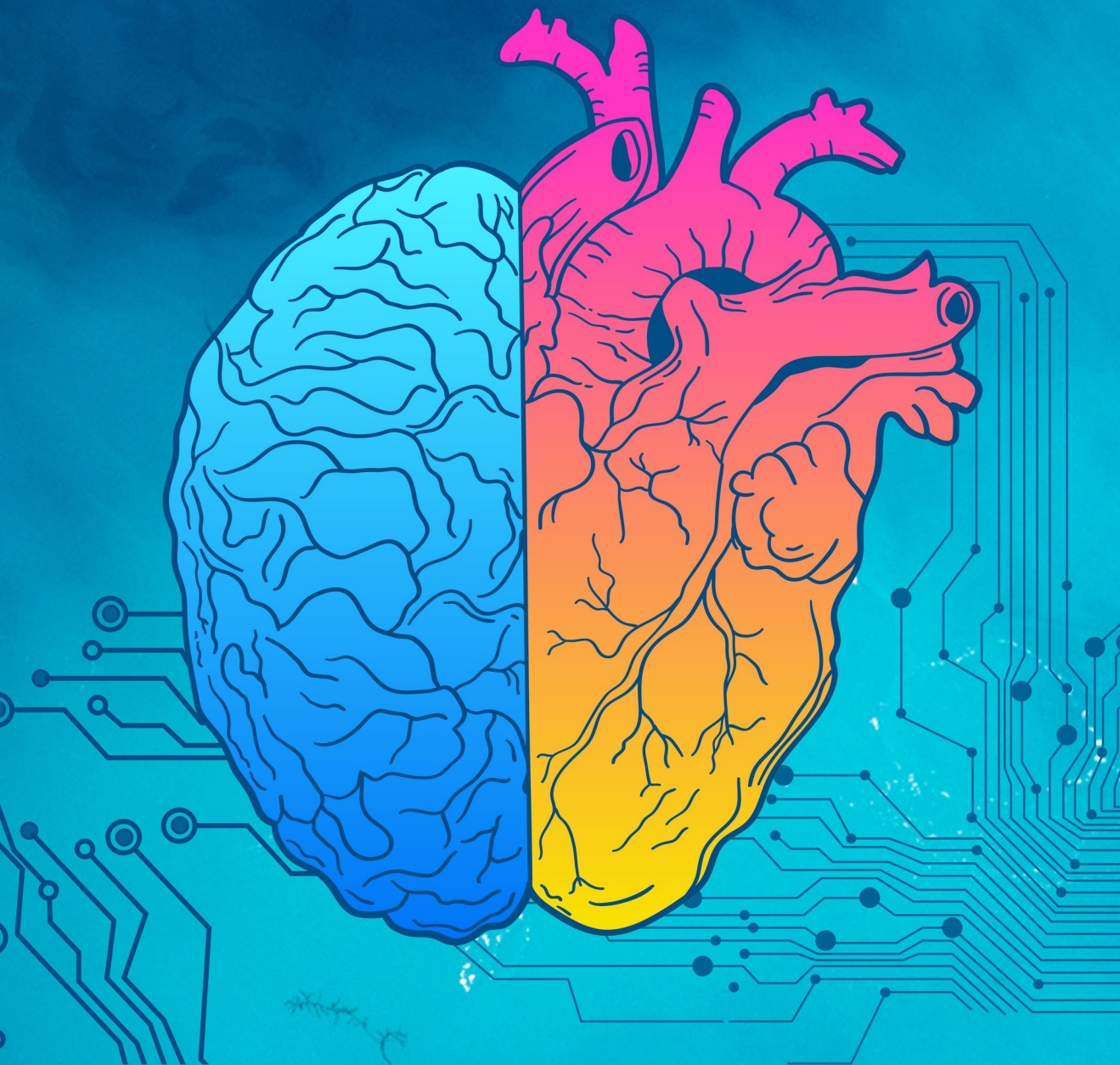
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

# ARTIFICIAL INTELLIGENCE IN THE PROGNOSTICATION AND CLASSIFICATION OF CARDIOVASCULAR DISEASES

Lucas Alexandre Ramos



Artificial intelligence in the  
prognostication and classification of  
cardiovascular diseases

Lucas Alexandre Ramos

**Design/Layout:** Ramos LA, da Silva MC

**Cover Design:** da Silva MC (Canva Pro, PhotoPhiltre, InDesign)

**Print:** Proefschriftmaken.nl

The work in this thesis was supported by ITEA3 Medolution: Project 14003

© Lucas Alexandre Ramos, 2021

All rights reserved. No part of this book may be reproduced, distributed, stored in a retrieval system, or transmitted in any form or by any means, without prior written permission of the author.

Artificial intelligence in the prognostication and classification of cardiovascular diseases

## ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor  
aan de Universiteit van Amsterdam  
op gezag van de Rector Magnificus  
prof. dr. ir. K.I.J. Maex

ten overstaan van een door het College voor Promoties ingestelde commissie,  
in het openbaar te verdedigen in de Agnietenkapel  
op donderdag 27 mei 2021, te 13.00 uur

door Lucas Alexandre Ramos  
geboren te Botucatu - SP

***Promotiecommissie***

*Promotores:*

prof. dr. A.H. Zwinderman AMC-UvA  
prof. dr. ir. G.J. Strijkers AMC-UvA

*Copromotores:*

dr. H.A. Marquering AMC-UvA  
dr. S.D. Olabariaga AMC-UvA

*Overige leden:*

prof. dr. ir. I. Išgum AMC-UvA  
prof. dr. W.J. Niessen Erasmus Universiteit Rotterdam  
prof. dr. A. Abu-Hanna AMC-UvA  
prof. dr. M. Hoogendoorn Vrije Universiteit Amsterdam  
prof. dr. ir. C.I. Sánchez  
Gutiérrez Universiteit van Amsterdam  
prof. dr. Y.M. Pinto AMC-UvA

Faculteit der Geneeskunde

# Table of Contents

<i>Chapter 1</i>	Introduction	7
<i>Chapter 2</i>	Machine learning improves prediction of delayed cerebral ischemia in patients with subarachnoid hemorrhage Journal of NeuroInterventional Surgery 2019;11:497-502	21
<i>Chapter 3</i>	Predicting outcome of endovascular treatment for acute ischemic stroke: Potential value of machine learning algorithms Front Neurol. 2018 Sep 25;9:784	49
<i>Chapter 4</i>	Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke Computers in Biology and Medicine 2019;115	73
<i>Chapter 5</i>	Combination of radiological and clinical baseline data for outcome prediction of patients with an acute ischemic stroke <i>Under submission</i>	101
<i>Chapter 6</i>	Predicting poor outcome prior to endovascular treatment in patients with acute ischemic stroke Front. Neurol. 2020;11- 1215	135
<i>Chapter 7</i>	Computer versus cardiologist: Is a machine learning algorithm able to outperform an expert in diagnosing phospholamban (PLN) p.Arg14del mutation on ECG? Heart Rhythm. 2021 Jan;18 (1):79-87	173
<i>Chapter 8</i>	Quantitative analysis of EEG reactivity for neurological prognostication after cardiac arrest <i>Submitted to Clinical Neurophysiology</i>	199
<i>Chapter 9</i>	Discussion	225



1



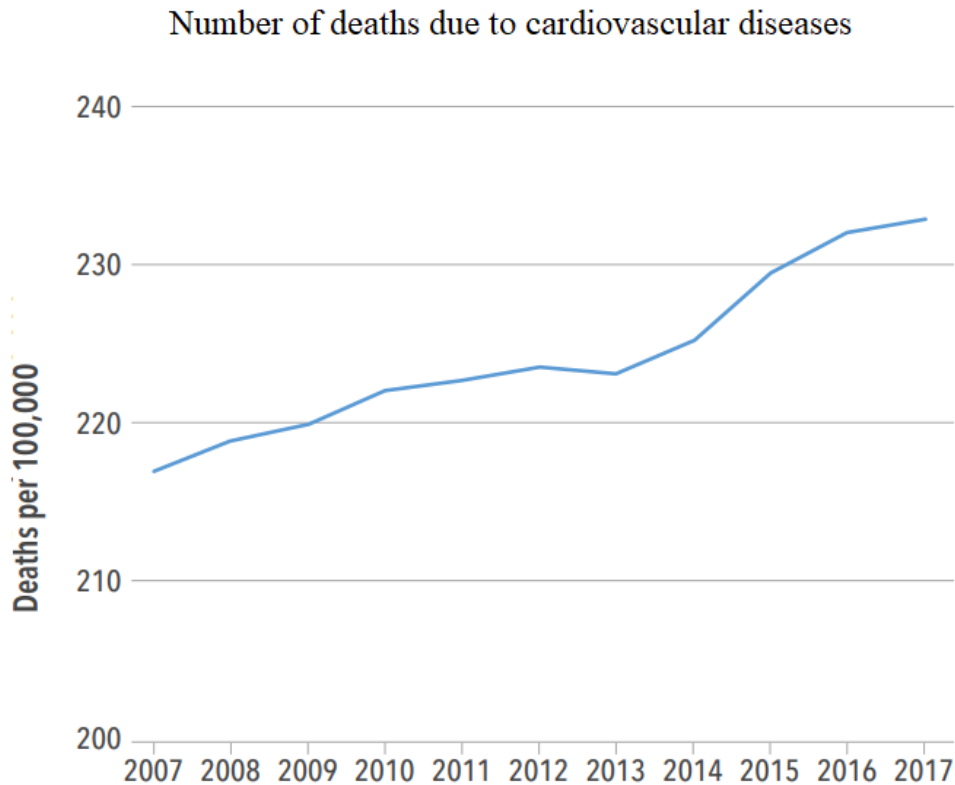
# CHAPTER 1.

## Introduction

## Cardiovascular diseases

Cardiovascular-related diseases are a major cause of death worldwide. Despite increasing efforts to develop new medications intended to prevent mortality from cardiovascular diseases, the number of deaths has been steadily increasing since 2007 (see Figure 1.1). Among the most common cardiovascular diseases is cardiac arrest. Cardiac arrest is the leading cause of mortality worldwide, and is responsible for almost half of all deaths related to cardiovascular diseases (1). Survival rates after cardiac arrest range from 1% to 10% (2–4), but vary greatly depending on the country and population (1). Cardiomyopathies and inherited arrhythmias are responsible for 15% and 2% of all cardiac arrest deaths, respectively. The Phospholamban (PLN) p.Arg14del mutation is a rare cardiomyopathy that appears in only 0.08% to 0.38% in cardiomyopathy cohorts (5). Despite being rare, patient that carry a PLN mutation are subjected to malignant arrhythmias and end-stage heart failure, which can affect patients from a very young age.

Stroke is another common cause of death worldwide (6,7). According to the World Health Organization (8), in 2016 stroke was the second leading cause of early death and disability, affecting millions of people worldwide, and it is expected to remain in this position until 2040 (8). Stroke can be divided into a hemorrhagic and an ischemic subtype.



**Figure 1.1.** The steady increase of number of deaths due to cardiovascular diseases. Adapted from (8).

## Hemorrhagic stroke

Hemorrhagic stroke is the less common subtype of stroke, and accounts for 10% to 20% of all stroke cases (9). It is most commonly caused by the ruptured of an artery that is weakened due to an aneurysm or malformation. Hemorrhagic stroke can be divided into intracranial and subarachnoid, the last being the less common one. Despite being less common, subarachnoid hemorrhagic (SAH) stroke is very severe, with a mortality rate of around 50%, and often affects patients younger than 55 (10), leading to severe loss of productivity and quality of life. Patients with SAH are initially treated with clipping or coiling to stop the hemorrhage. After initial treatment, these patients are at risk to several complications, with rebleeding being the most common one, usually occurring within 24 hours after stroke onset (11,12). Another complication is Delayed Cerebral Ischemia (DCI), which usually occurs from 4 to 14 days after stroke (9), and is one of the major causes of death and morbidity after SAH, requiring intensive patient monitoring (13). Currently, nimodipine administered after stroke onset is the only treatment

for SAH that has shown to improve patient outcome and prevent complications (11).

## Ischemic stroke

Ischemic stroke accounts for around 80% of all strokes (14). It occurs when a blood clot blocks the blood flow of an intracranial vessel, leading to extensive ischemia of the brain tissue. In around 30% of ischemic stroke patients, the occlusion occurs in one of the major intracranial arteries, including the middle cerebral artery, the anterior cerebral artery and the carotid cerebral artery (15). This type of occlusion is called a Large Vessel Occlusion (LVO). Despite recent advances in treatment options, around one-third of the patients who suffer an ischemic stroke die or remain dependent on nursing care (16). Therapy with intravenous alteplase (IV) has been proven to improve patient outcome, provided that it is administered within the effective time window (less than 4.5 hours after stroke onset) (14,17). Despite its proven success, recanalization (the restoration of cerebral blood flow in the affected region) is only achieved in 33% of the cases, due to several limitations in the treatment, which includes the appropriate time window for administration, low success rates in patients with LVO and large clot burden (14,18). Endovascular treatment (EVT) is a state-of-the-art treatment for patients with LVO. Combined with IVT, EVT has been shown to significantly improve patient outcome in many trials (16,19–21), becoming the standard treatment for ischemic stroke patients.

## Cardiomyopathies

Cardiac arrest is often caused by a heart malfunction that causes abnormal heart rhythms, such as tachycardia or fibrillation. Patients who survive a cardiac arrest are admitted to the intensive care unit, where they are medicated and intensively monitored for signs of complications. Responsiveness is also evaluated during patient monitoring and, based on treatment guidelines, after 72 hours it can already be decided to stop treatment and remove life support in case the patient remains unresponsive. Therefore, the accurate prediction of patient outcome can be of great assistance in decision support (22,23).

Phospholamban (PLN) is a phosphoprotein responsible for regulating calcium homeostasis in heart muscle cells. It was recently discovered in the Netherlands that a mutation in this gene can lead to severe ventricular arrhythmias, such as tachycardia and fibrillation, which increases the risk of developing heart-failure. It is estimated that, in the north of The Netherlands, 1 in 1500 people carry this mutation without knowing it. Despite the fact that there is currently no treatment available to mitigate the effects of the mutation, early diagnosis is of utmost importance for patients to properly treat symptoms and reduce the risk of complications (24–26).

## Computer Vision and Artificial intelligence

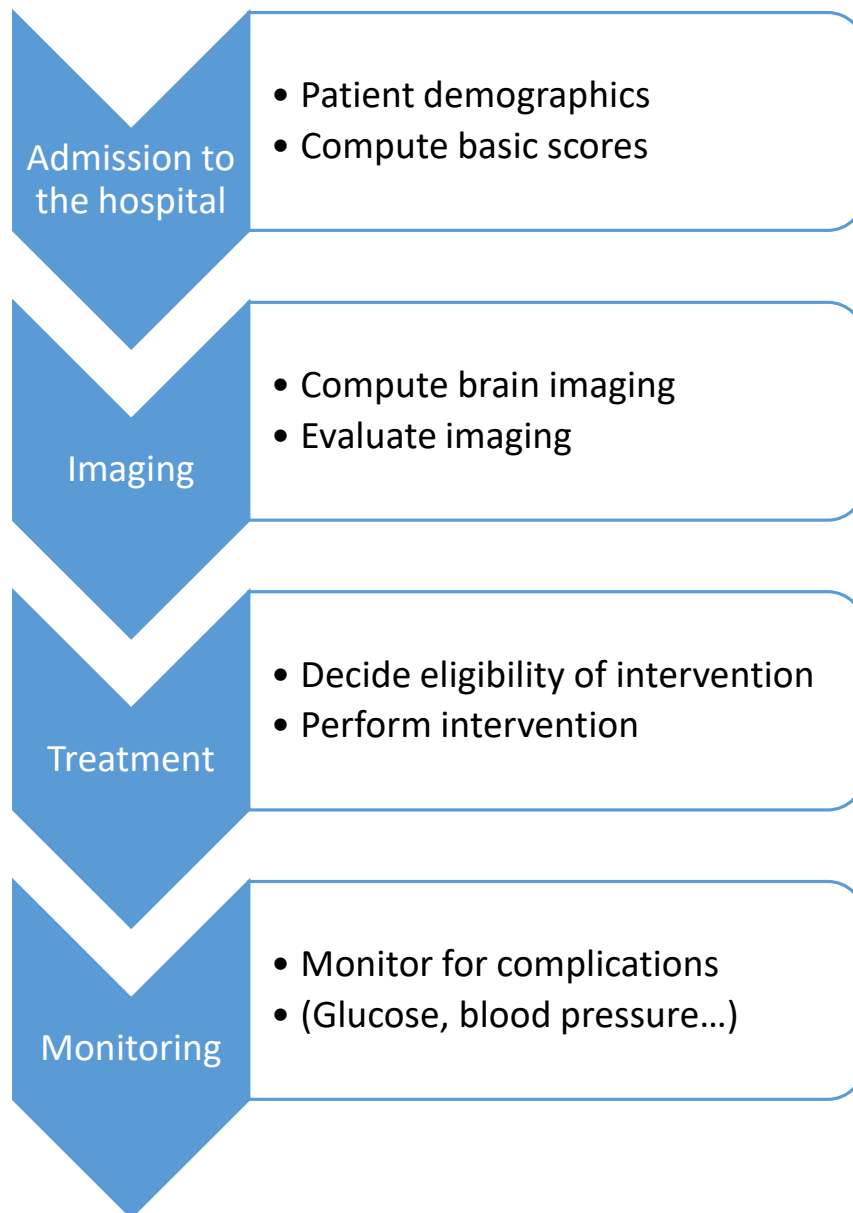
Computer Vision and Artificial intelligence (AI) have been broadly and successfully used to address and solve clinical problems (27–36). Clinical data is often composed by a large number of variables, images and signals, that can offer great predictive value if used correctly. We present in Figure 1.2, a diagram with an example of the data that is often collected during stroke care. As it can be observed in Figure 1.2, data from multiple sources and types are collected during patient care.

In recent years, the development and use of AI has gained significant traction fueled by advances in computer hardware and powerful Graphical Processing Units (GPUs). From the subfields of AI, machine learning and deep learning are amongst the most popular ones. Machine and deep learning can automatically learn patterns from the data, enabling the discovery of important feature interactions through linear and non-linear combinations. Provided that the right approach is used, machine learning models can handle a large number of variables and data from multiple sources, which are often available in clinical datasets. Regarding imaging data, in the past we relied on hand-crafted features derived from computer vision approaches (e.g. filtering) to solve computer vision tasks. Nowadays, deep learning offers the possibility of learning filters that better fit the image and task at hand, which has greatly increased results in multiple clinical and non-clinical applications.

In the field of cardiovascular diseases, machine/deep learning have been used broadly used for multiple types of tasks and data. For example, for electrocardiogram classification (37–39), for predicting cardiac arrest (40–

42), for computing radiological stroke scores (33,43,44), for predicting tissue outcome after stroke (27), and for predicting stroke patient outcome and complications (45–48), among many others. In all these works, machine/deep learning led to state-of-the-art results, greatly improving the performance in those tasks and leading to new discoveries.

Despite the success, there is still a lot of room for improvement in machine and deep learning models. Given the heterogeneous and complex nature of medical data, often being composed of demographics, images and signals, many approaches tend to focus on one type of data rather than exploring the combination of all data available, which can lead to information loss. Moreover, nonlinearities introduced by machine/deep learning models make model visualization and interpretation challenging, which can hamper the trust in results and their use in clinical practice.



**Figure 1.2.** Example of data collected during stroke care. During hospital admission patient demographics and evaluation scores are computed to assess stroke severity. Imaging is used to determine stroke subtype, location and confirm severity. Based on the data collected, the intervention is chosen and performed. Finally, the patient is intensively monitored for possible signs of complications.

## Thesis outline

This thesis focuses on the application of machine/deep learning into the field of cardiovascular-related diseases. I have explored various methods for the prediction of patient functional outcome and reperfusion after treatment of ischemic stroke, for the prediction of complications such as delayed cerebral ischemia in hemorrhagic stroke, the classification of patients with

cardiovascular-related gene mutations using electrocardiogram (ECG), and for the prediction of patient outcome after cardiac arrest. In all these topics, I aimed at combining heterogeneous data when possible and available, and to explore the best way of interpreting and visualizing the prediction models. My main contributions were: (a) applications of machine/deep learning methods to clinically relevant topics, (b) machine/deep learning approaches for heterogeneous data analysis (including image, signals and patient variables), (c) extensively optimized and validated model development pipelines, to prevent overoptimistic results and data leakage during training and testing, and (d) clinically interpretable models, to show the reasoning behind model predictions.

**Chapter 2** is about the prediction of delayed cerebral ischemia after subarachnoid hemorrhagic stroke, since previous models showed very low prediction accuracies, and DCI remains a major cause of death after SAH. We developed models that combined features automatically learned from NCCT, with patient demographics and radiological scores.

Regarding ischemic stroke, in **chapter 3** we investigated the prediction of good functional outcome at three months and good reperfusion (treatment outcome). We used all variables that were available at baseline from a large ischemic stroke registry, developed and validated several machine learning models, and compared them to previous prediction models found in literature.

After evaluating all baseline variables, **chapter 4** explored the predictive value from imaging data, where we adapted deep learning approaches to contrast enhanced computed tomography to predict functional outcome and reperfusion after acute ischemic stroke. We compared the added value from the automatically learned deep learning features to the common radiological scores used for patient condition assessment, and generated visualizations of important image regions.

**Chapter 5** focuses on the combination of clinical and imaging data for the prediction of good functional outcome and reperfusion in patients who suffered from an acute ischemic stroke. We implemented two approaches for extracting features from the images; a radiomics and a deep learning approach. For both, the whole 3D scans were used. Finally, we combined the



features learned with our approaches to all clinical variables available at baseline, evaluated performance and visualized feature importance.

The prediction of poor functional outcome was assessed in **chapter 6**, since around 30% of ischemic stroke patients either die or remain severely disabled. For these patients, treatment would be essentially futile. Therefore, we aimed at creating models for selecting poor functional outcome patients, while keeping the number of misclassified good outcome patients as low as possible.

In **chapter 7**, we present an approach for predicting the PLN gene mutation using ECG signals. We explored multiple approaches, using wavelets, convolutional neural networks and recurrent models and identified ECG regions that were deemed relevant for prediction.

**Chapter 8** is about predicting the outcome at 6 months of comatose patients after cardiac arrest using electroencephalograms (EEGs). We trained machine learning models on features extracted from the EEG signals and evaluated the differences in prediction accuracy between applying several kinds of stimuli to the patients and the signal background.

Finally, in **chapter 9** I discuss the main thesis findings, strengths and limitations of the applications of artificial intelligence to cardiovascular problems and suggest some areas of improvements for future research.

## References

1. Wong CX, Brown A, Lau DH, Chugh SS, Albert CM, Kalman JM, et al. Epidemiology of Sudden Cardiac Death: Global and Regional Perspectives. *Hear Lung Circ.* 2019;28 (1):6–14.
2. Beck B, Bray J, Cameron P, Smith K, Walker T, Grantham H, et al. Regional variation in the characteristics, incidence and outcomes of out-of-hospital cardiac arrest in Australia and New Zealand: Results from the Aus-ROC Epistry. *Resuscitation.* 2018;
3. Gräsner JT, Lefering R, Koster RW, Masterson S, Böttiger BW, Herlitz J, et al. EuReCa ONE—27 Nations, ONE Europe, ONE Registry: A prospective one month analysis of out-of-hospital cardiac arrest outcomes in 27 countries in Europe. *Resuscitation.* 2016;
4. Benjamin EJ, Virani SS, Callaway CW, Chamberlain AM, Chang AR, Cheng S, et al. Heart disease and stroke statistics - 2018 update: A report from the American Heart Association. *Circulation.* 2018;

5. Hof IE, van der Heijden JF, Kranias EG, Sanoudou D, de Boer RA, van Tintelen JP, et al. Prevalence and cardiac phenotype of patients with a phospholamban mutation. *Netherlands Hear J*. 2019;
6. Johnson CO, Nguyen M, Roth GA, Nichols E, Alam T, Abate D, et al. Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol*. 2019;
7. World Health Organization. WHO - The top 10 causes of death. 24 Maggio. 2018.
8. World Health Organization (WHO). Findings from the Global Burden of Disease Study 2017. *Lancet*. 2017;
9. Ikram MA, Wieberdink RG, Koudstaal PJ. International epidemiology of intracerebral hemorrhage. *Curr Atheroscler Rep*. 2012;14 (4):300–6.
10. Van-Gijn J, Kerr R, Rinkel G. Subarachnoid haemorrhage. *Lancet*. 2007;369:306–318.
11. Manoel AL de O, Loch Macdonald R. Neuroinflammation as a target for intervention in subarachnoid hemorrhage. *Front Neurol*. 2018;9 (MAY).
12. Vergouwen MDI, Vermeulen M, van Gijn J, Rinkel GJE, Wijdicks EF, Muizelaar JP, et al. Definition of delayed cerebral ischemia after aneurysmal subarachnoid hemorrhage as an outcome event in clinical trials and observational studies: proposal of a multidisciplinary research group. *Stroke*. 2010;41 (10):2391–5.
13. Macdonald RL. Delayed neurological deterioration after subarachnoid haemorrhage. *Nat Rev Neurol*. 2014;10 (1):44–58.
14. Blackham KA, Meyers PM, Abruzzo TA, Alberquerque FC, Fiorella D, Fraser J, et al. Endovascular therapy of acute ischemic stroke: Report of the Standards of Practice Committee of the Society of NeuroInterventional Surgery. *J Neurointerv Surg*. 2012;4 (2):87–93.
15. Chen C-J, Ding D, Starke RM, Mehndiratta P, Crowley RW, Liu KC, et al. Endovascular vs medical management of acute ischemic stroke. *Neurology*. 2015;85 (22):1980–90.
16. Jansen IGH, Mulder MJHL, Goldhoorn RJB. Endovascular treatment for acute ischaemic stroke in routine clinical practice: Prospective, observational cohort study (MR CLEAN Registry). *BMJ*. 2018;360.
17. Fransen PSS, Beumer D, Berkhemer OA, van den Berg LA, Lingsma H, van der Lugt A, et al. MR CLEAN, a multicenter randomized clinical trial of endovascular treatment for acute ischemic stroke in the Netherlands: Study protocol for a randomized controlled trial. *Trials*. 2014;15 (1).
18. Alqahtani SA, Steiner AB, McCullough MF, Bell RS, Mai J, Liu A-H, et al. Endovascular Management of Stroke Patients with Large Vessel Occlusion and Minor Stroke Symptoms. *Cureus*. 2017;9 (6):6–11.
19. Goyal M, Menon BK, Van Zwam WH, Dippel DWJ, Mitchell PJ, Demchuk AM, et al. Endovascular thrombectomy after large-vessel ischaemic stroke: A meta-analysis of individual patient data from five randomised trials. *Lancet*. 2016;387 (10029):1723–31.
20. Muir KW, Ford GA, Messow CM, Ford I, Murray A, Clifton A, et al. Endovascular therapy for acute ischaemic stroke: The Pragmatic Ischaemic Stroke Thrombectomy Evaluation (PISTE) randomised, controlled trial. *J Neurol Neurosurg Psychiatry*. 2017;88 (1):38–44.

21. Bracard S, Ducrocq X, Mas JL, Soudant M, Oppenheim C, Moulin T, et al. Mechanical thrombectomy after intravenous alteplase versus alteplase alone after stroke (THRACE): a randomised controlled trial. *Lancet Neurol.* 2016;15 (11):1138–47.
22. Ruijter BJ, Tjepkema-Cloostermans MC, Tromp SC, van den Bergh WM, Foudraine NA, Kornips FHM, et al. Early electroencephalography for outcome prediction of postanoxic coma: A prospective cohort study. *Ann Neurol.* 2019;
23. Rossetti AO, Urbano LA, Delodder F, Kaplan PW, Oddo M. Prognostic value of continuous EEG monitoring during therapeutic hypothermia after cardiac arrest. *Crit Care.* 2010;
24. Van Der Zwaag PA, Van Rijsingen IAW, Asimaki A, Jongbloed JDH, Van Veldhuisen DJ, Wiesfeld ACP, et al. Phospholamban R14del mutation in patients diagnosed with dilated cardiomyopathy or arrhythmogenic right ventricular cardiomyopathy: Evidence supporting the concept of arrhythmogenic cardiomyopathy. *Eur J Heart Fail.* 2012;
25. Van Rijsingen IAW, Van Der Zwaag PA, Groeneweg JA, Nannenbergh EA, Jongbloed JDH, Zwinderman AH, et al. Outcome in phospholamban R14del carriers results of a large multicentre cohort study. *Circ Cardiovasc Genet.* 2014;
26. Bosman LP, Verstraelen TE, van Lint FHM, Cox MGPI, Groeneweg JA, Mast TP, et al. The Netherlands Arrhythmogenic Cardiomyopathy Registry: design and status update. *Netherlands Hear J.* 2019;
27. Nielsen A, Hansen MB, Tietze A, Mouridsen K. Prediction of Tissue Outcome and Assessment of Treatment Effect in Acute Ischemic Stroke Using Deep Learning. *Stroke.* 2018;49 (6):1394–401.
28. Tjepkema-Cloostermans MC, da Silva Lourenço C, Ruijter BJ, Tromp SC, Drost G, Kornips FHM, et al. Outcome Prediction in Postanoxic Coma With Deep Learning. *Crit Care Med.* 2019;47 (10):1424–32.
29. Salem M, Taheri S, Yuan JS. ECG Arrhythmia Classification Using Transfer Learning from 2- Dimensional Deep CNN Features. 2018 IEEE Biomed Circuits Syst Conf BioCAS 2018 - Proc. 2018;
30. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal.* 2017;36:61–78.
31. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A Survey on Deep Learning in Medical Image Analysis. 2017; (1995). Available from: <http://arxiv.org/abs/1702.05747><http://dx.doi.org/10.1016/j.media.2017.07.005>
32. Shen D, Wu G, Suk H. Deep Learning in Medical Image Analysis. *Annu Rev Biomed Engeneerring.* 2017; (March):221–48.
33. Kuang H, Najm M, Chakraborty D, Maraj N, Sohn SI, Goyal M, et al. Automated aspects on noncontrast CT scans in patients with acute ischemic stroke using machine learning. *Am J Neuroradiol.* 2019;40 (1):33–8.
34. Alawieh A, Zaraket F, Alawieh MB, Chatterjee AR, Spiotta A. Using machine learning to optimize selection of elderly patients for endovascular thrombectomy. *J Neurointerv Surg.* 2019;1–6.

35. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: A 5-year multicentre prospective registry analysis. *Eur Heart J*. 2017;38 (7):500–7.
36. Giacobbe DR. Clinical interpretation of an interpretable prognostic model for patients with COVID-19. *Nat Mach Intell* [Internet]. 2020;42256. Available from: <http://dx.doi.org/10.1038/s42256-020-0207-0>
37. Sansone M, Fusco R, Pepino A, Sansone C. Electrocardiogram pattern recognition and analysis based on artificial neural networks and support vector machines: A review. *Journal of Healthcare Engineering*. 2013.
38. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. 2019;25 (January).
39. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med*. 2019;25 (1):70–4.
40. Kim J, Chae M, Chang H-J, Kim Y-A, Park E. Predicting Cardiac Arrest and Respiratory Failure Using Feasible Artificial Intelligence with Simple Trajectories of Patient Data. *J Clin Med*. 2019;8 (9):1336.
41. Kwon JM, Lee Y, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in-hospital cardiac arrest. *J Am Heart Assoc*. 2018;7 (13):1–11.
42. Sandroni C, D'Arrigo S, Cacciola S, Hoedemaekers CWE, Kamps MJA, Oddo M, et al. Prediction of poor neurological outcome in comatose survivors of cardiac arrest: a systematic review [Internet]. Vol. 46, *Intensive Care Medicine*. Springer Berlin Heidelberg; 2020. 1803–1851 p. Available from: <https://doi.org/10.1007/s00134-020-06198-w>
43. Lee E-J, Kim Y-H, Kim N, Kang D-W. Deep into the Brain: Artificial Intelligence in Stroke Imaging. *J Stroke*. 2017;19 (3):277–85.
44. Do LN, Baek BH, Kim SK, Yang HJ, Park I, Yoon W. Automatic assessment of ASPECTS using diffusion-weighted imaging in acute ischemic stroke using recurrent residual convolutional neural network. *Diagnostics*. 2020;10 (10).
45. Cui H, Wang X, Bian Y, Liu Y. Clinical Outcome Prediction Of Ischemic Stroke Based On Image Signature Selection From Multimodality Data. 40th Annu Int Conf IEEE Eng Med Biol Soc. 2018;722–5.
46. Forkert ND, Verleger T, Cheng B, Thomalla G, Hilgetag CC, Fiehler J. Multiclass support vector machine-based lesion mapping predicts functional outcome in ischemic stroke patients. *PLoS One*. 2015;10 (6):1–16.
47. Alaka SA, Menon BK, Brobbey A, Williamson T, Goyal M, Demchuk AM, et al. Functional Outcome Prediction in Ischemic Stroke: A Comparison of Machine Learning Algorithms and Regression Models. *Front Neurol*. 2020.
48. Nishi H, Oishi N, Ishii A, Ono I, Ogura T, Sunohara T, et al. Deep Learning-Derived High-Level Neuroimaging Features Predict Clinical Outcomes for Large Vessel Occlusion. *Stroke*. 2020;1484–92.





2

# CHAPTER 2.

## Machine learning improves prediction of delayed cerebral ischemia in patients with subarachnoid hemorrhage

Ramos LA, van der Steen WE, Sales Barros R, Majoie CB, van den Berg R, Verbaan D, Vandertop WP, Zijlstra IJA, Zwinderman H, Strijkers GJ, Delgado Olabarriaga S, Marquering HA. Machine learning improves prediction of delayed cerebral ischemia in patients with subarachnoid hemorrhage. *Journal of NeuroInterventional Surgery* 2019;11:497-502.

DOI: [10.1136/neurintsurg-2018-014258](https://doi.org/10.1136/neurintsurg-2018-014258)



## Abstract

*Background and Purpose:* Delayed Cerebral Ischemia (DCI) is a severe complication in patients with aneurysmal subarachnoid hemorrhage. Several associated predictors have been previously identified. However, their predictive value is generally low. We hypothesize that Machine Learning (ML) algorithms for the prediction of DCI using a combination of clinical and image data lead to higher predictive accuracy than previously applied logistic regressions.

*Materials and Methods:* Clinical and baseline CT image data from 317 patients with aneurysmal subarachnoid hemorrhage were included. Three types of analysis were performed to predict DCI. First, the prognostic value of known predictors was assessed with logistic regression models. Second, ML models were created using all clinical variables. Third, image features were extracted from the CT-images using an auto-encoder and combined with clinical data to create ML models. Accuracy was evaluated based on the Area Under the Curve (AUC), sensitivity and specificity with 95% CI.

*Results:* The best AUC of the logistic regression models for known predictors was 0.63 (0.62–0.63). For the ML algorithms with clinical data there was a small, but statistically significant, improvement in the AUC 0.68 (0.65–0.69). Notably, aneurysm width and height were included in many of the ML models. The area under the curve was the highest for ML models that also included image features 0.74 (0.72–0.75).

*Conclusion:* Machine Learning algorithms significantly improve the prediction of DCI in patients with aneurysmal subarachnoid hemorrhage, particularly when image features are also included. Our experiments suggest that aneurysm characteristics are also associated to the development of DCI.



## Introduction

Delayed Cerebral Ischemia (DCI) is one of the most severe complications in patients with aneurysmal Subarachnoid Hemorrhage (aSAH) and is related to worsening of functional outcome. DCI occurs in 20 to 30% of patients who suffered from aSAH (1). The selection of patients with a high risk of developing DCI may improve patient outcome as well as reduce the costs related to futile intensive care monitoring for DCI (2).

Several studies have identified risk factors associated with the development of DCI such as: World Federation of Neurosurgical Societies (WFNS) grade, age, aneurysm treatment (clipping or coiling), intraparenchymal and intraventricular hemorrhage, Total Blood Volume (TBV) (3), hypertension, diabetes mellitus, history of smoking, alcohol use, hyperglycemia and Hunt and Hess grade on admission (4).

Most studies searching for DCI predictors relied on univariable and multivariable logistic regression analysis. The accuracy of these regression models is generally low (AUC of 0.63 (5) and 0.65 (6)) and their approaches often do not correct for over-optimistic results by applying bootstrapping or cross-validation strategies (7).

The volume and availability of (digital) clinical and image data has enormously increased over the past years, opening up new possibilities for predictive modelling. The integration and interpretation of data from multiple sources of information can be quite challenging (8). Machine Learning (ML) is a field of computer science whose algorithms can learn patterns from large datasets with multiple variables. An advantage of ML algorithms is that, once the outcome label is defined, the algorithms can automatically optimize (learn) their parameters with minimal oversight (9). Differently from regression models, ML algorithms can handle large amounts of data and patient characteristics while taking all their interactions into account (9). Therefore, ML algorithms yield a potential predictive gain in accuracy over regression models (9,10).

Recent works that applied ML algorithms to heterogeneous data (data from different sources, such as image and clinical characteristics) presented

positive results for classifying Alzheimer's disease and predicting patients at risk for aortic stenosis (11,12).

We hypothesize that ML algorithms can increase the accuracy of DCI prediction compared to traditional logistic regression models. Moreover, since the TBV and blood location present in baseline CT scans have already been proven to be associated to DCI (3,4,6), we hypothesize that the addition of automatically extracted image features from baseline CT-scans to clinical data improves accuracy of DCI prediction. To test these hypotheses, we explored three approaches for predicting the development of DCI in aSAH patients: (1) using known predictors from the literature and logistic regression, (2) using ML algorithms with all available variables, and (3) combining imaging and clinical data.

## Materials and Methods

### Population

Patients were included from a prospectively collected cohort consisting of consecutive aSAH patients admitted to the (Academic Medical Center, Amsterdam, The Netherlands) between December 2011 and December 2015. Inclusion criteria were: 1) aSAH with subarachnoid blood visible on admission non-contrast CT, or confirmed by xanthochromic cerebrospinal fluid after lumbar puncture, and 2) causative aneurysm proven on angiographic imaging. Patients who were included in the ongoing Ultra-Early Tranexamic Acid After Subarachnoid Hemorrhage (ULTRA) trial were excluded from the analysis because data from ongoing trials should not be used prematurely. Furthermore, we excluded patients for whom the admission CT-scan presented severe artifacts. As a result, a total of 317 were used for analysis. From the included 317 patients, 97 (30%) developed DCI. DCI was strictly defined as the occurrence of new focal neurological impairment or a decrease of two points or more on the Glasgow Coma Scale (GCS) (with or without new hypodensity on CT) that could not be attributed to other causes, according to Vergouwen et al (13). All patients received nimodipine orally (6x 60mg daily) as prophylaxis of DCI. The diagnosis was assessed by the treating neurosurgeon and patients were treated with hypertension induction. The medical ethics committee of (Academic Medical Center, Amsterdam,

The Netherlands) waived ethics approval for this retrospective analysis of pseudonymised patient data. The database has been pseudonymised and patients have given consent for the use of data for research

Because of the sensitive nature of the data, it is available upon request to the corresponding author. All code used is publicly available at the authors Github page.

### Machine Learning Algorithms

We selected the following ML algorithms: Logistic Regression (LR), Support Vector Machine (SVM) (14), Random Forest Classifier (RFC) (15), Multi-layer Perceptron (MLP) (16), Stacked Convolutional Denoising Auto-encoders (17), and Principal Component Analysis (PCA). These algorithms have shown state-of-the-art results in several studies on disease prediction, image segmentation and image feature representation (17,18). The parameters used for these algorithms are presented in the Supplemental Tables I, II and III. For the development of ML models, datasets are generally split in two: a training and a testing dataset. Machine learning models are first trained using a training dataset to optimize the prediction. Subsequently, the accuracy of the ML algorithms is evaluated on the testing dataset. The separation of training and testing data adopted in cross-validation (7) is important to assess the model performance and generalization to unseen data. In this study, we used Monte-Carlo cross-validation with 100 random splits (with 75% for training and 25% for testing) of the dataset into training and testing data and 5-fold cross-validation for optimizing the parameters of each model.

### Clinical Data

A total of 48 variables were included in this study. The full list of available demographic and clinical variables is presented in the Supplemental Table IV. Collected radiological variables were: modified Fisher scale on admission, number, location, height and width of aneurysm were determined based on CTA image data. Furthermore, data on treatment (clipping, coiling or no treatment) was also collected.

The percentage of missing values per variable is presented in the Supplemental Table IV. Missing values in the dataset were imputed using the incremental attribute regression imputation with random forest. This

imputation technique has shown high accuracy rates in several datasets (19). After data imputation, data normalization was performed by subtracting the mean and scaling to unit variance. For the nominal data, dummies were created. It has been shown that data normalization increases convergence rates (time and number of iterations for training the models) and it is necessary for many ML algorithms (20).

### Image Data

The available baseline non-contrast CT image data consists of  $512 \times 512 \times N$  voxels (where  $N$  is the number of slices) with an average voxel spacing of  $0.45 \pm 0.05$  mm and an average slice thickness of  $4.9 \text{ mm} \pm 0.6$  mm. Some image-derived features that are well known for being associated with DCI are the TBV and the blood location (3). The manual extraction of features from medical images is a time-consuming task and these features might not be the only important ones available in the images (21).

Potentially, each voxel can be considered a feature, therefore the number of features is too large to be efficiently used in ML algorithms. If the number of training samples is small compared to the number of features, the accuracy of ML algorithms can be strongly reduced, this problem is known as the curse of dimensionality (22). To avoid this problem and to account for variations in the image data (rotation and translation), in this work we applied image data downsampling and data augmentation following the approach adopted in a previous study (23).

Therefore, since the number of image features (voxels) is very large and relevant unknown image features might still be present in the baseline CT scans, we opted for an unsupervised feature learning technique (24). Feature learning is a technique used to automatically extract useful information from image data when building ML models (21). The Stacked Denoising Convolutional Auto-encoder (SDCAE) (17) is an unsupervised feature learning technique designed to automatically learn the most relevant features of an image. The parameters used for the auto-encoder are presented in the Supplemental Table III.

## Prediction Models

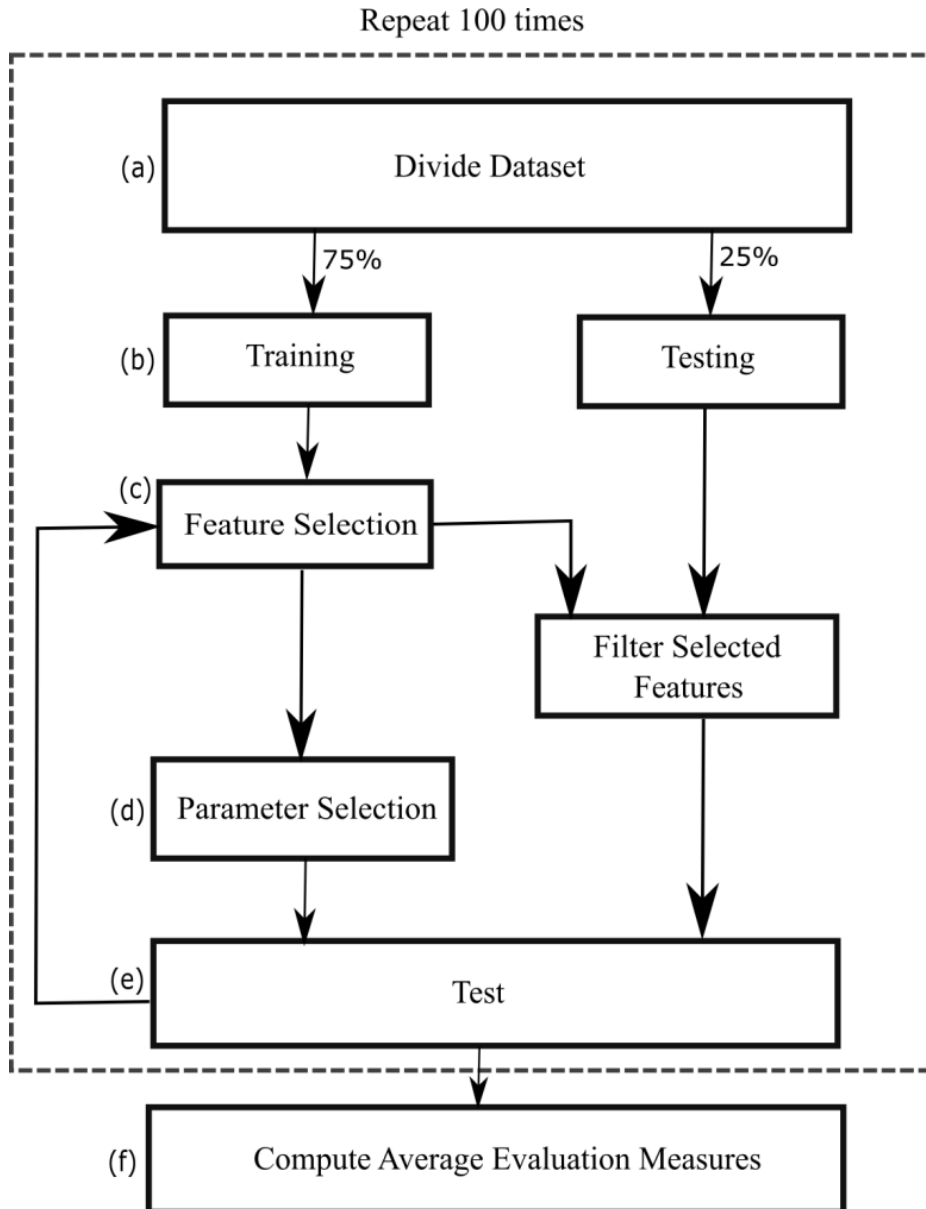
In our experiments, we used the implementations of LR, SVM, RFC and MLP algorithms available in the Scikit-learn toolkit (20). The parameters used for optimization are presented in the Supplemental Tables I and II. The Microsoft Cognitive Toolkit (CNTK) (25) was used for the auto-encoder algorithm. In this study we explored three approaches described below.

### Prior knowledge variables with logistic regression

We built 2 models using clinical variables for which the association with DCI has previously been established in literature (3,4) using multivariable logistic regression. The dataset is randomly split into training (75%) and testing (25%) set to prevent overoptimistic results. Model 1 included the following variables: WFNS, age, aneurysm treatment (clipping or coiling), intraparenchymal and intraventricular hemorrhage and TBV (3). Model 2 included the following variables: hypertension, diabetes mellitus, history of smoking, alcohol use, hyperglycemia and Hunt and Hess grade on admission (4).

### Clinical variables with Machine Learning

We built four predictive models using only clinical variables and ML algorithms (SVM, RFC, LR and MLP) and determined the most important variables. Figure 2.1 provides an overview of the workflow. First, the dataset is randomly split into training (75%) and testing (25%) set to prevent overoptimistic results and prevent overfitting. Subsequently, the training set is randomly split into training and validation using 5-fold cross-validation for feature selection and parameter optimization. Random Forest was used to assess feature importance since it is easily interpretable (20). Based on the RF feature importance variables were recursively eliminated. The variables left after each elimination were used to optimize the models. Finally, The ML models were applied to the testing set and their accuracy was measured. The steps c-e (Figure 2.1) were repeated until only one feature was left. The steps a-e were repeated 100 times using Monte-Carlo cross-validation (7). The averages and 95% CI of the accuracy measures were computed for the 100 cross validation iterations (Figure 2.1, step f).



**Figure 2.1.** Machine learning model creation workflow.

### Image features and clinical data with Machine Learning

We built four models using ML algorithms and a combination of the best clinical variables (determined with RFC) and features automatically extracted from CT-images using the auto-encoder (for implementation details see the Supplemental Section I. The number of features generated by the auto-encoder was much higher than the number of features in the clinical dataset (2048 vs 48). Therefore, to preserve the value of the clinical features, the dimension of image features was reduced using PCA, which transforms the data into a smaller set of features based on the variance as proposed by Zhang et al, (18). The number of PCA components was optimized based on the AUC.

The image features obtained with PCA were added to the clinical features (most relevant ones obtained from the ML approach) and the dataset containing the combination of features was used with the workflow presented in (Figure 2.1).

### Model predictive performance assessment

To evaluate the performance of each approach, we computed the average of the area under the curve (AUC) of the receiver operating characteristic curve (ROC) and the sensitivity and specificity with 95% CI. Differences in accuracy were considered significant if the CI did not overlap, and if the 95% CI of the difference between AUC distributions did not contain the null value. The specificity and sensitivity were calculated based on upper left corner of the ROC curve.

### Model interpretation

Machine Learning models are often seen as black boxes. However, for clinical decision making it is of utmost importance to understand what variables are considered important for the model and, in a deeper level, what variable influenced each individual prediction. To increase the interpretability of our results we explored the best performing model (Random Forest), by computing the average feature importance and ranking them (from most important to least important) to provide more insight into the impact of those features in the models.

For this purpose we applied a model explanation technique named Local Interpretable Model-agnostic Explanations (LIME) (26). LIME automatically creates an interpretable model locally around the prediction boundary of a given model (in our case the ML methods SVM, RF, LR, NN), providing an interpretation of each individual prediction and how the value of each variable affects it. To stress the importance of image features, we compared the models with and without image features and assessed the impact on DCI prediction using LIME. More details about LIME can be found in the Supplemental Section II.

## Results

The AUC values for the models built with variables manually chosen based on the prior knowledge approach are shown in Table 2.1. The combination of TBV, age, WFNS, treatment (clipping, coiling or no treatment), presence of intraparenchymal and intraventricular hemorrhage (Model 1) yielded the best average AUC of 0.63 (95% CI 0.62 - 0.63).

**Table 2.1.** Average AUC, Sensitivity and Specificity with 95% CI of DCI prediction models for all approaches. First two columns specify which Data (variables) were used to build each Model

Data	Model	AUC 95% CI	Sensitivity 95% CI	Specificity 95% CI
Prior knowledge variables*	LR (Model 1)	0.63 (0.62–0.63)	0.67 (0.64–0.70)	0.62 (0.59-0.65)
	LR (Model 2)	0.59 (0.57–0.60)	0.61 (0.57–0.65)	0.64 (0.60-0.68)
All clinical variables	SVM	0.64 (0.63–0.65)	0.67 (0.63–0.70)	0.64 (0.61-0.67)
	RFC	0.68 (0.65–0.69)	0.78 (0.75–0.81)	0.57 (0.54-0.61)
	LR	0.61 (0.60–0.63)	0.65 (0.61–0.68)	0.62 (0.59-0.67)
	MLP	0.63 (0.62–0.64)	0.59 (0.56–0.62)	0.79 (0.76-0.81)
All clinical variables see combined with extracted image features	SVM	0.68 (0.65–0.68)	0.63 (0.59–0.66)	0.73 (0.70-0.76)
	RFC	0.74 (0.72–0.75)	0.67 (0.65–0.70)	0.75 (0.72-0.78)
	LR	0.65 (0.64 - 0.67)	0.65 (0.62-0.67)	0.69 (0.66-0.71)
	MLP	0.67 (0.66 – 0.68)	0.64 (0.60-0.67)	0.72 (0.69-0.75)

AUC = Area under the curve; LR = Logistic Regression; SVM = Support Vector Machine; RFC = Random Forest; MLP = Multilayer Perceptron; All Variables = see Supplemental Table IV.

\* WFNS, age, treatment (clipping or coiling), intraparenchymal and intraventricular hemorrhage, (TBV) (3).

\*\* Hypertension, diabetes mellitus, history of smoking, alcohol use, hyperglycemia and Hunt and Hess grade on admission (4).

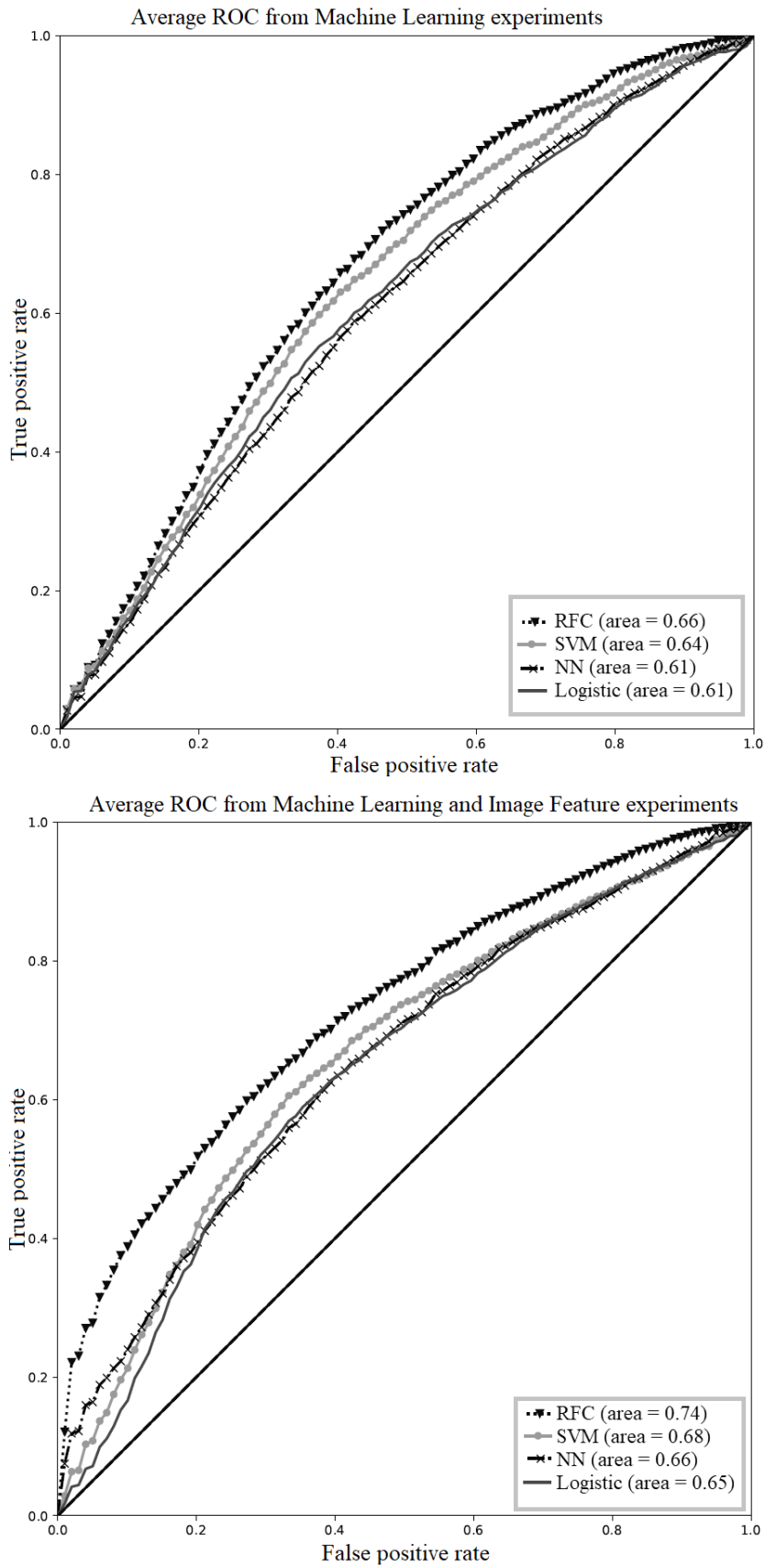
The most relevant clinical features (with the best AUC) selected by the ML models were, in order of relevance for the model: TBV, presence of intraparenchymal blood, time from ictus to CT, age, GCS, aneurysm height,



presence of subdural blood, aneurysm width, treatment (clipping, coiling or no treatment) and aneurysm location. The AUC measures for the ML methods are shown in Table 2.1 and the ROC curves are displayed in Figure 2.2 (top). The RFC had the highest accuracy with an AUC of 0.68 (95% CI 0.65 - 0.69).

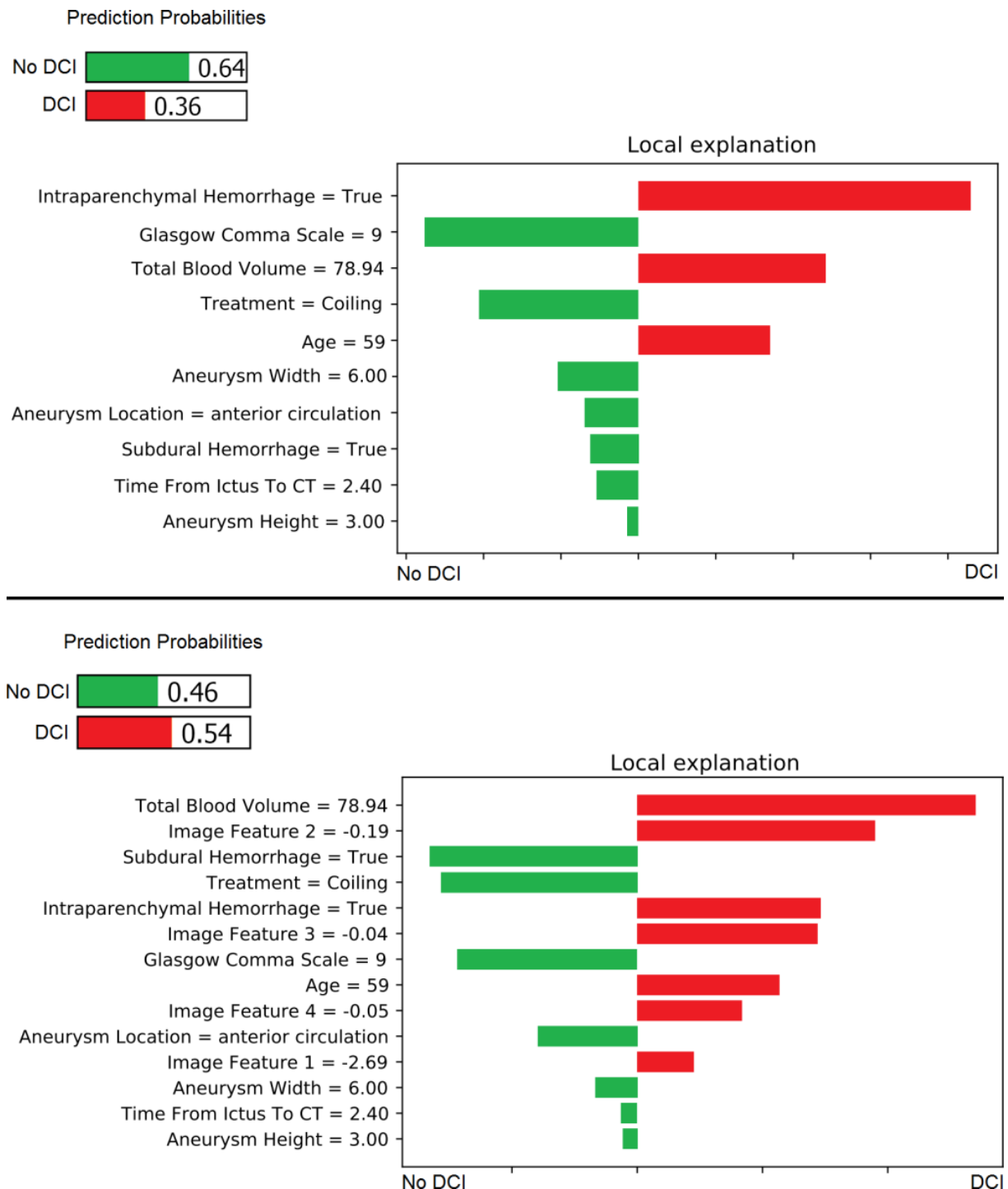
The AUCs for the Image features and clinical data with Machine Learning approach are shown in Table 2.1 and the ROC curves in Figure 2.2 (bottom). Again, the RFC had the highest accuracy with an AUC of 0.74 (95% CI 0.72 - 0.75), which was the highest accuracy obtained in our experiments. The most relevant features for this approach were, in order of relevance for the model: two automatically extracted image features, TBV, presence of intraparenchymal blood, time from ictus to CT, two other automatically extracted image features, age, aneurysm height, presence of subdural blood, aneurysm width, and GCS. The 95% CI of the difference between the AUC distributions of the Clinical variables with Machine Learning approach and the Image features and clinical data with Machine Learning approach were: 0.04 - 0.07. Therefore, we can conclude that there is statistically significant difference in the two distributions, suggesting that the image features extracted using an auto-encoder improved DCI prediction.

The Supplemental Figure II presents the feature importance for the best performing model (RF) using only clinical variables and using the combination of clinical variables and auto-encoder image features.



**Figure 2.2.** LIME model explanation of a DCI positive patient. The model built using the clinical variables suggest a lower risk of DCI (top). After including the image features (bottom), the model suggests a higher risk for DCI.

Figure 2.3 was created using the model explanation technique LIME using clinical features (top) and the combination of clinical and image features (bottom) to explain the decision of the RF model for a specific DCI patient.



**Figure 2.3.** Average ROC curve for only clinical variables with ML methods (top) and both clinical and image features with ML (bottom). RFC= Random Forest; SVM= Support Vector Machine; MLP = Multilayer Perceptron; LR= Logistic Regression.

We can note that the model without images suggests a lower risk of DCI (0.36), even though some variables point to a higher risk. This occurs because most of the variables point to a lower risk.

After combining the clinical and image features, many still point to a lower risk of DCI, though the majority of image features point to a higher risk of DCI. The combined features increase the total risk of DCI for this patient. More examples can be found in the Supplemental Figure I and II.

## Discussion

In this dataset, most ML methods showed higher accuracy in predicting DCI compared to logistic regression. The prediction accuracy of the models was improved when image features extracted automatically with auto-encoder were combined into the model. The highest average accuracy was obtained using the RFC and the combination of clinical and image features. Using LIME, we have shown how each feature used in the model affects an individual prediction, providing insight into the “black-box” ML models. This visualization provides insight into a model’s risk prediction. We have further provided a visualization of how the combination of image feature improved the accuracy of DCI risk prediction (Figure 2.3).

Our results suggest that the TBV, blood location, age, GCS, and treatment are associated with the occurrence of DCI, which is in accordance with previous studies (3,4,6). These previous studies relied mostly on multivariable LR. Notably, the accuracy obtained by LR models were the lowest in our study. With the use of ML algorithms, we found variables that increased the predictive accuracy, which have not been associated with DCI before (time from ictus to CT, presence of subdural blood, GCS, treatment, and aneurysm height, width and location). However, a causal relationship between these features and DCI was not further explored in this study. “In our analysis, some of the parameters with value in the prediction of DCI were not identified as risk factors in previous studies. For example, in contrast to previous studies (3,4), aneurysm width and height were included in our ML models. ML models use different mechanisms than commonly used linear regression techniques, which may put these parameters forward in the predictions. However, these parameters were not the most relevant ones in

our ML models, which is expressed by the relative low importance compared to other parameters (Supplemental Figure I).

In (27), the outcome of SAH patients using ML was the main topic. Their family of methods was restricted to decision trees, while in our work we included multiple families of methods such as Neural Networks, Support Vector Machines, Logistic Regression and Ensemble methods. Since the learning process of these families of methods differ from each other, a higher range of feature relationships could be explored in our set-up. The major contribution of our study comes from the combination of clinical and image data. With an automatic unsupervised feature extraction approach, image features were extracted from baseline CT scans and their combination with clinical features showed significant improvement in DCI prediction. Since our approach is unsupervised, the images do not require any sort of annotation and are less prone to bias from labels and overfitting. A downside of our approach is that the multiple downsampling steps hamper the interpretation of these image features.

There was a significant increase in sensitivity when comparing prior knowledge model 1 to RFC using all clinical variables. This shows that the RFC was better at identifying patients at risk of DCI than the other models. Though, the models were not statistically significant better at identifying patients not at risk. The combination of clinical data with image features increased the specificity of the models, making them more precise at identifying patients not at risk of developing DCI, which for clinical practice may be more useful to reduce the costs related to futile intensive care monitoring for DCI (2).

A limitation of common regression models is that the number of features that can be included is limited. Based on the Rule of Ten, one should have at least 10 events per feature included in the LR model. Note that it has already been proven that this rule is not so strict and that models with less events per feature (5-9) can still be used with good predictive results (28). In our dataset we had less than three events per feature, which makes the LR model prone to overfitting. The NCCT images contained a large number of voxels. Using the whole image for training the auto-encoder increases the risk of overfitting, due to the large number of input image features and parameters to optimize.

We reduced this risk by downscaling and augmenting the scans and applying cross-validation. The ML algorithms used in this study are able to handle such high dimensional feature spaces with less risk of overfitting, provided that proper approaches, such as data augmentation, cross-validation and regularization are taken into account (7,8). Even though Monte-Carlo cross-validation was used with 100 iterations, it does not replace the need for validation on an external dataset. Moreover, the loosely formulated definitions used for DCI makes external validation even harder, since two datasets with the same definition are needed. In our study, however, DCI was strictly defined according to the definition of Vergouwen et al (13) and consistently adopted throughout the dataset.

To determine the best parameter configurations to build the ML models can be computationally expensive and time consuming. Moreover, selecting the range of values used for fine-tuning is difficult. In this study, the selection of the range of values for the parameters was based on previous studies and the Scikit -learn toolkit implementation suggestions (20,29). Nevertheless, it may be worthwhile to study models with different (number of) parameters.

The interpretation of these 3D-image features is challenging, as discussed in other studies (30). This will be subject of future work which will investigate other feature extraction techniques for the image data that are easier to visualize, to provide insight in the interpretation of the image features.

## Conclusion

Our findings indicate that ML algorithms improves prediction of DCI in patients with aSAH in the population studied. We show that features that have not been considered before may increase the accuracy of DCI prediction. Feature visualization using LIME provides a better understanding of the models and might improve clinical decision making. Imaging features extracted automatically using ML techniques further improve the accuracy in predicting DCI.

## References

1. Van-Gijn J, Kerr R, Rinkel G. Subarachnoid haemorrhage. *Lancet*. 2007;369:306–318.
2. Macdonald RL. Delayed neurological deterioration after subarachnoid haemorrhage. *Nat Rev Neurol*. 2014;10 (1):44–58.
3. Zijlstra IA, Gathier CS, Boers AM, Marquering HA, Slooter AJ, Velthuis BK, et al. Association of automatically quantified total blood volume after aneurysmal subarachnoid hemorrhage with delayed cerebral ischemia. *Am J Neuroradiol*. 2016;37 (9):1588–93.
4. Ko SB, Choi HA, Carpenter AM, Helbok R, Schmidt JM, Badjatia N, et al. Quantitative analysis of hemorrhage volume for predicting delayed cerebral ischemia after subarachnoid hemorrhage. *Stroke*. 2011;42 (3):669–74.
5. De Oliveira Manoel AL, Jaja BN, Germans MR, Yan H, Qian W, Kouzmina E, et al. The VASOGRADE: A Simple Grading Scale for Prediction of Delayed Cerebral Ischemia after Subarachnoid Hemorrhage. *Stroke*. 2015;46:1826–31.
6. De Rooij NK, Greving JP, Rinkel GJE, Frijns CJM. Early prediction of delayed cerebral ischemia after subarachnoid hemorrhage: Development and validation of a practical risk chart. *Stroke*. 2013;44 (5):1288–94.
7. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Appear Int Jt Conf Artificial Intell*. 1995;5:1–7.
8. Lee E-J, Kim Y-H, Kim N, Kang D-W. Deep into the Brain: Artificial Intelligence in Stroke Imaging. *J Stroke*. 2017;19 (3):277–85.
9. Singal AG, Mukherjee A, Joseph Elmunzer B, Higgins PD, Lok AS et al. Machine Learning Algorithms Outperform Conventional Regression Models in Predicting Development of Hepatocellular Carcinoma. *Am J Gastroenterol*. 2016;1848 (11):3047–54.
10. Asadi H, Dowling R, Yan B, Mitchell P. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS One*. 2014;9 (2):14–9.
11. Daoqiang Zhanga, Yaping Wanga, b, Luping Zhoua, Hong Yuana, Dinggang Shena A, Initiative1 ADN. Multimodal Classification of Alzheimer’s Disease and Mild Cognitive Impairment. *Neuroimage*. 2011;55 (3):856–67.
12. B AS, Nutt DJ, Mcgonigle J. Identifying Patients at Risk for Aortic Stenosis Through Learning from Multimodal Data. *MICCAI 2011 14th Int Conf (Vision, Pattern Recognition, Graph*. 2016;9902.
13. Vergouwens MDI, Vermeulen M, van Gijn J, Rinkel GJE, Wijndicks EF, Muizelaar JP, et al. Definition of delayed cerebral ischemia after aneurysmal subarachnoid hemorrhage as an outcome event in clinical trials and observational studies: proposal of a multidisciplinary research group. *Stroke*. 2010;41 (10):2391–5.
14. Cortes C, Vapnik V. Support-Vector Networks. *Mach Learn*. 1995;
15. Breiman L. Random forests. *Mach Learn*. 2001;45 (1):5–32.
16. Bishop CM. Neural networks for pattern recognition. *J Am Stat Assoc*. 1995;92:482.
17. Du B, Xiong W, Wu J, Zhang L, Zhang L, Tao D. Stacked Convolutional Denoising Auto-Encoders for Feature Representation. *IEEE Trans Cybern*. 2017;47 (4):1017–27.

18. Zhang Y, Wang S, Dong Z. Classification of Alzheimer Disease Based on Structural Magnetic Resonance Imaging by Kernel Support Vector Machine Decision Tree. *Prog Electromagn Res.* 2014;144:171–84.
19. van Stein B, Wojtek K. An Incremental Algorithm for Repairing Training Sets with Missing Values. *Int Conf Inf Process Manag Uncertain Knowledge-Based Syst.* 2016;175–86.
20. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2012;12:2825–30.
21. Bengio Y, Courville A, Vincent P. Representation Learning : A Review and New Perspectives. 2012; (1993):1–30.
22. Duda RO, Hart PE. Pattern classification. *Pattern Recognit.* 2000;680.
23. Miki Y, Muramatsu C, Hayashi T, Zhou X, Hara T, Katsumata A, et al. Classification of teeth in cone-beam CT using deep convolutional neural network. *Comput Biol Med.* 2017;80 (September 2016):24–9.
24. Masci J, Meier U, Cireşan D, Schmidhuber J. Stacked convolutional auto-encoders for hierarchical feature extraction. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics).* 2011;6791 LNCS (PART 1):52–9.
25. Yu D, Eversole A, Seltzer M, Yao K, Huang Z, Guenter B, et al. An Introduction to Computational Networks and the Computational Network Toolkit. *Microsoft Tech Rep.* 2015;112 (MSR-TR-2014-112).
26. Ribeiro MT, Singh S, Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. 2016;1135–44. Available from: <http://arxiv.org/abs/1602.04938>
27. de Toledo P, Rios PM, Ledezma A, Sanchis A, Alen JF, Lagares A. Predicting the Outcome of Patients With Subarachnoid Hemorrhage Using Machine Learning Techniques. *Inf Technol Biomed IEEE Trans.* 2009;13 (5):794–801.
28. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and cox regression. *Am J Epidemiol.* 2007;165 (6):710–8.
29. Villanueva A, Hoshida Y, Battiston C, Tovar V, Sia D, Alsinet C, et al. Combining clinical, pathology, and gene expression data to predict recurrence of hepatocellular carcinoma. *Gastroenterology.* 2011;140 (5):1501–1512.e2.
30. Aubry M, Russell BC. Understanding deep features with computer-generated imagery. *Proc IEEE Int Conf Comput Vis.* 2015;2875–83.



## Supplemental material

**Supplemental Table I.** Hyper-parameters used for SVM

Classifier	Kernel Type	Penalty parameter $C$	Kernel coefficient $\gamma$	Degree of the Polynomial kernel
SVM	Linear	[0.001, 0.01, 0.1, 1, 10, 100]	n.a.	n.a.
	Radial basis function	[0.001, 0.01, 0.1, 1, 10, 100]	[1, 0.1, 0.01, 0.001, 0.0001]	n.a.
	Polynomial	[0.001, 0.01, 0.1, 1, 10, 100]	[1, 0.1, 0.01, 0.001, 0.0001]	[1,2,3,4]

**Supplemental Table II.** Hyper-parameters used for RFC and MLP

Classifier	Parameter Name	Parameter Value
RFC	Number of Trees	[100,200,400,600,800]
	Max features for split	auto, sqrt and log2
	Quality of split	Gini or Entropy
MLP	Hidden Layer sizes	[50,25], [60,30 ], [60,40,20], [50,30,10], [70,40,20], [70,30], [80,50,30], [80,60,30,10]
	Regularization parameter	[0.1, 0.01, 0.001, 0.0001]
	Batch size	[64, 128]
	Learning rate	[0.01, 0.05, 0.001, 0.005, 0.0001]

**Supplemental Table III.** Hyper-parameters used for the auto-encoder

Patch Size	Conv Layer	Max Pool	Conv Layer	Max pool	Conv Layer	Max pool
128x128x19	7x7x7	2x2x2	5x5x5	2x2x2	3x3x3	2x2x2
	feature maps=16	stride (2:2:1)	feature maps=16	stride (2:2:2)	feature maps=32	stride (2:2:1)
128x128x19	5x5x5	2x2x2	3x3x3	2x2x2	3x3x3	2x2x2
	feature maps=16	stride (2:2:1)	feature maps=16	stride (2:2:2)	feature maps=32	stride (2:2:1)

Supplemental Table IV. Patient characteristics

Variable	All (317)	no DCI (220)	DCI (97)	Missing %	p-value
Age (mean/SD)	57.66 (12.1)	57.68 (10.9)	57.62 (12.6)	0 (0.0)	0.964
Female sex (%)	211 (66.6)	143 (70.1)	68 (65.0)	0 (0.0)	0.448
History of aneurysmal subarachnoid hemorrhage (%)	5 (1.6)	4 (1.0)	1 (1.8)	31 (9.8)	1.0
History of intracerebral hemorrhage (%)	2 (0.6)	2 (0.0)	0 (0.9)	31 (9.8)	1.0
History of cardiovascular disorder (%)	58 (18.3)	43 (15.5)	15 (19.5)	24 (7.6)	0.50
History of diabetes mellitus (%)	21 (6.6)	15 (6.2)	6 (6.8)	29 (9.1)	0.939
History of hypertension (%)	104 (32.8)	75 (29.9)	29 (34.1)	27 (8.5)	0.521
History of hyper cholesterol (%)	53 (16.7)	35 (18.6)	18 (15.9)	32 (10.1)	0.617
History of Smoking (%)				83 (26.2)	0.628
<b>No</b>	55 (17.4)	37 (18.6)	18 (16.8)		
<b>Yes, but stopped</b>	60 (18.9)	44 (16.5)	16 (20.0)		
<b>Yes, still smokes</b>	119 (37.5)	79 (41.2)	40 (35.9)		
History of alcohol use (%)	134 (42.3)	89 (46.4)	45 (40.5)	83 (26.2)	0.349
Previous MRs (%)				89 (28.1)	0.741
<b>0</b>	158 (49.8)	108 (51.5)	50 (49.1)		
<b>1</b>	46 (14.5)	30 (16.5)	16 (13.6)		
<b>2</b>	16 (5.0)	11 (5.2)	5 (5.0)		
<b>3</b>	6 (1.9)	4 (2.1)	2 (1.8)		
<b>4</b>	1 (0.3)	0 (1.0)	1 (0.0)		
<b>5</b>	1 (0.3)	1 (0.0)	0 (0.5)		
Patient sedated (%)	64 (20.2)	43 (21.6)	21 (19.5)	112 (35.3)	0.631
Glasgow coma scale (mean/SD) on admission	13.17 (3.2)	13.17 (2.9)	13.14 (3.3)	99.00 (31.2)	0.946
WFNS on admission (%)				79 (24.9)	0.391
<b>1</b>	118 (37.2)	86 (33.0)	32 (39.1)		

Supplemental Table IV (continued)

	<b>2</b>	55 (17.4)	33 (22.7)	22 (15.0)		
	<b>3</b>	9 (2.8)	5 (4.1)	4 (2.3)		
	<b>4</b>	32 (10.1)	23 (9.3)	9 (10.5)		
	<b>5</b>	24 (7.6)	15 (9.3)	9 (6.8)		
<b>Hunt and Hess score (%)</b>					<b>7</b> (2.2)	0.238
	<b>1</b>	55 (17.4)	37 (18.6)	18 (16.8)		
	<b>2</b>	96 (30.3)	71 (25.8)	25 (32.3)		
	<b>3</b>	56 (17.7)	33 (23.7)	23 (15.0)		
	<b>4</b>	24 (7.6)	15 (9.3)	9 (6.8)		
	<b>5</b>	79 (24.9)	59 (20.6)	20 (26.8)		
<b>Fisher score (%)</b>					<b>1</b> (0.3)	0.228
	<b>1</b>	11 (3.5)	10 (1.0)	1 (4.5)		
	<b>2</b>	18 (5.7)	14 (4.1)	4 (6.4)		
	<b>3</b>	44 (13.9)	27 (17.5)	17 (12.3)		
	<b>4</b>	243 (76.7)	168 (77.3)	75 (76.4)		
<b>Modified Fisher score (%)</b>					<b>1</b> (0.3)	0.317
	<b>0</b>	12 (3.8)	11 (1.0)	1 (5.0)		
	<b>1</b>	17 (5.4)	13 (4.1)	4 (5.9)		
	<b>2</b>	2 (0.6)	2 (0.0)	0 (0.9)		
	<b>3</b>	69 (21.8)	45 (24.7)	24 (20.5)		
	<b>4</b>	216 (68.1)	148 (70.1)	68 (67.3)		
<b>Presence of intraventricular hemorrhage (%)</b>		214 (67.5)	148 (68.0)	66 (67.3)	<b>88</b> (27.8)	0.782
<b>Presence of intraparenchymal hemorrhage (%)</b>		83 (26.2)	54 (29.9)	29 (24.5)	<b>172</b> (54.3)	0.434
<b>Presence of subdural hemorrhage (%)</b>		19 (6.0)	13 (6.2)	6 (5.9)	<b>209</b> (65.9)	0.979
<b>Total hemorrhage volume (mean/SD)</b>		37.22 (30.3)	34.60 (31.0)	43.22 (29.6)	<b>2.00</b> (0.6)	0.023
<b>Time from Ictus to admission (mean/SD)</b>		30.93 (107.0)	31.60 (58.6)	29.39 (122.4)	<b>30.93</b> (107.0)	0.829

Supplemental Table IV (continued)

<b>Number of aneurysms (mean/SD)</b>	1.28 (0.7)	1.31 (0.6)	1.23 (0.7)	0.00 (0.0)	0.292
<b>Height of aneurysm (mean/SD)</b>	6.82 (5.4)	6.93 (3.8)	6.58 (6.0)	6.00 (1.9)	0.539
<b>Width of aneurysm (mean/SD)</b>	5.51 (4.3)	5.54 (3.4)	5.45 (4.6)	7.00 (2.2)	0.855
<b>Side of aneurysm (%)</b>				3 (0.9)	0.021
<b>Left</b>	156 (49.2)	99 (58.8)	57 (45.0)		
<b>Right</b>	131 (41.3)	94 (38.1)	37 (42.7)		
<b>Middle</b>	27 (8.5)	24 (3.1)	3 (10.9)		
<b>Shape of aneurysm (%)</b>				5 (1.6)	0.573
<b>Saccular</b>	284 (89.6)	195 (91.8)	89 (88.6)		
<b>Non-saccular (fusiform/ruptured)</b>	26 (8.2)	20 (6.2)	6 (9.1)		
<b>Other</b>	2 (0.6)	1 (1.0)	1 (0.5)		
<b>Aneurysm treatment (%)</b>				0 (0.0)	0.002
<b>No</b>	48 (15.1)	44 (4.1)	4 (20.0)		
<b>Coiling</b>	212 (66.9)	144 (70.1)	68 (65.5)		
<b>Clipping</b>	54 (17.0)	30 (24.7)	24 (13.6)		
<b>Coiling plus stent</b>	2 (0.6)	1 (1.0)	1 (0.5)		
<b>Flow diversion</b>	1 (0.3)	1 (0.0)	0 (0.5)		
<b>Rebleed number (%)</b>				91 (28.7)	0.628
<b>0</b>	178 (56.2)	120 (59.8)	58 (54.5)		
<b>1</b>	34 (10.7)	26 (8.2)	8 (11.8)		
<b>2</b>	12 (3.8)	7 (5.2)	5 (3.2)		
<b>3</b>	1 (0.3)	1 (0.0)	0 (0.5)		
<b>5</b>	1 (0.3)	1 (0.0)	0 (0.5)		
<b>Treatment for rebleed (%)</b>				0 (0.0)	0.998

Supplemental Table IV (continued)

<b>No</b>	254 (80.1)	176 (80.4)	78 (80.0)	
<b>Yes, based on both CT blood increase and clinical deterioration</b>	27 (8.5)	19 (8.2)	8 (8.6)	
<b>Yes, based on blood increase on CT scan</b>	6 (1.9)	4 (2.1)	2 (1.8)	
<b>Yes, based on clinical deterioration</b>	30 (9.5)	21 (9.3)	9 (9.5)	
<b>Location of aneurysm (%)</b>				0 (0.0)      0.044
<b>Posterior circulation</b>	66 (20.8)	53 (24.1)	13 (13.4)	
<b>Anterior circulation</b>	251 (79.2)	167 (75.9)	84 (86.6)	

### Supplemental Section I – Auto-encoder implementation

Stacked Convolutional Auto-encoder is a typical unsupervised feature learning algorithm that scales well to high-dimensional inputs and is robust to noise and variations. This method, learns features (characteristics) from the image by first encoding the input into a lower dimensional space using convolutional and pooling layers, and then reconstructs it using the inverse operations (deconvolution and unpooling)(1). The weights from this network are trained based on the difference between the input image and the reconstructed output image. The features learned by the auto-encoder are used to reconstruct the image. These features are usually an average representation of the images, which is unlikely to yield the discovery of a more useful representation than the input image. To solve this problem, we used the same approach from (2) which consists in applying noise to the input image and trying to reconstruct the normal scan using the SCAE. This approach is called Stacked Denoising Convolutional Auto-encoder (SDCAE) and will force the auto-encoder to extract more robust features.

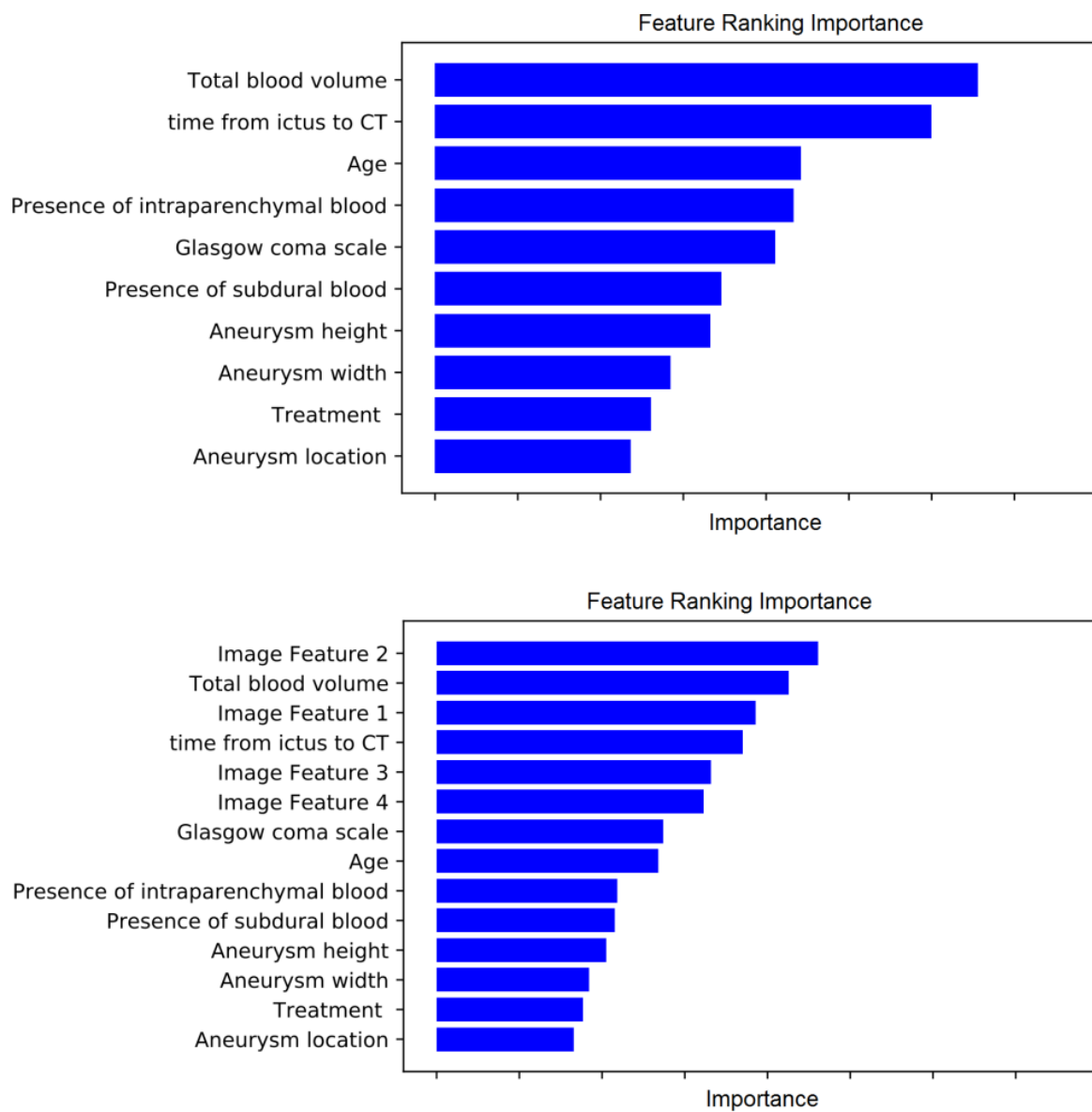
To speed up the training process (and allow the use of more samples per mini-batch) the images were downscaled by a factor of 4, resulting in scans of size 128x128x20. In order to account for variations in the data and increase the number of samples, we performed data augmentation using label-preserving transformations (translation, rotation and reflection), following the approach used in (3).

### Supplemental Section II - Lime explanation

Machine learning (ML) methods are often seen as black boxes, since explaining their predictions is usually not a trivial task. In order to build trust, it is important to

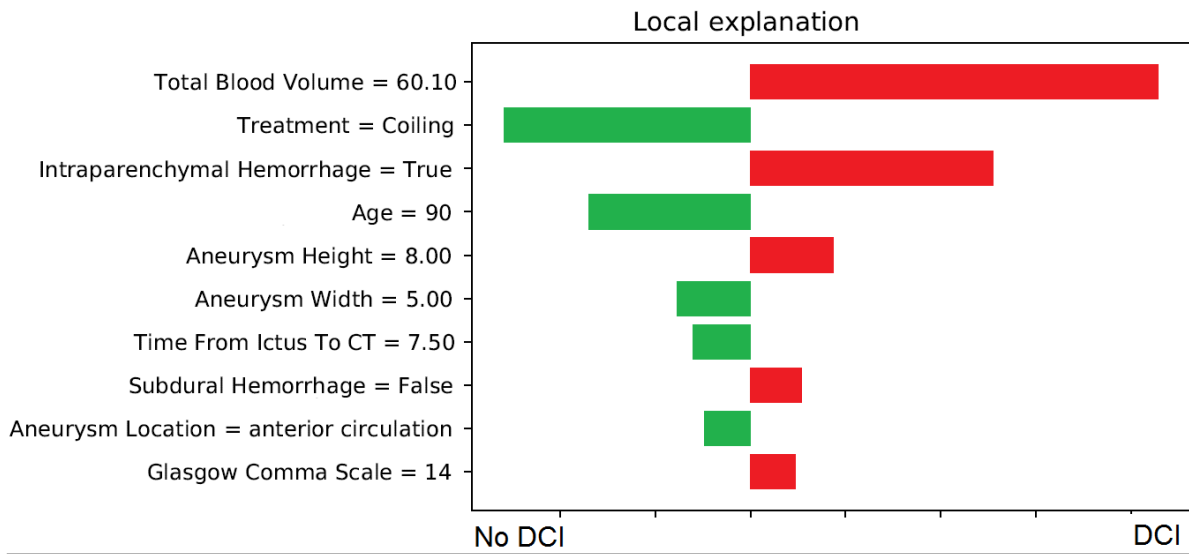
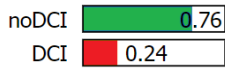
visualize which features influenced the model's prediction. LIME is a tool that can be used to locally explain the predictions of a given model. The explanation is based on visual representations that provide qualitative understanding of the model. While a model's decision boundary can be very complex globally, it can be easier to interpret the vicinity around a particular sample of this complex decision boundary. This particular sample is perturbed and a sparse linear model is built around it and used for explanation. In summary, LIME creates an explanation by approximating a "black box" model by a more interpretable one.

More examples of LIME explanations are shown in Supplemental Figures II and III. In Figure II (top), a patient that developed DCI received a low risk prediction when using only clinical features and a Random Forest model, even though some variables, such as total blood volume, strongly suggest a higher risk of DCI. After including image features (Figure II bottom), the risk for DCI increased (from 0.24 to 0.71), and most of the images features suggested a higher risk of DCI. In Figure III (top) a patient that **did not** develop DCI was assessed with LIME. First the risk for DCI and no DCI is similar (top). After including the image features (Figure III bottom), some image features strongly suggest a lower risk of DCI, which reduces the overall risk predicted by the ML model.

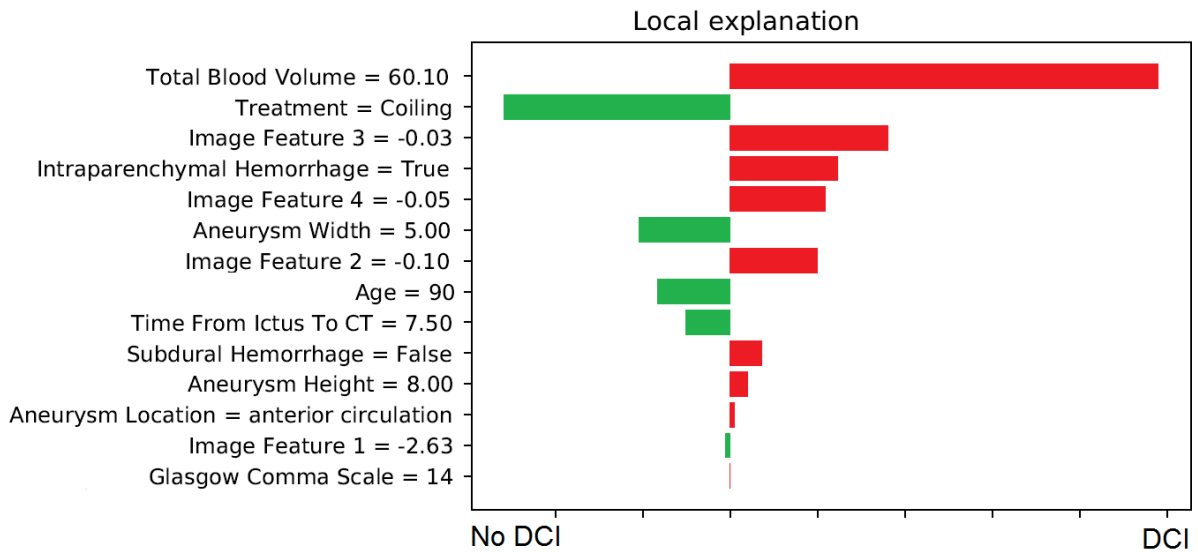
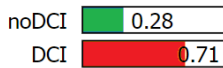


**Supplemental Figure I.** Feature Importance for RF classifier. Top using only the clinical data and bottom using a combination of clinical and image features.

Prediction probabilities

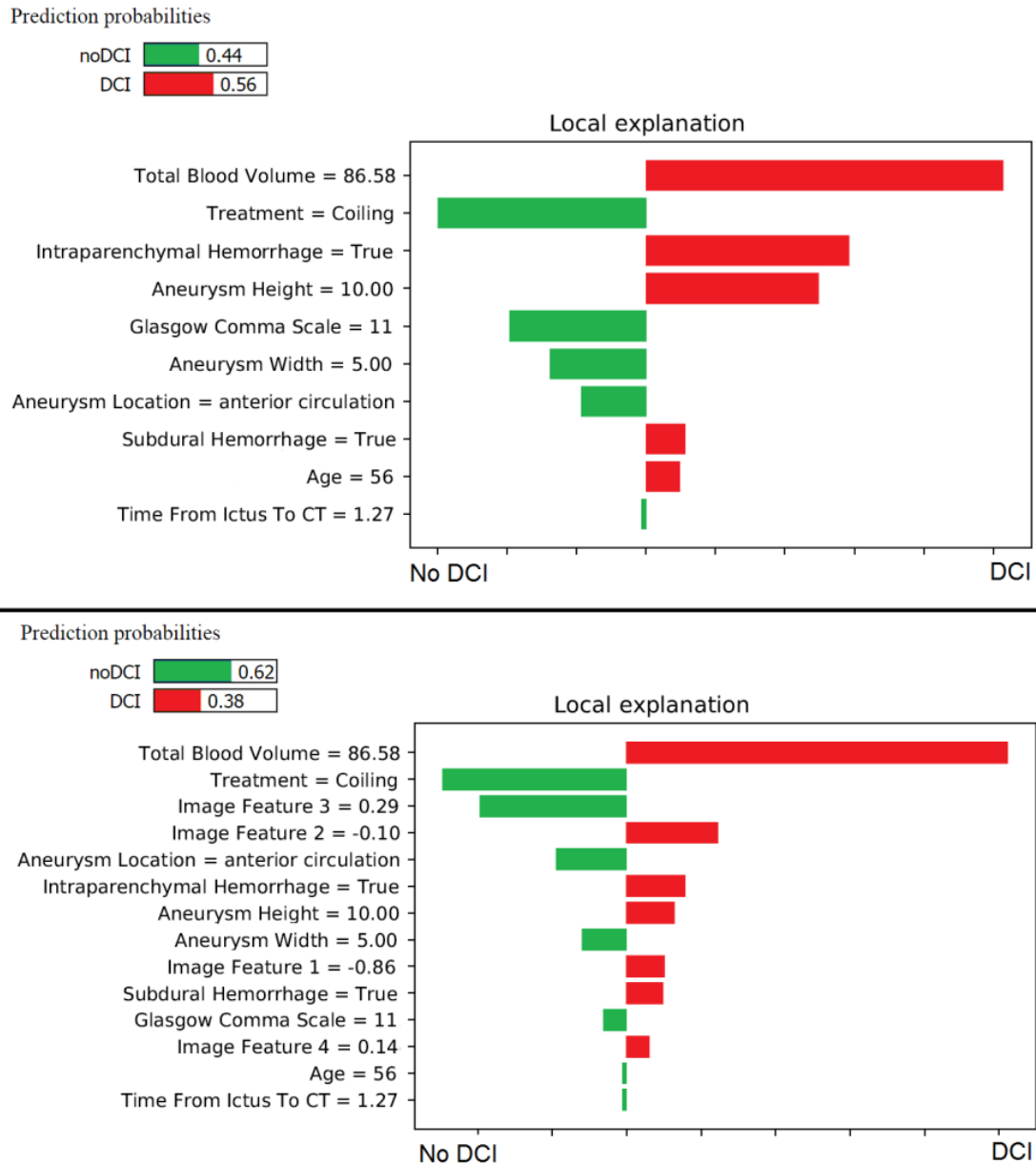


Prediction probabilities



**Supplemental Figure II.** LIME model explanation of a DCI positive patient. The model built using the clinical features suggest a lower risk of DCI (Top). After including the image features (bottom), the model suggests a higher risk for DCI.





**Supplemental Figure III.** LIME model explanation of a DCI Negative patient. The model built using the clinical features suggest a higher risk of DCI (Top). After including the image features (bottom), the model suggests a lower risk for DCI.

### Supplemental References

1. Masci J, Meier U, Cireşan D, Schmidhuber J. Stacked convolutional auto-encoders for hierarchical feature extraction. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2011;6791 LNCS:52-59.
2. Du B, Xiong W, Wu J, Zhang L, Zhang L, Tao D. Stacked Convolutional Denoising Auto-Encoders for Feature Representation. IEEE Trans Cybern. 2017;47:1017-1027.
3. Miki Y, Muramatsu C, Hayashi T, et al. Classification of teeth in cone-beam CT using deep convolutional neural network. Comput Biol Med. 2017;80:24-29.



3

# CHAPTER 3.

## Predicting outcome of endovascular treatment for acute ischemic stroke: Potential value of machine learning algorithms

van Os HJA, Ramos LA, Hilbert A, van Leeuwen M, van Walderveen MAA, Kruyt ND, Dippel DWJ, Steyerberg EW, van der Schaaf IC, Lingsma HF, Schonewille WJ, Majoie CBLM, Olabarriaga SD, Zwinderman KH, Venema E, Marquering HA, Wermer MJH; MR CLEAN Registry Investigators. Predicting outcome of endovascular treatment for acute ischemic stroke: Potential value of machine learning algorithms. *Front Neurol.* 2018 Sep 25;9:784.

DOI: 10.3389/fneur.2018.00784



## Abstract

*Background:* Endovascular treatment (EVT) is effective for stroke patients with a large vessel occlusion (LVO) of the anterior circulation. To further improve personalized stroke care, it is essential to accurately predict outcome after EVT. Machine learning might outperform classical prediction methods as it is capable of addressing complex interactions and non-linear relations between variables.

*Methods:* We included patients from the Multicenter Randomized Clinical Trial of Endovascular Treatment for Acute Ischemic Stroke in the Netherlands (MR CLEAN) Registry, an observational cohort of LVO patients treated with EVT. We applied the following machine learning algorithms: Random Forests, Support Vector Machine, Neural Network, and Super Learner and compared their predictive value with classic logistic regression models using various variable selection methodologies. Outcome variables were good reperfusion (post-mTICI  $\geq 2$ b) and functional independence (modified Rankin Scale  $\leq 2$ ) at 3 months using 1) only baseline variables and 2) baseline and treatment variables. Area under the ROC-curves (AUC) and difference of mean AUC between the models were assessed.

*Results:* We included 1383 EVT patients, with good reperfusion in 531 (38%) and functional independence in 525 (38%) patients. Machine learning and logistic regression models all performed poorly in predicting good reperfusion (range mean AUC:0.53-0.57), and moderately in predicting 3-month functional independence (range mean AUC:0.77-0.79) using only baseline variables. All models performed well in predicting 3-month functional independence using both baseline and treatment variables (range mean AUC:0.88-0.91) with a negligible difference of mean AUC (0.01;95%CI:0.00-0.01) between best performing machine learning algorithm (Random Forests) and best performing logistic regression model (based on prior knowledge).

*Conclusion:* In patients with LVO machine learning algorithms did not outperform logistic regression models in predicting reperfusion and 3-month functional independence after endovascular treatment. For all models at time

of admission radiological outcome was more difficult to predict than clinical outcome.

## Introduction

Endovascular treatment (EVT) is effective for ischemic stroke patients with a large vessel occlusion (LVO) of the anterior circulation. EVT results in a number needed to treat of 2.6 to reduce disability by at least one level on the modified Rankin Scale (mRS).<sup>1</sup> A recent meta-analysis showed a positive treatment effect of EVT across patient subgroups including different age groups, varying stroke severity, sex, and stroke localization.<sup>1</sup> However, many clinical and imaging predictors or their combinations were not considered in the subgroup analysis. Moreover, the RCTs that provided the data differed in their patient selection criteria. To further improve personalized stroke care, it is essential to accurately predict outcome and eventually differentiate between patients who will and will not benefit from EVT.

Machine learning belongs to the domain of artificial intelligence and provides a promising tool in pursuing personalized outcome prediction, which is increasingly used in medicine.<sup>2-7</sup> The machine learning methodology allows discovering empirical patterns in data through automated algorithms. In some clinical settings machine learning algorithms outperform classical regression models such as logistic regression, possibly through more efficient processing of non-linear relationships and complex interactions between variables,<sup>6, 8</sup> although poorer performance has also been observed.<sup>9</sup>

In this study, we used multiple machine learning algorithms and logistic regression with multiple variable selection methods to predict radiological and clinical outcome after EVT in a cohort of well-characterized stroke patients. We hypothesized that machine learning algorithms outperform classic multivariable logistic regression models in terms of discrimination between good and poor radiological and clinical outcome.

## Methods

### Patients

We included patients registered between March 2014 and June 2016 in the Multicenter Randomized Clinical Trial of Endovascular Treatment for Acute

Ischemic Stroke in the Netherlands (MR CLEAN) Registry. The MR CLEAN Registry is an ongoing, national, prospective, open, multicenter, observational monitoring study covering all 18 stroke intervention centers that perform EVT in the Netherlands, of which 16 participated in the MR CLEAN trial.<sup>10</sup> The registry is a continuation of the MR CLEAN trial collaboration and includes all patients undergoing EVT (defined as entry into the angiography suite and receiving arterial puncture) for acute ischemic stroke in the anterior and posterior circulation. In the current analysis we included those patients who adhered to the following criteria: age 18 years and older, treatment in a center that participated in the MR CLEAN trial, and LVO in the anterior circulation (internal carotid artery (ICA), internal carotid artery terminus (ICA-T), middle (M1/M2) cerebral artery, or anterior (A1/A2) cerebral artery), shown by CT angiography (CTA) or digital subtraction angiography (DSA).<sup>11</sup>

### Clinical baseline characteristics

We assessed the following clinical characteristics at admission: National Institutes of Health Stroke Scale (NIHSS), Glasgow Coma Scale, medical history (TIA, ischemic stroke, intracranial hemorrhage, subarachnoid hemorrhage, myocardial infarction, peripheral artery disease, diabetes mellitus, hypertension, hypercholesterolemia), smoking, laboratory tests (blood glucose, INR, creatinine, thrombocyte count, CRP), blood pressure, medication (thrombocyte aggregation inhibitors, oral anticoagulant drugs, anti-hypertensive drugs, statins), modified Rankin Score (mRS) before stroke onset, administration of intravenous tPA (yes/no), stroke onset to groin time, transfer from another hospital, and whether the patient was admitted during weekend or off hours.

### Radiological baseline parameters

All imaging in the MR CLEAN Registry was assessed by an imaging core laboratory.<sup>11</sup> On non-contrast CT, the size of initial lesion in the anterior circulation was assessed by the Alberta Stroke Program Early CT Score (ASPECTS). ASPECTS is a 10 point quantitative topographic score representing early ischemic change in the middle cerebral artery territory, with a scan without ischemic changes receiving an ASPECTS of 10 points.<sup>12</sup> In addition, presence of leukoaraiosis and old infarctions, hyperdense vessel

sign, and hemorrhagic transformation of the ischemic lesion were assessed on non-contrast CT.

On CTA, the core lab determined clot burden score, clot location, collaterals, and presence of intracranial atherosclerosis. The clot burden score evaluates the extent of thrombus in the anterior circulation by location scored on a 0–10 scale. A score of 10 is normal, implying clot absence; a score of 0 implies complete multi-segment vessel occlusion.<sup>12</sup> Presence of intracranial carotid artery stenosis, atherosclerotic occlusion, floating thrombus, pseudo-occlusion, and carotid dissection were scored on CTA of the carotid arteries. Collaterals were assessed using a 4 point scale, with 0 for absent collaterals (0% filling of the vascular territory downstream of the occlusion), 1 for poor collaterals (>0% and ≤50% filling of the vascular territory downstream of the occlusion), 2 for moderate collaterals (>50% and <100% filling of the vascular territory downstream of the occlusion), and 3 for excellent collaterals (100% filling of the vascular territory downstream of the occlusion).<sup>13</sup>

### Treatment specific variables

Variables collected during EVT were type of sedation during the procedure (general or conscious), use of a balloon guiding catheter, carotid stent placement, performed procedure (DSA only or thrombectomy), and type of EVT-device (stent retriever, aspiration device, or a combination of both). In addition, data were collected on adverse events during the procedure (perforation, dissection, distal thrombosis on DSA).

Interventional DSA parameters in our dataset were occluded vessel segment (ICA: origin, cervical, petrous, cavernous, supraclinoid, M1-M4, A1, A2), arterial occlusive lesion (AOL) recanalization score before and after EVT,<sup>14</sup> evidence of vascular injury (i.e. perforation, or dissection, vasospasm, new clot in different vascular territory or distal thrombus confirmed with imaging), and modified Thrombolysis in Cerebral Infarction (mTICI)-score before and after EVT. The mTICI-score grades the following categories of cerebral reperfusion: no reperfusion of the distal vascular territory (0), minimal flow past the occlusion but no reperfusion (1), minor partial reperfusion (2a), major partial reperfusion (2b), and complete reperfusion (3).<sup>15</sup> Further variables analyzed were time from stroke onset to recanalization, post-EVT stay on intensive care, high care or stroke care,

NIHSS after EVT (<48h), delta NIHSS (pre-treatment NIHSS subtracted from NIHSS <48h after EVT) and hemicraniectomy or symptomatic intracranial hemorrhage <48h after EVT.

### Outcome

The primary radiological outcome was good reperfusion defined as modified TICI-score directly post-procedure (post-mTICI)  $\geq 2$ .<sup>15</sup> The primary clinical outcome was functional independence at 3 months after stroke (mRS  $\leq 2$ ). We excluded patients in whom any of the main outcomes (3-month mRS and post-mTICI) were missing.

To investigate the full potential of Machine learning compared with conventional methods in different settings after stroke we defined two prediction settings:

First, we assessed the probability of good reperfusion and good 3-month functional independence in our cohort of patients that underwent EVT based only on variables that were available on admission before entry into the angiography suite. With this baseline prediction setting we are able to investigate the added value of machine learning for models that could potentially support future clinical decision making regarding the performance of EVT yes or no.

Second, we tested the models for predicting 3-month functional independence in patients after EVT was performed. For this analysis we used all variables collected up to 48 hours after the end of the endovascular procedure (baseline and treatment variables).

### Machine learning algorithms

The machine learning algorithms used in our study were Random Forests, Artificial Neural Network and Support Vector Machine, because they are among the algorithms that are currently most widely and successfully used for clinical data.<sup>2-7</sup> Each one of them represents a different algorithm ‘family’, each with radically different internal algorithm structures.<sup>16</sup> Since it was not known beforehand which kind of algorithm would perform best, we chose algorithms with different internal structures to increase the probability of good discriminative performance. We also used Super Learner, which is an ensemble method that finds the optimal weighted combination of predictions



of the Random Forests, Artificial Neural Network and Support Vector Machine algorithms used in this study. Ensemble methods such as Super Learner have been shown to increase predictive performance by increasing model flexibility.<sup>17</sup> For the implementation of all machine learning algorithms we used off-the-shelf methods in the Python module Scikit-Learn.<sup>18</sup>

### Super Learner

Super Learner is a stacking algorithm using cross-validated predictions of other models (i.e. a machine learning algorithm and logistic regression) and assigning weights to these predictions to optimize the final prediction. Super Learner's predictive performance has been found to surpass individual machine learning models in various clinical studies.<sup>17, 19, 20</sup>

### Random Forests

Random Forests consists of a collection of decision tree classifiers that are fit on random subsamples of patients and variables in the dataset. The variation of the subsampled variables creates a robust classifier. In the decision trees, each node represents a variable and splits the input data into branches based on an objective function that determines the optimal threshold for separating the outcome classes. The predictions from each tree are used as 'votes', and the outcome with the most votes is considered the predicted outcome for that specific patient.<sup>6, 21</sup> From the Random Forests algorithm variable importances can be derived, which are the sum of weights of nodes of the trees containing a certain variable, averaged over all trees in the forest.<sup>22</sup>

### Support Vector Machine

Support Vector Machine (SVM) is a kernel-based supervised machine learning classifier which can also be used to output probabilities. The SVM works by first mapping the input data into a high dimensional variable space. For binary classification, a hyperplane is subsequently determined to separate two classes such that the distance between the hyperplane and the closest data points is maximized.<sup>23</sup>

### Artificial Neural Network

In this study we use the multilayer perceptron, a popular class of artificial neural network architecture composed of one or more interconnected layers of neurons that process data from the input layer into predictions for the

output layer. The algorithm computes a weight for each neuron based on input activation. These weights are updated by backpropagation and stochastic gradient descent.<sup>24, 25</sup>

### Logistic regression

For logistic regression, generally a set of variables has to be selected to be included in the model. Since model performance can rely heavily on selecting the right variables, we tested five different variable selection methods prior to logistic regression. We first selected variables based on prior knowledge, a still widely used method that could be considered ‘classical’.<sup>26</sup> We selected 13 variables available at baseline that were included in a previous study for a similar purpose.<sup>27</sup> (Supplemental Table Ia) In addition, from baseline and treatment variables we selected 15 variables based on expert opinions of vascular neurologists and radiologists. (Supplemental Table Ib).

We further considered four automated variable selection methods: i) backward elimination, which is also considered to be a more classical approach,<sup>26</sup> and three state-of-the-art variable selection methods: ii) least absolute shrinkage and selection operator (LASSO)<sup>28</sup>, iii) Elastic Net, which is a modification of the LASSO found to outperform the former while still having the advantage of a similar sparsity of representation<sup>29</sup>, and iv) selection based on Random Forests variable importance.

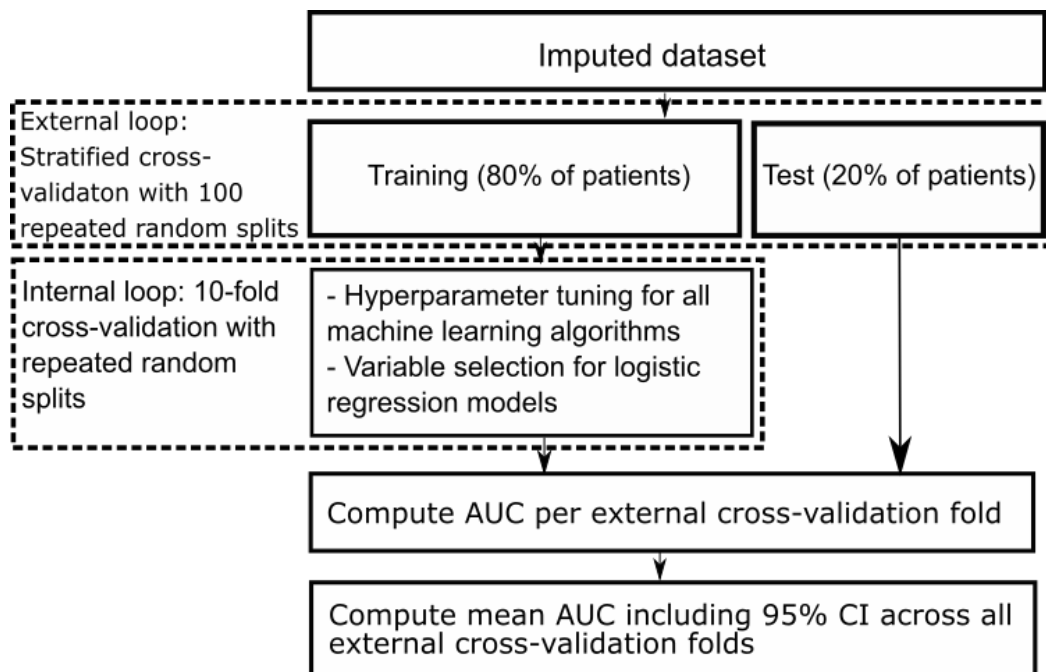
### Analysis pipeline

We imputed missing values using multiple imputations by chained equations (MICE).<sup>30</sup> Variables with 25% missing values or more were discarded from further analysis. All remaining variables used in this study are listed in Supplemental Table II and III. In total, 53 baseline variables and 30 treatment variables were used as input for machine learning algorithms and automated variable selection methods for logistic regression.

The ordinal clinical (NIHSS) and radiological (clot burden and ASPECTS) scores were presented as continuous scores in all models to increase model efficiency, and we assumed linear trends underlying the ordinal scores.

We used nested cross-validation (CV), consisting of an outer and an inner CV loop. In the outer CV loop we used stratified CV with 100 repeated random splits resulting in a training set including 80% and a test set including 20% of

all patients. Each training set was used as input for the inner CV loop, consisting of ten-fold CV.<sup>31, 32</sup> In the inner CV loop we selected variables for the logistic regression models using the different variable selection methods, and optimized hyperparameters of all machine learning models. Hyperparameters are tuning parameters specific to each machine learning algorithm whose values have to be preset and cannot be directly learned from the data. We optimized hyperparameters with the random grid search module from Scikit-Learn.<sup>18</sup> We selected those with highest area under the receiver operating characteristic (AUC) across all internal CV folds to find the best set of selected variables and hyperparameters. Figure 3.1 shows a schematic representation of our nested CV methodology.



**Figure 3.1.** Schematic representation of nested cross-validation methodology.

For all Random Forests models of both prediction settings we ranked variables by decreasing variable importance. For each variable we assessed the frequency of being among the 15 most important variables in a Random Forests model for each of the 100 external CV folds (Table 3.3).

### Model performance

We assessed model discrimination (the ability to differentiate between patients with good and poor outcome) with receiver operating characteristic (ROC) analyses. Because of our outer CV loop with 100 repeated random splits, we obtained 100 different AUCs from every model. We computed the

average ROC-curve and mean AUC with 95% confidence intervals (CI) for all models. We evaluated differences between mean AUCs of the best performing machine learning model and best performing logistic regression model by computing the difference of means including the associated 95% CI.

## Results

Of the 1627 patients registered between March 2014 and June 2016, we excluded 244 patients for this analysis because of age < 18 (n = 2), posterior circulation stroke (n = 79), missing MR CLEAN trial center (n = 20), and missing mRS or post-mTICI (n = 143). Mean age was 69.8 years (SD ± 14.4) and 738 (54%) of the 1383 included patients were men. In total, 531 (38%) patients had good reperfusion after EVT and 525 (38%) were functionally independent (mRS ≤ 2) three months after stroke. Baseline characteristics are shown in Table 3.1.

**Table 3.1.** Baseline characteristics of participants

<b>Characteristics</b>	<b>All patients (n = 1383)</b>
Mean age $\pm$ SD ( <i>years</i> )	69.8 $\pm$ 14.4
Men, <i>n</i> (%)	738 (53.5)
NIHSS score, <i>median</i> ( <i>IQR</i> )	16 (11 - 20)
Mean systolic blood pressure $\pm$ SD ( <i>mm Hg</i> )	150 $\pm$ 25
Medical history, <i>n</i> (%)	
Atrial fibrillation	411 (30.7)
Hypertension	697 (51.1)
Diabetes mellitus	235 (17.1)
Myocardial infarction	216 (15.9)
Peripheral artery disease	127 (9.4)
Ischaemic stroke	227 (16.5)
Hypercholesterolemia	411 (29.7)
Pre-stroke mRS > 2, <i>n</i> (%)	158 (11.6)
Smoking, <i>n</i> (%)	314 (22.9)
Medication use, <i>n</i> (%)	
DOAC**	35 (2.6)
Coumarine	179 (13.0)
Antiplatelet	461 (33.7)
Heparin	52 (3.8)
Blood pressure medication	707 (52.1)
Statin	490 (36.2)
Intravenous alteplase treatment, <i>n</i> (%)	1054 (76.2)
ASPECTS, <i>median</i> ( <i>IQR</i> )	9 (7 - 10)
Time from stroke onset to groin in minutes, <i>median</i> ( <i>IQR</i> )	210 (160 - 270)
Collateral score $\geq$ 2	764 (55)

\*National Institutes of Health Stroke Scale score

\*\*Direct Oral Anticoagulant drugs

## Prediction of good reperfusion after EVT in patients at time of admission

Discrimination between good and poor reperfusion of the best machine learning algorithm (Support Vector Machine, mean AUC: 0.55) and the best logistic regression model (using backward elimination, mean AUC: 0.57) was similar (difference of mean AUCs: 0.02; 95% CI: 0.01 – 0.03).

Discrimination was poor for all models, with a mean AUCs ranging from 0.53 to 0.57 (Table 3.2). Variable selection using LASSO or Elastic Net was not possible likely because the signal-to-noise ratio was insufficient.<sup>18</sup>

**Table 3.2.** Discrimination of machine learning algorithms and logistic regression models across the various prediction settings

Models, AUC (95% CI)*	Prediction setting (used variables: predicted outcome)		
	Baseline: post-mTICI	Baseline: mRS	All variables: mRS
Super Learner	0.55 (0.54 - 0.56)	0.79 (0.79 - 0.80)	0.90 (0.90 - 0.91)
Random Forests	0.55 (0.55 - 0.56)	0.79 (0.79 - 0.79)	0.91 (0.90 - 0.91)
Support Vector Machine	0.53 (0.53 - 0.54)	0.78 (0.77 - 0.78)	0.88 (0.88 - 0.89)
Neural Network	0.53 (0.53 - 0.54)	0.77 (0.76 - 0.77)	0.88 (0.88 - 0.89)
LR: automated selection**			
Random Forests	0.55 (0.55 - 0.56)	0.78 (0.78 - 0.78)	0.90 (0.90 - 0.90)
LASSO	NA <sup>‡</sup>	0.78 (0.78 - 0.79)	0.90 (0.89 - 0.90)
Elastic Net	NA <sup>‡</sup>	0.77 (0.77 - 0.78)	0.89 (0.88 - 0.89)
Backward elimination	0.57 (0.57 - 0.58)	0.78 (0.77 - 0.78)	0.90 (0.89 - 0.90)
LR: Prior knowledge <sup>†</sup>	0.55 (0.55 - 0.58)	0.78 (0.78 - 0.79)	0.90 (0.90 - 0.90)

\*Model discrimination is assessed by calculating mean Area Under the Curve (AUC) of the receiver operating characteristic across all outer cross-validation folds.

\*\*Logistic regression using automated variable selection methods.

<sup>‡</sup>Variable selection not possible, likely due to insufficient signal-to-noise ratio.

<sup>†</sup>Logistic regression using variables based on prior knowledge.

## Prediction of 3-month functional independence in patients at time of admission

Discrimination of good functional outcome of the best machine learning algorithm (Super Learner, mean AUC: 0.79) and the best logistic regression

model (using LASSO, mean AUC: 0.78) was similar (difference of mean AUCs: 0.01; 95% CI: 0.01 – 0.01).

Discrimination was moderate for all models, with a mean AUCs ranging from 0.77 to 0.79.

### Prediction of 3-month functional independence in patients after performance of EVT

Discrimination of good functional outcome of the best machine learning algorithm (Random Forests, mean AUC: 0.91) and the best logistic regression model (using prior knowledge, mean AUC: 0.90) was similar (difference of mean AUCs: 0.01; 95% CI: 0.00 – 0.01).

Discrimination was good for all models, with mean AUCs ranging from 0.88 to 0.91.

We performed a post hoc analysis in patients with good reperfusion as defined by post-mTICI  $\geq 2$ , predicting 3-month functional outcome both at time of admission and after performance of EVT. We did not find significant differences in performance between machine learning algorithms and logistic regression models in this patient subset (data not shown).

In Table 3.3 we show the top 15 variables based on the frequency of being among the 15 most important variables in a Random Forests model for each of the 100 external CV folds.

**Table 3.3.** Variable importance of Random Forests for various prediction settings (used variables: predicted outcome)

<b>Baseline: post-mTICI</b>	<b>Freq*</b>	<b>Baseline: mRS</b>	<b>Freq</b>	<b>All variables: mRS</b>	<b>Freq</b>
RR systolic at admission	100	Age	100	NIHSS after 24-48 hours	100
Duration stroke onset to groin	100	NIHSS at baseline	100	Delta NIHSS: follow-up minus baseline	100
RR diastolic at admission	100	Duration stroke onset to groin	100	Age	100
Thrombocyte count	100	Glasgow Coma Scale	100	NIHSS at baseline	100
Age	100	RR systolic at admission	100	Duration from onset to recanalization	100
Creatinine	100	CRP	100	Duration of procedure	100
CRP	100	Creatinine	100	Delta NIHSS $\geq$ 4 points higher after EVT	100
NIHSS at baseline	100	Thrombocyte count	100	Duration stroke onset to groin	100
Clot burden score	100	RR diastolic at admission	100	Glasgow Coma Scale	100
Glasgow ComaScale	100	mRS prior to stroke	100	Creatinine	100
ASPECTS score at baseline	100	ASPECTS score at baseline	100	CRP	100
Glucose	100	Glucose	100	Thrombocyte count	100
Location: proximal M1**	74	Clot burden score	99	RR systolic at admission	100
Hyperdense artery sign on NCCT	50	Presence of leukoaraiosis	96	mRS prior to stroke	91
History of atrial fibrillation	32	Collateral score	77	RR diastolic at admission	93

NCCT = non-contrast CT; CRP = C-Reactive Protein; RR = blood pressure; NIHSS = National Institutes of Health Stroke Scale score.

\*Frequency of being among the 15 most important variables in a Random Forests model for each of the 100 external CV folds. \*\*Location of intracranial occlusion on CTA.

## Discussion

We found no difference in performance between best performing machine learning algorithms and best performing logistic regression models in predicting radiological or clinical outcome in stroke patients treated with EVT. For prediction of good reperfusion using variables available at baseline, all models showed a poor discriminative performance. This could indicate that reperfusion after EVT depends on characteristics not present in our variables available at time of admission, such as vascular anatomy or interventionalist related factors. Prediction of 3-month functional independence using variables known at baseline was moderate, predicting 3-month functional independence using baseline and treatment variables resulted in good performance.



We hypothesized that machine learning would outperform logistic regression models due to simultaneous assessment of a large number of variables, and more efficient processing of non-linear relations and interactions between them. Although a large number of variables (83 in total, see Supplemental Table II and III) were available for analysis in the MR CLEAN Registry database, performance of best machine learning algorithms and best logistic regression models were similar. This could indicate that interactions and non-linear relationships in our dataset were of limited importance.

To interpret our results, several methodological limitations have to be considered. First, due to their great flexibility machine learning algorithms are prone to overfitting, which results in optimistic prediction performance. To account for overfitting we used nested CV, which is considered to be an effective method for this aim.<sup>33</sup> Second, our outer CV loop resulted in 100 AUCs per model leading to relatively small confidence intervals of mean AUCs. Although this increases the probability of statistically significant differences between mean AUCs of various models, the clinical relevance of these mean AUC differences is difficult to interpret. Because in our study mean AUC differences between models are minimal, clinical relevance of these differences is also negligible. Third, we used data from a registry. Registries might be prone to selection bias. However, we expect that selection bias in our study was minimal because the MRCLEAN Registry in principle covers all patients treated with EVT in the Netherlands. In addition, in all centers patients were treated according to national guidelines, and registration of treatment was a prerequisite for reimbursement.<sup>11</sup>

Strong points of this study include the large sample size and standardized collection of patient data. Moreover, because of extensive hyperparameter tuning and state-of-the art variable selection methods, machine learning and logistic regression models were compared at their best performance. In several other studies that compared machine learning algorithms with only logistic regression methods using variables based on prior knowledge, machine learning outperformed logistic regression.<sup>6, 7, 34</sup> Variable selection based on prior knowledge has the major drawback that predictive patterns in the data may be missed, as variable selection is strictly based on the literature and expert opinion.<sup>26</sup> In our study however, logistic regression using variables

based on prior knowledge performed similarly to logistic regression using automated variable selection methods.

The distinction between machine learning and ‘classical’ regression methods is largely artificial. However, a clear distinction between various machine learning algorithms and logistic regression exists in terms of model transparency, which could be seen as the understanding of the mechanism by which the model works.<sup>35</sup> Logistic regression has the advantage of transparency at the level of individual variable coefficients, since from these coefficients odds ratios can be derived. However, variable importances derived from the Random Forests algorithm also offer insight in the importance of individual variables for prediction performance.<sup>22</sup> These variable importances take interaction between variables into account and have a similar interpretation for continuous and discrete variables, unlike odds ratios which constitute an effect per unit change of a predictor. Hence, Random Forests could be used as an efficient screening tool to pick up predictive patterns in the data that could potentially lead to further hypothesis-driven research. In Table 3.3 we show the top 15 variables from either the baseline or baseline and treatment variable set, based on Random Forests variable importance. The majority of variables in Table 3.3 do not overlap with the selection of variables based on prior knowledge, potentially providing researcher with additional information.

In this dataset we found no clinically relevant differences in prediction of reperfusion and 3-month functional independence across all models. However, since it is generally not known on beforehand which type of model will result in the best predictive performance in a new dataset, our methodology could be of importance in future studies. We present an analysis pipeline with both machine learning algorithms and logistic regression models including state-of-the-art variable selection methods. Assessing predictive performance of all models simultaneously enables the researcher to make the proper trade-off between predictive performance and model transparency. As our analysis pipeline is fully automated and input variables and outcome label can be altered at will, it is relatively easy to reuse in future studies. The Python code of our pipeline has been made publicly available in an online repository (<http://bit.ly/mrcleanml>).

## References

1. Goyal M, Menon BK, van Zwam WH, Dippel DW, Mitchell PJ, Demchuk AM, et al. Endovascular thrombectomy after large-vessel ischaemic stroke: A meta-analysis of individual patient data from five randomised trials. *Lancet*. 2016;387:1723-1731
2. Ferroni P, Zanzotto FM, Scarpato N, Riondino S, Nanni U, Roselli M, et al. Risk assessment for venous thromboembolism in chemotherapy-treated ambulatory cancer patients: A machine learning approach. *Med. Decis. Making*. 2016
3. Konerman MA, Zhang Y, Zhu J, Higgins PD, Lok AS, Waljee AK. Improvement of predictive models of risk of disease progression in chronic hepatitis c by incorporating longitudinal data. *Hepatology*. 2015;61:1832-1841
4. Lambin P, van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, et al. Predicting outcomes in radiation oncology--multifactorial decision support systems. *Nat. Rev. Clin. Oncol*. 2013;10:27-40
5. Mani S, Chen Y, Li X, Arlinghaus L, Chakravarthy AB, Abramson V, et al. Machine learning for predicting the response of breast cancer to neoadjuvant chemotherapy. *J. Am. Med. Inform. Assoc*. 2013;20:688-695
6. Singal AG, Mukherjee A, Elmunzer BJ, Higgins PD, Lok AS, Zhu J, et al. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. *Am. J. Gastroenterol*. 2013;108:1723-1730
7. Kop R, Hoogendoorn M, Teije AT, Buchner FL, Slottje P, Moons LM, et al. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. *Comput. Biol. Med*. 2016;76:30-38
8. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N. Engl. J. Med*. 2016;375:1216-1219
9. van der Ploeg T, Nieboer D, Steyerberg EW. Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. *J. Clin. Epidemiol*. 2016;78:83-89
10. Berkhemer OA, Fransen PS, Beumer D, van den Berg LA, Lingsma HF, Yoo AJ, et al. A randomized trial of intraarterial treatment for acute ischemic stroke. *N. Engl. J. Med*. 2015;372:11-20
11. Jansen IGH, Mulder M, Goldhoorn RB, investigators MCR. Endovascular treatment for acute ischaemic stroke in routine clinical practice: Prospective, observational cohort study (mr clean registry). *BMJ*. 2018;360:k949
12. Pexman JH, Barber PA, Hill MD, Sevick RJ, Demchuk AM, Hudon ME, et al. Use of the alberta stroke program early ct score (aspects) for assessing ct scans in patients with acute stroke. *AJNR Am. J. Neuroradiol*. 2001;22:1534-1542
13. Tan IY, Demchuk AM, Hopyan J, Zhang L, Gladstone D, Wong K, et al. Ct angiography clot burden score and collateral score: Correlation with clinical and radiologic outcomes in acute middle cerebral artery infarct. *AJNR Am. J. Neuroradiol*. 2009;30:525-531
14. Khatri P, Neff J, Broderick JP, Khoury JC, Carrozzella J, Tomsick T, et al. Revascularization end points in stroke interventional trials: Recanalization versus reperfusion in ims-i. *Stroke*. 2005;36:2400-2403

15. Zaidat OO, Yoo AJ, Khatri P, Tomsick TA, von Kummer R, Saver JL, et al. Recommendations on angiographic revascularization grading standards for acute ischemic stroke: A consensus statement. *Stroke*. 2013;44:2650-2663
16. Fernandez-Delgado M CE, Barro S, et al. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? 2014. *Journal of Machine Learning Research*. 15; 3133-3181.
17. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat. Appl. Genet. Mol. Biol.* 2007;6:Article25
18. Scikit-learn: Machine learning in python. <http://scikit-learn.org/stable/> (accessed October 17, 2017). 2017
19. Kreif N, Grieve R, Diaz I, Harrison D. Evaluation of the effect of a continuous treatment: A machine learning approach with an application to treatment for traumatic brain injury. *Health Econ.* 2015;24:1213-1228
20. Petersen ML, LeDell E, Schwab J, Sarovar V, Gross R, Reynolds N, et al. Super learner analysis of electronic adherence data improves viral prediction and may provide strategies for selective hiv rna monitoring. *J. Acquir. Immune Defic. Syndr.* 2015;69:109-118
21. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. *Artif. Intell. Med.* 2005;34:113-127
22. Breiman L. Random forests. *Machine Learning*. 2001;45: 5–32
23. B. SAJS. A tutorial on support vector regression. *Stat. Comput.* 2004;vol. 14, no. 3, pp. 199–222
24. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436-444
25. Bishop CM. Neural networks for pattern recognition. *J. Am. Stat. Assoc.* 1995;vol. 92, p. 482
26. Walter S, Tiemeier H. Variable selection: Current practice in epidemiological studies. *Eur. J. Epidemiol.* 2009;24:733-736
27. Venema E, Mulder M, Roozenbeek B, Broderick JP, Yeatts SD, Khatri P, et al. Selection of patients for intra-arterial treatment for acute ischaemic stroke: Development and validation of a clinical decision tool in two randomised trials. *BMJ*. 2017;357:j1710
28. Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Stat Soc B Met.* 1996;58:267-288
29. Zou H, Hastie, T. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* 2005;67, Part 62, pp. 301–320
30. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? *Int. J. Methods Psychiatr. Res.* 2011;20:40-49
31. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. U. S. A.* 2002;99:6562-6566
32. Borra S, Di Ciaccio A. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Comput Stat Data An.* 2010;54:2976-2989

33. Krstajic D BL, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*. 2014;6 (1):10. doi:10.1186/1758-2946-6-10.
34. Decruyenaere A, Decruyenaere P, Peeters P, Vermassen F, Dhaene T, Couckuyt I. Prediction of delayed graft function after kidney transplantation: Comparison between logistic regression and machine learning methods. *BMC Med. Inform. Decis. Mak.* 2015;15:83.
35. Lipton ZC. The mythos of model interpretability. *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, NY, USA. 2017.

## Supplemental material

**Supplemental Table I.** Variables selected for logistic regression based on prior knowledge

<b>a. Baseline variables (n = 13)</b>	<b>b. Baseline and treatment variables (n = 15)</b>
Age	Age
mRS prior to stroke	mRS prior to stroke
History of diabetes mellitus	History of diabetes mellitus
History of previous ischemic stroke	History of hypertension
History of atrial fibrillation	History of previous ischemic stroke
Systolic blood pressure	Systolic blood pressure
Intravenous thrombolysis	Intravenous thrombolysis
Collateral score on CTA	Collateral score on CTA
Location of intracranial occlusion on CTA	Time from onset stroke to groin
ASPECTS score on baseline	Duration of EVT procedure
NIHSS at baseline	Location of intracranial occlusion on DSA
Duration stroke onset stroke to groin	mTICI post EVT
Clot burden score on CTA	NIHSS post EVT (24-48 hours)
	General anesthesia during EVT
	Symptomatic intracerebral hemorrhage

mRS = modified Rankin Scale; CTA = CT angiography; DSA = Digital Substraction Angiography; mTICI = modified Thrombolysis in Cerebral Infarction score; AOL = Arterial Occlusive Lesion recanalization score; NIHSS = National Institutes of Health Stroke Scale score; DSA = Digital Substraction Angiography.

**Supplemental Table II.** Variables available at baseline

<b>Variables (n = 53)</b>	
Age	Non-contrast CT
Sex	Hyperdense artery sign
Medical history	Relevant (new) ischemia/ hypodensity
Stroke	Hemorrhagic transformation
Myocardial infarction	Leukoariosis
Peripheral artery disease	Old infarcts in same ASPECTS region
Diabetes mellitus	ASPECTS score
Hypertension	CT angiography
Atrial fibrillation	Intracranial atherosclerosis
Hypercholesterolemia	Vascular malformation/ aneurysm
mRS prior to stroke	Most proximal occlusion segment
Medication use	Collateral score
Antiplatelet use	Clot burden score
DOAC use	Symptomatic carotid bifurcation
Coumarine use	Stenosis
Heparin use	Atherosclerotic occlusion
Blood pressure medication	Floating thrombus
Statin use	Pseudo-occlusion
RR systolic	Carotid dissection
RR diastolic	NIHSS at baseline
Laboratory parameters	Admission on weekend
INR	Admission during off hours
Thrombocyte count	Transfer from other hospital
Creatinine	Intravenous thrombolysis
CRP	Glasgow Coma Scale
Glucose	Duration from onset to groin in minutes
Smoking	

NIHSS = National Institutes of Health Stroke Scale score; CRP: C-Reactive Protein; INR: International Normalized Ratio; mRS = modified Rankin Scale; DOAC = Direct Oral Anticoagulant drugs.

**Supplemental Table III.** Variables available during and after EVT

<b>Variables (n = 30)</b>	
Most proximal occlusion segment on DSA	Stent placement in ICA
Occlusion other territories	Balloon angioplasty
Used balloon guiding	Evidence of vascular injury on DSA
Complication during intervention	Duration from onset to recanalization
Performed procedure	Duration of procedure
First used EV treatment	General anesthesia
Attempts with MERCI as first choice	Conscious sedation
Administration of EVT medication	Reperfusion during EVT
Hemicraniectomy	IC stay
mTICI score pre EVT	High care stay
mTICI score post EVT	Stroke care stay
Occlusion side on DSA	Delta NIHSS: follow-up minus baseline
Pre-EVT AOL score	Delta NIHSS $\geq 4$ points higher after EVT
Post-EVT AOL score	Symptomatic intracranial hemorrhage
Total attempts	NIHSS after 24-48 hours

EVT = Endovascular Treatment; DSA = Digital Subtraction Angiography; mTICI = modified Thrombolysis in Cerebral Infarction score; AOL = Arterial Occlusive Lesion recanalization score; NIHSS = National Institutes of Health Stroke Scale score.







4

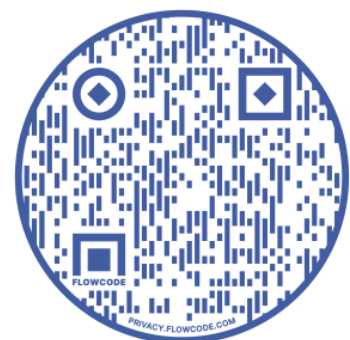
# CHAPTER 4.

## Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke

Hilbert A, Ramos LA\*, van Os HJA, Olabbarriaga SD, Tolhuisen ML, Wermer MJH, Sales Barros R, van der Schaaf I, Dippel D, Roos YBWEM, van Zwam WHm Yoo AJ, Emmer BJ, Lucklama à Nijeholt GJ, Zwinderman AH, Strijkers GJ, Majoie CBLM, Marquering HA. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. *Computers in Biology and Medicine* 2019;115.

\*Shared first author

DOI: 10.1016/j.combiomed.2019.103516



## Abstract

Treatment selection is becoming increasingly more important in acute ischemic stroke patient care. Clinical variables and radiological image biomarkers (old age, pre-stroke mRS, NIHSS, occlusion location, ASPECTS, among others) have an important role in treatment selection and prognosis. Radiological biomarkers require expert annotation and are subject to inter-observer variability. Recently, Deep Learning has been introduced to reproduce these radiological image biomarkers. Instead of reproducing these biomarkers, in this work, we investigated Deep Learning techniques for building models to directly predict good reperfusion after endovascular treatment (EVT) and good functional outcome using CT angiography images. These models do not require image annotation and are fast to compute. We compare the Deep Learning models to Machine Learning models using traditional radiological image biomarkers. We explored Residual Neural Network (ResNet) architectures, adapted them with Structured Receptive Fields (RFNN) and auto-encoders (AE) for network weight initialization. We further included model visualization techniques to provide insight into the network's decision-making process. We applied the methods on the MR CLEAN Registry dataset with 1301 patients. The Deep Learning models outperformed the models using traditional radiological image biomarkers in three out of four cross-validation folds for functional outcome (average AUC of 0.71) and for all folds for reperfusion (average AUC of 0.65). Model visualization showed that the arteries were relevant features for functional outcome prediction. The best results were obtained for the ResNet models with RFNN. Auto-encoder initialization often improved the results. We concluded that, in our dataset, automated image analysis with Deep Learning methods outperforms radiological image biomarkers for stroke outcome prediction and has the potential to improve treatment selection.

## Introduction

Stroke is ranked among the leading causes of death and permanent disability in the last 15 years worldwide [1], [2]. Approximately 80% of all stroke patients with untreated large vessel occlusion in the anterior circulation do not regain functional independence or die within 90 days after stroke onset [3]. The Multicenter Randomized Clinical Trial of Endovascular Treatment for Acute Ischemic Stroke in the Netherlands (MR CLEAN Registry) has shown that this patient population can be effectively treated with endovascular treatment (EVT) [4].

Accurate prediction of reperfusion and functional outcome has the potential to improve stroke care, as it could lead to selecting the most beneficial treatment option for the individual patient: to perform or to withhold EVT. Recent studies on outcome prediction strategies in ischemic stroke patients after EVT utilized clinical variables and radiological image biomarkers [5], [6]. In favor of standardized prognosis, various radiological stroke imaging biomarkers have been defined by specific, visually observable phenomena that imply stroke severity and functional outcome. These biomarkers include the extent of tissue damage characterized by edema (e.g. ASPECTS [7]) and extent of blood flow through the collateral circulation (e.g. Collateral Score [8]), and they have been proven to be associated with functional outcome. The number of proposed radiological image biomarkers for prognosis in acute ischemic stroke is quite large. In MR CLEAN Registry, for example, 20 biomarkers have been assessed. These biomarkers are commonly scored manually, may demand a considerable time effort and suffer from observer variability. For the collateral score, observer agreement as low as 50% with kappa's ranging from 0.49 to 0.60 has been reported [9], and for the ASPECTS score a mean deviation close to one point has been found, with above 25% of the cases deviating more than two points [10]. Details about the other image biomarkers are shown in Supplemental Table I.

Machine Learning (ML) enables the discovery of empirical patterns and linear/non-linear relationships in data through automated algorithms. Regarding imaging data, Deep Learning (DL) algorithms are particularly able to learn important predictive patterns, which may lead to increased prediction accuracy [11]. For example, an encoder-decoder CNN (inspired by the

SegNet [12]) was developed to predict final lesion volume and outcome in acute ischemic stroke patients using magnetic resonance imaging, with an AUC of 0.88 (10% higher than linear models) [13]. E-ASPECTS, a machine learning-based commercial software developed to automate ASPECTS scoring in CT scans, has recently been used to predict functional recovery and adverse outcome in acute ischemic stroke patients [14]. In the stroke lesion segmentation challenge (ISLES), a multi-scale 3D-CNN was the best performing model, with an average Dice score of 0.59 [15]. This multi-scale 3D-CNN (named DeepMedic [15]) was also successfully applied to CTA to detect acute ischemic stroke (and segment the lesion) with an AUC of 0.93 [16]. However, these specific DL approaches are generally limited by the manual determination of many of the biomarkers that are considered ground truth but suffer from high inter-observer variability. Besides, DL methods generally come at the cost of high complexity models with low interpretability [17], hampering the applicability in clinical settings.

Due to the recent success of DL approaches in stroke medical imaging, we hypothesize that data-efficient DL methods trained on CT Angiography (CTA) imaging data might outperform well-known radiological image biomarkers in predicting good reperfusion after EVT and good functional outcome in patients with acute ischemic stroke. Next to assessing the accuracy, we adopt visualization techniques, since these prediction systems can be of more assistance when they provide direct insights into their decision-making process beyond generating a probability distribution.

## Methods

### Clinical data and pre-processing

We included 1526 ischemic stroke patients registered between March 2014 and June 2016 in the MR CLEAN Registry part1 [18]. The MR CLEAN Registry is an ongoing, prospective, observational, multicenter study at 16 intervention hospitals in the Netherlands. Imaging data (CTA scans) available before EVT were used to develop the DL models and to determine radiological image biomarkers by expert radiologists.

The raw CTA scans were of size (512x512xS), where S was the number of axial slices. Due to limited computational resources, we opted to reduce

sparsity and dimensionality of images by computing Maximum Intensity Projections (MIPs) from the CTA data in the axial plane. First, CTA scans were co-registered to a reference scan (a scan with no abnormalities) using rigid registration with the Elastix software [19] and the skull was removed from the images with a region growing algorithm since it is of high intensity and can hamper the quality of the MIPs [20]. The attenuation of the MIPs was clipped between +50 and +400 Hounsfield Units (HU) and normalized to the interval of [0,1]. The surrounding air was removed to reduce image size. The final image data size used as input for the DL models was 368x432 pixels (voxel size of 0.52,0.52 mm). After the pre-processing steps, 225 patients were excluded due to failure during registration, poor image quality, and noise or artifacts, leaving a total of 1301 patients to be used for model development. Table 4.1 contains the baseline characteristics of the patients used in our models.

We created models to predict two outcome measures. First, good functional outcome after ischemic stroke - defined by the dichotomized modified Rankin Scale ( $mRS \leq 2$ ) at 90 days - mRS is a scale commonly used to assess the disability of stroke patients in daily activities. mRS ranges from 0 to 6, where zero means no disability, progressing to five (severe disability) and six (death).

Second, good reperfusion - defined by the dichotomized modified Thrombolysis In Cerebral Infarction score ( $mTICI \geq 2b$ ). mTICI is a score that ranges from 0 (no antegrade reperfusion of the occluded vascular territory) to 3 (complete reperfusion). mTICI was assessed by 20 neuroradiologists and one neurologist at an imaging core laboratory. The observers were blinded to all clinical findings, except occlusion location. mTICI was scored on digital subtraction angiography images [18].

**Table 4.1.** Characteristics of patients in MR CLEAN Registry. Values correspond to the percentages of participants unless stated otherwise

<b>Characteristics</b>	<b>All patients N=1301</b>	<b>mRS 0 – 2 N=463</b>	<b>mRS 3 – 6 N=838</b>	<b>mTICI 0- 2a N = 552</b>	<b>mTICI 2b-3 N = 749</b>
<b>Age (years) (median/IQR)</b>	71 (59 - 79)	66 (55 - 74)	74 (63 - 82)	72 (60 - 80)	70 (59 - 78)
<b>Men (%)</b>	695 (53.4)	262 (56.6)	433 (51.7)	290 (52.5)	405 (54.1)
<b>NIHSS at baseline (median/IQR)</b>	16 (11 - 20)	1 (9 - 17)	17 (13 - 21)	16 (12 - 20)	15 (11 - 19)
<b>Onset to groin puncture time (mins) (IQR)</b>	210 (160-270)	19 (145-253)	220 (170-279)	215 (157-281)	205 (160-265)
<b>Systolic blood pressure (mm Hg) (Mean/STD)</b>	150 (1.89)	146 (4.86)	152 (3.03)	152 (4.63)	148 (3.16)
<b>Intravenous alteplase treatment (%)</b>	1014 (77.9)	380 (82.1)	634 (75.7)	411 (74.5)	603 (80.5)
<b>ASPECTs at Baseline (subgroups)</b>					
<b>0-4</b>	81 (6.2)	16 (3.5)	6 (7.8)	36 (6.5)	45 (6.0)
<b>5-7</b>	310 (23.8)	95 (20.5)	215 (25.7)	131 (23.7)	179 (23.9)
<b>8-10</b>	880 (67.6)	340 (73.4)	540 (64.4)	370 (67.0)	510 (68.1)

### Structured Receptive Field Neural Networks (RFNNs)

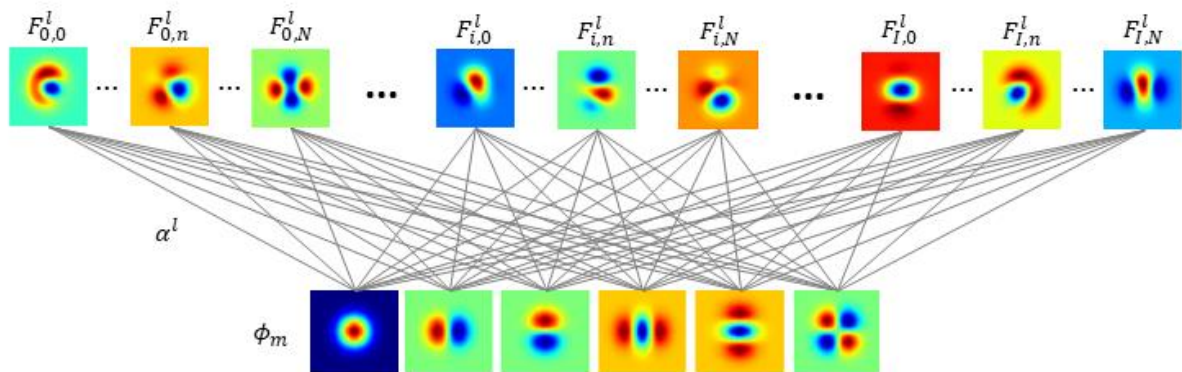
Conventional CNNs (Convolutional Neural Networks) hardly excel in the presence of relatively small datasets, which presents a common challenge for many medical applications. To this end, in this work we explore a data-efficient CNN formulation that builds on the structure of biological receptive fields.

Structured Receptive Field Neural Networks (RFNNs) were proposed in [21] and have been shown to outperform CNNs on small- and medium-sized



datasets. RFNNs redefine convolutional kernels as linear combinations of Gaussian derivative filters. Contrary to traditional kernels, only the combination weights are trained, whereas the set of Gaussian derivatives is fixed. In this way, the number of parameters to train is potentially decreased, and prior knowledge about the spatial properties of local features is introduced. Gaussian filters and the Scale-space theory in computer vision [22] have been broadly explored in the medical imaging domain. Scale-space approaches have been successfully applied to medical imaging classification and segmentation with great performance improvements since they often assist the classifiers by revealing low and high-level features without introducing artifacts [23]–[25]. Furthermore, the interpretability of CNNs is enhanced by explicitly connecting classical image processing methods with the data-driven paradigm.

Figure 4.1 illustrates the computation of  $I \times N$  kernels of a RFNN convolutional layer  $l$ , where  $I$  is the number of input feature maps and  $N$  is the number of output feature maps. This RFNN formulation can be used in any convolutional layer to replace the conventional convolutional kernels while keeping the architecture of a CNN intact.



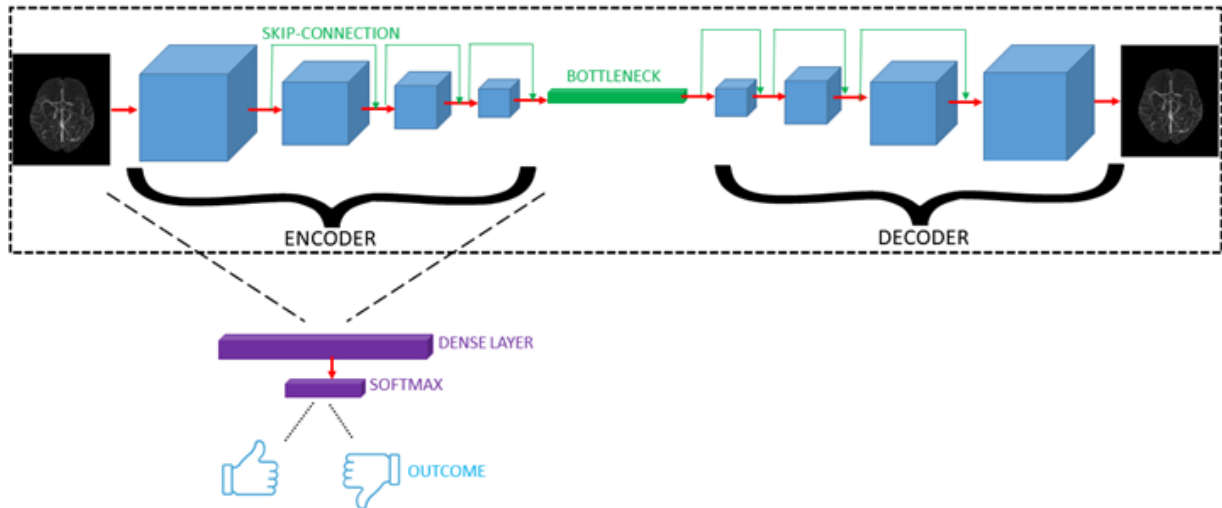
**Figure 4.1.** Construction of RFNN convolutional kernels.  $\Phi_m$  denotes the fixed set of Gaussian derivatives,  $\alpha^l$  the combination weights in the  $l^{\text{th}}$  convolutional layer, and  $F_{(i,n)}^l$  the convolutional kernel producing the  $n^{\text{th}}$  output feature map of  $l^{\text{th}}$  layer from  $i^{\text{th}}$  input feature-map.

### Unsupervised pre-training

A random initialization scheme can place the parameters of a CNN in regions that do not generalize well, while the limitations in training data and computational resources create a burden in improving generalization during

training (e.g. increasing batch size). These problems make the training of deep architectures unstable and can lead to lower model accuracy [26]. Moreover, supervised training of CNNs is influenced by the ground truth labels, even though learning effective image features does not necessarily rely on image annotation.

To face these challenges, we included unsupervised pre-training in the experiments using stacked denoising convolutional Auto-Encoders (AE) [26]. AEs learn a feature representation by compressing the input into a latent space and subsequently reconstruct the input using this representation. By optimizing the reconstruction from the input data, the AE is able to learn features that best represent the image. We constructed AEs from each CNN model by using their convolutional layers as the encoder part and extending with a corresponding sequence of transposed convolutional blocks as the decoder. Transposed convolutional blocks are comprised of the same Batch-Normalization, ReLU, convolutional sequence but convolutions are replaced by up-sampling transposed convolutions. Using the training data, we trained an AE until the loss between the output and the input images stopped decreasing (depicted by the dashed lines in Figure 4.2). The learned encoding part of the network was subsequently used to train a dense layer using the labels (in a transfer learning fashion) [27], [28]. The weights from the encoding part were used in two approaches: keeping them frozen during the training or fine-tuning them during the training of the dense layer [27].



**Figure 4.2.** Unsupervised and supervised training pipeline. First the AE is trained based on the reconstruction loss and then the trained encoder is used to train a dense layer for prediction.

## Baseline models

To assess the added value of DL methods compared to existing radiological image biomarkers, we created ML-based models using radiological image biomarkers that have shown state-of-the-art results on the same dataset [5]. In these baseline prediction models, we used 20 radiological imaging biomarkers, which have been manually scored by designated experts of the core-lab of the MR CLEAN Registry. These radiological image biomarkers have shown predictive value for functional outcome and thus are commonly considered in clinical practice [4], [29], [30]. Supplemental Table I lists all included radiological imaging biomarkers. We developed two clinical baseline prediction models for both outcome measures using only radiological image biomarkers. The first is a Logistic Regression (LR) model, following the most common approach in clinical research. The second is a Random Forest Classifier (RFC), which has earlier been successfully used for the same patient population [5].

To assess the benefits of the proposed application of RFNN convolutional layers and unsupervised pre-training, we compared them to a standard DL model. We developed and optimized a ResNet architecture, and used it as the standard DL model. The ResNet architecture was composed of four blocks with two consecutive convolutions in each block, and skip-connections connecting the input and the output of blocks [31]. Inspired by [28], we

followed the Batch-Normalization, ReLU, convolution sequence in each layer. Further details about the ResNet architecture can be found in Supplemental Tables II-III and Supplemental Figure II. Details of the ResNet-AE implementation for unsupervised pre-training are shown in Supplemental Table IV. RFNN models had the same ResNet (ResNet-AE in case of unsupervised pre-training) architecture, but the conventional convolutional kernels were replaced with structured receptive field kernels.

### Experimental Setup

The 1301 patients were split into four balanced folds for cross-validation. In our data, class imbalance for functional outcome (mRS) was 463 good outcome (35.6%) and 838 (64.4%) unfavorable outcome. For reperfusion (mTICI), class imbalance was smaller, with 749 (60.9%) as good reperfusion and 552 (39.1%) as poor reperfusion. Since our data was slightly imbalanced, we opted for balancing the classes using random under-sampling. For each iteration, three folds were used to train and optimize the models, and one fold was used for testing the models. Area Under the Curve (AUC) was used to assess model accuracy. For each CNN model, we created three training schemes: (1) training models from scratch; (2) using unsupervised pre-training to initialize convolutional weights (encoder part of the AE) and keeping them unchanged during supervised training; and (3) keeping pre-trained encoder weights unchanged for 50 epochs, then releasing and fine-tuning the whole architecture. We selected the cut-off of 50 epochs (from 25, 50 or 75) by monitoring convergence of convolutional filters and training loss. Further details about the experimental setup and the hyper-parameters used for optimization can be found in Supplemental Table V. All experiments were run on a PC with a single Titan X Pascal GPU, AMD Ryzen 7 1700X CPU, 16GB of RAM memory and Windows 10.

### Model Visualization

Deep Neural Networks are commonly referred to as black boxes because of their complex structure utilizing millions of parameters, in contrast to classical image processing techniques. In medical applications, a good prediction system, in addition to high accuracy, also needs to deliver interpretable predictions. Even though RFNN convolutional layers increase interpretability, here we further investigate the explanation of neural

predictions of our models. We hypothesize that the best way to explain outcome predictions is to visualize traits of input scans that led the model to the prediction. Various visualization techniques that analyze prediction models have been developed [32], [33]. Here, we explored the visualizations with Gradient-weighted Class Activation Mapping (Grad-CAM) [33]. Grad-CAM is a popular technique for generating visual interpretations of CNN-based networks, which fuses the localization and class-discriminative properties of Class Activation Mapping [34] and the precision of Guided Backpropagation [35]. Grad-CAM explains predictions by unveiling the gradient-weighted contribution of convolutional feature maps in the input space. In practice, we used the implementation of a slightly improved version of the technique, namely Grad-CAM++ [36].

We created two visualizations for each of the best mRS and mTICI prediction models. The first, coined GCAM, was created with the gradient-weighted CAM method. It reveals the parts of a certain input scan that were the most influential in making a prediction as a heat-map. The second, coined GWGBP (gradient weighted guided backpropagation), was created using the output of Guided Backpropagation throughout the whole network multiplied pixel-wise by the output of GCAM. GWGBP shows how the network interprets an input scan in terms of the most relevant imaging features utilized for a prediction. We thresholded GCAM heat-maps at 0.5 significance level to facilitate interpretation by highlighting the most contributing areas only.

## Results

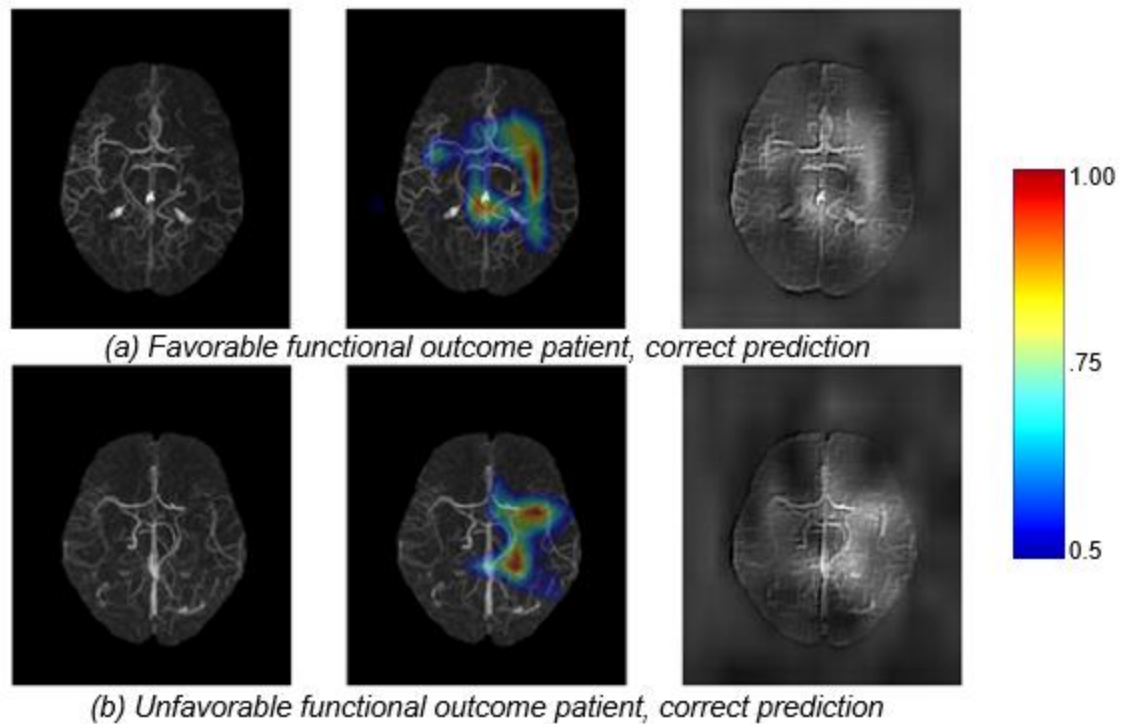
The average and range of AUC values for predicting good functional outcome ( $mRS \leq 2$ ) and good reperfusion ( $mTICI \geq 2b$ ) are reported in Table 4.2. The LR and RFC methods used with the radiological image biomarkers for predicting good functional outcome resulted in an AUC of 0.68 for LR and 0.66 for RFC. For predicting reperfusion, the AUC was 0.52 for both methods. The best average AUC for mRS prediction was obtained using the RFNN-ResNet model without AE pre-training (trained from scratch). The best average AUC for mTICI prediction was obtained with RFNN-ResNet-AE fine-tuned (with AE initialization and fine-tuning). Note that all models benefit from the AE pre-training for mTICI prediction. However, this is not

the case for the prediction of mRS, where RFNN-ResNet yielded the best result. Also, there was no difference in AUC between LR and RFC, as shown in previous studies [5]. Most importantly, the best performing data-efficient models outperformed the radiological image biomarkers baseline as well as standard CNN models for both mRS and mTICI outcome predictions. The AUC results for each fold are shown in Supplemental Tables VI and VII.

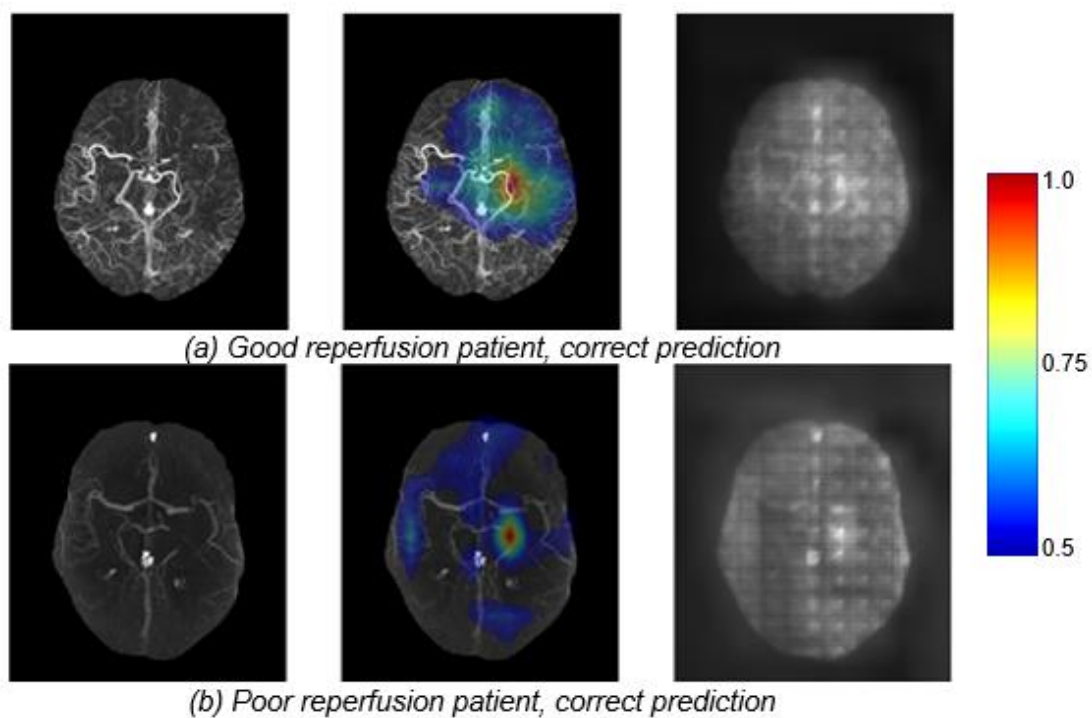
**Table 4.2.** AUC using 4-fold cross-validation. Standard *ResNet* and *RFNN-ResNet* trained with scheme (1), *ResNet-AE* and *RFNN-ResNet-AE* with scheme (2) and *ResNet-AE* fine-tuned and *RFNN-ResNet-AE* fine-tuned with scheme (3).

<b>Method</b>	<b>mRS AUC Avg (range)</b>	<b>mTICI AUC Avg (range)</b>
<b>LR Baseline</b>	0.68 (0.66 – 0.69)	0.52 (0.51 – 0.54)
<b>RFC Baseline</b>	0.66 (0.64 – 0.69)	0.52 (0.50 – 0.55)
<b>Standard ResNet</b>	0.56 (0.54 – 0.58)	0.51 (0.41 – 0.56)
<b>ResNet-AE</b>	0.58 (0.53 – 0.61)	0.57 (0.55 – 0.58)
<b>ResNet-AE fine-tuned</b>	0.57 (0.51 – 0.66)	0.57 (0.54 – 0.60)
<b>RFNN-ResNet</b>	0.71 (0.62 – 0.75)	0.57 (0.55 – 0.59)
<b>RFNN-ResNet-AE</b>	0.65 (0.60 – 0.69)	0.55 (0.53 – 0.57)
<b>RFNN-ResNet-AE fine-tuned</b>	0.67 (0.59 – 0.73)	0.65 (0.55 – 0.72)

In Figures 4.3 and Figure 4.4 we present the model visualization for the best models, RFNN-ResNet and RFNN-ResNet-AE fine-tuned, respectively. We can observe in the center column (GCAM), that the affected side of the brain (right in this case) contributes the most for the predictions (Figure 4.3). Even though the relevant regions are relatively large, the most important regions (depicted in red), are usually more specific. In the right column (GWGBP), we can see that the arteries are highlighted as important features learned by the model in Figure 4.3, while in Figure 4.4 we observe a noisier pattern. Additional model visualization examples are shown in Supplemental Figure I.



**Figure 4.3.** Visualization of predictions for mRS using the *RFNN-ResNet* model. Original MIP scans shown in the left column, the GCAM heat-map in overlay in the center and the GWGBP visualization on the right. Colors indicate the level of contribution of each region (GCAM). Most contributing regions (1.0) are represented in red, less contributing (0.5) in blue.



**Figure 4.4.** Visualization of the predictions for mTICI using the *RFNN-ResNet-AE* fine-tuned model. Order of visualizations corresponds to Figure 4.3.

## Discussion

We have shown that data-efficient DL analysis of CTA images outperformed prediction models with commonly used radiological image biomarkers in predicting reperfusion and functional outcome for patients with acute ischemic stroke. With model visualization tools, we have shown that the arteries are amongst the most common and influential features extracted by the DL models when predicting outcomes.

Recent studies on DL learning applied to stroke focused mostly on reproducing radiological image biomarkers and image segmentation tasks. In [37], SegNet [12] (an encoder-decoder architecture for image segmentation) was applied to MRI stroke imaging focused on predicting tissue outcome after acute ischemic stroke. DeepMedic [38], an open-source 3D CNN, was applied to CTA to detect ischemic stroke and segment lesions with high sensitivity and specificity in [16] and [14], showed that e-ASPECTS, a commercially available artificial intelligence software for ASPECTS scoring, is statistically non-inferior to neuroradiologists in scoring ASPECTS. A method using 3D CNNs was proposed in [39] for automatic assessment of DWI-ASPECTS with high accuracy. Finally, [13] presents a CNN named Stroke U-Net (SUNet) that was developed to segment and predict outcome of acute stroke lesions, and showed better results than 3D U-Net and uResNet.

Our results indicate an improvement by using structured convolutional kernels and unsupervised pre-training to predict good reperfusion ( $mTICI \geq 2b$ ) when compared to baseline models. The performance improvement derived from the use of Gaussian filters confirms the effect that has already been previously seen in previous medical imaging classification tasks [24]. The best performing method was the RFNN-ResNet-AE fine-tuned model (with auto-encoder initialization and fine-tuning after 50 epochs). For good functional outcome prediction ( $mRS \leq 2$ ), the RFNN-ResNet model (trained from scratch) achieved the highest AUC, slightly outperforming the baseline prediction models, with higher AUC scores for three out of four testing folds (Supplemental Tables VI and VII). Interestingly, for good mRS prediction, the unsupervised pre-training and supervised fine-tuning strategy resulted in a lower AUC when compared to the RFNN-ResNet. This could be caused by the nature of the output labels. mTICI is derived directly from imaging thus



more general image features (most effectively learnt by unsupervised learning) appeared to have the potential to be more predictive. One of the strongest predictors of mRS is age [5], [18], [40], which naturally reflects on many attributes of the brain detectable on CTA, e.g. atherosclerosis, structural abnormalities of the vasculature, changes in white matter and brain volume [41]. Recognition of any of these properties directly or indirectly on images can potentially lead to more confident mRS prediction in contrast to more general image features learnt during unsupervised training. We believe, for mRS, that supervised training from scratch enabled RFNN-ResNet models to grasp such complex features and discriminate subsets of patients better than more general image features learned in an unsupervised fashion. Our experiments suggest that image features learned directly from MIP images using RFNN-ResNet models can predict patients with good mTICI and mRS with higher accuracy than prediction models using well-known radiological image biomarkers. However, it should be mentioned that the predictive performance of the models is still limited.

MIP images are either present in organized databases or can be computed quickly and efficiently in seconds, making our method suitable for clinical practice. A prediction from our DL models takes only a few seconds, although the pre-processing steps (registration, skull-stripping and MIP computation) might take up to a couple of minutes (around 2 minutes for a scan with 400 slices). Consequently, an important advantage of our approach is that it is orders of magnitude faster and it does not require any manual image annotation, even during pre-processing, while delivering comparable prediction accuracy as existing radiological imaging biomarkers.

We selected the ResNet architecture based on state-of-the-art performance on natural image classification and refined it for our task. Our main aim was to evaluate the advantages of RFNN kernels over the standard ones. Despite optimizing some hyper-parameters, deeper and wider architectures (such as DenseNets) could potentially yield better results in our experiments and should be explored in the future. CNNs intrinsically contain many parameters that need to be optimized. In determining these parameters, we were restricted by limited computational resources – single GPU –, thus deeper models trained with larger batches or with 3D images might further improve

predictions. Also, note that we did not explore transfer learning from another domain (e.g. ImageNet [42]), because the input images would have to be scaled down to a lower resolution. This could lead to the loss of relevant information and the depths of layers would be restricted by the chosen architecture. Another potential limitation is the use of MIPs. Even though a lot of information is lost in using MIPs to represent the 3D images, the MIPs retain important artery structures, while keeping the input size feasible for training the DL models and reducing sparsity.

Even though k-fold cross-validation was applied, validation on an external dataset should be considered in future studies. Furthermore, we opted for a small number of folds for cross-validation due to the limited number of samples and class imbalance, as a high number of folds would lead to a test set with few samples, causing severe variance in our results. Also, due to the high number of experiments and hyper-parameters (from the optimizers and the RFNN), increasing the number of folds would have increased the duration of experiments radically. Besides, 4-fold cross-validation does not provide enough AUC values to compute the statistical significance of differences in accuracy between models. With more data available, more cross-validation folds could be performed to assess if the difference in AUC between models is statistical significant.

Other important clinical factors that are predictive for good mRS and mTICI, such as age, National Institutes of Health Stroke Scale (NIHSS), time from stroke onset to groin puncture, among others, should also be included in future prediction models[5], [6]. Finally, we selected the cut-off of  $\leq 2$  for mRS to make our models comparable to previous mRS prediction modeling research [5], [6]. However, other cut-offs should be explored, for instance  $mRS \leq 5$ , where the patients are severely disabled. Besides, provided that more data is available in the future, experiments comparing models trained on the full dataset and models trained on smaller portions could be used to quantify the extent of RFNN improvements over standard ResNets.

Understanding predictions is of utmost importance to improve reliable decision support for individual prospective patients and to further assist in discovering relevant-yet-unseen image features. From our visualizations, one can observe that the highest contribution for good mRS prediction comes

mostly from one – the occluded – side of the brain. For predicting reperfusion, however, information from both sides of the brain is taken into account. For good mRS prediction, it is clear from the GWGBP, that arteries – i.e., the extent of blood flow – were important for prediction, since such patterns were extracted in all cases. The important role of arteries is well known in clinical practice. For example, the collateral score (which is highly dependent on the visualization of the arteries in the brain) is commonly used to assess the alternative blood flow and is strongly associated with the size of infarction. The occlusion location is also an important predictor of functional outcome and reperfusion. Regarding mTICI, in cases of poor reperfusion, little is known about the reasons despite a successful recanalization after EVT, though many aspects of the artery have an effect on the efficacy of EVT treatment [43], [4], [44]. Further study is necessary to evaluate and properly quantify the visual explanations of the networks in more depth, since the most important regions are diverse, arteries are not always at the same location and stroke can occur in various locations of both sides of the brain. Given the important role of arteries and the feature pattern extracted by the networks, we suggest that future research on mTICI and mRS prediction should include the artery pattern as a feature. Also, one could create quantitative measures of interpretability of models, sensitivity and specificity of detection of certain features could greatly facilitate the understanding and improvement of DL models, which can help to identify new relevant image regions and patterns.

## Conclusion

We have shown that, in our dataset, automated radiological image analysis with data-efficient DL methods outperforms the combination of multiple radiological image biomarkers for good stroke outcome prediction. Our approach does not require image annotation and is faster to compute than any radiological image biomarker considered in this study. We also improved the interpretability of our models using model visualization tools, which is valuable in clinical practice. Even though DL has shown improvement for outcome prediction, the predictive value is still relatively low and clinical characteristics should be included in future prediction models.

# References

1. World Health Organization, “WHO - The top 10 causes of death,” 24 Maggio, 2018. .
2. O. Website, “Global Burden of Disease Study 2017 , Country Profile :,” pp. 1–7, 2018.
3. S. A. Alqahtani et al., “Endovascular Management of Stroke Patients with Large Vessel Occlusion and Minor Stroke Symptoms,” *Cureus*, vol. 9, no. 6, pp. 6–11, 2017.
4. M. Goyal et al., “Endovascular thrombectomy after large-vessel ischaemic stroke: A meta-analysis of individual patient data from five randomised trials,” *Lancet*, vol. 387, no. 10029, pp. 1723–1731, 2016.
5. H. J. A. Van Os et al., “Predicting outcome of endovascular treatment for acute ischemic stroke: Potential value of machine learning algorithms,” *Front. Neurol.*, vol. 9, no. SEP, pp. 1–8, 2018.
6. E. Venema et al., “Selection of patients for intra-arterial treatment for acute ischaemic stroke: development and validation of a clinical decision tool in two randomised trials,” *Bmj*, p. j1710, 2017.
7. P. A. Barber, A. M. Demchuk, J. Zhang, and A. M. Buchan, “Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy,” *Lancet*, vol. 355, no. 9216, pp. 1670–1674, 2000.
8. I. Y. L. Tan et al., “CT angiography clot burden score and collateral score: Correlation with clinical and radiologic outcomes in acute middle cerebral artery infarct,” *Am. J. Neuroradiol.*, vol. 30, no. 3, pp. 525–531, 2009.
9. O. A. Berkhemer et al., “Collateral Status on Baseline Computed Tomographic Angiography and Intra-Arterial Treatment Effect in Patients with Proximal Anterior Circulation Stroke,” *Stroke*, vol. 47, no. 3, pp. 768–776, 2016.
10. B. C. Stoel et al., “Automated brain computed tomographic densitometry of early ischemic changes in acute stroke,” *J. Med. Imaging*, vol. 2, no. 1, p. 014004, 2015.
11. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
12. V. Badrinarayanan, A. Kendall, R. Cipolla, and S. Member, “SegNet : A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.
13. A. Clèrigues, S. Valverde, J. Bernal, J. Freixenet, A. Oliver, and X. Lladó, “SUNet: a deep learning architecture for acute stroke lesion segmentation and outcome prediction in multimodal MRI,” 2018.
14. S. Nagel et al., “e-ASPECTS software is non-inferior to neuroradiologists in applying the ASPECT score to computed tomography scans of acute ischemic stroke patients,” *Int. J. Stroke*, vol. 12, no. 6, pp. 615–622, 2017.
15. K. Kamnitsas et al., “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation,” *Med. Image Anal.*, vol. 36, pp. 61–78, 2017.
16. O. Oman, T. Makela, E. Salli, S. Savolainen, and M. Kangasniemi, “3D convolutional neural networks applied to CT angiography in the detection of acute ischemic stroke.,” *Eur. Radiol. Exp.*, vol. 3, no. 1, p. 8, 2019.

17. Q. Zhang and S.-C. Zhu, "Visual Interpretability for Deep Learning: a Survey," vol. 19, no. 1423305, pp. 27–39, 2018.
18. I. G. H. Jansen, M. J. H. L. Mulder, and R. J. B. Goldhoorn, "Endovascular treatment for acute ischaemic stroke in routine clinical practice: Prospective, observational cohort study (MR CLEAN Registry)," *BMJ*, vol. 360, 2018.
19. S. Klein, M. Staring, K. Murphy, M. a. Viergever, and J. Pluim, "elastix: A Toolbox for Intensity-Based Medical Image Registration," *IEEE Trans. Med. Imaging*, vol. 29, no. 1, pp. 196–205, 2010.
20. J. K. Berge and R. A. Bergman, "Variations in size and in symmetry of foramina of the human skull," *Clin. Anat.*, vol. 14, no. 6, pp. 406–413, 2001.
21. J.-H. Jacobsen, J. van Gemert, Z. Lou, and A. W. M. Smeulders, "Structured Receptive Fields in CNNs," 2016.
22. T. Lindeberg, *Scale-Space Theory in Computer Vision*. 1994.
23. R. Manniesing, M. A. Viergever, and W. J. Niessen, "Vessel enhancing diffusion. A scale space representation of vessel structures," *Med. Image Anal.*, vol. 10, no. 6, pp. 815–825, 2006.
24. R. Zhang, J. Shen, F. Wei, X. Li, and A. K. Sangaiah, "Medical image classification based on multi-scale non-negative sparse coding," *Artif. Intell. Med.*, vol. 83, pp. 44–51, 2017.
25. W. S. Oliveira, J. V. Teixeira, T. I. Ren, G. D. C. Cavalcanti, and J. Sijbers, "Unsupervised retinal vessel segmentation using combined filters," *PLoS One*, vol. 11, no. 2, pp. 1–21, 2016.
26. D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why Does Unsupervised Pre-training Help Deep Learning?," *J. Mach. Learn. Res.*, vol. 11, pp. 625–660, 2010.
27. B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, and D. Tao, "Stacked Convolutional Denoising Auto-Encoders for Feature Representation," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1017–1027, 2017.
28. K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9908 LNCS, pp. 630–645, 2016.
29. O. Ozdemir, S. Giray, Z. Arlier, D. F. B. G, Y. Inanc, and E. Colak, "Predictors of a Good Outcome after Endovascular Stroke Treatment with Stent Retrievers," vol. 2015, no. Iv, 2015.
30. V. Nambiar et al., "CTA Collateral Status and Response to Recanalization in Patients with Acute Ischemic Stroke," no. May 2004, 2014.
31. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016.
32. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," pp. 1135–1144, 2016.
33. M. Cogswell, A. Das, and D. Batra, "Visual Explanations from Deep Networks via Gradient-based Localization."

34. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," pp. 2921–2929.
35. J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," pp. 1–14, 2014.
36. A. Chattopadhyay, A. Sarkar, and P. Howlader, "Grad-CAM ++ : Improved Visual Explanations for Deep Convolutional Networks," IEEE Winter Conf. Appl. Comput. Vis., pp. 839–847, 2018.
37. A. Nielsen, M. B. Hansen, A. Tietze, and K. Mouridsen, "Prediction of Tissue Outcome and Assessment of Treatment Effect in Acute Ischemic Stroke Using Deep Learning," *Stroke*, vol. 49, no. 6, pp. 1394–1401, 2018.
38. K. Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, 2017.
39. L. Do, I. Park, H. Yang, B. Baek, and S. Kee, "Automatic Assessment of DWI-ASPECTS for Assessment of Acute Ischemic Stroke using 3D Convolutional Neural Network," *IEEE Trans. Med. Imaging*, 2018.
40. M. J. H. L. Mulder et al., "Towards personalised intra-arterial treatment of patients with acute ischaemic stroke: a study protocol for development and validation of a clinical decision aid," *BMJ Open*, vol. 7, no. 3, p. e013699, 2017.
41. A. Gupta et al., "Neuroimaging of cerebrovascular disease in the aging brain," *Aging Dis.*, vol. 3, no. 5, pp. 414–425, 2012.
42. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2012.
43. H. Leischner et al., "Reasons for failed endovascular recanalization attempts in stroke patients," *J. Neurointerv. Surg.*, vol. 11, no. 5, pp. 439–442, 2019.
44. E. Qazi, F. S. Al-Ajlan, M. Najm, and B. K. Menon, "The Role of Vascular Imaging in the Initial Assessment of Patients with Acute Ischemic Stroke," *Curr. Neurol. Neurosci. Rep.*, vol. 16, no. 4, pp. 1–9, 2016.
45. U. Jensen-Kondering, "Hyperdense artery sign on computed tomography in acute ischemic stroke," *World J. Radiol.*, vol. 2, no. 9, p. 354, 2010.

## Supplemental material

**Supplemental Table I.** Description of Imaging biomarkers available before EVT

<b>Variable</b>	<b>Description</b>
Hyperdense artery sign (HAS) on baseline non-contrast CT [45]	Indicator of clot occlusion in case of acute ischemic stroke.
Relevant (new) ischemia or hypodensity	Evident
Hemorrhagic transformation at baseline	The conversion of a bland infarction into an area of hemorrhage. Potentially severe complication of ischemic stroke.
Leukoaraiosis	Neuroimaging abnormalities of the white matter, which appear as hypodense or hyperintense areas, are located predominantly in the periventricular area.
Presence of old infarcts in same ASPECTS region	Evident
ASPECTS at baseline	A 10-point quantitative topographic CT scan score where segmental assessment of the MCA vascular territory is made and 1 point is deducted from the initial score of 10 for every region involved.
Intracranial atherosclerosis on CTA	Evident
Intracranial vascular malformation or aneurysm visible on CTA	Evident
Most proximal occlusion segment on CTA	Evident
Collateral score on CTA	Scoring system that allows quick evaluation of collateral filling delay in acute ischemic stroke. A score on a scale of 0 to 5 is given, with 5 being the mildest and 0 the most severe.
Clot burden score	Scoring system to define the extent of thrombus found in the proximal anterior circulation by location and is scored on a scale of 0–10. A score of 10 is normal, implying clot absence.
Less than 50% atherosclerotic stenosis at symptomatic carotid bifurcation on CTA	Narrowing of the symptomatic bifurcation <50%
50% or more atherosclerotic stenosis at symptomatic carotid bifurcation on CTA	Narrowing of the symptomatic bifurcation ≥50%
Atherosclerotic occlusion at symptomatic carotid bifurcation on CTA baseline	Evident

Supplemental Table I (continued)

Floating thrombus at symptomatic carotid bifurcation on CTA baseline	Elongated thrombus attached to the arterial wall with circumferential blood flow with cyclical motion relating to cardiac cycles
Pseudo-occlusion at symptomatic carotid bifurcation on CTA baseline	Entity where flow-related artifact leads to an appearance of complete carotid bifurcation occlusion on computed tomographic angiography (CTA) or digital subtraction angiography (DSA), where in reality the carotid bifurcation is patent
Carotid dissection at symptomatic carotid bifurcation on CTA baseline	A condition where the layers of the carotid artery are spontaneously separated, which can potentially lead to ischemic stroke
Carotid web at symptomatic carotid bifurcation on CTA baseline	A rare condition where a thin, linear, membrane extends from the posterior aspect of the internal carotid artery bulb into the lumen, located just beyond the carotid bifurcation
Other lesion at symptomatic carotid bifurcation on CTA baseline	Lesions at the symptomatic carotid bifurcation on CTA, other than the above

**Supplemental Table II.** Parameters of the ResNet architecture (same for RFNN-ResNet)

Dropout (in blocks / after dense layer)	0.1 / 0.2
Batch normalization momentum	0.7
Scale of Gaussians in initial conv layer (RFNN-ResNet)	2.0
Scale of Gaussians in composite layers (RFNN-ResNet)	1.0
Order of derivatives of Gaussians (RFNN-ResNet)	3



**Supplemental Table III.** Output dimensions of ResNet architecture (same for RFNN-ResNet)

<b>Layers</b>	<b>Output</b>	<b>Type</b>	
Initial convolution	215x187x16	7 x 7 conv, stride 2 x 2	
Block (1)	108x94x16	3 x 3 conv, stride 1 x 1 3 x 3 conv, stride 2 x 2	x 1
Block (2)	54x47x32	3 x 3 conv, stride 1 x 1 3 x 3 conv, stride 2 x 2	x 1
Block (3)	27x24x64	3 x 3 conv, stride 1 x 1 3 x 3 conv, stride 2 x 2	x 1
Block (4)	14x12x128	3 x 3 conv, stride 1 x 1 3 x 3 conv, stride 2 x 2	x 1
Dense Layer	1x2x128	14 x 6 avg pool, stride 14 x 6	
	128	256D Fully-connected	
Classification Layer	2	128D Fully-connected, 2D softmax	

**Supplemental Table IV.** Output dimensions of ResNet-AE architecture (same for RFNN-ResNet-AE)

Layers	Output	Type		
Initial convolution	215x187x16	7 x 7 conv, stride 2 x 2		E n c o d e r
Block (1a)	108x94x16	3 x 3 conv, stride 2 x 2 3 x 3 conv, stride 1 x 1	x 1	
Block (2a)	54x47x32	3 x 3 conv, stride 2 x 2 3 x 3 conv, stride 1 x 1	x 1	
Block (3a)	27x24x64	3 x 3 conv, stride 2 x 2 3 x 3 conv, stride 1 x 1	x 1	
Block (4a)	14x12x128	3 x 3 conv, stride 2 x 2 3 x 3 conv, stride 1 x 1	x 1	
Block (4b)	27x24x64	3 x 3 conv, stride 2 x 2 3 x 3 conv, stride 1 x 1	x 1	D e c o d e r
Block (3b)	54x47x32	3 x 3 conv, stride 2 x 2 3 x 3 conv, stride 1 x 1	x 1	
Block (2b)	108x94x16	3 x 3 conv, stride 2 x 2 3 x 3 conv, stride 1 x 1	x 1	
Block (1b)	215x187x16	3 x 3 conv, stride 2 x 2 3 x 3 conv, stride 1 x 1	x 1	
Final convolution	430x374x1	7 x 7 conv, stride 2 x 2		

**Supplemental Table V.** Parameters of training algorithm

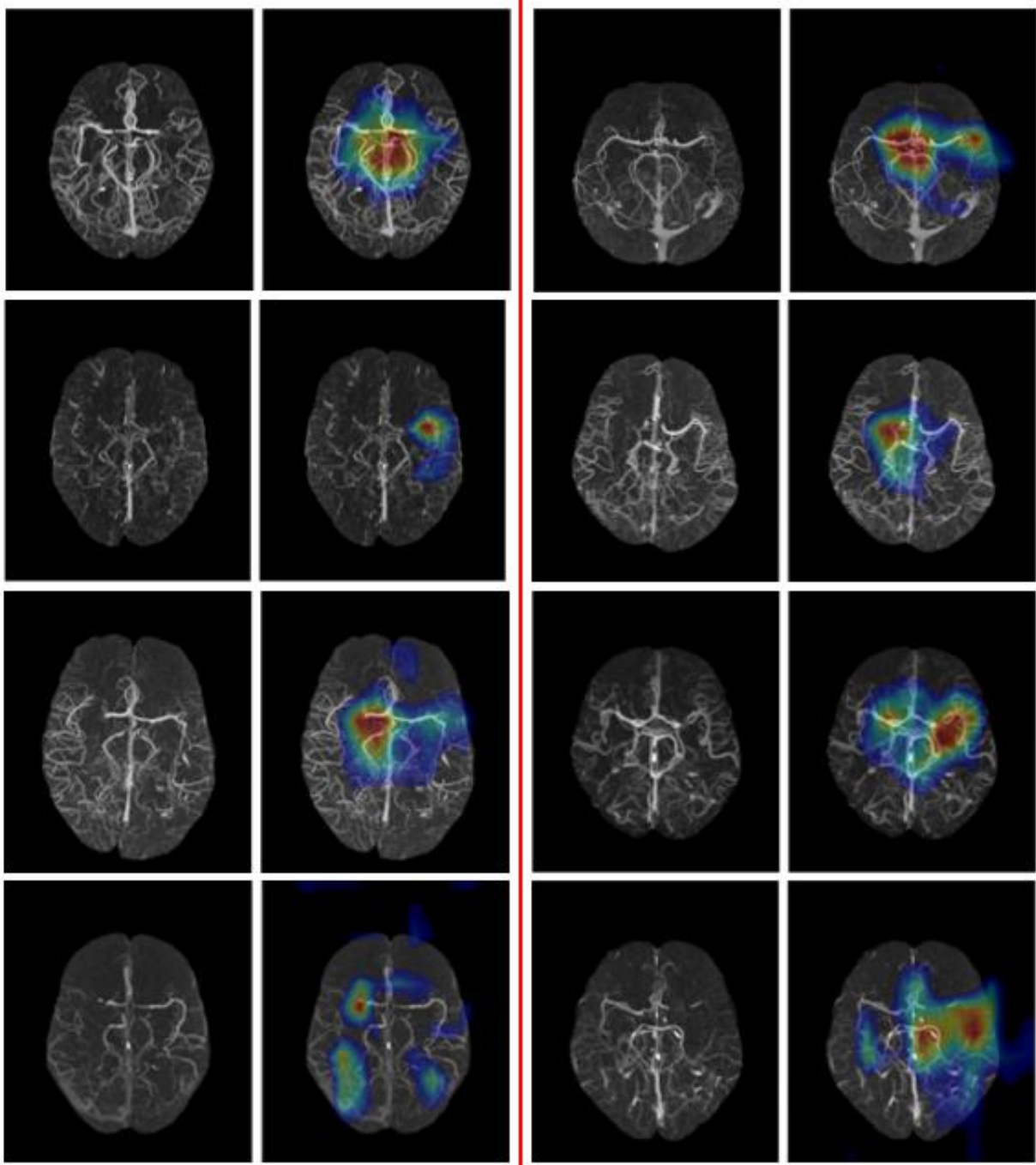
	Unsupervised training	Supervised training
Algorithm	RMS Prop	SGD with Nesterov mom
Momentum	0.9	0.9
Batch size	128	128
Learning rate	0.01	0.01
Number of epochs	50	100
Number of epochs before learning division by 10	-	50
Weight decay	Adaptive to equal initial cross entropy loss	1e-5

**Supplemental Table VI.** AUC results for each fold from the 4-fold cross-validation for good functional outcome prediction ( $mRS \leq 2$ ). ResNet and RFNN-ResNet pertain to training scheme (1), ResNet-AE and RFNN-ResNet-AE to (2) and ResNet-AE fine-tuned and RFNN-ResNet-AE fine-tuned to (3).

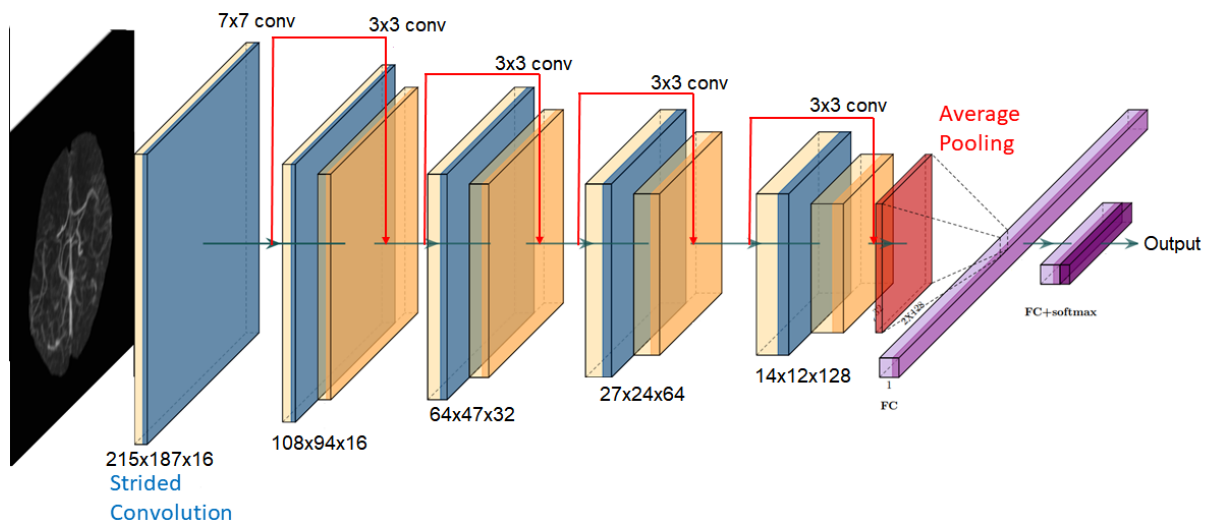
<b>Method</b>	<b>Fold-0</b>	<b>Fold-1</b>	<b>Fold-2</b>	<b>Fold-3</b>	<b>Avg</b>
LR Baseline	0.64	0.64	0.64	0.67	0.65
RFC Baseline	0.65	0.66	0.63	0.66	0.65
Standard ResNet	0.56	0.58	0.56	0.54	0.56
ResNet-AE	0.61	0.53	0.58	0.60	0.58
ResNet-AE fine-tuned	0.55	0.56	0.51	0.66	0.57
RFNN-ResNet	0.62	0.75	0.71	0.75	0.71
RFNN-ResNet-AE	0.61	0.60	0.68	0.69	0.65
RFNN-ResNet-AE fine-tuned	0.59	0.67	0.67	0.73	0.67

**Supplemental Table VII.** AUC results for each fold from the 4-fold cross-validation for good reperfusion prediction ( $mTICI \geq 2b$ ). ResNet and RFNN-ResNet pertain to training scheme (1), ResNet-AE and RFNN-ResNet-AE to (2) and ResNet-AE fine-tuned and RFNN-ResNet-AE fine-tuned to (3).

<b>Method</b>	<b>Fold-0</b>	<b>Fold-1</b>	<b>Fold-2</b>	<b>Fold-3</b>	<b>Avg</b>
LR Baseline	0.52	0.51	0.51	0.54	0.52
RFC Baseline	0.50	0.53	0.51	0.55	0.52
Standard ResNet	0.41	0.52	0.56	0.54	0.51
ResNet-AE	0.57	0.58	0.57	0.55	0.57
ResNet-AE fine-tuned	0.54	0.57	0.55	0.60	0.57
RFNN-ResNet	0.55	0.58	0.59	0.55	0.57
RFNN-ResNet-AE	0.53	0.55	0.57	0.54	0.55
RFNN-ResNet-AE fine-tuned	0.55	0.72	0.65	0.66	0.65



**Supplemental Figure I.** Visualization of predictions using the RFNN-ResNet model. Original MIP scans and their respective GCAM heat-map in overlay. Each pair represents a different scan. Colors indicate the level of contribution of each region (GCAM). Most contributing regions (1.0) are represented in red, less contributing (0.5) in blue.



Supplemental Figure II. ResNet architecture used in our approaches.



5

# CHAPTER 5.

Combination of radiological and clinical baseline data  
for outcome prediction of patients with an acute ischemic  
stroke

Ramos LA, van Os H, Hilbert A, Ernst M, Olabarriaga SD, Wermer M, Majoie C, Algra A, van der Schaaf I, Lingsma H, Dippel D, Roos YBWEM, van Zwam WH, van der Lugt A, van Walderveen MAA, Marquering H.

Combination of radiological and clinical baseline data for outcome prediction of patients with an acute ischemic stroke. *Under submission.*

## Abstract

*Background:* Accurate prediction of acute stroke patient outcome is of utmost importance for stroke patient care. Recent studies on prediction modeling for stroke focused mostly on clinical characteristics and radiological scores available at baseline. Radiological images are composed of millions of voxels, and a lot of information can be lost when representing this information by a single value. Therefore, in this study we aimed at developing prediction models that take into account the whole imaging data combined with clinical data available at baseline.

*Methods:* We included 3279 patients from the MR CLEAN Registry; a prospective, observational, multicenter registry of ischemic stroke patients treated with EVT. We developed two approaches to combine the imaging data to the clinical data. The first approach was based on radiomics features, extracted from 70 atlas regions combined to the clinical data to train machine learning models. For the second approach, we trained 3D deep learning models using the whole images and the clinical data. We compared models trained with only the clinical data to models trained with the combination of clinical and image data. Finally, we explored feature importance plots for the best models and identified many known variables and image features/brain regions that were relevant in the model decision process.

*Results:* From 3279 patients included, 1241 (37%) patients had a good functional outcome ( $mRS \leq 2$ ) and 1954 (60%) patients had good reperfusion ( $eTICI \geq 2b$ ). There was no significant improvement by combining the image data to the clinical data for mRS prediction (mean AUC of 0.81 vs 0.80), regardless of the approach used. Regarding predicting reperfusion, there was a significant improvement when image features were combined to the clinical ones (mean AUC of 0.54 vs 0.61), with the highest AUC obtained by the deep learning approach.

*Conclusions:* The combination of radiomics and deep learning image features with clinical data significantly improved the prediction of good reperfusion. We found no improvement in the prediction of good functional outcome, despite many image features appearing as important in the model visualization.



## Introduction

Around one third of patients who suffer from acute ischemic stroke die or remain severely disabled, making stroke a very severe condition worldwide (1). Occlusions in one of the major cerebral arteries are present in one third of patients, and is often referred as large vessel occlusion (LVO) (2). Endovascular treatment (EVT) is the standard treatment for LVO, and its great benefits have been proven extensively (3–6). However, despite successful treatment, around 30% of patients still present a poor outcome at 3 months. Outcome after treatment is dependend on multiple factors, from patient characteristics and condition to severity and location of the occlusion (7). Accurate prediction of patient outcome has been explored in several studies and it is of utmost importance to correctly identify patients who will and who will not have a good outcome. This information can be used to further personalize acute stroke care (7,8).

Most prediction models found in literature focused on small subsets of clinical features (7), although some recent studies have explored a broader set of variables (8,9). In most cases, information in radiological images was included in the form of visual scores, such as ASPECTS and the collateral score. However, translating millions of voxels in a radiological image to one or several visual scores potentially can result in significant loss of information.

In this study we explored a more extensive feature representation of radiological images, including a multitude of handcrafted features (radiomics) and automatically learned features using deep learning approaches. We hypothesized that a more extensive image feature representation can lead to improved outcome prediction of ischemic stroke patients through leveraging information that is complementary to or more detailed than radiological scores. We performed the combination of the automatically extracted images features with patient data available at baseline, and evaluated their impact on prediction accuracy of both clinical and radiological outcomes. Finally, we presented the feature importance for the best models and their impact in the predictions.

## Methods

### Study population

We included 3279 patients from the MR CLEAN Registry, a prospective, observational, multicenter study, which consecutively included all EVT-treated acute ischemic stroke patients in the Netherlands since the completion of the MR CLEAN trial (10) in March 2014. The central medical ethics committee of the Erasmus Medical Centre Rotterdam, the Netherlands, evaluated the study protocol and granted permission (MEC-2014–235) to carry out the data collection as a registry. (11) Patients provided permission for study participation through an opt-out procedure. Data that have been used for this study are available upon reasonable request from the MR CLEAN Registry committee (mrclean@erasmusmc.nl).

### Variables and outcome

All variables available at baseline were included in the models, including radiological scores. In total, 58 variables were selectively included. Ordinal variables such as pre-stroke mRS, collaterals, ASPECTS, National Institutes of Health Stroke Scale (NIHSS), Clot Burden Score (CBS), and Glasgow Coma Scale were treated as linear continuous scores. We created dummies for variables with multiple categories, therefore, the final input size for the models consisted of 58 features. A complete list of variables available can be found in the Supplemental Table I.

We created prediction models for two outcome variables, (1) favorable functional outcome after 3 months, defined by the modified Rankin Scale ( $mRS \leq 2$ ) and (2) good reperfusion defined by the modified Thrombolysis in Cerebral Infarction (eTICI)-score after EVT ( $post-eTICI \geq 2b$ ).

### Image data pre-processing

We included CT angiography (CTA) scans from all patients available in the dataset, following the approach from (12), where the added value of CTA for outcome predictions has already been proven. The first step in pre-processing the images was to strip the skull, since it contains voxels that are not relevant for the prediction tasks (13). For this segmentation task, we used a U-Net (14), (a convolutional neural network designed for segmentation of

biomedical images) trained on skull segmentations that were created using the approach described in (15) and subsequently manually corrected.

Since the slice thickness and the orientation of the head varied significantly between different scans, we registered the images to a reference scan using rigid and affine transformations. For this, we used an atlas as reference scan (16), with a size of 256x256x90 voxels. This atlas was developed using the Laboratory of Neuro Imaging Probabilistic Brain Atlas (LPBA40), which is publicly available (17). The atlas served not only as reference for registration, but it also contains the annotation of 70 brain regions, which allowed region-based feature extraction.

### Radiomics approach

For the first approach, we computed radiomics features for each brain region of the scans that were registered to the atlas. An advantage of crafting features from specific regions is that we can easily trace back which regions of the brain were the most important for prediction. We used all 70 regions contained in the atlas. For each region, 18 first-order features were computed using the Pyradiomics library (18). This resulted in a total of 1260 features. Since some regions overlapped and others were relatively small, we checked the correlation between the features. We only kept features that were less than 50% correlated to the others (19) reducing the number of radiomics features to 68. This was necessary since highly correlated features (multicollinearity) can hamper learning (19). For example, in case the of Logistic Regression, multicollinearity can lead severe variations in the coefficients, making the results less robust and trustworthy. Moreover, the large number of features would also be a problem because of the limited sample size  $n=3001$  (20). A complete list of the computed features is presented in the Supplemental Table II. These features were subsequently combined (concatenated) with the clinical data available at baseline and used to create prediction models for the outcomes. We selected the following state-of-the-art machine learning models: Random forest classifier (RFC) (21), Support vector machine (SVM) (22), Artificial neural networks (NN) (23), Gradient boosting (XGB) (24) and Logistic regression (LR).

### Deep learning approach

In the deep learning approach, the skull-stripped scans were used to train Convolutional Neural Networks (CNNs) to predict outcome. We opted for using skull-stripped scans prior to image registration to keep the information in the scans as raw as possible, and avoid changes in the Hounsfield units caused by the registration process. Moreover, we included transformations for data augmentation, such as rotation and flipping, which makes registration a futile step. Therefore, the registered scans were only used in *radiomics* approach. Since the input for deep learning models has to be uniform, and the voxel size among scans was not, we resampled the scans in all directions. The final input size to the network was 256x256x30 voxels. To increase the number of training samples and account for possible variations in the data, we performed data augmentation (vertical flipping and rotation). We trained and optimized 3D CNNs to predict favorable functional outcome and good reperfusion. For each network implemented, we added a Squeeze and Excitation (SE) module before the fully-connected layers (25), since it has been shown to greatly improve results in diverse Resnet models in multiple prediction tasks. The SE module models interdependencies between channels by adding learnable weights channels-wise. This way, the contribution of certain features in a given channel can have more or less impact than the others in the final prediction. Finally, we combined the clinical features with the image features by concatenating the clinical features to the image features before the fully-connected layers of the CNN (26–28).

We selected the ResNet10 architecture since it has shown comparable results to deeper architectures in medical imaging related tasks (29) while keeping training time feasible. The models were trained from scratch with the SE module. Finally, we used a Transfer Learning based approach, using the model developed in (29). In that model, the aim was to develop a robust ResNet by training it in a large amount of medical data (by putting together multiple datasets), including CT and MRI scans. By doing so, the CNN was able to learn filters that can extract relevant image features and generalized well to other tasks, making it ideal for transfer learning to other datasets with less images.

The ResNet model was trained from scratch for 75 epochs, and for 50 epochs when Transfer Learning was used. Following the results presented in (30), we opted to train the models built from scratch for longer than those using Transfer Learning to allow for a more fair comparison. We optimize our models using the Focal Loss (31), since there was some class imbalance in our labels (around 0.4/0.6 for both labels). Finally, we used the Adam optimizer (32), with a learning rate of 0.001 and the weight decay to 0.00006. The other hyper-parameters were left unchanged. The mini-batch size varied from two (for the 3D ResNets) to eight (for the 2D ones).

### Pipeline and experiment setup

Several of the 58 clinical variables included in our experiments had missing values. We imputed the missing values using Multiple Imputation with Chained Equations (33), since it has shown state-of-the-art results in several datasets. After imputation, we created dummies for the categorical variables. The data was then scaled by subtracting the mean and dividing by the standard deviation, for optimal performance of ML models. We used an inner and outer k-fold cross-validation strategy to train, validate and test our models. First, in the outer cross-validation loop, the dataset was split into training and testing using a 5-fold cross-validation strategy (4 folds are used for training and 1 for testing). The training set was then split again (inner-cross-validation) into training and validation, with 20% being used for validation. The validation set is used to assess model performance during training, for early-stopping and for hyper-parameter optimization. The list of the hyper-parameters used to optimize the models is shown in the Supplemental Table III.

For each approach, we designed four experiments to predict the outcomes: first (coined *clinical*), using all clinical features available at baseline, therefore including patient demographics and image derived scores such as ASPECTS; second (coined *image*), using only the features hand-crafted or learned (radiomics or deep learning features) from the CTA scans; third (coined *combination*), by combining all the features from the first experiment with the features of the second (all clinical data available at baseline, including image scores, and features learned from CTA scans), and fourth (coined *no image score*), by repeating the first and third experiments, but without any image derived scores such as ASPECTS or collateral scores.

To assess statistically significant differences between the models, we reported the Confidence Intervals for all the cross-validation iterations and used the McNemars test (34).

All code used for the development of the models and data analysis is available at: [https://github.com/L-Ramos/mrclean\\_combination](https://github.com/L-Ramos/mrclean_combination).

### Feature importance

To visualize feature importance of our models we used SHAP (SHapley Additive exPlanations), which is based on game theory to explain the output of any machine learning model (35). In SHAP, the contribution of a feature is given by computing the average contribution of all features by permuting all of them. SHAP has many advantages over other feature visualization techniques, such as LIME (Local Interpretable Model-agnostic Explanations) (36). SHAP provides global explanations instead of sample-oriented ones, offers tools to evaluate feature dependence and interactions and, the output explanations are generated based on the trained model provided by the user, instead of training a new model to explain feature importance, which is the case for LIME. Finally, with SHAP, the impact of low and high values of a given feature have in the final outcome can be more clearly evaluated with the plots, along with how important the feature is in predicting the correct class (35).

## Results

### Study population

Of the 3279 patients that were eligible, 278 were excluded due failure during skull-stripping (133 patients), because of incomplete scans, severe artifacts or due to failure during image registration (145 patients). In total 3001 patients were included, the mean age was 72 years old, and median baseline NIHSS was 16 (Supplemental Table I). At 90 days, 1241 (37%) patients had a good functional outcome ( $mRS \leq 2$ ) with 214 missing values (7%). Regarding reperfusion (post-eTICI  $\geq 2b$ ), 1954 (60%) patient had good reperfusion after treatment, with 90 missing values (3%).

### Radiomics approach

We present the results of good functional outcome prediction using the *radiomics* approach in Table 5.1. The AUC value was the highest for the *clinical* experiment (0.81), though it does not differ significantly from the *combination* experiment. The AUC is the lowest for the *image* experiment (0.69). For the *combination* experiment, the best AUC was 0.80. Sensitivity was the highest for the clinical experiment (0.79), while Specificity was the highest for the *combination* (0.77). The difference between the *clinical* and the *combination* experiments was not statistically significant (p-value=0.12) for the RFC.

There was no significant difference (drop of 0.01 or at most 0.02 in the average of all measures) in the results of the *no image score* experiment, see Supplemental Table IV.

In Table 5.2 we show the results for good reperfusion prediction (post-eTICI  $\geq 2b$ ). The highest AUC was for the *combination* experiment (0.57), while the lowest was for the clinical experiment (0.51). Sensitivity was the highest (0.91) for the image experiment, but specificity was also the lowest (0.11), showing that the RFC model might be biased towards one of the classes in this experiment. The same does not occur for all models in the image experiment, LR for instance, shows a good balance between sensibility and specificity values for all experiments. The difference between the *clinical* and the *combination* experiments was statistically significant, p-value=0.008 for the RFC.

Finally, there was no significant difference in the measures for the *no image scores* experiment, see Supplemental Table V.

**Table 5.1.** Results of the *clinical*, *image* and *combination* for the *radiomics* approach for predicting good functional outcome (mRS $\leq$ 2). Average over 5-fold cross-validation

Method	AUC 95% CI	F1-Score	Sensitivity	Specificity	PPV	NPV
<b>Clinical Experiment</b>						
<b>RFC</b>	0.81 (0.79-0.82)	0.69 (0.67-0.72)	0.72 (0.68-0.76)	0.75 (0.72-0.77)	0.67 (0.63-0.71)	0.79 (0.76-0.82)
<b>SVM</b>	0.81 (0.80-0.83)	0.71 (0.68-0.74)	0.79 (0.75-0.82)	0.70 (0.68-0.72)	0.65 (0.61-0.69)	0.82 (0.79-0.85)
<b>LR</b>	0.81 (0.80-0.82)	0.71 (0.68-0.73)	0.77 (0.74-0.80)	0.71 (0.69-0.73)	0.65 (0.62-0.69)	0.81 (0.78-0.84)
<b>XGB</b>	0.81 (0.80-0.82)	0.71 (0.68-0.74)	0.77 (0.74-0.81)	0.71 (0.70-0.72)	0.66 (0.62-0.69)	0.82 (0.79-0.84)
<b>NN</b>	0.81 (0.80-0.82)	0.69 (0.68-0.71)	0.73 (0.66-0.80)	0.74 (0.67-0.81)	0.67 (0.60-0.73)	0.79 (0.75-0.84)
<b>Image Experiment</b>						
<b>RFC</b>	0.68 (0.65-0.70)	0.50 (0.42-0.58)	0.45 (0.33-0.57)	0.77 (0.71-0.83)	0.58 (0.53-0.62)	0.66 (0.61-0.71)
<b>SVM</b>	0.69 (0.66-0.71)	0.60 (0.54-0.65)	0.64 (0.58-0.71)	0.64 (0.62-0.66)	0.56 (0.50-0.62)	0.72 (0.67-0.76)
<b>LR</b>	0.68 (0.66-0.70)	0.58 (0.53-0.63)	0.60 (0.53-0.66)	0.67 (0.65-0.69)	0.56 (0.53-0.60)	0.70 (0.65-0.74)
<b>XGB</b>	0.67 (0.65-0.69)	0.55 (0.52-0.58)	0.56 (0.51-0.61)	0.67 (0.63-0.71)	0.55 (0.51-0.59)	0.68 (0.64-0.72)
<b>NN</b>	0.65 (0.59-0.71)	0.49 (0.45-0.52)	0.45 (0.37-0.52)	0.72 (0.61-0.83)	0.54 (0.48-0.61)	0.65 (0.60-0.69)
<b>Combination Experiment</b>						
<b>RFC</b>	0.80 (0.79-0.81)	0.67 (0.64-0.70)	0.66 (0.60-0.73)	0.77 (0.72-0.82)	0.67 (0.63-0.72)	0.76 (0.73-0.80)
<b>SVM</b>	0.79 (0.78-0.81)	0.70 (0.67-0.73)	0.78 (0.73-0.82)	0.68 (0.66-0.71)	0.64 (0.60-0.67)	0.81 (0.78-0.84)
<b>LR</b>	0.80 (0.78-0.81)	0.70 (0.66-0.73)	0.76 (0.72-0.80)	0.70 (0.68-0.73)	0.65 (0.60-0.69)	0.80 (0.78-0.83)
<b>XGB</b>	0.80 (0.78-0.81)	0.69 (0.67-0.71)	0.76 (0.72-0.79)	0.69 (0.66-0.72)	0.64 (0.61-0.67)	0.80 (0.77-0.83)
<b>NN</b>	0.78 (0.77-0.79)	0.67 (0.65-0.68)	0.64 (0.60-0.68)	0.74 (0.70-0.75)	0.66 (0.62-0.68)	0.74 (0.68-0.76)

RFC, random forest classifier; SVM, support vector machine; LR, logistic regression; XGB, gradient boosting; NN, neural networks. AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.



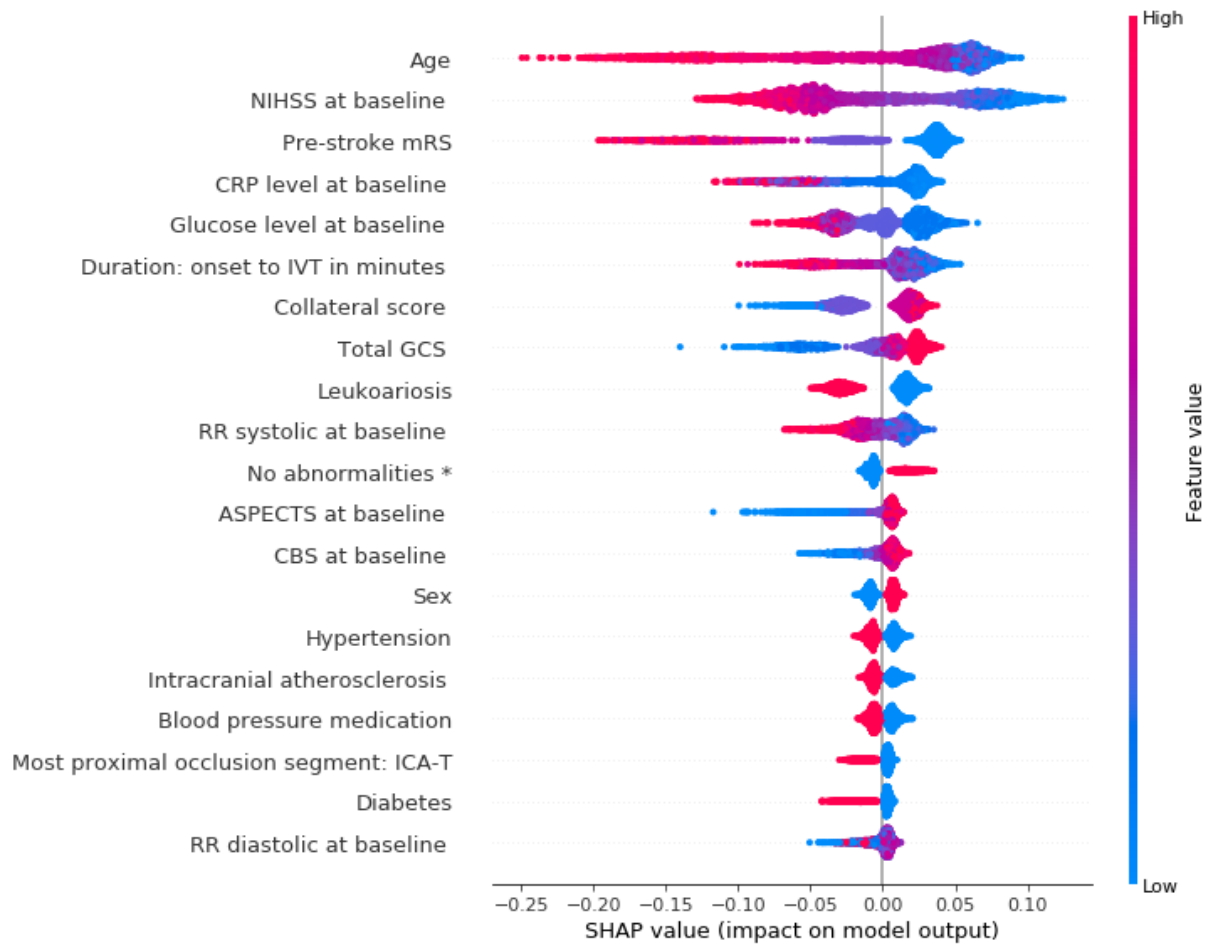
**Table 5.2.** Results of the *clinical*, *image* and *combination* experiments for the *radiomics* approach for predicting good reperfusion (post- eTICI  $\geq 2$ b). Average over 5-fold cross-validation

Methods	AUC 95% CI	F1-Score	Sensitivity	Specificity	PPV	NPV
<b>Clinical Experiment</b>						
<b>RFC</b>	0.53 (0.51-0.55)	0.71 (0.68-0.74)	0.79 (0.74-0.84)	0.26 (0.19-0.32)	0.64 (0.60-0.69)	0.42 (0.36-0.48)
<b>SVM</b>	0.54 (0.53-0.56)	0.39 (0.08-0.70)	0.32 (0.01-0.64)	0.73 (0.44-1.02)	0.68 (0.65-0.72)	0.39 (0.35-0.43)
<b>LR</b>	0.54 (0.51-0.56)	0.61 (0.57-0.66)	0.59 (0.54-0.64)	0.44 (0.39-0.50)	0.64 (0.61-0.68)	0.39 (0.34-0.43)
<b>XGB</b>	0.51 (0.50-0.54)	0.63 (0.57-0.69)	0.63 (0.55-0.71)	0.37 (0.30-0.45)	0.63 (0.58-0.68)	0.37 (0.33-0.41)
<b>NN</b>	0.51 (0.50-0.53)	0.70 (0.62-0.79)	0.81 (0.60-1.03)	0.19 (0.03-0.41)	0.63 (0.59-0.67)	0.37 (0.32-0.43)
<b>Image Experiment</b>						
<b>RFC</b>	0.54 (0.52-0.56)	0.75 (0.74-0.75)	0.91 (0.81-1.01)	0.11 (0.01-0.22)	0.64 (0.59-0.68)	0.42 (0.35-0.50)
<b>SVM</b>	0.55 (0.53-0.57)	0.70 (0.61-0.79)	0.79 (0.55-1.03)	0.25 (0.03-0.53)	0.64 (0.60-0.69)	0.41 (0.37-0.46)
<b>LR</b>	0.53 (0.50-0.57)	0.61 (0.57-0.64)	0.57 (0.53-0.61)	0.47 (0.40-0.54)	0.65 (0.61-0.69)	0.39 (0.32-0.46)
<b>XGB</b>	0.53 (0.50-0.56)	0.64 (0.60-0.68)	0.65 (0.54-0.75)	0.39 (0.28-0.49)	0.64 (0.61-0.68)	0.40 (0.33-0.46)
<b>NN</b>	0.53 (0.50-0.56)	0.67 (0.65-0.69)	0.69 (0.66-0.73)	0.36 (0.32-0.40)	0.65 (0.61-0.69)	0.41 (0.35-0.46)
<b>Combination Experiment</b>						
<b>RFC</b>	0.57 (0.55-0.59)	0.75 (0.71-0.78)	0.89 (0.85-0.93)	0.15 (0.10-0.19)	0.64 (0.60-0.68)	0.45 (0.38-0.52)
<b>SVM</b>	0.57 (0.54-0.61)	0.63 (0.59-0.66)	0.58 (0.55-0.61)	0.52 (0.46-0.58)	0.68 (0.63-0.72)	0.42 (0.38-0.47)
<b>LR</b>	0.57 (0.54-0.60)	0.63 (0.60-0.66)	0.59 (0.57-0.62)	0.50 (0.46-0.55)	0.67 (0.63-0.72)	0.42 (0.38-0.46)
<b>XGB</b>	0.57 (0.55-0.58)	0.59 (0.55-0.64)	0.54 (0.46-0.61)	0.55 (0.46-0.63)	0.67 (0.63-0.71)	0.41 (0.36-0.46)
<b>NN</b>	0.53 (0.51-0.55)	0.66 (0.62-0.70)	0.68 (0.63-0.72)	0.37 (0.32-0.43)	0.65 (0.60-0.70)	0.40 (0.37-0.43)

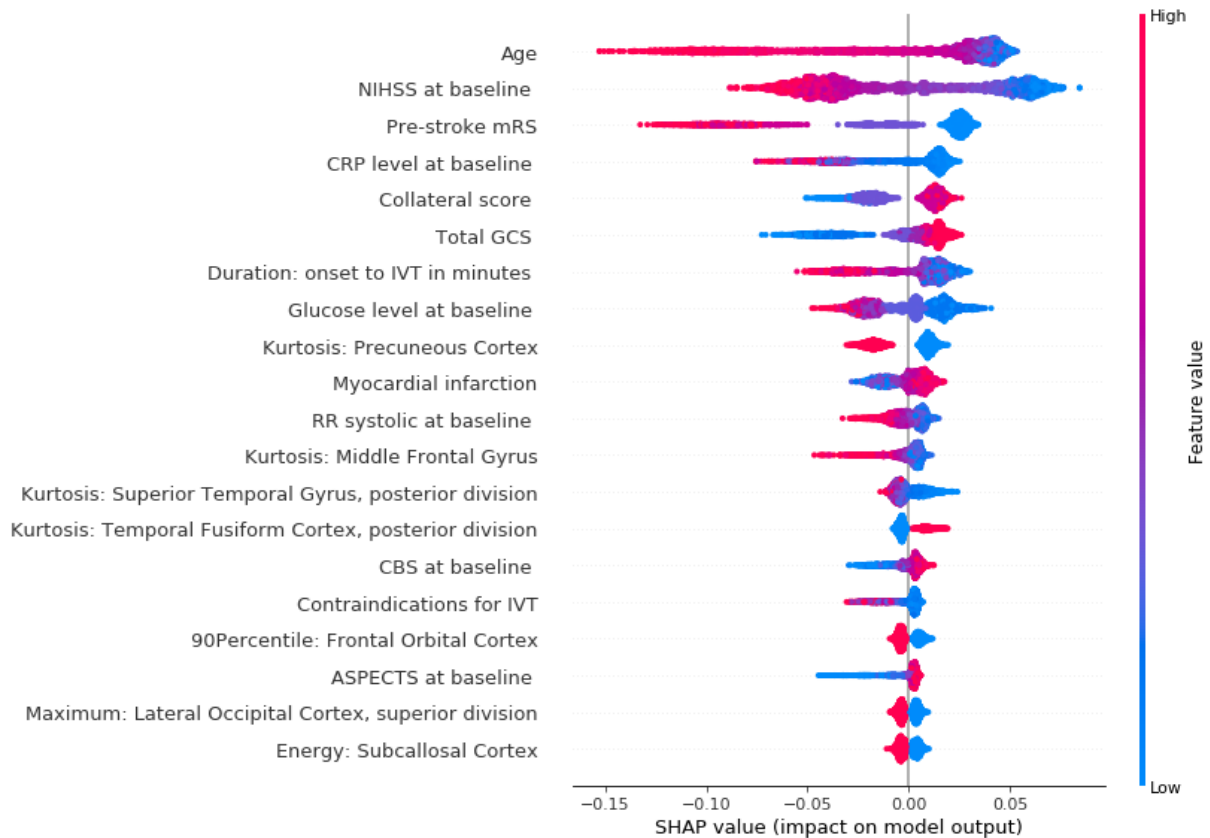
RFC, random forest classifier; SVM, support vector machine; LR, logistic regression; XGB, gradient boosting; NN, neural networks. AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.

### Feature importance – radiomics approach

In Figure 5.1 and Figure 5.2 we present feature importance using SHAP for the *clinical* (Figure 5.1) and the *combination*( Figure 5.2) experiments for RFC model and mRS prediction (for all 5-fold cross-validation iterations). The performance measures were the same across the models, therefore, we present feature importance for RFC model only, since SHAP has extensive support for three based models. Despite the addition of multiple radiomics features in the *combination* experiment, the top three most important features remain the same for both experiments (Age, NIHSS at baseline and pre-stroke mRS). It is also clear that low values of these three features is associated to good functional outcome. Collateral score and the Glasgow Comma Scale (GCS), become more important when the radiomics features are combined to the clinical data. Other features, like leukoariorosis and sex seem to lose importance when the radiomics features are combined. Finally, radiomics features from the following regions seem to have a significant impact in the prediction model, despite no improvements in the performance measures: precuneous cortex, middle frontal gyrus, superior temporal gyrus, temporal fusiform cortex, frontal orbital cortex, lateral occipital cortex and subcallosal cortex.

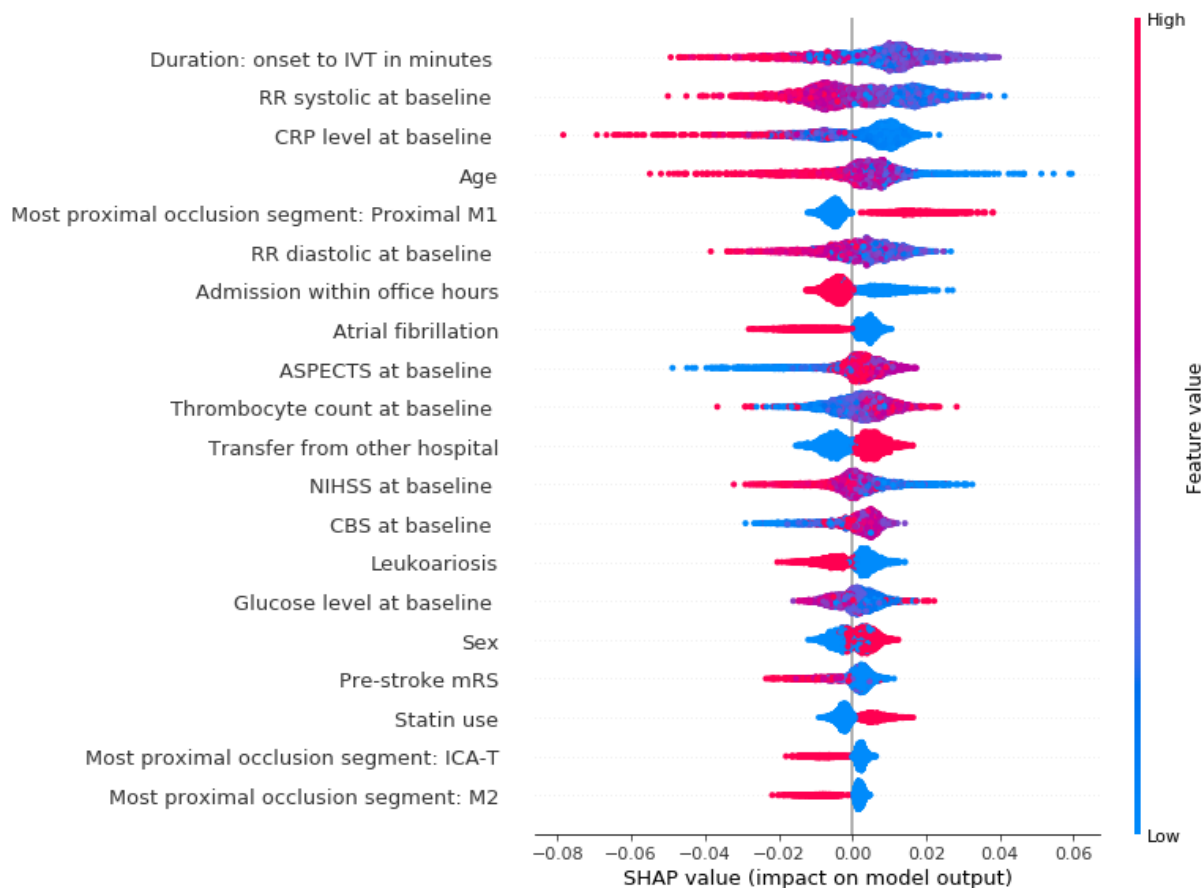


**Figure 5.1.** SHAP feature importance for the *clinical* experiment for mRS prediction using the RFC model. For visualization purposes we included only the top 20 features. Features are shown in order of importance, from most important (top) to less important (bottom). The color legend on the right shows how the feature values influence outcome: high values are depicted in red, while low values are presented in blue. Positive SHAP values (above zero in the x-axis) mean that the feature values are associated to the positive outcome (in this case good functional outcome), while SHAP values below zero indicate the opposite. \* at symptomatic carotid bifurcation on CTA at baseline.

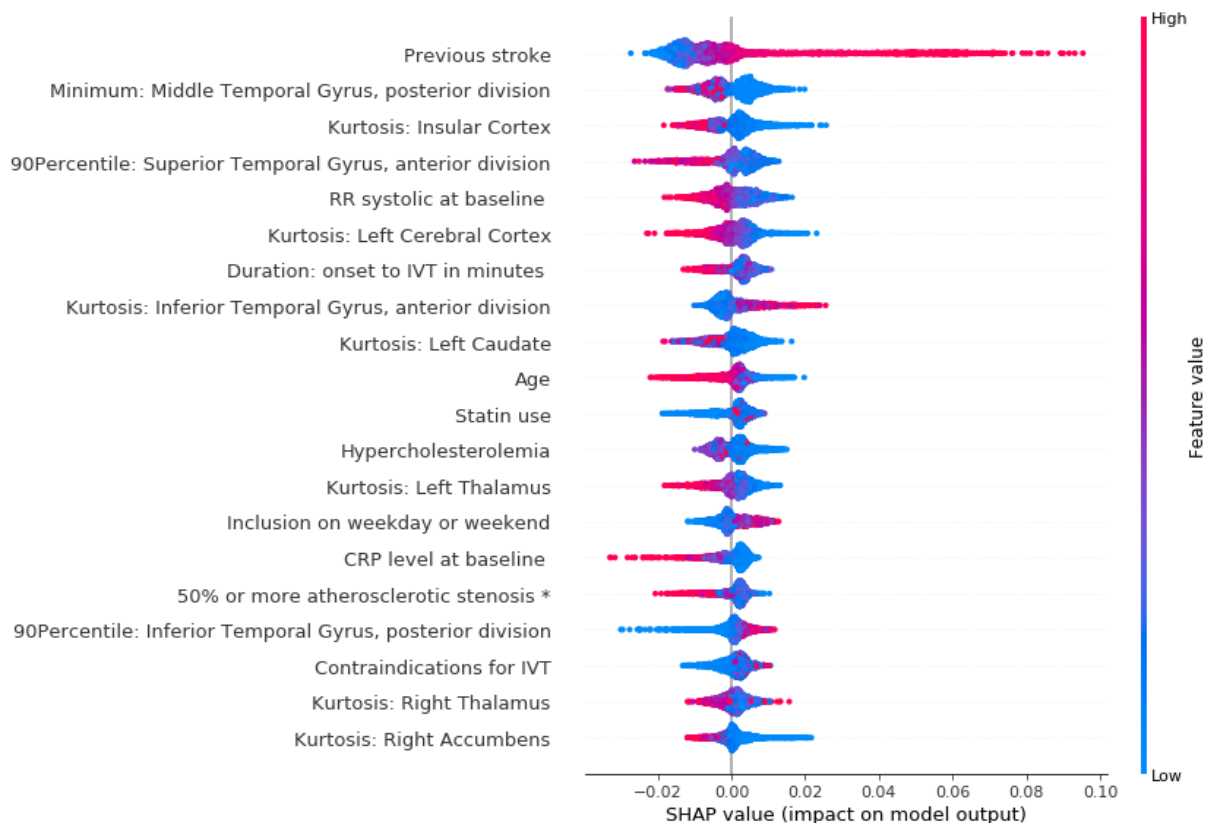


**Figure 5.2.** SHAP feature importance for the *combination* experiment for mRS prediction using the RFC model. For visualization purposes we included only the top 20 features. Features are shown in order of importance, from most important (top) to less important (bottom). The color legend on the right shows how the feature values influence outcome: high values are depicted in red, while low values are presented in blue. Positive SHAP values (above zero in the x-axis) mean that the feature values are associated to the positive outcome (in this case good functional outcome), while SHAP values below zero indicate the opposite. \* at symptomatic carotid bifurcation on CTA at baseline.

In Figure 5.3 and Figure 5.4 we show the feature importance using SHAP for the *clinical* (Figure 5.3) and *combination* (Figure 5.4) experiments for the prediction of good reperfusion using the RFC model. In this case, the most important features are different from each other when comparing both experiments. While the duration from onset to IVT, the RR systolic, CRP level and age seem to be the most important for the *clinical* experiment, these features are all replaced by many radiomics features from multiple brain regions in the *combination* experiment. This difference can also explain the slightly increased performance of the *combination* experiment when compared to the *clinical* and *image* ones.



**Figure 5.3.** SHAP feature importance for the *clinical* experiment for the prediction of good reperfusion using the RFC model. For visualization purposes we included only the top 20 features. Features are shown in order of importance, from most important (top) to less important (bottom). The color legend on the right shows how the feature values influence outcome: high values are depicted in red, while low values are presented in blue. Positive SHAP values (above zero in the x-axis) mean that the feature values are associated to the positive outcome (in this case good r), while SHAP values below zero indicate the opposite. \* at symptomatic carotid bifurcation on CTA at baseline. \* at symptomatic carotid bifurcation on CTA at baseline.



**Figure 5.4.** SHAP feature importance for the *combination* experiment for the prediction of good reperfusion using the RFC model. For visualization purposes we included only the top 20 features. Features are shown in order of importance, from most important (top) to less important (bottom). The color legend on the right shows how the feature values influence outcome: high values are depicted in red, while low values are presented in blue. Positive SHAP values (above zero in de x-axis) mean that the feature values are associated to the positive outcome (in this case good r), while SHAP values below zero indicate the opposite. \* at symptomatic carotid bifurcation on CTA at baseline. \* at symptomatic carotid bifurcation on CTA at baseline.

## Deep learning approach

In Table 5.3 we present the results for predicting good functional outcome using the *deep learning* approach. To keep the number of experiments feasible, we present results for the *clinical* and *combination* experiments. All the measures were similar for both the *clinical* and *combination* experiments. Regardless, for the *combination* experiment, training the ResNet10 models from scratch resulted in a worse performance than when using Transfer Learning from other image datasets. Similar to the *radiomics* approach, there seems to be no improvement in the performance measures when combining

the image learned features to the clinical data. The difference between the *clinical* and the *combination* was not significant, p-value=0.285 for the ResNet10 trained using Transfer Learning.

**Table 5.3.** Results of all experiments from the deep learning approach for predicting good functional outcome (mRS $\leq$ 2). Average over 5-fold cross-validation

Methods	AUC	F1-Score	Sensitivity	Specificity	PPV	NPV
<b>Clinical Experiment</b>						
<b>Feed Forward</b>	0.77 (0.76-0.78)	0.66 (0.61-0.71)	0.70 (0.66-0.73)	0.70 (0.67-0.74)	0.63 (0.57-0.70)	0.76 (0.72-0.80)
<b>Combination Experiment</b>						
<b>ResNet10 From Scratch</b>	0.54 (0.45-0.64)	0.29 (0.13-0.70)	0.30 (0.20-0.79)	0.78 (0.39-1.00)	0.58 (0.25-0.91)	0.61 (0.51-0.72)
<b>Combination Experiment</b>						
<b>ResNet10 Transfer Learning</b>	0.77 (0.75-0.78)	0.66 (0.62-0.70)	0.70 (0.67-0.73)	0.70 (0.68-0.73)	0.63 (0.57-0.68)	0.76 (0.72-0.80)

Finally, we present in Table 5.4 the results for predicting good reperfusion. All evaluation measures are higher for the combination experiment, and, despite confidence intervals are often overlapping, the average AUC is 0.08 higher. Again, the Transfer Learning approach for the Resnet10 model yielded better results than training from scratch. The difference between the *clinical* and the *combination* experiments was statistically significant, p-value<0.005 for the ResNet10 trained using Transfer Learning.

**Table 5.4.** Results of all experiments from the deep learning approach for predicting good reperfusion (post- eTICI  $\geq$  2b). Average over 5-fold cross-validation

Methods	AUC	F1-Score	Sensitivity	Specificity	PPV	NPV
<b>Clinical Experiment</b>						
<b>Feed Forward</b>	0.53 (0.50-0.55)	0.57 (0.54-0.61)	0.51 (0.48-0.54)	0.53 (0.52-0.55)	0.65 (0.59-0.71)	0.38 (0.32-0.43)
<b>Combination Experiment</b>						
<b>ResNet10 From Scratch</b>	0.50 (0.50-0.52)	0.13 (0.00-0.56)	0.12 (0.00-0.51)	0.87 (0.46-1.00)	0.15 (0.00-0.62)	0.36 (0.31-0.40)
<b>Combination Experiment</b>						
<b>ResNet10 Transfer Learning</b>	0.61 (0.50-0.72)	0.63 (0.54-0.71)	0.57 (0.50-0.64)	0.57 (0.50-0.64)	0.69 (0.60-0.80)	0.43 (0.40-0.45)

## Discussion

Our results suggest that there is a statistically significant improvement in performance for the prediction of good reperfusion (post-eTICI  $\geq 2b$ ) when data driven image features were combined with the clinical data, regardless of the radiomics or deep learning approach. In contrary, the addition of image features does not improve the prediction of good functional outcome (mRS $\leq 2$ ), regardless of the approach. Despite the lack of improvement in prediction accuracy for good functional outcome, radiomics features were relatively important for the models, when viewing the 20 features with highest feature importance.

In terms of prediction accuracy, our results are in line with previous works on mRS and reperfusion prediction, where AUCs around 0.80 and 0.57 respectively were reported (8,37). The current study is among the first to assess the combination of clinical and image features using both a radiomics and deep learning approaches for the prediction of good functional outcome and reperfusion. A previous study (38) explored a combination of clinical and image data using deep learning approaches to predict good functional outcome at baseline, and found a significant improvement in the AUC, despite presenting AUC values lower than the ones reported in this study and in the literature (8). Despite not finding the same improvements as reported in (38), our work included a much larger population (3279 patients vs 500 respectively) and we performed extra cross-validation iterations, while (38) reported the results for only one fold, which might be due to chance.

Strengths of our study include the large and heterogeneous population compared to previous studies that aimed at predicting good functional outcome and reperfusion (8,12,37). A heterogeneous dataset is important since we aimed to develop models on data that is as close to the clinical practice as possible. Besides, we employed two different approaches for combining the data, a *radiomics* approach, offering a more interpretable and visual solution, and a *deep learning* approach, which is a more state-of-the-art solution, increasing our chances of finding significant improvements. Another strength of our study is the use of inner and outer cross-validation for optimizing and testing the models, which can help identifying overfitting and reduces the risk of reporting overoptimistic results.



Several methodological challenges need to be considered interpreting the results of this study. First, only a relatively small number of Deep Learning models were used, and they were all based on the same architecture (ResNet10). Second, while other, deeper architectures are available and could yield better results, training a deeper architecture often requires more data and computer power. Since each validation iteration of our experiments takes around 24 hours to compute on a single GPU (and deeper 3D architectures would not fit the GPU memory), optimizing other models was out of the scope of this study. Also, given the size of our dataset, one could perform more cross-validation iterations, which would make our results more robust. Third, another limitation is the use of CTA modality, while other modalities could also be of added value like non contrast CT (NCCT). We chose to use CTA instead of NCCT because previous deep learning studies (12) already found a significant added value from CTAs for predicting good outcome, but did not explore a 3D approach for the images or their combination with clinical data. Fourth, the large number of variables included can also be a downside, since some are not readily available at baseline, despite all being possible to compute before treatment (either from patient history or recent imaging).

This study contributes to the understanding of imaging and clinical features that are associated good functional outcome and reperfusion. Age, NIHSS at baseline and pre-stroke mRS were found to be the top most important variables for functional outcome prediction, regardless of the data experiment, and have also been found to be relevant in previous studies (8,37). Besides that, with our imaging approaches we identified many relevant brain regions that have also been reported to be significantly associated to functional outcome (16). Regarding the prediction of good reperfusion, our findings suggest that image features (radiomics or deep learning) can significantly improve prediction and may replace many common clinical predictors in the models. For future research, one should consider computing more complex radiomics such as the gray level co-occurrence matrix or shape based features, since these have already been shown to be significantly associated to stroke patient outcome (39). Regarding deep learning, deeper ResNet networks could be considered, provided that enough data is available to train such models, and a Transfer

Learning approach should often be explored, since this can greatly surpass models trained from scratch as shown in this study.

## Conclusion

We found a significant improvement in the prediction of good reperfusion when combining image to clinical features. Regarding functional outcome, the addition of image features had no impact in the prediction accuracy. Nevertheless, the prediction accuracy of our models is still rather limited to be considered in clinical practice. The visualization of prediction feature importance showed both known and novel features with predictive value. Finally, we found that Transfer Learning can be of great assistance when training deep learning models for prediction tasks.

## References

1. World Health Organization. WHO - The top 10 causes of death. 24 Maggio. 2018.
2. Feigin VL, Lawes CMM, Bennett DA, Anderson CS. Stroke epidemiology: A review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century. *Lancet Neurology*. 2003.
3. Goyal M, Demchuk AM, Menon BK, Eesa M, Rempel JL, Thornton J, et al. Randomized assessment of rapid endovascular treatment of ischemic stroke. *N Engl J Med*. 2015;372(11):1019–30.
4. Jovin TG, M.D. AC, M.D. EC, Ph.D., María A. de Miquel M.D. CAM, M.D. AR, M.D. LSR, et al. Thrombectomy within eight hours after symptom onset in ischemic stroke. *N Engl J Med*. 2016;22(1):36.
5. Saver JL, Goyal M, Bonafe A, Diener H-C, Levy EI, Pereira VM, et al. Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke. *N Engl J Med*. 2015;372(24):2285–95.
6. Campbell BCV, Mitchell PJ, Kleinig TJ, Dewey HM, Churilov L, Yassi N, et al. Endovascular Therapy for Ischemic Stroke with Perfusion-Imaging Selection. *N Engl J Med*. 2015;372(11):1009–18.
7. Venema E, Mulder MJHL, Roozenbeek B, Broderick JP, Yeatts SD, Khatri P, et al. Selection of patients for intra-arterial treatment for acute ischaemic stroke: Development and validation of a clinical decision tool in two randomised trials. *BMJ*. 2017;357.
8. Van Os HJA, Ramos LA, Hilbert A, Van Leeuwen M, Van Walderveen MAA, Kruyt ND, et al. Predicting outcome of endovascular treatment for acute ischemic stroke: Potential value of machine learning algorithms. *Front Neurol*. 2018;9 (SEP):1–8.
9. Alaka SA, Menon BK, Brobbey A, Williamson T, Goyal M, Demchuk AM, et al. Functional Outcome Prediction in Ischemic Stroke: A Comparison of Machine Learning Algorithms and Regression Models. *Front Neurol*. 2020;
10. Berkhemer OA, Fransen PSS, Beumer D, van den Berg LA, Lingsma HF, Yoo AJ, et al. A Randomized Trial of Intraarterial Treatment for Acute Ischemic Stroke. *N Engl J Med*. 2015;372(1):11–20.
11. Jansen IGH, Mulder MJHL, Goldhoorn RJB. Endovascular treatment for acute ischaemic stroke in routine clinical practice: Prospective, observational cohort study (MR CLEAN Registry). *BMJ*. 2018;360.
12. Hilbert A, Ramos LA, van Os HJA, Olabariaga SD, Tolhuisen ML, Wermer MJH, et al. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. *Comput Biol Med*. 2019;115 (June):103516.
13. Sales Barros R, Tolhuisen ML, Boers AMM, Jansen I, Ponomareva E, Dippel DWJ, et al. Automatic segmentation of cerebral infarcts in follow-up computed tomography images with convolutional neural networks. *J Neurointerv Surg*. 2019;848–52.
14. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Miccai*. 2015;234–41.

15. Berge JK, Bergman RA. Variations in size and in symmetry of foramina of the human skull. *Clin Anat.* 2001;14 (6):406–13.
16. Ernst M, Boers AMM, Aigner A, Berkhemer OA, Yoo AJ, Roos YB, et al. Association of Computed Tomography Ischemic Lesion Location with Functional Outcome in Acute Large Vessel Occlusion Ischemic Stroke. *Stroke.* 2017;48 (9):2426–33.
17. Shattuck DW, Mirza M, Adisetiyo V, Hojatkashani C, Salamon G, Narr KL, et al. Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage.* 2008;
18. Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* 2017;77 (21):e104–7.
19. Yu L, Liu H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In: *Proceedings, Twentieth International Conference on Machine Learning.* 2003.
20. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and cox regression. *Am J Epidemiol.* 2007;165 (6):710–8.
21. Breiman L. Random forests. *Mach Learn.* 2001;45 (1):5–32.
22. Cortes C, Vapnik V. Support-Vector Networks. *Mach Learn.* 1995;
23. Bishop CM. *Pattern Recognition and Machine Learning.* Information Science and Statistics. 2006.
24. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min.* 2016;Pages 785-794.
25. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks. *IEEE Trans Pattern Anal Mach Intell.* 2020;
26. A Study on the Fusion of Pixels and Patient Metadata in CNN-Based Classification of Skin Lesion Images.
27. Ellen JS, Graff CA, Ohman MD. Improving plankton image classification using context metadata. *Limnol Oceanogr Methods.* 2019;
28. Nguyen LD, Lin D, Lin Z, Cao J. Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. In: *Proceedings - IEEE International Symposium on Circuits and Systems.* 2018.
29. Chen S, Ma K, Zheng Y. Med3D: Transfer Learning for 3D Medical Image Analysis. 2019;1–12.
30. He K, Girshick R, Dollár P. Rethinking imageNet pre-training. In: *Proceedings of the IEEE International Conference on Computer Vision.* 2019.
31. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. *Proc IEEE Int Conf Comput Vis.* 2017;2017-Octob:2999–3007.
32. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2014;1–15.
33. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and howdoes it work. 2012;20 (1):40–9.
34. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika.* 1947;

35. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. Explainable AI for Trees: From Local Explanations to Global Understanding. 2019;1–72.
36. Ribeiro MT, Singh S, Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. 2016;1135–44.
37. Venema E, Mulder MJHL, Roozenbeek B, Broderick JP, Yeatts SD, Khatri P, et al. Selection of patients for intra-arterial treatment for acute ischaemic stroke: Development and validation of a clinical decision tool in two randomised trials. *BMJ*. 2017;357.
38. Samak ZA, Mirmehdi M. Prediction of Thrombectomy Functional Outcomes using Multimodal Data. In: *MIUA 2020: Medical Image Understanding and Analysis*. 2020. p. 267–79.
39. Frindel C, Rouanet A, Giacalone M, Cho TH, Østergaard L, Fiehler J, et al. Validity of Shape as a Predictive Biomarker of Final Infarct Volume in Acute Ischemic Stroke. *Stroke*. 2015;46 (4):976–81.

## Supplemental material

**Supplemental Table I.** Details of included variables. A1, first segment of anterior cerebral artery; ASPECTS, Alberta stroke programme early CT score; cat, categorical; CBS, clot burden score; cont, continuous; CRP, C-reactive protein; CTA, CT angiography; DOAC, direct oral anticoagulant; ER, emergency room; HAS, hyperdense artery sign; IQR, interquartile range; M1/M2/M3, first/second/third segment of middle cerebral artery; mRS, modified Rankin Scale; NCCT, non-contrast CT; NIHSS, National Institutes of Health stroke scale; RR, blood pressure (Riva-Rocci).

Name	Occurrence (%) N=3279	Missing n (%)	Analyzed as
Previous stroke		27 (1)	cat
0 – no	2706 (83)		
1 – yes	546 (17)		
Myocardial infarction		67 (2)	cat
0 – no	2759 (84)		
1 – yes	453 (14)		
Peripheral arterial disease		68 (2)	cat
0 – no	2910 (89)		
1 – yes	301 (9)		
Diabetes		24 (1)	cat
0 – no	2723 (83)		
1 – yes	532 (16)		
Hypertension		66 (2)	cat
1 – yes	1688 (51)		
0 – no	1525 (47)		
Atrial fibrillation		43 (1)	cat
0 – no	2464 (75)		
1 – yes	772 (24)		
Hypercholesterolemia		143 (4)	cat
0 – no	2169 (66)		
1 – yes	967 (29)		
Antiplatelet use		41 (1)	cat
0 – no	2227 (68)		
1 – yes	1011 (31)		
DOAC use		40 (1)	cat
0 – no	3132 (96)		
1 – yes	107 (3)		
Coumarin use		24 (1)	cat
0 – no	2839 (87)		
1 – yes	416 (13)		
Heparin use		43 (1)	cat
0 – no	3135 (96)		
1 – yes	101 (3)		
Blood pressure medication		62 (2)	cat

Supplemental Table I (continued)

	1 – yes	1739 (53)		
	0 – no	1478 (45)		
Statin use			74 ( 2)	cat
	0 – no	2070 (63)		
	1 – yes	1135 (35)		
HAS on baseline NCCT			131 ( 4)	cat
	1 – yes	1704 (52)		
	0 – no	1444 (44)		
Relevant (new) ischemia / hypodensity			157 ( 5)	cat
	1 – yes	1908 (58)		
	0 – no	1214 (37)		
Hemorrhagic transformation			137 ( 4)	cat
	0 – no	3098 (94)		
	1 – yes	44 ( 1)		
Leukoariosis			128 ( 4)	cat
	0 – no	1903 (58)		
	1 – yes	1248 (38)		
Old infarcts in same ASPECTS region?			126 ( 4)	cat
	0 – no	2721 (83)		
	1 – yes	432 (13)		
Intracranial atherosclerosis on CTA scored by core lab			132 ( 4)	cat
	1 – yes	1886 (58)		
	0 – no	1261 (38)		
Sex			0 (0)	cat
	Male	1696 (52)		
	Female	1583 (48)		
Most proximal occlusion segment on CTA scored by core lab, based on CBS			151 ( 5)	cat
	Distal M1	1061 (32)		
	Proximal M1	754 (23)		
	ICA-T	663 (20)		
	M2	455 (14)		
	Intracranial ICA	161 ( 5)		
	None	13 ( 0)		
	M3	9 ( 0)		
	A2	6 ( 0)		
	A1	6 ( 0)		
Smoking			758 (23)	cat
	0 – no	1813 (55)		
	1 – yes	708 (22)		
Inclusion on weekday or weekend			0 (0)	cat

Supplemental Table I (continued)			
	0 – weekday	2415 (74)	
	1 – weekend	864 (26)	
Admission between 17.00-08-00 (weekday)/weekend or holiday. Based on ER time.			0 (0) cat
	1 – office hours	2088 (64)	
	0 – outside office hours	1191 (36)	
Transfer from other hospital			1 (0) cat
	1 – transfer	1783 (54)	
	0 – no transfer	1495 (46)	
Contraindications for IVT			2461 (75) cat
	0 – no	772 (24)	
	1 – yes	46 (1)	
No abnormalities at symptomatic carotid bifurcation on CTA baseline by core lab			400 (12) cat
	0 – no abnormalities	2110 (64)	
	1 – any abnormalities	769 (23)	
50% or more atherosclerotic stenosis at symptomatic carotid bifurcation on CTA baseline			400 (12) cat
	0 – no	2615 (80)	
	1 – yes	264 (8)	
Atherosclerotic occlusion at symptomatic carotid bifurcation on CTA baseline by core lab			400 (12) cat
	0 – no	2564 (78)	
	<b>1 – yes</b>	315 (10)	
Floating thrombus at symptomatic carotid bifurcation on CTA baseline by core lab			400 (12) cat
	0 – no	2826 (86)	
	1 – yes	53 (2)	
Pseudo-occlusion at symptomatic carotid bifurcation on CTA baseline by core lab			400 (12) cat
	0 – no	2684 (82)	
	1 – yes	195 (6)	



Supplemental Table I (continued)

Carotid dissection at symptomatic carotid bifurcation on CTA baseline by core lab		400 (12)	cat
0 – no	2777 (85)		
1 – yes	102 (3)		
Occlusion side on CTA scored by core lab		2 (0)	cat
Left hemisphere	1745 (53)		
Right hemisphere	1515 (46)		
Neither	17 (1)		
In-hospital stroke		534 (16)	cat
0 – no	2416 (74)		
1 – yes	329 (10)		
Second occlusion in other territory present on CTA scored by core lab		546 (17)	cat
0 – no	2454 (75)		
1 – yes	279 (9)		
Collateral score on CTA scored by core lab		207 (6)	cont
100% of occluded area	595 (18)		
>50% but less <100% filling <50% of occluded area	1190 (36)		
Absent collaterals	1100 (34)		
187 (6)			
Pre-stroke mRS		72 (2)	cont
0	2170 (66)		
1	424 (13)		
2	241 (7)		
3	211 (6)		
4	133 (4)		
5	28 (1)		
90-day mRS		214 (7)	cat
6	886 (27)		
2	561 (17)		
1	471 (14)		
3	404 (12)		
4	366 (11)		
0	209 (6)		
5	168 (5)		
Post-eTICI		90 (3)	cat
3	905 (28)		
2b	702 (21)		
2a	597 (18)		
0	543 (17)		

Supplemental Table I (continued)

	2c			
	1	347 (11)		
		95 (3)		
ASPECTS baseline scored by core lab – median (IQR)		9 (7- 10)	109 (3)	cont
CBS at baseline – median (IQR)		6 (4- 8)	766 (23)	cont
NIHSS at baseline – median (IQR)		16 (11- 20)	55 (2)	cont
Glucose level at baseline – median (IQR)		7 (6- 8)	371 (11)	cont
RR systolic at baseline – median (IQR)		150 (131-165)	89 (3)	cont
RR diastolic at baseline – median (IQR)		80 (71- 91)	97 (3)	cont
INR at baseline – median (IQR)		1 (1- 1)	608 (19)	cont
Thrombocyte count at baseline – median (IQR)		234 (194-289)	445 (14)	cont
CRP level at baseline – median (IQR)		4 (2- 10)	651 (20)	cont
Age – median (IQR)		72 (61- 80)	0 (0)	cont
Total glasgow coma scale at baseline – median (IQR)		13 (11- 15)	113 (3)	cont
Duration from onset to groin in minutes – median (IQR)		195 (150-260)	15 (0)	cont
Duration: onset to IVT in minutes in first hospital – median (IQR)		24 (18- 33)	1353 (41)	cont

**Supplemental Table II.** List of radiomics features computed per atlas region in the radiomics approach. Adapted from the PyRadiomics documentation (17). ROI: Region of Interest.

Feature Name	Explanation
10 <sup>th</sup> Percentile	The 10 <sup>th</sup> percentile of the ROI
90 <sup>th</sup> Percentile	The 90 <sup>th</sup> percentile of the ROI
Energy	Energy is a measure of the magnitude of voxel values in an image. A larger values implies a greater sum of the squares of these values.
Entropy	Entropy specifies the uncertainty/randomness in the image values. It measures the average amount of information required to encode the image values.

Supplemental Table 2 (continued)

Interquartile Range	The 75 <sup>th</sup> minus the 25 <sup>th</sup> percentiles of the image array, respectively.
Kurtosis	It is a measure related to the peak of the distribution of values in the image ROI. A higher kurtosis implies that the mass of the distribution is concentrated towards the tail(s) rather than towards the mean. A lower kurtosis implies the reverse: that the mass of the distribution is concentrated towards a spike near the Mean value.
Maximum	Maximum gray level intensity within the ROI
Mean	Mean gray level intensity within the ROI
Mean Absolute Deviation	The mean distance of all intensity values from the Mean Value of the image array).
Median	Median gray level intensity within the ROI.
Minimum	Minimum gray level intensity within the ROI
Range	The range of gray values in the ROI given by: maximum – minimum.
Robust Mean Absolute Deviation	mean distance of all intensity values from the Mean Value calculated on the subset of image array with gray levels in between, or equal to the 10th and 90th percentile
Root Mean Squared	Square-root of the mean of all the squared intensity values. It is another measure of the magnitude of the image values
Skewness	Measures the asymmetry of the distribution of values about the Mean value. Depending on where the tail is elongated and the mass of the distribution is concentrated, this value can be positive or negative.
Total Energy	Total Energy is the value of Energy feature scaled by the volume of the voxel in cubic mm.
Uniformity	Measure of the sum of the squares of each intensity value. This is a measure of the homogeneity of the image array, where a greater uniformity implies a greater homogeneity or a smaller range of discrete intensity value
Variance	Variance is the mean of the squared distances of each intensity value from the Mean value. This is a measure of the spread of the distribution about the mean.

---

**Supplemental Table III.** Hyper-parameters used for optimizing the Machine Learning models using grid-search.

<b>Classifier</b>	<b>Parameter Name</b>	<b>Parameter Value</b>
RFC	Number of Trees	[100,200,400,600,800,1000,1200,1400]
	Max features for split	auto, sqrt and log2
	Max depth of trees	[10,20,30,40, 50, 60, 70, 80, 90, 100, None]
	Quality of split	Gini or Entropy
	Minimum number of samples required to split an internal node	[2,4,6,8]
	Minimum number of samples required to be at a leaf node	[2,4,6,8,10]
SVM	Kernel type	Linear, Radial basis function, Polynomial
	Penalty parameter C	[0.001, 0.01, 0.1, 1, 10, 100]
	Kernel coefficient $\gamma$ (gamma)	[1, 0.1, 0.01, 0.001, 0.0001]
	Degree of the Polynomial kernel	[1,2,3,4,5,6]
LR	Regularization	[0.001, 0.01, 0.1, 1, 10, 100]
	Optimization algorithm	[newton-cg, lbfgs, liblinear, sag, saga]
NN	Hidden Layer sizes	[90,180,90], [90,120,90], [90,90], [90,180], [90], [180]
	Activation	ReLU, logistic
	Regularization parameter	[0.1, 0.01, 0.001, 0.0001]
	Batch size	[32, 64, 128]
	Learning rate	[0.01, 0.001, 0.005]
	Optimization algorithm	Adam
XGB	Learning rate	[0.1, 0.01, 0.001, 0.005]
	Minimum sum of instance weight (hessian) needed in a child	[1, 5, 10]
	Minimum loss reduction required to make a further partition on a leaf node of the tree	[0, 0.5, 1, 1.5, 2, 5]
	Subsample ratio of the training instances	[0.7, 0.8, 0.9, 1.0]
	Parameters for subsampling the columns	[0.3,0.4,0.5,0.6,0.7,0.8]
	Maximum depth of a tree	[3, 5, 7, 9, 10]

**Supplemental Table IV.** Result of the forth experiment (no image score) for the radiomics approach for predicting good functional outcome (mRS $\leq$ 2). All image-related scores were removed from the clinical data. The average of 5 cross validation iterations is presented. RFC, random forest classifier; SVM, support vector machine; LR, logistic regression; XGB, gradient boosting; NN, neural networks. AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.

Methods	AUC	F1-Score	Sensitivity	Specificity	PPV	NPV
<b>Clinical (no image scores)</b>						
<b>RFC</b>	0.80 (0.78-0.82)	0.68 (0.65- .72)	0.72 (0.68-0.76)	0.73 (0.71-0.75)	0.65 (0.61-0.70)	0.78 (0.76-0.81)
<b>SVM</b>	0.80 (0.78-0.82)	0.70 (0.66-0.73)	0.79 (0.73-0.84)	0.66 (0.65-0.68)	0.63 (0.58-0.67)	0.82 (0.78-0.85)
<b>LR</b>	0.80 (0.78-0.81)	0.69 (0.66-0.73)	0.77 (0.73-0.81)	0.68 (0.67-0.69)	0.63 (0.59-0.67)	0.81 (0.77-0.84)
<b>XGB</b>	0.79 (0.77-0.81)	0.68 (0.65-0.72)	0.75 (0.70-0.80)	0.69 (0.67-0.70)	0.63 (0.60-0.66)	0.79 (0.76-0.83)
<b>NN</b>	0.80 (0.78-0.81)	0.68 (0.65-0.71)	0.72 (0.68-0.76)	0.72 (0.67-0.76)	0.64 (0.59-0.70)	0.78 (0.75-0.81)
<b>Combination (no image scores)</b>						
<b>RFC</b>	0.79 (0.78-0.81)	0.67 (0.64-0.70)	0.68 (0.65-0.72)	0.75 (0.72-0.78)	0.66 (0.62-0.70)	0.77 (0.74-0.79)
<b>SVM</b>	0.79 (0.78-0.80)	0.69 (0.67-0.72)	0.77 (0.74-0.80)	0.68 (0.66-0.69)	0.63 (0.60-0.66)	0.81 (0.78-0.83)
<b>LR</b>	0.79 (0.78-0.80)	0.68 (0.65-0.71)	0.75 (0.70-0.81)	0.68 (0.66-0.70)	0.63 (0.59-0.66)	0.79 (0.76-0.83)
<b>XGB</b>	0.79 (0.77-0.81)	0.69 (0.66-0.72)	0.76 (0.74-0.79)	0.67 (0.65-0.70)	0.62 (0.58-0.67)	0.80 (0.78-0.82)
<b>NN</b>	0.72 (0.69-0.75)	0.60 (0.55-0.64)	0.60 (0.54-0.66)	0.71 (0.69-0.73)	0.59 (0.55-0.64)	0.71 (0.68-0.75)

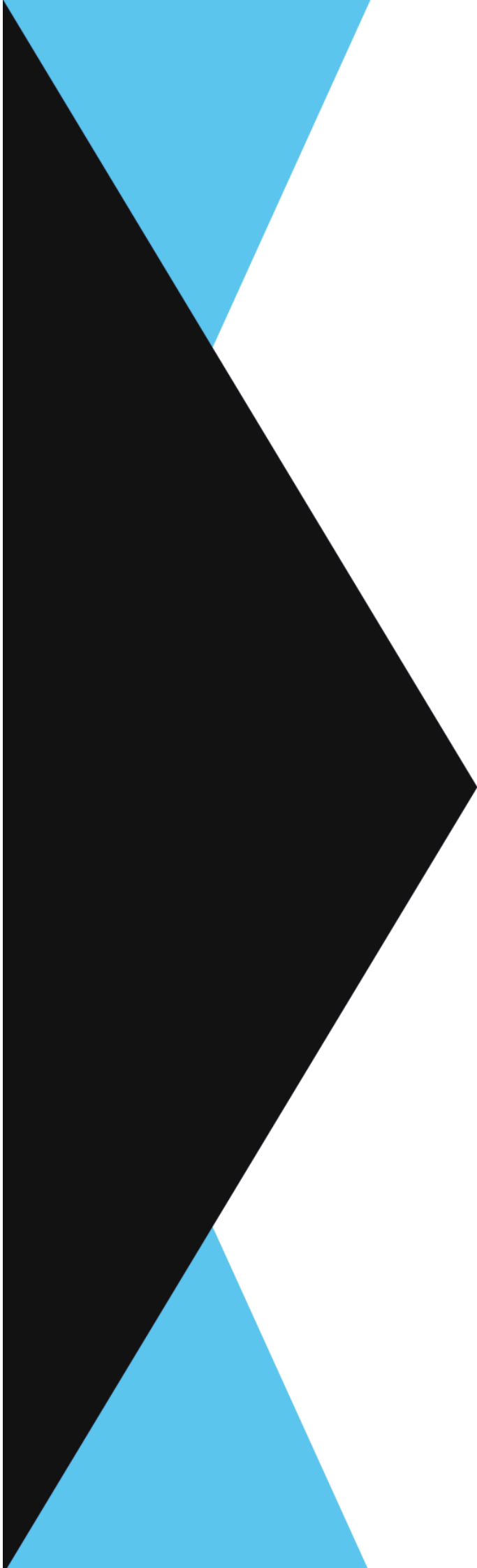
**Supplemental Table V.** Results of the forth experiment (no image score) for the radiomics approach for predicting good reperfusion (post-eTICI  $\geq 2b$ ). All image-related scores were removed from the clinical data. The average of 5 cross validation iterations is presented. RFC, random forest classifier; SVM, support vector machine; LR, logistic regression; XGB, gradient boosting; NN, neural networks. AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.

Methods	AUC	F1-Score	Sensitivity	Specificity	PPV	NPV
<b>Clinical (no image scores)</b>						
<b>RFC</b>	0.51 (0.48-0.54)	0.70 (0.67-0.73)	0.77 (0.72-0.82)	0.25 (0.21-0.30)	0.64 (0.59-0.68)	0.39 (0.33-0.46)
<b>SVM</b>	0.53 (0.51-0.55)	0.75 (0.66-0.83)	0.92 (0.72-1.13)	0.09 (0.16-0.33)	0.64 (0.60-0.67)	0.43 (0.19-0.61)
<b>LR</b>	0.53 (0.52-0.55)	0.63 (0.58-0.68)	0.61 (0.56-0.66)	0.42 (0.38-0.46)	0.64 (0.60-0.69)	0.39 (0.36-0.42)
<b>XGB</b>	0.51 (0.50-0.52)	0.64 (0.58-0.71)	0.65 (0.54-0.76)	0.39 (0.31-0.47)	0.64 (0.60-0.69)	0.40 (0.36-0.44)
<b>NN</b>	0.52 (0.49-0.55)	0.76 (0.71-0.80)	0.94 (0.83-1.05)	0.07 (0.06-0.20)	0.63 (0.59-0.67)	0.43 (0.23-0.64)
<b>Combination (no image scores)</b>						
<b>RFC</b>	0.57 (0.54-0.60)	0.74 (0.71-0.76)	0.87 (0.84-0.89)	0.17 (0.13-0.21)	0.64 (0.60-0.68)	0.42 (0.33-0.52)
<b>SVM</b>	0.56 (0.53-0.60)	0.65 (0.54-0.75)	0.66 (0.42-0.90)	0.41 (0.12-0.70)	0.66 (0.62-0.70)	0.42 (0.35-0.49)
<b>LR</b>	0.56 (0.53-0.59)	0.61 (0.59-0.64)	0.57 (0.55-0.60)	0.49 (0.44-0.54)	0.66 (0.62-0.70)	0.40 (0.35-0.46)
<b>XGB</b>	0.56 (0.54-0.59)	0.60 (0.57-0.64)	0.55 (0.52-0.59)	0.53 (0.48-0.57)	0.67 (0.63-0.70)	0.41 (0.36-0.46)
<b>NN</b>	0.52 (0.50-0.55)	0.66 (0.64-0.69)	0.70 (0.60-0.79)	0.34 (0.22-0.46)	0.64 (0.59-0.69)	0.39 (0.33-0.45)

## Supplemental References

1. Bergstra JAMESBERGSTRA J, Yoshua Bengio YOSHUABENGLIO U. Random Search for HyperParameter Optimization. *J Mach Learn Res.* 2012;13:281305.
2. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2012;12:2825–30.
3. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min.* 2016;Pages 785-794.





6



# CHAPTER 6.

## Predicting poor outcome prior to endovascular treatment in patients with acute ischemic stroke

Ramos LA, Kappelhof M, van Os HJA, Chalos V, van Kranendonk K, Kruyt ND, Roos YBWEM, van der Lugt A, van Zwam WH, van der Schaaf IC, Zwinderman AH, Strijkers GJ, van Walderveen MAA, Wermer MJH, Olabarriaga SD, Majoie CBLM, Marquering HA. Predicting poor outcome prior to endovascular treatment in patients with acute ischemic stroke.

*Front. Neurol.* 2020;11- 1215.

DOI: [10.3389/fneur.2020.580957](https://doi.org/10.3389/fneur.2020.580957)



## Abstract

*Background:* Although endovascular treatment (EVT) has greatly improved outcomes in acute ischemic stroke, still one third of patients die or remain severely disabled after stroke. If we could select patients with poor clinical outcome despite EVT, we could prevent futile treatment, avoid treatment complications, and further improve stroke care. We aimed to determine the accuracy of poor functional outcome prediction, defined as 90-day mRS $\geq$ 5, despite EVT treatment.

*Methods:* We included 1526 patients from the MR CLEAN Registry; a prospective, observational, multicenter registry of ischemic stroke patients treated with EVT. We developed machine learning prediction models using all variables available at baseline prior to treatment. We optimized the models for both maximizing the AUC reducing the number of false positives.

*Results:* From 1526 patients included, 480 (31%) of patients showed poor outcome. The highest area under the curve was 0.81 for random forest. The highest area under the precision recall curve was 0.69 for the SVM. The highest achieved specificity was 95% with a sensitivity of 34% for neural networks, indicating that all models contained false positives in their predictions. From 921 mRS 0-4 patients, 27 to 61 (3 to 6%) were incorrectly classified as poor outcome. From 480 poor outcome patients in the registry, 99 to 163 (21 to 34%) were correctly identified by the models.

*Conclusions:* All prediction models showed a high area under the curve. The best-performing models correctly identified 34% of the poor outcome patients at a cost of misclassifying 4% of non-poor outcome patients. Further studies are necessary to determine whether these accuracies are reproducible before implementation in clinical practice.

## Introduction

Over the past four years, endovascular thrombectomy (EVT) unquestionably proved its value in anterior circulation acute ischemic stroke. (1–8) Despite the encouraging results however, still approximately 30% of patients die or remain dependent of daily nursing care after EVT, making their treatment benefit essentially minimal. (5,6)

If we could reliably select patients with poor outcome after stroke despite EVT, we could spare patients a futile treatment with a needless risk of complications and enable a more efficient use of resources. (9) Unfortunately, so far, no studies have been able to definitively identify a subgroup of patients that should not be treated with EVT. (9)

In patient selection, it could be useful to predict poor outcome. Many previous studies focused on predicting functional independence after EVT. (10) However, the use of such models would raise an ethical question. If a model predicts a zero percent chance of functional independence with EVT for a patient, one might advise to not treat. Untreated, the patient likely has a worse outcome, possibly needing continuous care in a nursing home. Treated, the patient may be able to function with some assistance in daily activities. Should we not treat this patient? A more valuable argument could be a reliable prediction of death or complete dependence of continuous care, even after EVT.

Some studies, such as MR PREDICTS, used data from randomized trials to predict treatment benefit as a Rankin Scale (mRS) score shift, using ordinal logistic regression. (11) Predicting treatment benefit can be useful: if a patient is predicted to benefit from EVT in addition to regular care, one would proceed with EVT. However, data from randomized trials are necessary for such a model, since predicted outcomes need to be based on a sufficient number of patients that did or did not receive EVT without indication bias. The amount of available data from randomized trials on EVT is limited. No new data after the HERMES trials will be available to train and validate models. (5) An outcome measure that can enable long-term model improvement such as poor functional outcome could be of added value to models predicting treatment benefit.

Only a few studies have used poor outcome as their outcome measure, however, they had a limited amount of data and focused on linear classifiers. (12) Machine learning (ML) may be of added value in predicting outcome after EVT. The number of relevant prognostic factors in stroke patients is high, and their effects on outcome may be indirect, combined, or otherwise complicated. With the ability to identify relevant prognostic variables through linear and non-linear relationships, ML may have added value in poor outcome prediction.

ML belongs to the artificial intelligence domain, where algorithms are designed to automatically learn patterns from data. In the work by (10), ML methods predicted functional independence after acute ischemic stroke in a large population (1383 patients), with reasonable certainty (area under the curve [AUC] 0.79).

Since the addition of EVT to standard care, the amount of available outcome data has greatly increased, now allowing for more powerful and elaborate prediction modelling. In the current study, we aim to assess the accuracy of pre-procedural prediction of poor functional outcome after EVT using ML models, in patients from the MR CLEAN Registry.

## Methods

### Study population

We included patients from the MR CLEAN Registry, which is a prospective, observational, multicenter study, consecutively including all EVT-treated acute ischemic stroke patients in the Netherlands since the completion of the MR CLEAN trial (13) in March 2014. The MR CLEAN registry contains data from 16 centers distributed across The Netherlands. The current study is a retrospective report on patients included in the MR CLEAN Registry between March 2014 and June 2016 with intracranial proximal occlusions of the anterior arterial circulation (internal carotid artery (ICA) or internal carotid artery terminus (ICA-T), middle (M1/M2) or anterior (A1/A2) cerebral artery); aged  $\geq 18$  years; and treated in a MR CLEAN trial center. Patients were treated with intravenous thrombolysis (IVT) prior to EVT, if eligible. The central medical ethics committee of the Erasmus Medical Centre Rotterdam, the Netherlands, evaluated the study protocol and granted

permission (MEC-2014–235) to carry out the data collection as a registry. (6) The procedures followed were in accordance with institutional guidelines. Patients provided permission for study participation through an opt-out procedure. The data can be made available upon reasonable request from the MR CLEAN Registry committee (mrclean@erasmusmc.nl). All code used for the development of the models and data analysis is available at: [https://github.com/L-Ramos/MrClean\\_Poor](https://github.com/L-Ramos/MrClean_Poor).

All imaging was assessed by an independent core laboratory, composed of 21 observers (20 interventional neuro- and/or interventional radiologists and one interventional neurologist) who were blinded to all clinical findings, except for symptom side. Assessed baseline imaging modalities were non-contrast CT (dense vessel sign, ASPECTS, hemorrhage, old infarcts, leukoaraiosis), CT angiography (CTA; occlusion location, clot burden score, collateral grade), and digital subtraction angiography (DSA; successful reperfusion, defined as extended thrombolysis in infarction score 2B-3). Other imaging variables that have proven to be predictive for outcome such as stroke lesion shape and size, are difficult to observe on CT scans and were therefore, not included in our models (14).

### Study variables, Outcome, Missing data

Provided the correct methodology is used, ML methods allow the analysis of a large number of features. Therefore, we analyzed all 51 patient variables collected at baseline before treatment. Ordinal variables such as pre-stroke mRS, collaterals, ASPECTS, NIHSS, clot burden score, and Glasgow Coma Scale were treated as linear continuous scores. Some variables like time to groin puncture, despite not being readily available at baseline, can be estimated. If groin puncture is estimated to be possible within 6 hours, patients can be treated within the regular EVT time window. In addition, achievable door-groin time of <60 minutes is currently used as inclusion criterion for several acute stroke trials (such as MR CLEAN-NO IV; ISRCTN80619088). More details about the included variables, distributions and how they were included in the models are listed in Supplemental Table I.

The outcome measure of interest of this study was as poor functional outcome, defined as a modified Rankin Scale (mRS) score of  $\geq 5$  at 90 days

after stroke. Data on the mRS were collected by the MR CLEAN Registry hospitals as part of usual care. (6)

Missing baseline and outcome data (mRS, n=125 [8%]) were imputed using two approaches: a multiple imputation approach using Multiple Imputation by Chained Equations (MICE) (15), which is the most commonly used in literature (and the standard for MR CLEAN Registry-based studies) and a single imputation approach using Random-Forest Imputation (RFI) (16), which is a more recent, state-of-the-art imputation method. Variables with more than 40% missing were excluded from the analysis.

### Machine learning methods

We applied the following ML methods:

Random forest classifier (RFC) (17), an ensemble classifier that combines many decision trees trained individually. Each decision tree is trained on random samples from the dataset, which reduces the variance of the prediction without increasing the bias;

Support vector machine (SVM) (18), which separates classes by constructing hyperplanes and maximizing the margin in a multidimensional space;

Artificial neural networks (NN) (19), which is composed of many interconnected nodes arranged in layers, where information is propagated from the first input layer up to a final output layer that delivers a prediction; and

Gradient boosting (XGB) (20), which is also an ensemble classifier that uses decision trees, but instead of training the trees individually, Gradient Boosting trains the trees sequentially, gradually improving them based on the previous ones.

Logistic regression (LR), which models the probability of a binary outcome using a linear function of the predictor variables;

Since there are many ML methods described in the literature, for which learning occurs in very different ways, we selected models that differ in learning procedure to increase the chance of developing models that generalize well (21). These methods have shown state-of-the-art results in

several stroke-related applications. (10,22,23) For the Gradient Boosting method, we used the implementation from <https://github.com/dmlc/xgboost>. (20) For all the other methods, we used the implementations from *Scikit Learn* toolkit version 0.21.3. (24)

### Machine learning pipeline

We used a nested cross-validation (CV) strategy for model optimization and evaluation. In the outer CV loop, the dataset was split into ten equally sized folds. For each CV iteration, nine folds were used as training set and one was used as test set. In the inner CV loop, the training set was again divided into five folds (four used for training and one for validation), used for training the RFI imputer and determining the best hyper-parameters for all ML models. Hyper-parameters are parameters specific to each ML method. Their values cannot be automatically learned by the methods. The hyper-parameters were optimized using the random grid search function available on *Scikit Learn* (24), for maximizing the AUC. A list of the hyper-parameters used can be found in Supplemental Table II, together with a description of the optimization procedure and choice of values.

For the LR models, we used feature selection using LASSO to define a subset of relevant variables. Creating a subset avoids diluting the coefficients of the model, which can form a challenge in interpreting variable importance. (25)

Since the outcome variable was slightly imbalanced, and class imbalance can bias some classifiers, we applied balanced class weights during training of all models (24,26). Class weights change the way the loss is calculated. The individual errors are multiplied by a sample weight, which shifts the minimum of the loss function. This way, when the error is high for a sample from a less prominent class, its impact will be higher in the loss, leading to a larger penalization in the whole model We chose this approach since it has shown to work well even when class imbalanced in severe (up to thousands of times fewer samples from a given class). (26)

### Model performance

Model performance was evaluated on the testing sets.. We evaluated model performance using AUC, sensitivity (poor outcome patients correctly classified as poor outcome), specificity (percentage of non-poor outcome

patients correctly classified as non-poor outcome), positive predictive value (PPV) (predicted poor outcome patients actually having poor outcome), negative predictive value (NPV) (predicted non-poor outcome patients actually having non-poor outcome), Matthews Correlation Coefficient (MCC) (correlation coefficient between the observed and predicted classes that is robust to class imbalance) (27), and the Area under the Precision Recall curve (AUPRC). A high AUPRC relates to high precision (low false positive rate) and recall (low false negative rate), and is also a robust measure for class imbalance. (24) We built ten models for each ML method through cross-validation. Therefore, the measures were averaged over all iterations and 95% confidence intervals (CI) were computed. To limit the number of false positives (and, consequently, the risk of withholding treatment from patients who may still have good functional outcome), we optimized the predictions from the models (probability of poor outcome) to maximize specificity; above or equal to 0.95, 0.98 and 1.00, using the validation dataset to determine a threshold for the probabilities. This threshold was determined based on incremental search, by continuously increasing the threshold in 0.01 units until specificity was equal or higher than 0.95.

To assess model performance, we used Grotta bars to visualize the mRS distribution of patients that were classified by the models into poor outcome versus non-poor outcome. Per ML method, three Grotta bars were computed for a specificity threshold of 0.95, 0.98 and 1.00, to assess the impact of reducing the number of false positive predictions. Finally, we investigated the variables with the most predictive value for the best performing models (high PPV and small number of FP) using odds ratio for LR and permutation feature importance. (28) In permutation feature importance, each variable is individually shuffled before training and the decrease in accuracy (or in our case, AUC) is computed. The more the AUC decreases, the more important the variable is for the model.



## Results

### Study population

A total of 1526 patients were included (Supplemental Figure I). Mean age was 71 years, median baseline NIHSS was 14 (Table 6.1). Successful reperfusion was achieved in 863/1505 patients (57%), and 753/1092 (69%) of patients with complete post-EVT DSA runs available. At 90 days, 480 (31%) patients had a poor functional outcome (mRS 5-6), whereas 921 (61%) did not (outcome missing in n=125 [8%]).

**Table 6.1.** Baseline characteristics; overall compared to mRS 5-6 versus 0-4

Characteristics	Total study sample N=1526	mRS 5 – 6 N=480	mRS 0 – 4 N=921
Age (years) – median (IQR)	71 (60-79)	77 (69-84)	67 (55-75)
Male sex – n (%)	809 (53.0)	245 (51.0)	502 (54.5)
Diabetes – n (%)	145 (13.9)	117 (24.4)	262 (17.2)
Pre-stroke mRS - n (%)			
	0-2 1327 (86.9)	370 (77.1)	957 (91.5)
	3-5 172 (11.3)	95 (19.8)	77 (7.4)
NIHSS at baseline – median (IQR)	14 (9-18)	16 (12-20)	13 (8-16)
Systolic blood pressure (mmHg) – mean (SD)	150 (24.6)	154 (25.8)	147(23.7)
Glucose level before EVT median (IQR)	6.7 (8.0-5.9)	7.2 (8.8-6.1)	6.6 (7.8-5.8)
Intravenous alteplase – n (%)	1170 (76.7)	327 (68.1)	743 (80.7)
Onset to groin puncture time (minutes) – median (IQR)	210 (160-270)	219 (170-273)	200 (155-266)
Hyperdense artery sign- n (%)	773 (50.7)	248 (51.7)	459 (49.8)
ASPECTS subgroups-n (%)			
	0-4 95 (6.2)	39 (8.13)	51 (5.5)
	5-7 351 (23.0)	120 (25.0)	198 (21.5)
	8-10 1013 (66.4)	292 (60.8)	639 (69.4)
Occlusion location – n (%)			
	ICA-T 322 (21.1)	128 (26.7)	194 (18.6)
	M1 842 (55.2)	242 (50.4)	600 (57.4)
	M2 181 (11.9)	52 (10.8)	129 (12.3)
	Intracranial ICA 85 (5.6)	21 (4.4)	64 (6.2)
	Other (M3 or anterior) 19 (1.3)	6 (1.3)	13 (1.2)
Clot Burden Score-median (IQR)	6 (4-8)	6 (4-8)	6 (4-8)
Collateral score-n (%)			
	0 98 (6.4)	57 (11.9)	35 (3.8)
	1 467 (30.6)	188 (39.2)	246 (26.7)
	2 547 (35.8)	135 (28.1)	361 (39.2)
	3 305 (20.0)	61 (12.7)	218 (23.7)

IQR, interquartile range; mRS, modified Rankin Scale; NIHSS, National Institutes of Health stroke scale; SD, standard deviation.

### Prediction accuracy

For all models trained the best average AUC was 0.81 (Table 6.2) for NN, and the best AUPRC was 0.69 for the SVM. In the test sets, the highest PPV was 0.69 for the NN and the highest NPV was 0.87 for SVM: from all non-poor outcome predictions, 79% of the patients indeed had a non-poor outcome. All models but the SVM showed higher values of specificity than sensitivity, with 0.89 being the highest specificity (for the NN). The NN showed also the highest MCC (0.45) and LR the highest balanced accuracy (0.73) (Supplemental Table III).

**Table 6.2.** Evaluation measures in validation data for all poor outcome prediction models, trained to maximize the AUC. The average of 10 cross validation iterations is presented

Method	Specificity	Sensitivity	PPV	NPV	AUC	AUPRC
<b>RFC</b>	0.84 (0.81-0.86)	0.56 (0.51-0.62)	0.62 (0.56-0.68)	0.80 (0.78-0.83)	0.80 (0.77-0.82)	0.66 (0.61-0.72)
<b>SVM</b>	0.67 (0.61-0.72)	0.78 (0.75-0.81)	0.53 (0.48-0.57)	0.87 (0.84-0.89)	0.77 (0.74-0.76)	0.69 (0.65-0.74)
<b>NN</b>	0.89 (0.87-0.90)	0.53 (0.49-0.57)	0.69 (0.65-0.74)	0.80 (0.78-0.83)	0.81 (0.79-0.83)	0.68 (0.64-0.73)
<b>XGB</b>	0.79 (0.76-0.83)	0.63 (0.60-0.67)	0.59 (0.54-0.65)	0.82 (0.80-0.84)	0.78 (0.76-0.81)	0.64 (0.59-0.69)
<b>LR</b>	0.75 (0.73-0.78)	0.71 (0.68-0.73)	0.57 (0.53-0.62)	0.85 (0.83-0.86)	0.80 (0.78-0.82)	0.68 (0.63-0.74)

RFC, random forest classifier; SVM, support vector machine; LR, logistic regression; XGB, gradient boosting; NN, neural networks. AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.

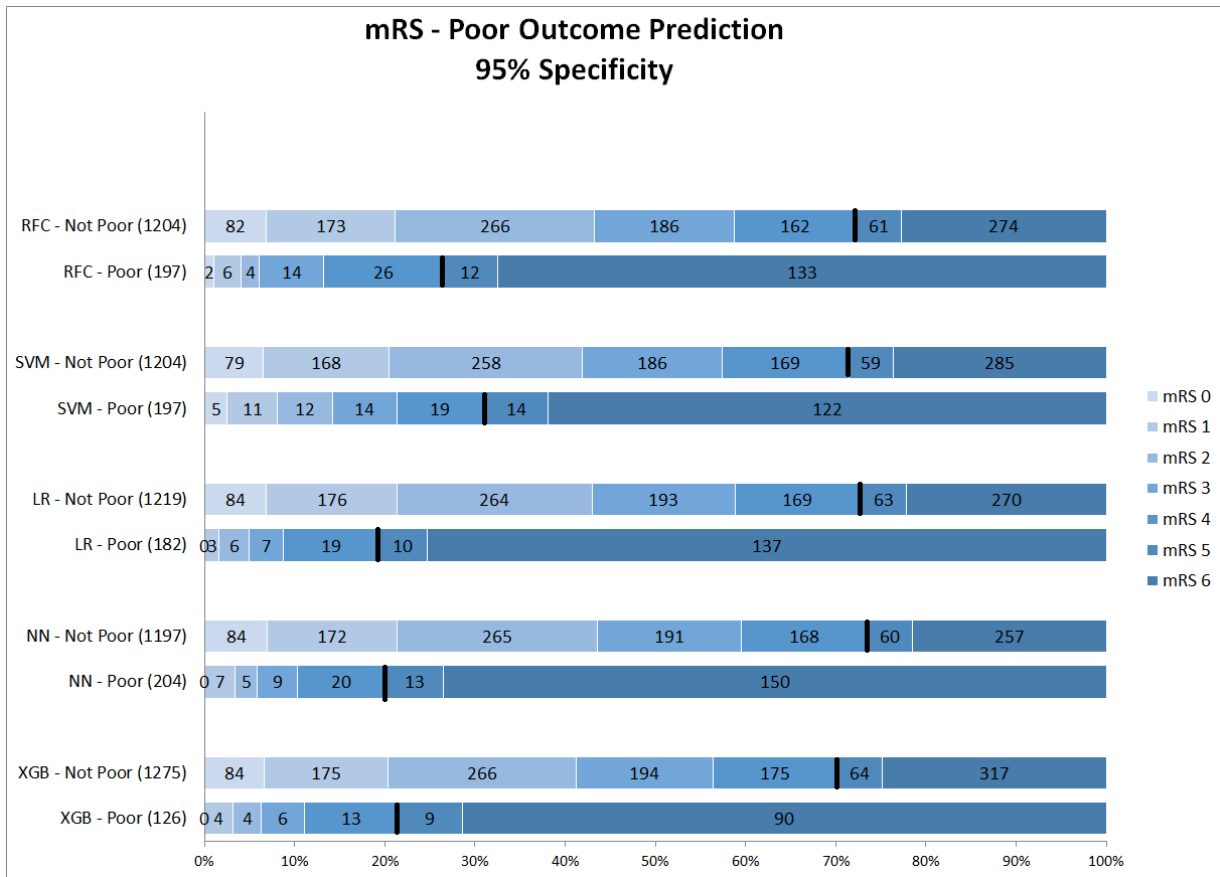
Supplemental Table IV the results for the probability threshold of 95% specificity are shown. Note that since the specificity is based on the training set, the actual specificity in the validation set is somewhat lower than 0.95. Since the probability thresholds were optimized for high specificity, values for sensitivity were low (highest 0.34 for NN), indicating a relatively high number of false negatives (poor outcome patients classified as non-poor).

NN, was considered the most accurate model, since it showed the highest PPV values (Table 6.2). For the probability threshold of 95% specificity, NN, XGB and LR showed the best PPV results, and NN and LR showed the highest AUPRC results (Supplemental Table IV). They also had the highest

NPV values among the other models. We did not find any difference between single imputation using Random Forest and multiple imputation using MICE, therefore we used Random Forest imputation as default. The results for the MICE imputation approach are shown in Supplemental Table V.

### Model performance

Figures 6.1, 6.2 and 6.3 show the mRS distribution outcome of patients classified as poor outcome and non-poor outcome in the testing data, for different specificity thresholds. For each ML method on the y-axis, we show how many patients were classified as poor and as non-poor outcome along the y-axis. Along the x-axis, the percentage of patients per mRS value is presented. In each graph, the black bar separates mRS 0-4 (non-poor outcome) from 5-6 (poor outcome). In Figure 6.1, the probability threshold was optimized to reach 95% specificity, and for some classifiers the rate of correct poor outcome prediction was higher than 80%. This is the case for LR, where from all poor outcome predictions, the total of mRS 0-4 patients is lower than 20%. However, all models still mistakenly predicted some mRS 0-4 patients as poor outcome (27 patients for the best model; less than 3% of all mRS 0-4 patients; Table 6.3). More patients were classified as non-poor outcome than poor outcome. For the NN model for example (Supplemental Figure II), 11% (163/1526) of all patients were classified as poor outcome, whereas 31% actually had a poor outcome.



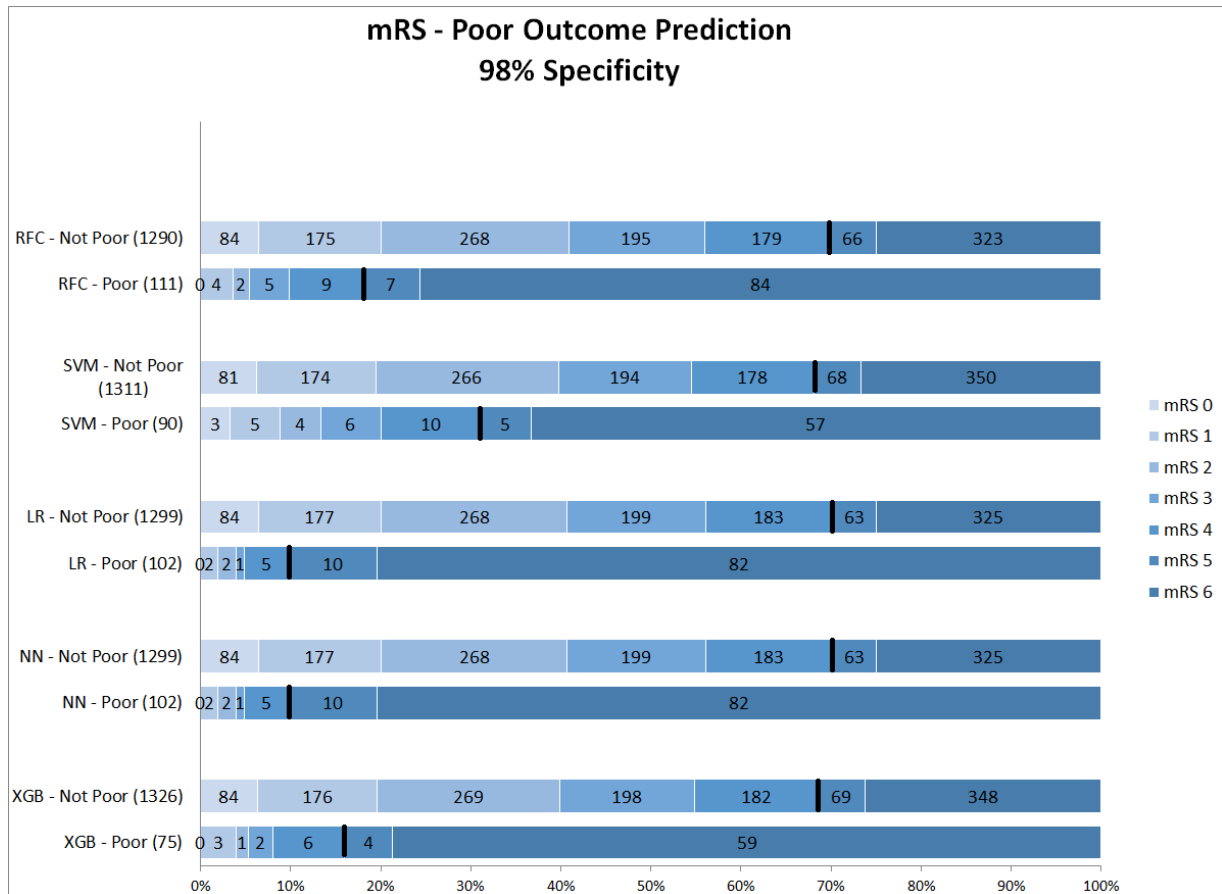
**Figure 6.1.** Distribution of mRS for the predictions of each model as poor vs non-poor outcome with 95% specificity threshold. Along the y axis, the various ML method are presented including the number of patients that were classified as poor and non-poor outcome. Along the x axis, the percentage of patients per mRS value is presented. In each graph, the black bar separates mRS 0-4 from 5. RFC, random forest classifier; SVM, support vector machine; LR, logistic regression; NN, neural network; XGB, gradient boosting. mRS, modified Rankin Scale. Numbers in bars represent absolute number of patients.

**Table 6.3.** Number of false positives (mRS 0-4 classified as poor) and true positives (mRS 5-6 classified as poor) per specificity threshold for each ML method

Method	Optimized specificity	<u>True positives</u>	<u>False positives</u>
		mRS 5-6 patients classified as poor of total mRS 5-6 patients (n=480)	mRS 0-4 patients classified as poor of total mRS 0-4 patients (n=921)
<b>RFC</b>	95%	145 (30.2%)	52 (5.6%)
	98%	91 (19.0%)	20 (2.2%)
	100%	39 (8.1%)	8 (0.9%)
<b>SVM</b>	95%	136 (28.3%)	61 (6.2%)
	98%	62 (12.9%)	28 (3.0%)
	100%	33 (6.9%)	10 (1.1%)
<b>NN</b>	95%	163 (34.0%)	41 (4.5%)
	98%	92 (19.2%)	10 (1.1%)
	100%	21 (4.4%)	2 (0.2%)
<b>XGB</b>	95%	99 (20.6%)	27 (2.8%)
	98%	63 (13.1%)	12 (1.3%)
	100%	21 (4.4%)	6 (0.7%)
<b>LR</b>	95%	147 (30.6%)	35 (3.8%)
	98%	92 (19.2%)	10 (1.1%)
	100%	23 (4.8%)	1 (0.1%)

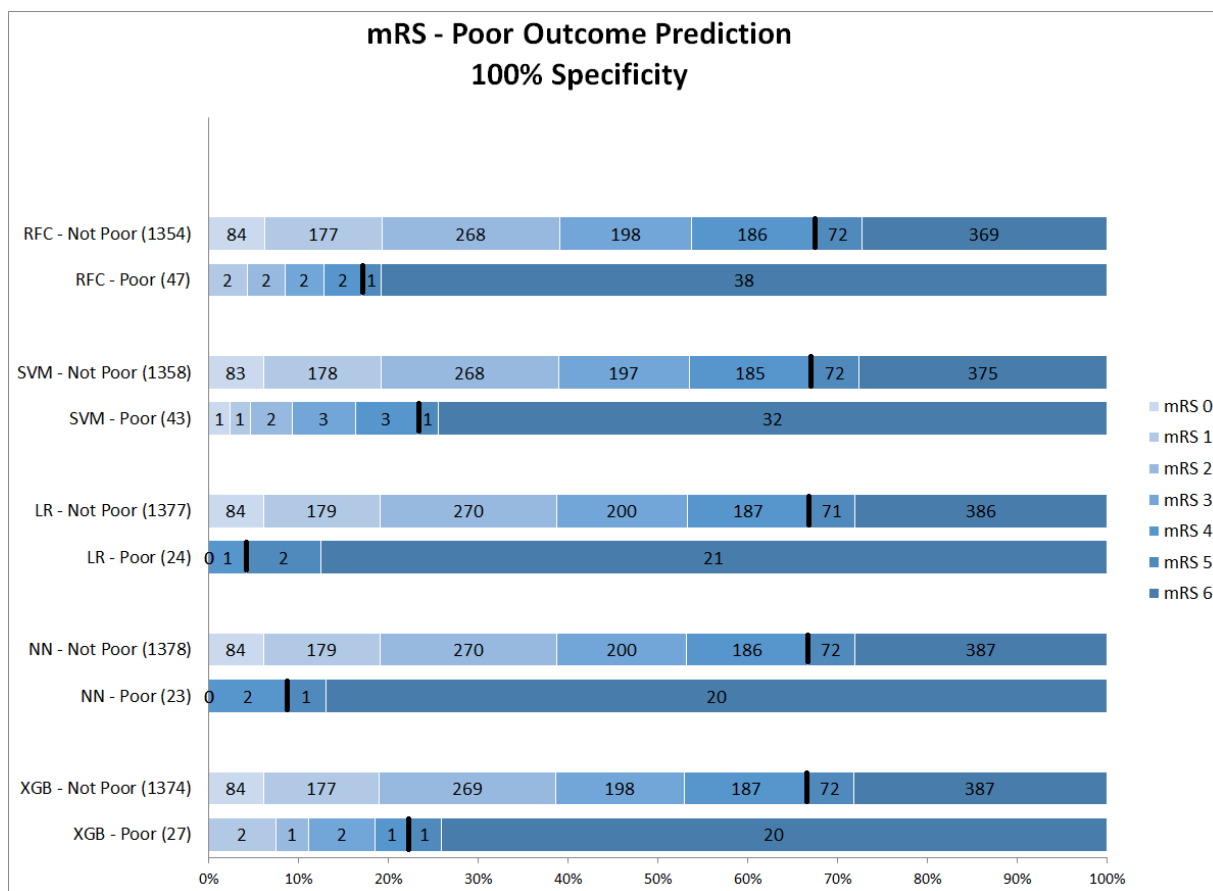
RFC, random forest classifier; SVM, support vector machine; LR, logistic regression; NN, neural network; XGB, gradient boosting. mRS, modified Rankin Scale.

Figure 6.2 shows mRS distributions for the probability threshold optimized to 98% specificity. The numbers of both correct and incorrect poor outcome predictions were reduced compared to the 95% threshold. Ten (1.1%) of mRS 0-4 patients were still misclassified as poor in the best-performing models (NN and LR); 92 poor outcome patients were correctly classified (Supplemental Figure III). Different colors are shown for each specificity value.



**Figure 6.2.** Distribution of mRS for the predictions of each model as poor vs non-poor outcome with 98% specificity threshold. Along the y axis, the various ML method are presented including the number of patients that were classified as poor and non-poor outcome. Along the x axis, the percentage of patients per mRS value is presented. In each graph, the black bar separates mRS 0-4 from 5-6. RFC, random forest classifier; SVM, support vector machine; LR, logistic regression; mRS, modified Rankin Scale; NN, neural network; XGB, gradient boosting. Numbers in bars represent absolute number of patients.

Figure 6.3 shows the mRS distribution of patients that were classified as poor outcome versus non-poor outcome in the validation data, for the probability threshold optimized to reach 100% specificity. Again, both correct and incorrect poor outcome predictions were reduced compared to the 95% and 98% thresholds. One (0.1%) patient was misclassified as poor outcome by LR, and two (0.2%) by NN (Supplemental Figure IV). However, the ability to correctly identify poor outcome patients was reduced with 8.1% (n=39) of poor outcome patients being correctly identified (RFC).



**Figure 6.3.** Distribution of mRS for the predictions of each model as poor vs non-poor outcome with 100% specificity threshold. Along the y axis, the various ML method are presented including the number of patients that were classified as poor and non-poor outcome. Along the x axis, the percentage of patients per mRS value is presented. In each graph, the black bar separates mRS 0-4 from 5-6. RFC, random forest classifier; SVM, support vector machine; LR, logistic regression; NN, neural network; XGB, gradient boosting. mRS, modified Ranking Scale. Numbers in bars represent absolute number of patients.

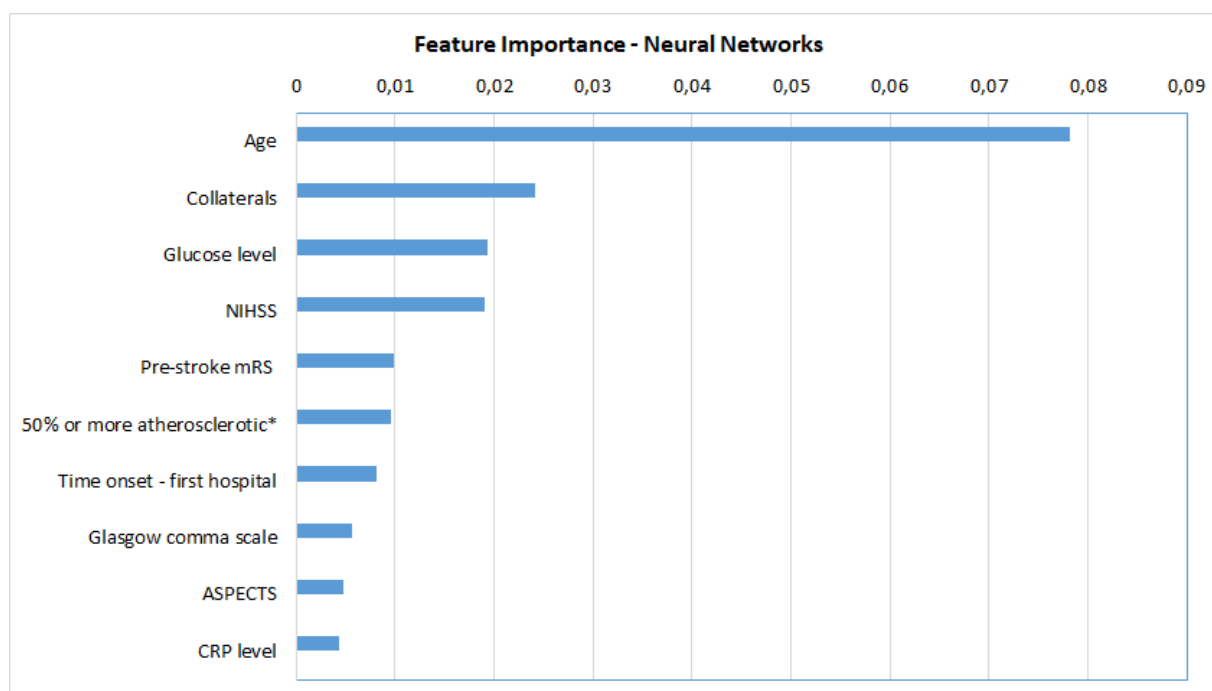
Table 6.4 shows the odds ratios for each variable included in the LR model. Baseline NIHSS, glucose level before EVT, age, 50% or more atherosclerotic stenosis at symptomatic carotid bifurcation on CTA, pre-stroke mRS, collateral score, leukoaraiosis, atrial fibrillation, and Glasgow coma scale were significantly associated with poor outcome.

**Table 6.4.** Odds ratio of each variable included in the logistic regression model

Variable	Odds Ratio (95% CI)
Age (years)	1.05 (1.04 - 1.06)
Pre-stroke mRS	1.35 (1.21 - 1.50)
Atrial fibrillation	1.37 (1.01 - 1.85)
NIHSS at baseline	1.06 (1.03 - 1.09)
Glucose level	1.16 (1.10 - 1.22)
Glasgow coma scale	0.90 (0.84 - 0.97)
Time: onset to groin puncture	1.00 (1.00 - 1.01)
50% or more atherosclerotic stenosis at symptomatic carotid bifurcation on CTA	0.61 (0.38 - 0.99)
ASPECTS on baseline NCCT	0.94 (0.88 - 1.01)
Leukoaraiosis	1.69 (1.28 - 2.24)
Collaterals	0.60 (0.51 - 0.70)

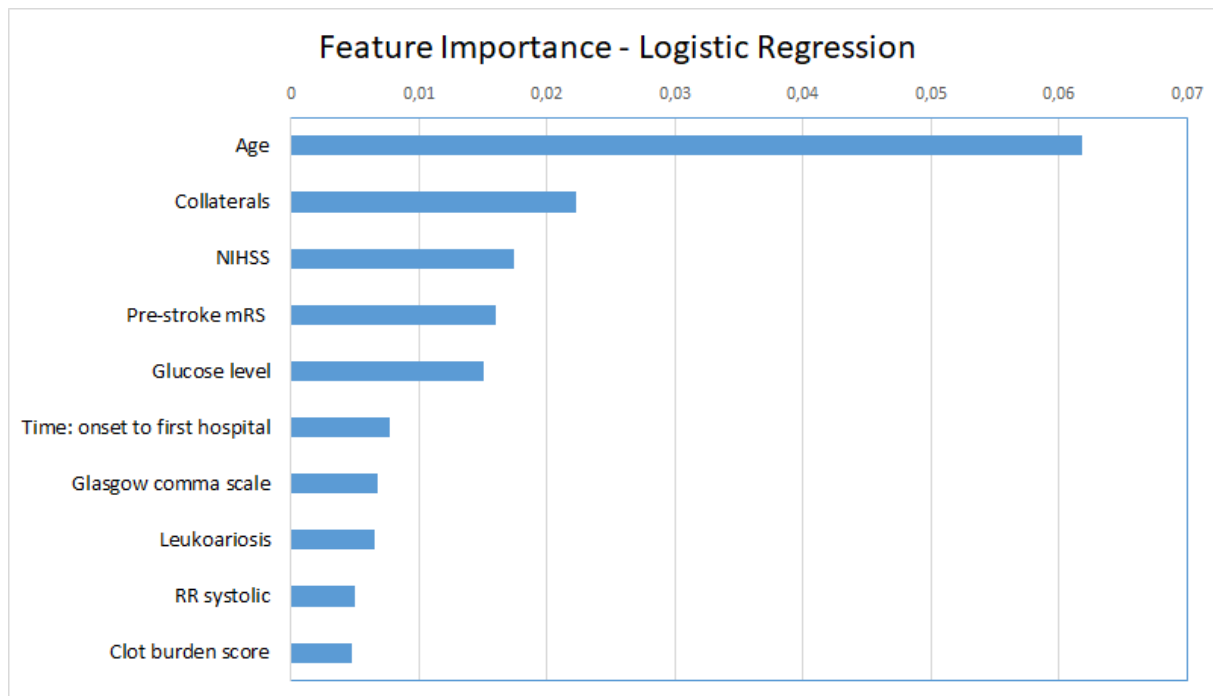
CI, confidence interval; CTA, CT angiography; mRS, modified Rankin Scale; NIHSS, National Institutes of Health stroke scale; NCCT, non-contrast CT.

For the ML models, we show the permutation feature importance for the models with the least number of FP (LR and NN – Table 6.3) in Figures 6.4 and 6.5. Permutation feature importance for the remaining ML methods is shown in Supplemental Figures V-VII. Age consistently shows the highest impact on the average AUC in all ML models. For both LR and NN, Age, collaterals, glucose level, NIHSS and pre-stroke mRS are ranked in the top 5 of the most important variables. In addition, RR diastolic at baseline and time from onset to first hospital were important variables for in other ML models.

**Figure 6.4.** Permutation feature importance for the Neural Network models.



Average impact on the AUC. \*50% or more atherosclerotic stenosis at symptomatic carotid bifurcation on CTA baseline. ASPECTS, Alberta Stroke Programme Early CT Score; CRP, C-reactive protein; mRS, modified Rankin Scale; NIHSS, National Institutes of Health stroke scale.



**Figure 6.5.** Permutation feature importance for the Logistic Regression models. Average impact on the AUC. mRS, modified Rankin Scale; NIHSS, National Institutes of Health stroke scale; RR, blood pressure (Riva-Rocci)

## Discussion

We have shown that poor outcome for acute ischemic stroke patients who were treated with EVT from the MR CLEAN Registry can be predicted with a high specificity. Although the models were optimized for high AUC and the thresholds optimized to high specificity, all models still classified some non-poor outcome patients as poor outcome, suggesting that these models are not yet accurate enough to be included in clinical practice.

To our knowledge, this is the first study to use multiple ML models and a large dataset for the prediction of poor functional outcome in acute ischemic stroke patients. Besides, our study included a larger number of variables than most stroke prediction models to date, so our study can be considered quite extensive (29). The presented accuracy was similar to the results of studies focusing on good functional outcome prediction, though a different cut-off

for dichotomization could have impacted prediction accuracy and relevance of variables. (10,11)

All models showed similar performance in terms of AUC and AUPRC, though NN was the method with the highest PPV and Specificity and was deemed the best performing model in multiple experiments. ML models can also be compared in terms of complexity (training time, number of hyper-parameters and interpretability). (30) However, this was beyond the scope of this study. Regarding the number of hyper-parameters, training time and interpretability, LR is the best method, being simpler to handle while showing accuracies similar to the other more complex models. In (30), it was shown that ML methods can greatly outperform each other in different datasets, though this was not the case in our study.

The most important features in our models were: age (in all models), collaterals, glucose level, baseline NIHSS, onset-to-first hospital time, and pre-stroke mRS (in LR and NN, the models with the lowest false-positive rates). Recent studies that used the MR CLEAN Registry dataset found similar variables with the highest relevance for functional outcome using logistic regression (mRS  $\leq 2$ ): age, NIHSS, diabetes, and time from stroke onset to treatment. (9,10,13) Interestingly, these studies also identified ASPECTS, location of occlusion, smoking, and hypertension as relevant, which were less frequently marked as important in our study. (9,10,13) This may be related to the different dichotomization of mRS we used. Alternatively, it may have to do with the manual selection of variables for these logistic regression models, as opposed to the selection of included variables by the ML methods we used. Despite none of the currently known prognostic factors being selective enough to base any EVT exclusion decision on, the more or less intuitive importance of age, baseline stroke severity, and workflow times for a patient's outcome is confirmed in both our data and the mentioned previous studies.

Regarding poor outcome prediction, some studies have identified groups of patients that show poor outcome after IVT regardless of reperfusion using diffusion-weighted MRI and CT perfusion respectively. (31,32)

## Strengths and limitations

Strengths of our study include the large sample size and heterogeneity (coming from multi-centers) of the data, which includes patients from all over the Netherlands. One of the possible downsides of a heterogeneous dataset is that the models could learn the differences between centers instead of focusing on the task at hand (predicting poor outcome). Nevertheless, we made sure that no variables related to the individual centers were included and shuffled the dataset to prevent pre-determined patient clusters. Despite this downside, the benefits of having a heterogeneous dataset outweigh this risk since we aim to develop models on data that is closer to the clinical practice setting.

Furthermore, we explored distinct state-of-the-art ML methods and optimized their hyper-parameters using an inner CV loop, while testing the optimized model on the test sets in the outer CV, which helps to prevent overoptimistic results, increasing stability and reliability. We did not separate a unique test set due to risk of, by change, separating a dataset with easier or harder samples. We used several evaluation measures that allow the models to be assessed from different points of view, highlighting their differences. Our results show that there is little difference in AUC values between models. By using other measures, such as PPV, differences in performance between models became clearer.

Some limitations to the current study should be noted. Even though we used imputation to account for missing data, a bias in the imputed values can never fully be excluded, since the estimates are always based on the available data. No difference between imputation using Random Forest and imputation using MICE was found. This can be due to the fact that the disadvantages of single imputation are mostly relevant in small datasets (with less than 100 events), which is not the case in the MR CLEAN Registry. (33) Despite MICE being a more common imputation approach, RFI imputation is often more efficient than MICE as shown in previous studies (34), and we therefore, only present the results for this approach. The large number of variables included can also be a limitation since some of the variables are not readily available or easily assessed and its assessment may delay treatment decision. However, all variables included can be derived before treatment decision (either by local

radiologists or automated tooling). Another limitation lies in the models' performance. The number of patients classified as poor outcome became very low when specificity was set to a very high value, and models still had false positives. Furthermore, we used class-weights to deal with data imbalance, since other approaches, such as under-sampling, would lead to a distribution that is not realistic when compared to the real-life scenario of acute ischemic stroke. Finally, we did not use test sets during cross-validation or imputation, preventing information leakage between datasets, which could lead to overoptimistic results.

Part of the goals of this study was to study to what extent the ratio of correctly and falsely classified patients was with ML models. In the prediction of poor outcome, high specificity and PPV are important to avoid withholding treatment from patients that may still have a non-poor outcome after EVT. The ML models investigated in the current study had relatively high AUC, PPV, and specificity, although not all patients were correctly classified, even with a specificity threshold of 100%.

The ML methods applied in this study highlighted the relevance of several baseline factors in the prediction of poor functional outcome. For future research datasets, inclusion of variables such as glucose level should be considered. In daily practice, knowledge of the relevance of these variables could support decision making by clinicians, when combined with other relevant factors such as time from symptom onset and the patient's or family's wishes. Although the prognostic models included many baseline characteristics, other data of prognostic relevance derived from CTP imaging were not included because these were not commonly available in the data from the MR CLEAN Registry. The inclusion of these parameters have the potential to improve prediction in future studies. Besides, the more extensive follow-up NIHSS could be used to define poor functional outcome in future studies. Furthermore, for future research, ML models could be created using the raw imaging data (CT or CT angiography or both) and combined with the models created in this study (35–38). However, the large number of data points has to be taken into account when developing such approaches, since imaging data is often of high dimensionality, and medical datasets have often a very limited number of samples.

Finally, we used poor outcome as our primary outcome. Poor functional outcome could be a valuable outcome measure for further studies, since the certainty of death or severe disability even after EVT could, to our expectations, form a relatively solid, ethically justifiable ground to refrain from EVT. That way, rates of futile treatment could be lowered. Poor outcome prediction may be useful as an outcome measure as an addition to the prediction of EVT benefit (mRS shift), since it does not require data from randomized trials, and can hence be used to train models on future new data.

## Conclusion

Poor outcome can be predicted with high specificity, though all of the prediction models incorrectly classified some patients as poor outcome. The percentage of misclassified non-poor outcome patients was low, while more than one third of the poor-outcome patients were correctly identified. However, lowering false-positive rates came at the cost of decreased sensitivity. It has to be studied further whether these accuracies are reproducible before implementation in clinical practice could be considered, or could be improved further. Age, NIHSS, baseline glucose levels, pre-stroke mRS and collaterals were consistently ranked as important variables in all prediction methods.

## References

1. Goyal M, Demchuk AM, Menon BK, Eesa M, Rempel JL, Thornton J, et al. Randomized assessment of rapid endovascular treatment of ischemic stroke. *N Engl J Med.* 2015;372 (11):1019–30.
2. Jovin TG, M.D. AC, M.D. EC, Ph.D., María A. de Miquel M.D. CAM, M.D. AR, M.D. LSR, et al. Thrombectomy within eight hours after symptom onset in ischemic stroke. *N Engl J Med.* 2016;22 (1):36.
3. Saver JL, Goyal M, Bonafe A, Diener H-C, Levy EI, Pereira VM, et al. Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke. *N Engl J Med.* 2015;372 (24):2285–95.
4. Campbell BCV, Mitchell PJ, Kleinig TJ, Dewey HM, Churilov L, Yassi N, et al. Endovascular Therapy for Ischemic Stroke with Perfusion-Imaging Selection. *N Engl J Med.* 2015;372 (11):1009–18.
5. Goyal M, Menon BK, Van Zwam WH, Dippel DWJ, Mitchell PJ, Demchuk AM, et al. Endovascular thrombectomy after large-vessel ischaemic stroke: A meta-analysis of individual patient data from five randomised trials. *Lancet.* 2016;387 (10029):1723–31.

6. Jansen IGH, Mulder MJHL, Goldhoorn RJB. Endovascular treatment for acute ischaemic stroke in routine clinical practice: Prospective, observational cohort study (MR CLEAN Registry). *BMJ*. 2018;360.
7. Muir KW, Ford GA, Messow CM, Ford I, Murray A, Clifton A, et al. Endovascular therapy for acute ischaemic stroke: The Pragmatic Ischaemic Stroke Thrombectomy Evaluation (PISTE) randomised, controlled trial. *J Neurol Neurosurg Psychiatry*. 2017;88 (1):38–44.
8. Bracard S, Ducrocq X, Mas JL, Soudant M, Oppenheim C, Moulin T, et al. Mechanical thrombectomy after intravenous alteplase versus alteplase alone after stroke (THRACE): a randomised controlled trial. *Lancet Neurol*. 2016;15 (11):1138–47.
9. Goyal M, Almekhlafi MA, Cognard C, McTaggart R, Blackham K, Biondi A, et al. Which patients with acute stroke due to proximal occlusion should not be treated with endovascular thrombectomy? *Neuroradiology*. 2018;3–8.
10. Van Os HJA, Ramos LA, Hilbert A, Van Leeuwen M, Van Walderveen MAA, Kruyt ND, et al. Predicting outcome of endovascular treatment for acute ischemic stroke: Potential value of machine learning algorithms. *Front Neurol*. 2018;9 (SEP):1–8.
11. Venema E, Mulder MJHL, Roozenbeek B, Broderick JP, Yeatts SD, Khatri P, et al. Selection of patients for intra-arterial treatment for acute ischaemic stroke: Development and validation of a clinical decision tool in two randomised trials. *BMJ*. 2017;357.
12. Sarraj A, Albright K, Barreto AD, Boehme AK, Sitton CW, Choi J, et al. Optimizing prediction scores for poor outcome after intra-arterial therapy in anterior circulation acute ischemic stroke. *Stroke*. 2013;44 (12):3324–30.
13. Berkhemer OA, Fransen PSS, Beumer D, van den Berg LA, Lingsma HF, Yoo AJ, et al. A Randomized Trial of Intraarterial Treatment for Acute Ischemic Stroke. *N Engl J Med*. 2015;372 (1):11–20.
14. Frindel C, Rouanet A, Giacalone M, Cho TH, Østergaard L, Fiehler J, et al. Validity of Shape as a Predictive Biomarker of Final Infarct Volume in Acute Ischemic Stroke. *Stroke*. 2015;46 (4):976–81.
15. Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate imputation by chained equations in R. *J Stat Softw* [Internet]. VV (II). Available from: <http://www.jstatsoft.org/>
16. Stekhoven DJ, Bühlmann P. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28 (1):112–8.
17. Breiman L. Random forests. *Mach Learn*. 2001;45 (1):5–32.
18. Cortes C, Vapnik V. Support-Vector Networks. *Mach Learn*. 1995;
19. Bishop CM. *Pattern Recognition and Machine Learning*. Information Science and Statistics. 2006.
20. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min*. 2016;Pages 785-794.
21. Fernández-Delgado M, Cernadas E, Barro S, Amorim D, Amorim Fernández-Delgado D. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *J Mach Learn Res*. 2014;15:3133–81.
22. Asadi H, Dowling R, Yan B, Mitchell P. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS One*. 2014;9 (2):14–9.

23. Monteiro MAB, Fonseca AC, Freitas AT, Pinho e Melo T, Francisco AP, Ferro JM, et al. Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients. *IEEE/ACM Trans Comput Biol Bioinforma*. 2018;5963 (c):1–8.
24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2012;12:2825–30.
25. Ranganathan P, Pramesh C, Aggarwal R. Common pitfalls in statistical analysis: Logistic regression. *Perspect Clin Res*. 2017;
26. King G, Zeng L. Logistic regression in rare events data. *J Stat Softw*. 2003;8:137–63.
27. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA - Protein Struct*. 1975;
28. Yang JB, Shen KQ, Ong CJ, Li XP. Feature selection for MLP neural network: The use of random permutation of probabilistic outputs. *IEEE Trans Neural Networks*. 2009;20 (12):1911–22.
29. Venema E, Mulder MJHL, Roozenbeek B, Broderick JP, Yeatts SD, Khatri P, et al. Selection of patients for intra-arterial treatment for acute ischaemic stroke: Development and validation of a clinical decision tool in two randomised trials. *BMJ*. 2017;357.
30. Fatima M, Pasha M. Survey of Machine Learning Algorithms for Disease Diagnostic. *J Intell Learn Syst Appl*. 2017;09 (01):1–16.
31. Mlynash M, Lansberg MG, De Silva DA, Lee J, Christensen S, Straka M, et al. Refining the definition of the malignant profile: Insights from the DEFUSE-EPITHET pooled data set. *Stroke*. 2011;42 (5):1270–5.
32. Nakatsukasa K, Kamura T, Brodsky JL. Patients With the Malignant Profile Within 3 Hours of Symptom Onset Have Very Poor Outcomes After Intravenous Tissue-Type Plasminogen Activator Therapy. *Curr Opin Cell Biol*. 2014;43 (9):82–91.
33. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. *Int Stat Rev*. 2009;
34. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *Am J Epidemiol*. 2014;179 (6):764–74.
35. Ramos LA, Van Der Steen WE, Sales Barros R, Majoie CBLM, Van Den Berg R, Verbaan D, et al. Machine learning improves prediction of delayed cerebral ischemia in patients with subarachnoid hemorrhage. *J Neurointerv Surg*. 2018;1–7.
36. Hilbert A, Ramos LA, van Os HJA, Olabbariaga SD, Tolhuisen ML, Wermer MJH, et al. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. *Comput Biol Med [Internet]*. 2019;115 (June):103516. Available from: <https://doi.org/10.1016/j.combiomed.2019.103516>
37. Choi Y, Kwon Y, Lee H, Kim BJ, Paik MC, Won JH. Ensemble of deep convolutional neural networks for prognosis of ischemic stroke. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2016.

38. Bacchi S, Zerner T, Oakden-Rayner L, Kleinig T, Patel S, Jannes J. Deep Learning in the Prediction of Ischaemic Stroke Thrombolysis Functional Outcomes: A Pilot Study. *Acad Radiol.* 2020;27 (2):e19–23.



## Supplemental material

Supplemental Table I. Details of included variables

Name	Occurrence (%) N=1526	Missing n (%)	Analyzed as
<b>Previous stroke</b>		9 (1)	cat
<b>0 – no</b>	1264 (83)		
<b>1 – yes</b>	253 (17)		
<b>Myocardial infarction</b>		31 (2)	cat
<b>0 – no</b>	1262 (83)		
<b>1 – yes</b>	233 (15)		
<b>Peripheral arterial disease</b>			30 (2)
<b>0 – no</b>	1358 (89)		
<b>1 – yes</b>	138 (9)		
<b>Diabetes</b>		9 (1)	cat
<b>0 – no</b>	1255 (82)		
<b>1 – yes</b>	262 (17)		
<b>Hypertension</b>		19 (1)	cat
<b>1 – yes</b>	765 (50)		
<b>0 – no</b>	742 (49)		
<b>Atrial fibrillation</b>		22 (1)	cat
<b>0 – no</b>	1169 (77)		
<b>1 – yes</b>	335 (22)		
<b>Hypercholesterolemia</b>		49 (3)	cat
<b>0 – no</b>	1035 (68)		
<b>1 – yes</b>	442 (29)		
<b>Antiplatelet use</b>		19 (1)	cat
<b>0 – no</b>	1001 (66)		
<b>1 – yes</b>	506 (33)		
<b>DOAC use</b>		26 (2)	cat
<b>0 – no</b>	1463 (96)		
<b>1 – yes</b>	37 (2)		
<b>Coumarin use</b>		11 (1)	cat
<b>0 – no</b>	1321 (87)		
<b>1 – yes</b>	194 (13)		
<b>Heparin use</b>		19 (1)	cat
<b>0 – no</b>	1452 (95)		
<b>1 – yes</b>	55 (4)		
<b>Blood pressure medication</b>			28 (2)
<b>1 – yes</b>	781 (51)		
<b>0 – no</b>	717 (47)		

Supplemental Table I (continued)

<b>Statin use</b>		32 (2)	cat
<b>0 – no</b>	958 (63)		
<b>1 – yes</b>	536 (35)		
<b>HAS on baseline NCCT</b>		87 (6)	cat
<b>1 – yes</b>	773 (51)		
<b>0 – no</b>	666 (44)		
<b>Relevant (new) ischemia / hypodensity</b>			113 (7)
<b>1 – yes</b>	928 (61)		
<b>0 – no</b>	485 (32)		
<b>Hemorrhagic transformation</b>			95 (6)
<b>0 – no</b>	1400 (92)		
<b>1 – yes</b>	31 (2)		
<b>Leukoariosis</b>		87 (6)	cat
<b>0 – no</b>	941 (62)		
<b>1 – yes</b>	498 (33)		
<b>Old infarcts in same ASPECTS region?</b>			76 (5)
<b>0 – no</b>	1247 (82)		
<b>1 – yes</b>	203 (13)		
<b>Intracranial atherosclerosis on CTA scored by core lab</b>			91 (6)
<b>1 – yes</b>	853 (56)		
<b>0 – no</b>	582 (38)		
<b>Sex</b>			cat
<b>Male</b>	809 (53)		
<b>Female</b>	717 (47)		
<b>Most proximal occlusion segment on CTA scored by core lab, based on CBS</b>			68 (4)
<b>Distal M1</b>	471 (31)		
<b>Proximal M1</b>	371 (24)		
<b>ICA-T</b>	322 (21)		
<b>M2</b>	181 (12)		
<b>Intracranial ICA</b>	85 (6)		
<b>None</b>	13 (1)		
<b>M3</b>	9 (1)		
<b>A2</b>	3 (0)		
<b>A1</b>	3 (0)		
<b>Smoking</b>		348 (23)	cat
<b>0 – no</b>	827 (54)		
<b>1 – yes</b>	351 (23)		
<b>Inclusion on weekday or weekend</b>			
<b>0 – weekday</b>	1133 (74)		

Supplemental Table I (continued)

<b>1 – weekend</b>	393 (26)	
<b>Admission between 17.00-08.00 (weekday)/ weekend or holiday. Based on ER time.</b>		
<b>1 – office hours</b>	982 (64)	
<b>0 – outside office hours</b>	544 (36)	
<b>Transfer from other hospital</b>		
<b>1 – transfer</b>	822 (54)	
<b>0 – no transfer</b>	704 (46)	
<b>Intravenous alteplase treatment</b>		3 (0)
<b>1 – yes</b>	1170 (77)	
<b>0 – no</b>	353 (23)	
<b>No abnormalities at symptomatic carotid bifurcation on CTA baseline by core</b>		
		264 (17)
<b>0 – no abnormalities</b>	943 (62)	
<b>1 – any abnormalities</b>	319 (21)	
<b>50% or more atherosclerotic stenosis at symptomatic carotid bifurcation on CTA baseline</b>		
		264 (17)
<b>0 – no</b>	1140 (75)	
<b>1 – yes</b>	122 (8)	
<b>Atherosclerotic occlusion at symptomatic carotid bifurcation on CTA baseline by core lab</b>		
		264 (17)
<b>0 – no</b>	1132 (74)	
<b>1 – yes</b>	130 (9)	
<b>Floating thrombus at symptomatic carotid bifurcation on CTA baseline by core lab</b>		
		264 (17)
<b>0 – no</b>	1241 (81)	
<b>1 – yes</b>	21 (1)	
<b>Pseudo-occlusion at symptomatic carotid bifurcation on CTA baseline by core lab</b>		
		264 (17)
<b>0 – no</b>	1180 (77)	
<b>1 – yes</b>	82 (5)	
<b>Carotid dissection at symptomatic carotid bifurcation on CTA baseline by core lab</b>		
		264 (17)
<b>0 – no</b>	1206 (79)	
<b>1 – yes</b>	56 (4)	
<b>Occlusion side on CTA scored by core lab</b>		
<b>Left hemisphere</b>	820 (54)	
<b>Right hemisphere</b>	694 (45)	
<b>Neither</b>	12 (1)	
<b>In-hospital stroke</b>	525 (34)	cat

Supplemental Table I (continued)			
	<b>0 – no</b>	857 (56)	
	<b>1 – yes</b>	144 (9)	
<b>Contraindications for IVT</b>			12 (1)
	<b>0 – no</b>	1178 (77)	
	<b>1 – yes</b>	336 (22)	
<b>Second occlusion in other territory present on CTA scored by core lab</b>			479 (31)
	<b>0 – no</b>	822 (54)	
	<b>1 – yes</b>	225 (15)	
<b>Collateral score on CTA scored by core lab</b>			109 (7) cont
	<b>100% of occluded area</b>	305 (20)	
	<b>&gt;50% but less &lt;100% filling &lt;50%</b>	547 (36)	
	<b>Absent collaterals</b>	467 (31)	
		98 (6)	
<b>Pre-stroke mRS</b>			27 (2) cont
	<b>0</b>	1017 (67)	
	<b>1</b>	195 (13)	
	<b>2</b>	115 (8)	
	<b>3</b>	98 (6)	
	<b>4</b>	62 (4)	
	<b>5</b>	12 (1)	
<b>90-day mRS</b>			125 (8) cat
	<b>6</b>	407 (27)	
	<b>2</b>	270 (18)	
	<b>3</b>	200 (13)	
	<b>4</b>	188 (12)	
	<b>1</b>	179 (12)	
	<b>0</b>	84 (6)	
	<b>5</b>	73 (5)	
<b>ASPECTS baseline scored by core lab – median (IQR)</b>			9 (7 - 10) 67 (4) cont
<b>CBS at baseline – median (IQR)</b>			6 (4 - 8) 255 (17) cont
<b>NIHSS at baseline – median (IQR)</b>			16 (11 - 20) 30 (2) cont
<b>Glucose level at baseline – median (IQR)</b>			7 (6 - 8) 173 (11) cont

Supplemental Table I (continued)

<b>RR systolic at baseline – median (IQR)</b>	150 (131 - 165)	43 (3)	cont
<b>RR diastolic at baseline – median (IQR)</b>	80 (70 - 91)	48 (3)	cont
<b>INR at baseline – median (IQR)</b>	1 (1 - 1)	276 (18)	cont
<b>Thrombocyte count at baseline – median (IQR)</b>	236 (194 - 290)	189 (12)	cont
<b>CRP level at baseline – median (IQR)</b>	5 (2 - 11)	307 (20)	cont
<b>Age – median (IQR)</b>	71 (60 - 79)	0 (0)	cont

A1, first segment of anterior cerebral artery; ASPECTS, Alberta stroke programme early CT score; cat, categorical; CBS, clot burden score; cont, continuous; CRP, C-reactive protein; CTA, CT angiography; DOAC, direct oral anticoagulant; ER, emergency room; HAS, hyperdense artery sign; IQR, interquartile range; M1/M2/M3, first/second/third segment of middle cerebral artery; mRS, modified Rankin Scale; NCCT, non-contrast CT; NIHSS, National Institutes of Health stroke scale; RR, blood pressure (Riva-Rocci).

## Hyper-parameter optimization

In Supplemental Table II, we present the range of values used for hyper-parameter optimization. We strived to include the largest range of values possible for all hyper-parameters while keeping the search computationally efficient. For example, for the neural network architecture, we started with a small number of hidden layers with fewer nodes and gradually increased to deeper networks with more nodes per layers. We selected this approach since (1), has shown that random grid search outperforms normal grid search and manual search, and to make the results between models more comparable since they went through the same optimization pipeline.

Values in bold were the ones selected during grid search. Some chosen values, like the number of trees in a Random Forest (RFC), are in the extreme of the range. For the number of trees, since the Random Forest is an ensemble classifier, the more trees the higher the accuracy. However, the benefit becomes smaller as the number of trees grows, while the computation time continuously increases. The chosen number of trees is already quite extreme, and the gain from adding more trees is minimum, especially because many of the trees will be quite similar given the limited number of samples available. For tree depth in the Gradient boosting (XGB), the deeper the tree, the higher the risk of overfitting, therefore we set a maximum of 10 to prevent overfitting. (2,3)

**Supplemental Table II.** Hyper-parameters used for optimizing the Machine Learning models using grid-search

<b>Classifier</b>	<b>Parameter Name</b>	<b>Parameter Value</b>
RFC	Number of Trees	[100,200,400,600,800,1000,1200, <b>1400</b> ]
	Max features for split	<b>auto</b> , sqrt and log2
	Max depth of trees	[10,20,30,40, 50, 60, <b>70</b> , 80, 90, 100, None]
	Quality of split	<b>Gini</b> or Entropy
	Minimum number of samples required to split an internal node	[2, <b>4</b> ,6,8]
	Minimum number of samples required to be at a leaf node	[2,4,6,8,10]
SVM	Kernel type	Linear, <b>Radial basis function</b> , Polynomial
	Penalty parameter C	[0.001, 0.01, 0.1, 1, 10, <b>100</b> ]
	Kernel coefficient $\gamma$ (gamma)	[1, <b>0.1</b> , 0.01, 0.001, 0.0001]
	Degree of the Polynomial kernel	[1,2,3,4,5, <b>6</b> ]
LR	Regularization	[0.001, <b>0.01</b> , 0.1, 1, 10, 100]
	Optimization algorithm	[newton-cg, lbfgs, liblinear, sag, <b>saga</b> ]
NN	Hidden Layer sizes	[90,180,90], [ <b>90,120,90</b> ], [90,90], [90,180], [90], [180]
	Activation	ReLU, <b>logistic</b>
	Regularization parameter	[0.1, <b>0.01</b> , 0.001, 0.0001]
	Batch size	[32, <b>64</b> , 128]
	Learning rate	[0.01, <b>0.001</b> , 0.005]
	Optimization algorithm	<b>Adam</b>
XGB	Learning rate	[ <b>0.1</b> , 0.01, 0.001, 0.005]
	Minimum sum of instance weight (hessian) needed in a child	[ <b>1</b> , 5, 10]
	Minimum loss reduction required to make a further partition on a leaf node of the tree	[0, <b>0.5</b> , 1, 1.5, 2, 5]
	Subsample ratio of the training instances	[ <b>0.7</b> , 0.8, 0.9, 1.0]
	Parameters for subsampling the columns	[ <b>0.3</b> ,0.4,0.5,0.6,0.7,0.8]
	Maximum depth of a tree	[3, 5, <b>7</b> , 9, <b>10</b> ]

Values in bold indicate hyper-parameters chosen by the best model.

**Supplemental Table III.** Extra evaluation measures in the testing data for all poor outcome prediction models, trained to maximize the AUC. The average of 10 cross validation iterations is presented.

<b>Method</b>	<b>Balanced Accuracy</b>	<b>MCC</b>
RFC	0.70 (0.68-0.72)	0.41 (0.37-0.45)
SVM	0.72 (0.70-0.74)	0.42 (0.38-0.46)
NN	0.71 (0.69-0.73)	0.45 (0.41-0.50)
XGB	0.71 (0.69-0.73)	0.42 (0.37-0.46)
LR	0.73 (0.71-0.75)	0.44 (0.40-0.48)

RFC, random forest classifier; SVM, support vector machine; LR, logistic regression; XGB, gradient boosting; NN, neural networks. AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.

**Supplemental Table IV.** Evaluation measures in the testing data for all poor outcome prediction models, with the probability threshold optimized to 95% specificity. The average of 10 cross validation iterations is presented.

<b>Method</b>	<b>Specificity</b>	<b>Sensitivity</b>	<b>PPV</b>	<b>NPV</b>	<b>Balanced Accuracy</b>	<b>MCC</b>	<b>AUPRC</b>
RFC	0.95 (0.93-0.96)	0.31 (0.23-0.39)	0.71 (0.63-0.79)	0.74 (0.71-0.78)	0.63 (0.59-0.67)	0.34 (0.25-0.42)	0.62 (0.56-0.68)
SVM	0.93 (0.89-0.97)	0.27 (0.20-0.42)	0.63 (0.57-0.66)	0.73 (0.69-0.77)	0.56 (0.51-0.61)	0.16 (0.07-0.26)	0.61 (0.53-0.69)
NN	0.95 (0.94-0.97)	0.34 (0.29-0.39)	0.77 (0.70-0.84)	0.77 (0.70-0.84)	0.65 (0.62-0.67)	0.39 (0.34-0.44)	0.66 (0.62-0.70)
XGB	0.97 (0.95-0.99)	0.21 (0.16-0.26)	0.79 (0.71-0.87)	0.79 (0.71-0.87)	0.59 (0.58-0.61)	0.30 (0.27-0.33)	0.63 (0.59-0.66)
LR	0.96 (0.94-0.97)	0.31 (0.26-0.37)	0.78 (0.71-0.86)	0.78 (0.71-0.86)	0.63 (0.61-0.66)	0.38 (0.33-0.42)	0.66 (0.62-0.70)

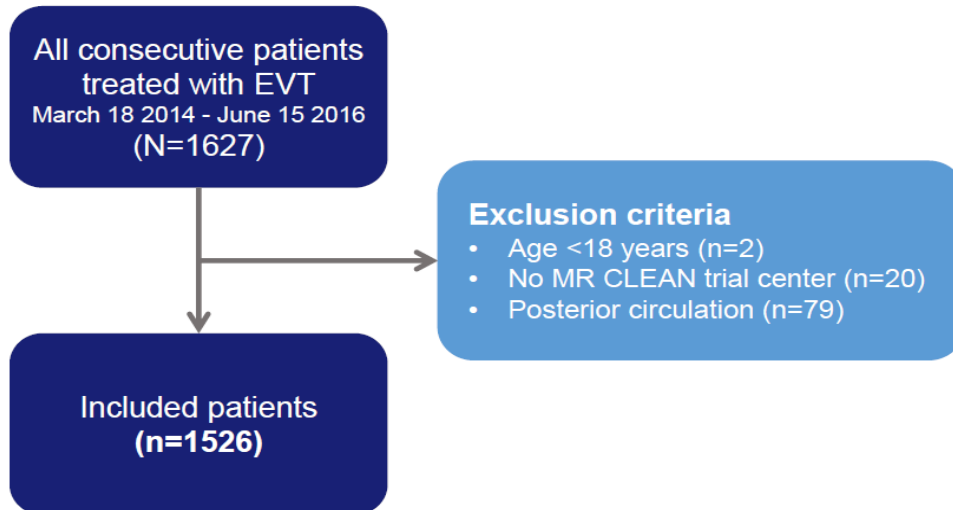
RFC, random forest classifier; SVM, support vector machine; LR, logistic regression; XGB, gradient boosting; NN, neural networks. AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.

**Supplemental Table V.** Evaluation measures in the testing data for all poor outcome prediction models, using the RFI and MICE imputation approaches. The average of 10 cross validation iterations is presented

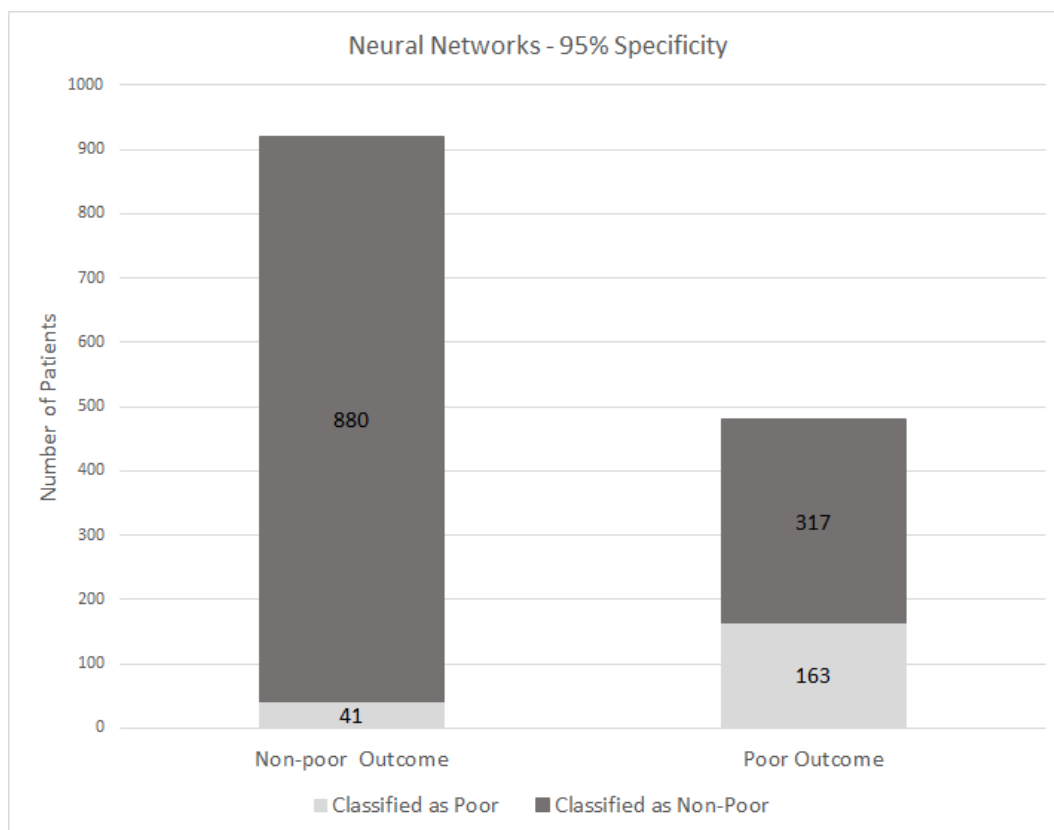
<b>Method</b>	<b>AUC</b>	<b>Specificity</b>	<b>Sensitivity</b>	<b>PPV</b>	<b>NPV</b>	<b>MCC</b>	<b>AUPRC</b>
RFC - RFI	0.80 (0.77- 0.82)	0.84 (0.81- 0.86)	0.56 (0.51- 0.62)	0.62 (0.56- 0.68)	0.80 (0.78- 0.83)	0.41 (0.37- 0.45)	0.66 (0.61- 0.72)
RFC - MICE	0.80 (0.77- 0.82)	0.81 (0.80- 0.83)	0.64 (0.61- 0.68)	0.63 (0.59- 0.67)	0.82 (0.80- 0.84)	0.41 (0.37- 0.45)	0.67 (0.63- 0.72)
SVM- RFI	0.77 (0.74- 0.76)	0.67 (0.61- 0.72)	0.78 (0.75- 0.81)	0.53 (0.48- 0.57)	0.87 (0.84- 0.89)	0.42 (0.38- 0.46)	0.69 (0.65- 0.74)
SVM - MICE	0.78 (0.76- 0.80)	0.67 (0.59- 0.75)	0.77 (0.73- 0.81)	0.55 (0.48- 0.60)	0.85 (0.83- 0.87)	0.42 (0.37- 0.47)	0.69 (0.65- 0.73)
NN - RFI	0.81 (0.79- 0.83)	0.89 (0.87- 0.90)	0.53 (0.49- 0.57)	0.69 (0.65- 0.74)	0.80 (0.78- 0.83)	0.45 (0.41- 0.50)	0.68 (0.64- 0.73)
NN - MICE	0.81 (0.78- 0.83)	0.88 (0.85- 0.90)	0.55 (0.50- 0.60)	0.69 (0.64- 0.74)	0.80 (0.77- 0.83)	0.46 (0.41- 0.50)	0.68 (0.64- 0.72)
XGB- RFI	0.78 (0.76- 0.81)	0.79 (0.76- 0.83)	0.63 (0.60- 0.67)	0.59 (0.54- 0.65)	0.82 (0.80- 0.84)	0.42 (0.37- 0.46)	0.64 (0.59- 0.69)
XGB - MICE	0.79 (0.77- 0.81)	0.75 (0.73- 0.77)	0.69 (0.65- 0.73)	0.58 (0.54- 0.62)	0.83 (0.81- 0.85)	0.43 (0.39- 0.46)	0.65 (0.61- 0.69)
LR - RFI	0.80 (0.78- 0.82)	0.7 (0.73- 0.78)	0.71 (0.68- 0.73)	0.57 (0.53- 0.62)	0.85 (0.83- 0.86)	0.44 (0.40- 0.48)	0.68 (0.63- 0.74)
LR - MICE	0.80 (0.78- 0.82)	0.73 (0.70- 0.76)	0.72 (0.69- 0.76)	0.57 (0.54- 0.60)	0.84 (0.82- 0.87)	0.43 (0.39- 0.48)	0.68 (0.63- 0.72)

RFC, random forest classifier; SVM, support vector machine; LR, logistic regression; XGB, gradient boosting; NN, neural networks. AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value, RFI: random forest imputation, MICE: multiple imputation by chained equations.

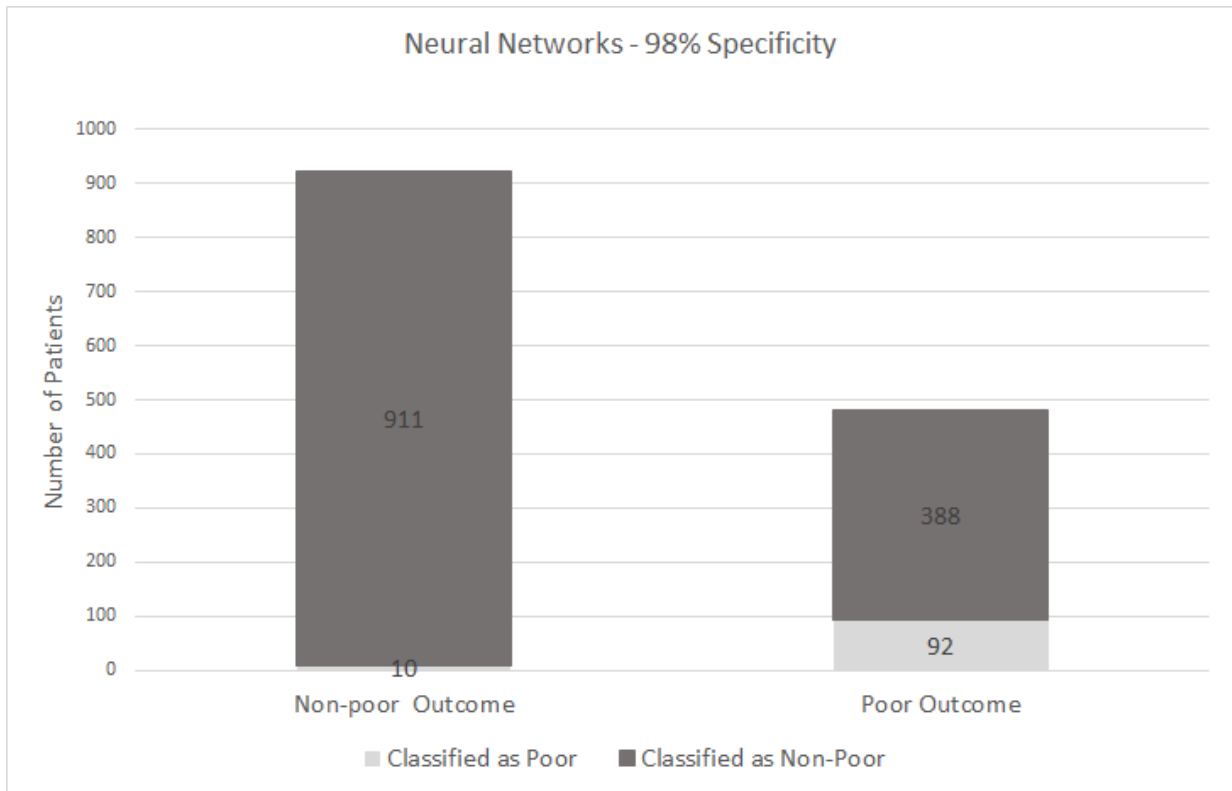




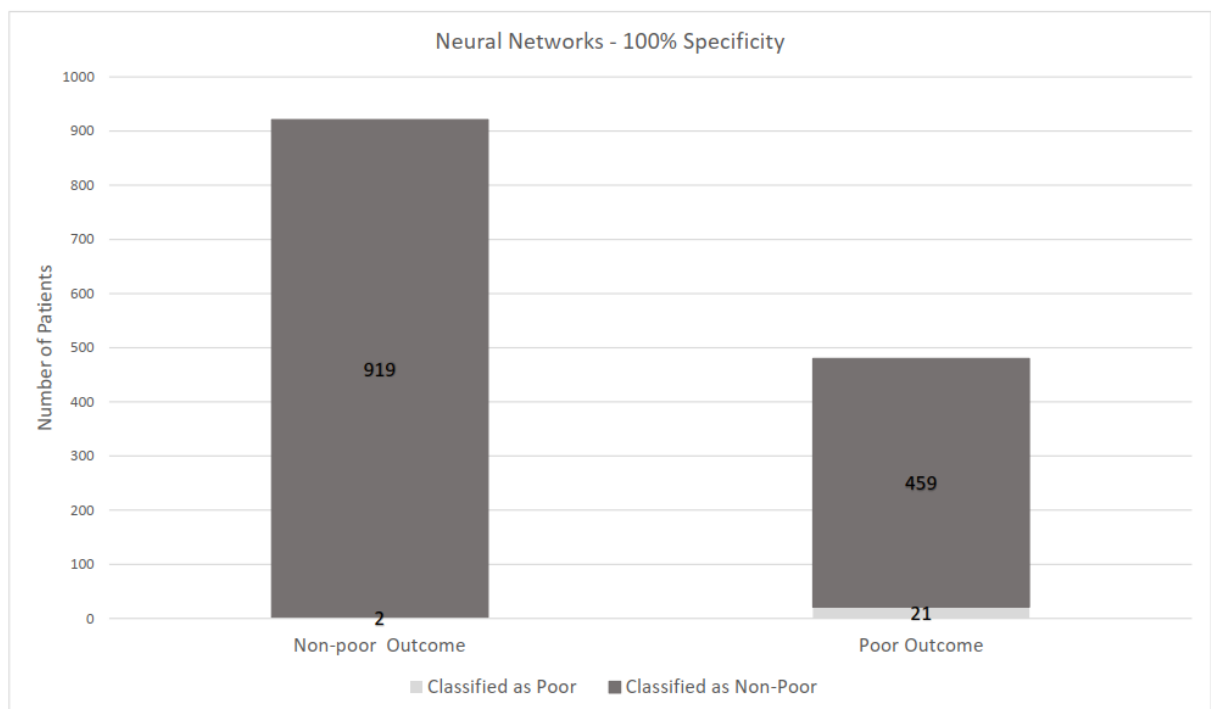
**Supplemental Figure I.** Patient inclusion flowchart. N denotes number of patients. EVT, endovascular treatment; MR CLEAN, multicenter randomized clinical trial for endovascular treatment of acute ischemic stroke in the Netherlands.



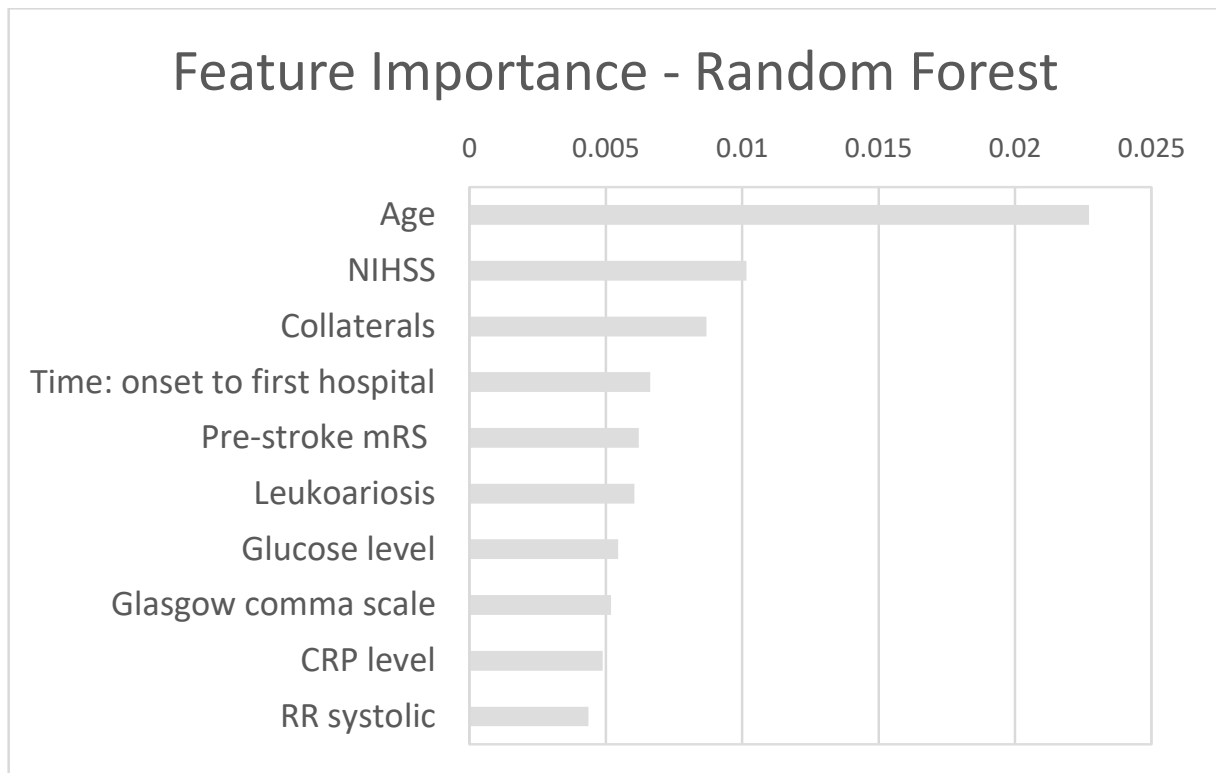
**Supplemental Figure II.** Performance of poor outcome prediction neural network model trained for 95% specificity, in validation data. Numbers in bars represent absolute number of patients.



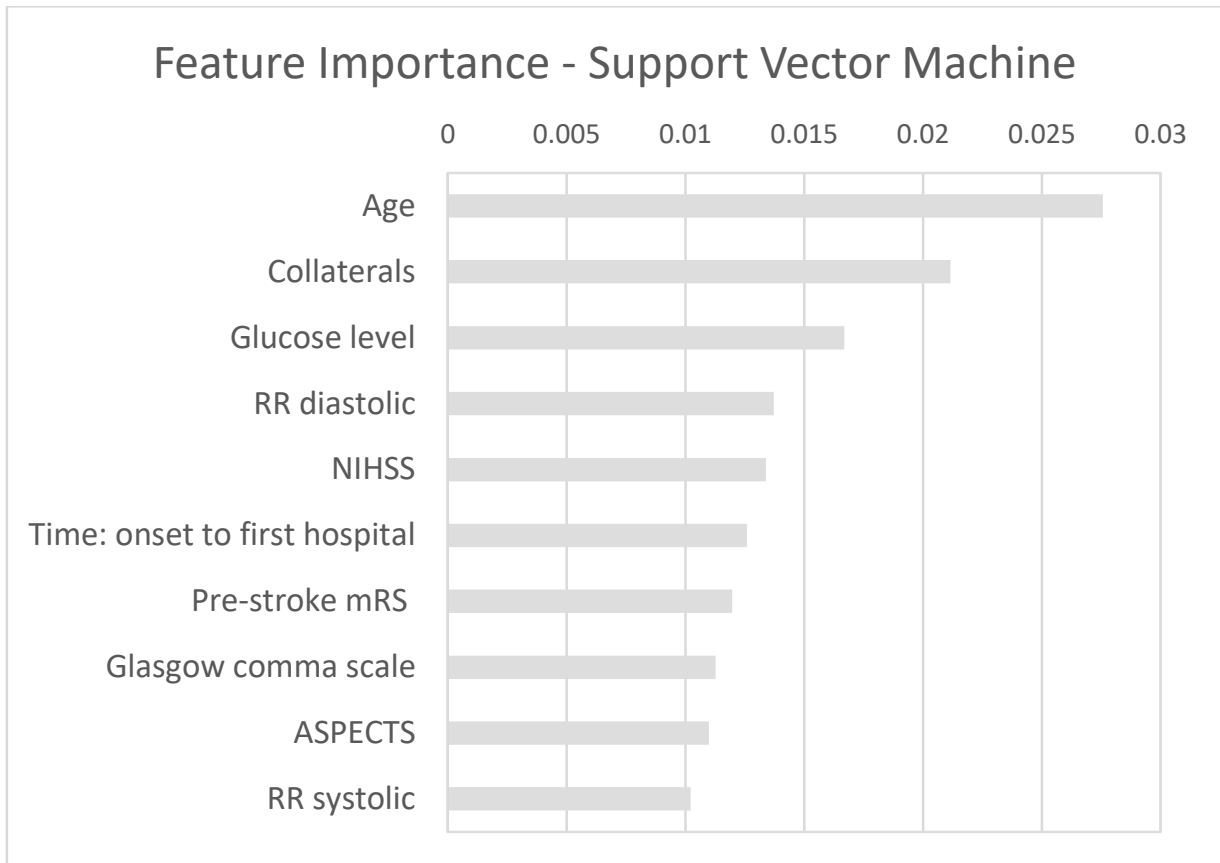
**Supplemental Figure III.** Performance of poor outcome prediction neural network model trained for 98% specificity, in validation data. Numbers in bars represent absolute number of patients.



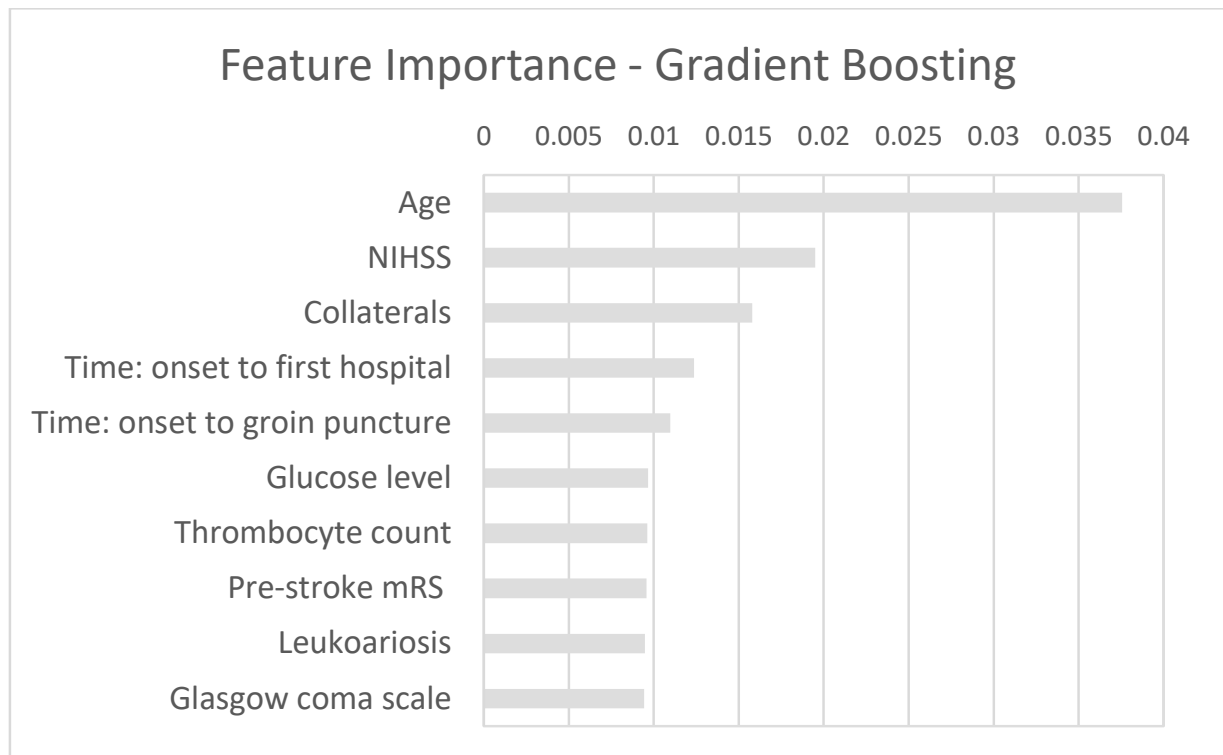
**Supplemental Figure IV.** Performance of poor outcome prediction neural network model trained for 100% specificity, in validation data. Numbers in bars represent absolute number of patients.



**Supplemental Figure VII.** Permutation feature importance for the Random Forest models. Average impact on the AUC. CRP, C-reactive protein; mRS, modified Rankin Scale; NIHSS, National Institutes of Health stroke scale; RR, blood pressure (Riva-Rocci).



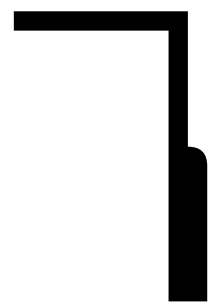
**Supplemental Figure VIII.** Permutation feature importance for the Support Vector Machine models. Average impact on the AUC. ASPECTS, Alberta Stroke Programme Early CT Score; CRP, C-reactive protein; mRS, modified Rankin Scale; NIHSS, National Institutes of Health stroke scale; RR, blood pressure (Riva-Rocci).



**Supplemental Figure IX.** Permutation feature importance for the Gradient Boosting models. Average impact on the AUC. mRS, modified Rankin Scale; NIHSS, National Institutes of Health stroke scale.

### Supplemental References

1. Bergstra JAMESBERGSTRA J, Yoshua Bengio YOSHUABENGIO U. Random Search for HyperParameter Optimization. *J Mach Learn Res.* 2012;13:281305.
2. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2012;12:2825–30.
3. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min.* 2016;Pages 785-794.



# CHAPTER 7.

Computer versus cardiologist: Is a machine learning algorithm able to outperform an expert in diagnosing phospholamban (PLN) p.Arg14del mutation on ECG?

Bleijendaal H, Ramos LA \*, Lopes RR, Verstraelen TE, Baalman SWE, Oudkerk Pool MD, Tjong FVY, Melgarejo-Meseguer FM, Gimeno-Blanes FJ, Gimeno-Blanes JR, Amin AS, Winter M, Marquering HA, Kok WEM, Zwiinderman AH, Wilde AAM, Pinto YM. Computer versus cardiologist: Is a machine learning algorithm able to outperform an expert in diagnosing phospholamban (PLN) p.Arg14del mutation on ECG? *Heart Rhythm*. 2021 Jan;18 (1):79-87.\*

\*Shared first author

DOI: 10.1016/j.hrthm.2020.08.021



## Abstract

*Background:* Phospholamban (PLN) p.Arg14del mutation carriers are known to develop dilated and/or arrhythmogenic cardiomyopathy and typical electrocardiographic (ECG) features have been identified for diagnosis. Machine learning is a powerful tool used in ECG analysis and has shown to outperform cardiologists.

*Objective:* We aimed to develop machine learning and deep learning models to diagnose PLN p.Arg14del cardiomyopathy using ECGs and evaluate their accuracy compared to an expert cardiologist.

*Methods:* We included 155 adult PLN mutation carriers and 155 age- and sex matched control subjects. 21 (13.4%) PLN mutation carriers were classified as symptomatic (symptoms of heart failure or malignant ventricular arrhythmias). The dataset was split into training and testing sets using 4-fold cross-validation. Multiple models were developed to discriminate between PLN mutation carrier or control subject. For comparison, expert cardiologists classified the same dataset. The best performing models were validated using an external PLN p.Arg14del mutation carriers dataset from Murcia, Spain (n= 50). We applied occlusion maps to visualize the most contributing ECG regions.

*Results:* In terms of specificity, the expert cardiologists (0.99) outperformed all models (range 0.53-0.81). In terms of accuracy and sensitivity the experts (0.28 and 0.64) was outperformed by all models (sensitivity range 0.65-0.81). T-wave morphology was most important for classification of PLN p.Arg14del. External validation showed comparable results, with the best model outperforming the experts.

*Conclusion:* This study shows that ML can outperform experienced cardiologists in the diagnosis of PLN p.Arg14del cardiomyopathy and suggests that the shape of the T-wave is of added importance to this diagnosis.



## Introduction

Phospholamban (PLN) is a transmembrane sarcoplasmic reticulum phosphoprotein and is a major regulator of calcium homeostasis in cardiomyocytes. Mutations in the gene encoding this protein are known to cause cardiomyopathy, including arrhythmogenic cardiomyopathy and dilated cardiomyopathy.<sup>1</sup> Carriers of mutations in PLN are at increased risk of developing malignant ventricular arrhythmias and end-stage heart failure, leading to high mortality.<sup>2-4</sup>

Low QRS voltages have been reported as the ECG hallmark in PLN mutation carriers.<sup>5</sup> Two large Dutch cohort studies reported low voltage ECGs in 46% and 41% in respectively 52 (van der Zwaag et al.<sup>1</sup>) and 295 patients (van Rijsingen et al.<sup>2</sup>). Additionally, repolarization changes on the ECG, in particular T wave inversions in the lateral leads, are frequently seen in PLN p.Arg14del mutation carriers. Van Rijsingen et al.<sup>2</sup> reported T-wave inversion in 40%, while van der Zwaag et al.<sup>1</sup> reported T-wave inversions in 57%. A Canadian cohort study by Cheung et al.<sup>6</sup> reported 53% in 50 patients. Additionally PLN is known to cause ARVC and one of the diagnostic criteria for ARVC is frequent ventricular extrasystoles (>500/24hours).<sup>7</sup> This was present in 48% of the carriers in the van Rijsingen cohort<sup>2</sup> and in 65% of the Holter that were evaluated by van der Zwaag et al.<sup>1</sup>

PLN p.Arg14del cardiomyopathy is a rare disease, with a prevalence of 0.08% to 0.38% in selected cardiomyopathy cohorts.<sup>8</sup> Other PLN gene mutations have been described, mostly in case reports and small cohorts, while Hof et al.<sup>8</sup> reported data of over a thousand p.Arg14del mutation carriers in the Netherlands alone, making p.Arg14del the most common PLN mutation in literature to this date.<sup>4</sup> Most general cardiologists do not routinely see patients with PLN cardiomyopathy, and consequently may not recognize the ECG features associated with this disease. The standard for diagnosing a PLN p.Arg14del mutation is genetic testing. However, when a patient is suspected of having a gene mutation causing structural heart disease, the electrocardiogram can increase (or decrease) the probability of having a mutation, assisting the clinician in early decision making regarding the diagnosis and possible therapy. Early diagnosis is of major importance,

because PLN associated cardiomyopathy is among the most malignant cardiomyopathies necessitating early ICD implants.<sup>2,7</sup>

In the past few years, the use of machine learning (ML) and, more specific, deep learning (DL) methods in medicine has increased significantly.<sup>9</sup> An advantage of DL is that it can automatically learn features from raw data, allowing the discovery of previously unknown relationships.<sup>10</sup> Within cardiology, DL is used for detection of a variety of cardiac arrhythmias, such as atrial fibrillation, in which the models outperform cardiologists, thereby positioning DL as a powerful tool for ECG analysis.<sup>9,11</sup> The increased accuracy of DL models often comes with the downside of lack of interpretability. However, new techniques have been developed, making it possible to visualize the features a deep learning model uses, and thus can be used to identify new features.<sup>12,13</sup>

In this study we aimed to develop ML and DL models and study their accuracy compared to an expert cardiologist in diagnosing PLN p.Arg14del cardiomyopathy on an ECG. We aimed to present a proof-of-concept to show how ML enabled ECG analysis is of added value, specifically when it concerns a very rare disease which is often missed simply because it is rarely seen.

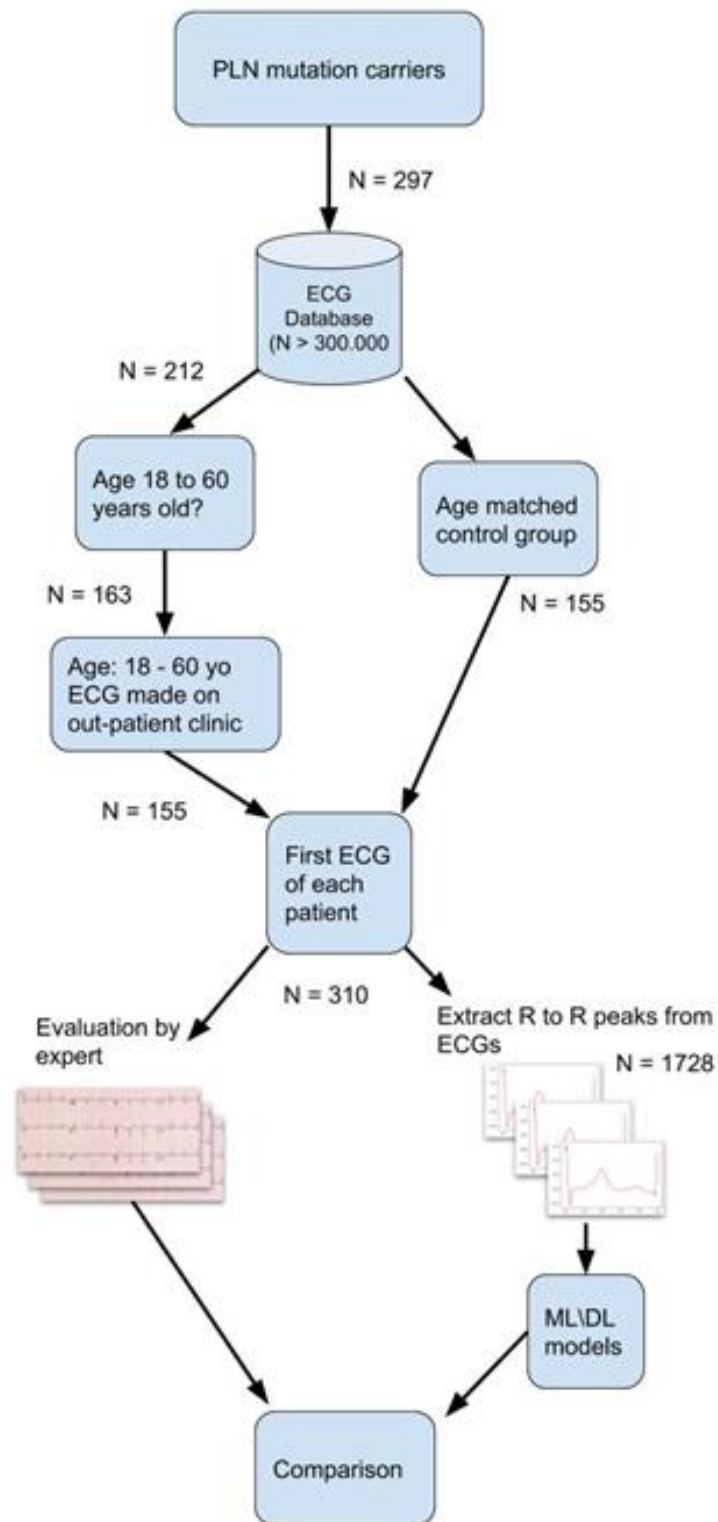
Moreover, we aimed to identify specific regions ECG that could give insights for improving diagnosis of this disease and be used for better understanding of PLN mutation cardiomyopathy in general.

## Methods

### Data collection and labelling

We collected ECGs from all patients which were stored in the ECG database (MUSE, GE Healthcare) of the Amsterdam University Medical Centers (UMC), location Academic Medical Center (AMC), during the period from 1998 up to and including 2018. To minimize the amount of non-PLN mutation related cardiovascular pathology which could potentially influence the ECG, we included only ECGs from patients aged 18 to 60 years old. From this database, we extracted all patients known to have a PLN p.Arg14del mutation. A mutation carrier was defined as symptomatic when they suffered

from either an arrhythmic event (sustained ventricular tachycardia or ventricular fibrillation) or symptomatic episode caused by heart failure (New York Heart Association (NYHA) class 2 or higher, as defined by clinical staff). This information was provided by the national PLN registry, and informed consent for re-use of patient information has been obtained. ECGs were excluded if they were made on the emergency ward or during hospitalization on a clinical ward, to exclude the possible effect of acute pathology on the ECG. As a control group, we selected ECGs from patients between 18 and 60 years of age who underwent general, non-cardiovascular pre-operative screening at the out-patient clinic of the Amsterdam UMC, location AMC, after which we randomly selected a subgroup to match the PLN population according to age and sex, to ensure the same distribution for each group. For both groups, only the first recorded ECG for each patient was used. Figure 7.1 contains a diagram with the PLN and control group selection process.



**Figure 7.1.** Data cleaning process from patient selection to model development.

We excluded all ECGs that were considered technically inadequate according to an experienced investigator (HB) (limb lead reversal, loss of signal on one or more leads and high amount of noise of two or more leads, making analysis impossible), or which had any other rhythm than sinus rhythm. ECGs were

labelled as “PLN” or “control” based on the presence of a PLN Arg14.del gene mutation. This dataset was named the Amsterdam Dataset, to discriminate from the external validation set. External validation was performed on a population of PLN p.Arg14del mutation carriers from the Virgen de Arrixaca Hospital in Murcia, Spain. From the local ECG database, a random set of non-PLN mutation carriers in this hospital was selected as a control group. This external validation set was named the Murcia Dataset.

The study was approved and the requirement for informed consent was waived by the Medical Ethics Commission of the Amsterdam UMC on 22-11-2018 (registration number W18\_371#18.425).

### Evaluation by expert cardiologists

All ECGs included were anonymized and visually evaluated separately by two cardiologists with expertise PLN cardiomyopathy (A.A.M.W and W.E.M.K). The experts classified the ECGs in PLN or non-PLN and were not informed of the ratio between carriers and non-carriers. For ECG classification they used known ECG features, as described in the introduction (low QRS voltages, T-wave inversion and frequent extrasystoles).

### Data pre-processing and development of ML models

To increase the amount of training data, we extracted all beats from each 10-second ECG available and used them as individual samples during training. Details about the data pre-processing are shown in Supplemental Methods I and Supplemental Figure I. The patients were randomly split into training, validation and testing sets using 4-fold cross validation stratified for carriers and controls. Initially, 3 folds are separate for training and 1 is left aside for testing. From the 3 folds used for training, 20% is separated as validation set to be used to assess network performance during training and hyperparameter optimization. All heartbeats from each individual patient were kept either in the training or the test set in the initial split, to prevent data leakage. For testing, only one beat was used per patient as reference. We did not choose a beat on one of the edges of the ECG, due to high probability of it containing noise. For creating the models we followed two approaches, defined below.

Our first approach, the wavelet-ML based approach, consisted of applying a wavelet transform for each individual beat, since wavelets have been broadly and successfully used in multiple ECG applications.<sup>14,15</sup> More details about wavelets and their implementation can be found in the Supplemental Methods II. The output of the wavelet transformation (of size (64x8)) was flattened and used as input to train Machine Learning classifiers: Logistic Regression (LR), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), and Gradient Boosting XGB), following the approach of Kumar et al.<sup>15</sup>

In our second, DL-based approach, we implemented 1 and 2D Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, using the R to R peak as input. For each type of network (CNN and LSTM), we implemented two approaches (using 1 and 2D convolutions), namely approach A and B. Details about these approaches and their implementations are available in the Supplemental Methods III.

### Statistical Analysis

For model evaluation we reported the average accuracy, sensitivity, specificity and the area under the receiver operating characteristic curve (AUC). We deemed the best performing models the ones with the highest accuracy and sensitivity due to greater importance of missing true positive (PLN) patients. We used the McNemar's test to check if the difference between the models and the experts was statistically significant.<sup>16</sup>

### Visualization of ECG features

For our best performing model, we created visualization plots to visualize the parts of the ECG that were most relevant for classification of PLN patients in our deep learning model, we used 'occlusion maps' for this purpose.<sup>17</sup> We generated occlusion maps by systematically occluding parts of the heart-beat signal. We split the (8x256) input signal into 16 parts of (8x16) and occluded a region by setting all its values to zero. Then we applied the trained model to the signal with the occluded region and evaluated the loss in model performance. The higher the loss in performance, the more important the occluded region is.

## Reproducibility and Open Access

Given its sensitive nature, the data used in this study is not publicly available. All the code used in this paper, however, is available at the following GitHub page: <https://github.com/L-Ramos/CardiologyAI>

## Results

From the 297 known PLN p.Arg14del mutations carriers in the Amsterdam UMC, 155 were eligible for inclusion in this study (see Figure 7.1 for a flow diagram). From the PLN carriers, 13.5% were symptomatic at the time the ECG was made. The mean age in this group was 39 years (IQR 28-50) and 63 (41%) were male. Baseline ECG characteristics are shown in Table 7.1.

**Table 7.1.** Description of the Amsterdam Data

Variable Name	PLN n=155	Control n=155
Age *	39 (28–50)	39 (28–50)
Sex (Male %)*	63 (41)	63 (41)
Ventricular Rate (bpm)	68 (60–75)	65 (57–73)
Atrial Rate (bpm)	68 (60–75)	66 (57–73)
QRS Duration (ms)	86 (80–94)	94 (84–104)
QT Interval (ms)	388 (368–406)	400 (374–426)
QT Corrected (ms)	407 (394–424)	410 (401–429)
P wave axis	55 (37–66)	48 (33–61)
R wave axis	48 (2–75)	34 (3–63)
T wave axis	46 (1–63)	38 (20–58)

Values shown represent the median and interquartile range, unless stated otherwise. Variables with \* were used to match samples from the control group.

## Performance of ML and DL models compared to expert cardiologists

In Table 7.2A, the results for both the experts, ML and DL models averaged over the 4 folds are displayed. Expert 1 and 2 had an accuracy of 0.65 and 0.63 respectively, a sensitivity of 0.32 and 0.27, and a specificity of 0.97 and 0.99 (Table 7.2A). Despite showing slightly higher accuracy, Expert 1 had also larger standard deviation when compared to Expert 2. Figure 7.2 shows receiver operating curves for a selection of the best performing models and the results of the best performing expert, the ROCs for the other models are shown in the Supplemental Figure II. Figure 7.3 shows an example of an ECG correctly classified as PLN by both the 1D CNN and the experts. For accessing inter-rater reliability between the two cardiologists we computed

the Cohen's Kappa score<sup>18</sup> which was  $\kappa=0.65$  for the Amsterdam database, indicating a substantial agreement between the experts. For the Murcia dataset,  $\kappa$  was 0.27, which indicates a fair agreement between the experts.

**Table 7.2A.** Performance (in terms of accuracy, sensitivity and specificity) of the experts, ML and DL models in the Amsterdam data

<b>Model</b>	<b>Sensitivity</b>	<b>Specificity</b>	<b>Accuracy</b>	<b>AUC</b>
<b>Expert 1</b>	0.32±0.01	0.97±0.02	0.65±0.06	0.65±0.06
<b>Expert 2</b>	0.25±0.05	1.0±0.00	0.63±0.03	0.63±0.02
<b>1D CNN-Approach A</b>	0.65±0.02	0.67±0.07	0.65±0.04	0.74±0.03
<b>2D CNN-Approach B</b>	0.77±0.03	0.67±0.09	0.72±0.03	0.78±0.03
<b>1D LSTM-Approach A</b>	0.65±0.13	0.59±0.18	0.62±0.05	0.72±0.09
<b>2D LSTM-Approach B</b>	0.81±0.08	0.53±0.12	0.67±0.08	0.74±0.09
<b>Wavelet-MLP</b>	0.70±0.05	0.76±0.03	0.73±0.02	0.78±0.02
<b>Wavelet-SVM</b>	0.71±0.05	0.81±0.06	0.76±0.05	0.80±0.06
<b>Wavelet-LR</b>	0.72±0.07	0.79±0.06	0.76±0.05	0.80±0.06
<b>Wavelet-KNN</b>	0.69±0.05	0.77±0.07	0.74 ±0.06	0.76±0.06
<b>Wavelet-RFC</b>	0.69±0.5	0.80±0.07	0.75±0.06	0.83±0.03
<b>Wavelet-XGB</b>	0.69±0.03	0.81±0.00	0.75±0.02	0.82±0.02

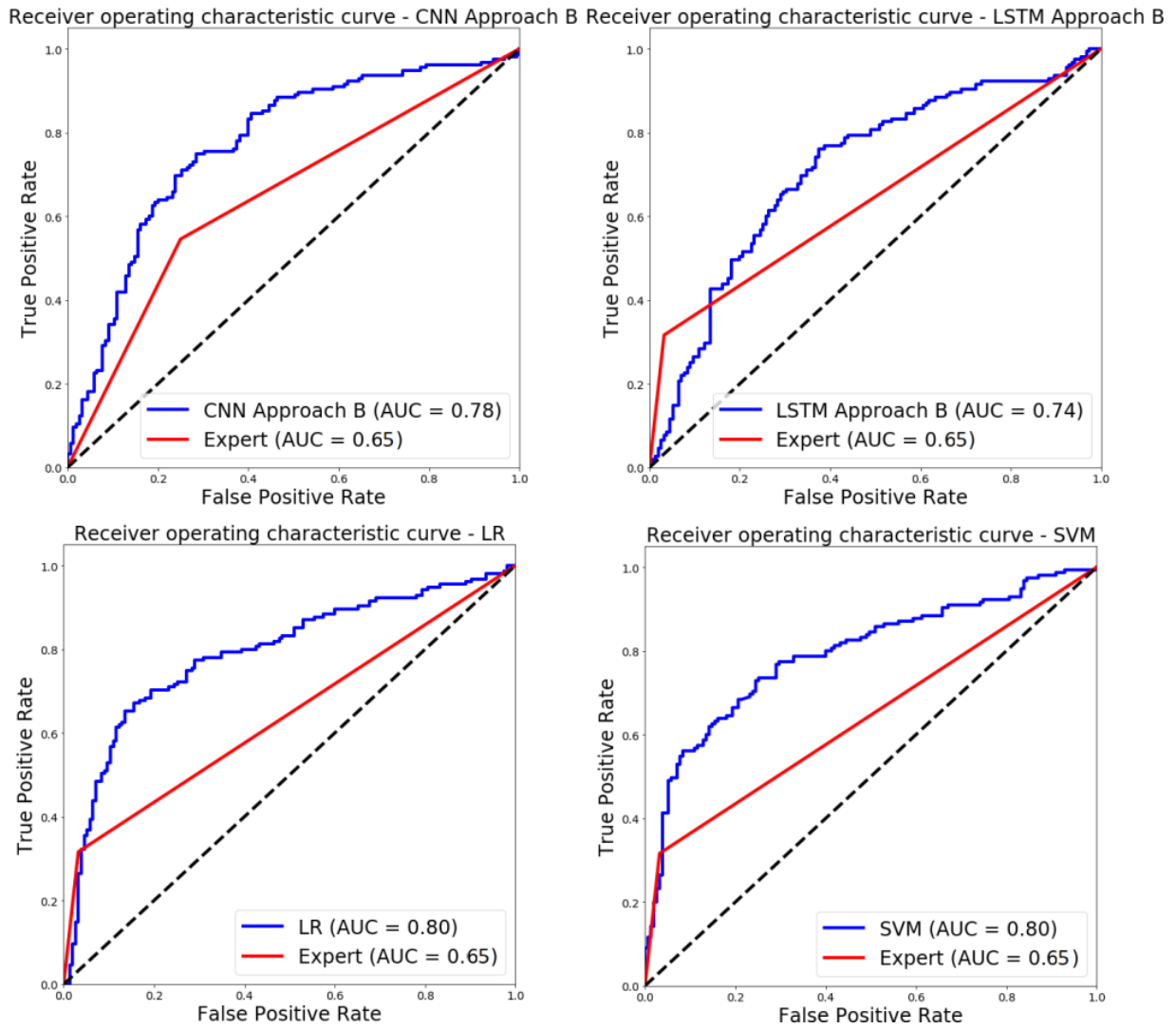
The value shown is the average over 4 folds. CNN=Convolutional Neural Network, LSTM=Long Short-Term Memory network, MLP=Multilayer Perceptron, SVM=Support Vector Machine, LR=Logistic Regression, KNN=K-Nearest Neighbors, RFC=Random Forest Classifier, XGB=Gradient Boosting.

**Table 7.2B.** External Validation in the Murcia Data

<b>Model</b>	<b>Sensitivity (%)</b>	<b>Specificity (%)</b>	<b>Accuracy (%)</b>	<b>AUC (%)</b>
<b>Expert 1</b>	0.55	0.75	0.65	0.65
<b>Expert 2</b>	0.18	0.91	0.56	0.55
<b>2D CNN-Approach B</b>	0.64	0.72	0.68	0.70
<b>Wavelet-LR</b>	0.96	0.20	0.58	0.58
<b>Wavelet-SVM</b>	0.96	0.20	0.58	0.58
<b>2D LSTM-Approach B</b>	0.48	0.68	0.58	0.63

The value shown is the average over 4 folds. CNN=Convolutional Neural Network, LSTM=Long Short-Term Memory network, , SVM=Support Vector Machine, LR=Logistic Regression.





**Figure 7.2.** Repeater operating curves (ROC) for the best performing expert and the four best performing models on the Amsterdam Data.



**Figure 7.3.** Example of an ECG which both the experts and the CNN labelled correctly as ‘PLN’. This example shows the typical ECG features which both the expert use to detect PLN: low QRS voltages on the limb leads and T-wave inversion in leads V3-V6.

### Machine learning based approach

The different wavelet-ML models showed comparable results. The Wavelet SVM (accuracy 0.76) and Wavelet LR (accuracy 0.76) can be marked as the two best performing models. In terms of sensitivity, wavelet-ML also outperformed the cardiologist (0.72 versus 0.31). In terms of specificity, the cardiologist outperformed the wavelet-ML (0.99 versus 0.81).

### Deep learning based approach

The DL model performing best on test data was the 2D CNN with approach B, with an accuracy of 0.72, outperforming both the expert cardiologists, with a standard deviation comparable to the experts. In terms of sensitivity, this CNN also outperformed both experts (0.77 versus 0.31 of the expert with highest sensitivity). In terms of specificity, the experts outperformed the CNNs (0.99 versus 0.67). Using the McNemar's test, we compared the best

performing model (CNN with approach B) with the expert with the highest accuracy. The chi-square statistic was 2.125 and p-value=0.145 for the Amsterdam dataset.

Since ML and DL models have multiple hyper-parameters to be optimized, we report in Supplemental Table I all the parameters used.

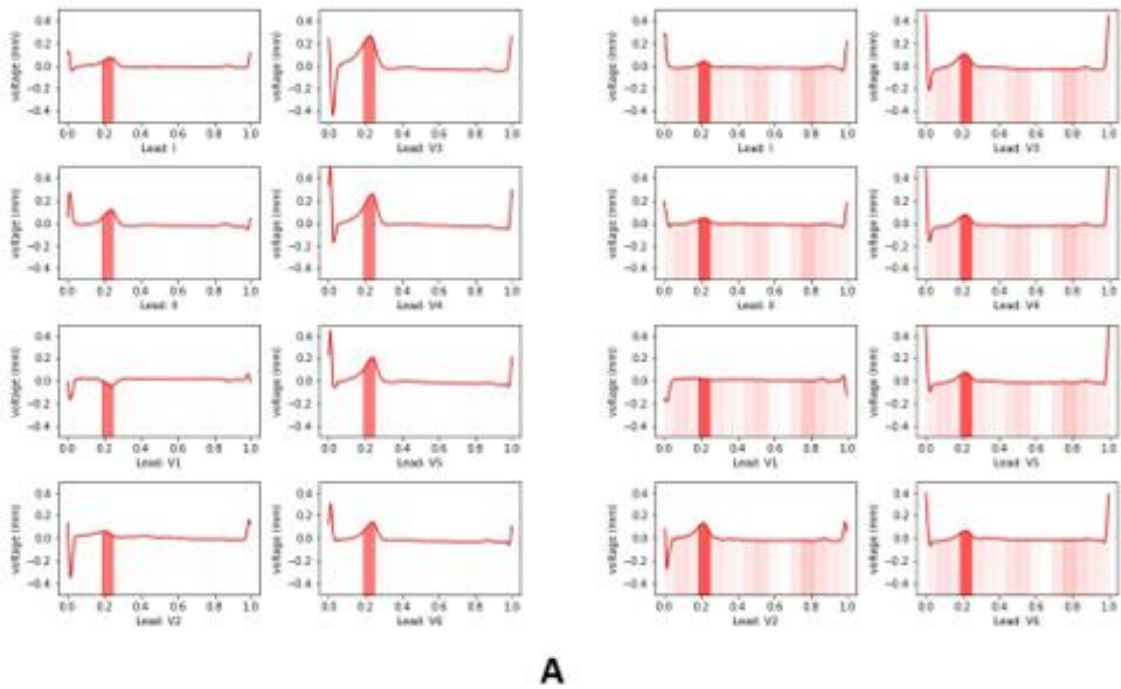
### External validation on the Murcia Data

Results of the external validation for the expert cardiologists, the two best performing Wavelet-ML and the two best performing DL models can be found in Table 7.2B (standard deviation is not available since the trained models were used for inference in the whole set). In terms of accuracy the CNN with approach B performed slightly better on the Murcia Dataset, compared to expert with highest accuracy (0.68 versus 0.65). Our Wavelet based ML models showed the highest sensitivity (0.96) versus the CNN and the LTSM (0.64 and 0.48), however both Wavelet ML models showed poor specificity (0.20). A comparison between the ROC of the best performing expert and the best ML/DL approaches for the Murcia dataset is presented in Supplemental Figure III. Using the McNemar's test for comparison of CNN with approach B and expert 1, the chi-square statistic was 4.114, p-value=0.043.

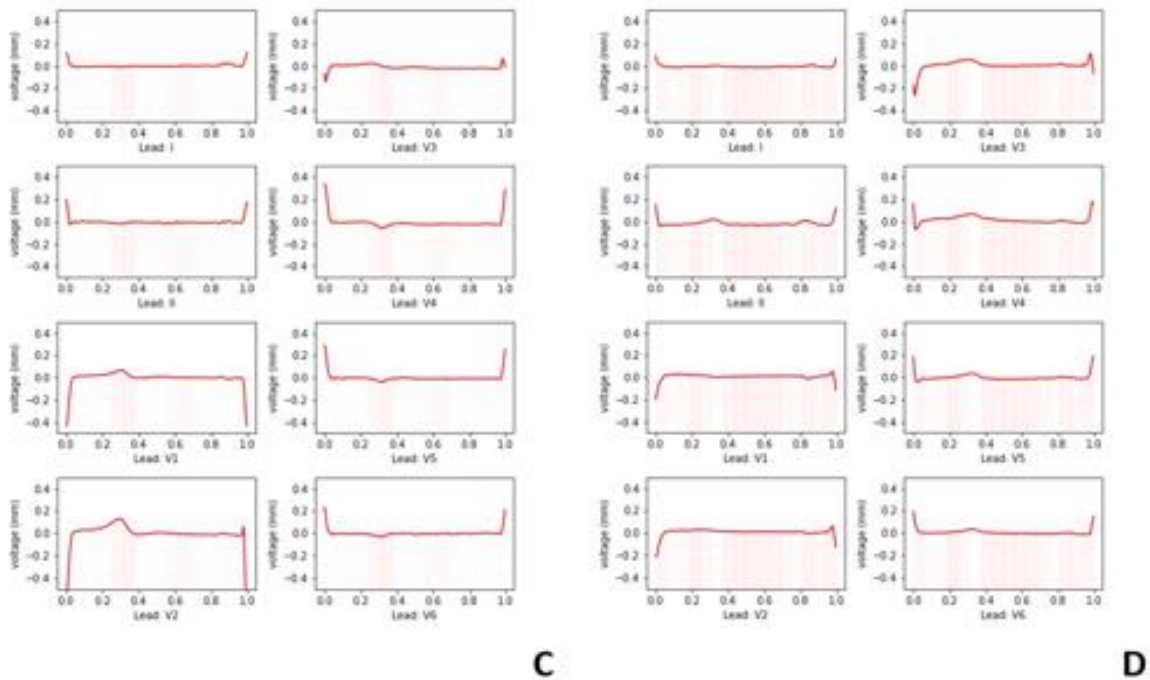
### Visualization of ECG features

Figure 7.4 shows four examples of the ECG regions the model extracted features from to classify the specific ECG sample in either PLN or control patients. In 63% of the true positives, our results showed that the T-wave was the most important part for the model, an example is shown in Figure 7.4A. In 14,2% the model did not use a specific part of the signal but used the whole signal (see Figure 7.4B). For the majority of the True Negatives (TN), the model used the whole signal for classification (56%) and in only 3%, the T-wave was the most prominent ECG feature. An overview of the ECG features used, is shown in Supplemental Table II.

## TRUE POSITIVE



## TRUE NEGATIVE



**Figure 7.4.** Examples of visualization of ECG features which the CNN used for classification using Occlusion Maps on unique ECGs. A + B are PLN ECGs correctly classified by the model as PLN. C + D are ECG form control while were correctly classified as non PLN. The red highlighted areas are the parts of the signal which the model used to classify. If no specific area was highlighted, this means the model used the whole signal for classification.

## Discussion

From all of our models, the 2D CNN approach B, outperformed the expert cardiologists in accuracy and sensitivity, both on the Amsterdam and the Murcia data. In terms of specificity, the cardiologists were superior in the identification of PLN mutation carriers on ECG. This suggests neither using ML/DL nor the assessment of an expert cardiologist for the diagnosis of PLN p.Arg14del mutation on ECG is superior to each other.

### Performance of the models

On the Amsterdam data, the Wavelet based SVM showed the highest accuracy (0.76), whilst the 2D CNN approach B, which had an accuracy of 0.72 on the Amsterdam data, performed best on the Murcia Dataset, with an accuracy of 0.68, compared to 0.58 from the Wavelet based SVM. It is clear from our results that the wavelet based models did not generalize well for a different population, which might indicate that the features extracted by the discrete wavelet transform might not be informative enough across different datasets. 2D CNN approach B resulted in the best DL models, where the learned convolutional kernels were shared among all leads, instead of learning individual kernels per lead (approach A). The standard deviation for accuracy and sensitivity for the 2D CNN approach B was also one of the lowest, showing that the model generalized well across different folds.

### Comparison with previous studies

This study is the first study to evaluate the diagnosis of PLN p.Arg14del mutation by solely using the ECG. Also, our study is the first to use both machine – and deep learning to prove this concept. In the field of cardiogenetics and ECG analysis, Hermans et al.<sup>19</sup> recently added T-wave morphology characterizations to age, gender and QTc in an SVM and improved the diagnosis of Long-QT syndrome on ECG. In the Amsterdam Dataset of our study, Wavelet based ML models also proved to perform with the highest accuracy. However, Hermans et al.<sup>19</sup> did not use a deep learning approach, in which the model learns features by itself.

A recent study from the Mayo Clinic developed a ML model to diagnose hypertrophic cardiomyopathy.<sup>20</sup> Their model outperform ours, however their dataset is much larger and furthermore, hypertrophic cardiomyopathy is a

diagnosis based on a (for example) echocardiographic phenotype, and not solely based on the prevalence of a gene mutation.<sup>21</sup> Therefore, to our knowledge, this study is the first to use deep learning to detect a genetically proven structural heart disease by solely looking at the ECG in a dataset including asymptomatic mutation carriers.

Several studies have used ML or DL based ECG analysis to diagnose cardiac arrhythmia, with a higher accuracy than our ML models. An example is the DL model of Hannun et al.<sup>11</sup> which is a neural network for the automatic detection of cardiac arrhythmia on ECG, and which was trained on a much higher number of ECGs than we used in our study. PLN is a rare disease worldwide, and it would be impossible to reach the same number of patients as they have included.

### Visualization of ECG features in the deep learning model

Many techniques have been developed to interpret these models and give insight into their decision process. To our knowledge we are the first to use a DL based approach to identify ECG features associated with genetic structural heart disease. In the majority of our correctly classified PLN ECGs, the model used the T-wave as its most important ECG feature. Although low QRS voltages are seen as the main ECG feature, T-wave inversions are also common in PLN p.Arg14del mutation carriers.<sup>1,2,5,6</sup> More focused research will be needed to further elaborate these findings and to identify more specific and potentially new ECG features.

### Clinical interpretation

To implement our models in a clinical setting, first their performance has to increase. The golden standard for diagnosis of a PLN p.Arg14del mutation is genetic testing. Models like ours are unlikely to replace genetic testing as a whole, but can serve as a risk stratification tool to predict which patients do need genetic testing. This is currently done by (expert) physicians, and, now that our models have shown to outperform the sensitivity of expert cardiologists, our results could contribute to improved and earlier diagnosis of this progressive genetic cardiomyopathy. Because sensitivity of our models is higher than that of the experts, the models are better at diagnosing PLN mutation carriers, compared with the assessment of the expert cardiologist, in the current setting. When looking at specificity it is the other way around, the

experts outperform the models, with almost a maximum specificity, therefore often correctly classifying an ECG as a PLN p.Arg14del mutation carrier.

PLN is not the only genetic heart disease which has a “typical” phenotype on ECG. Other diseases like the Long QT syndrome , hypertrophic cardiomyopathy and Brugada syndrome, are only a few examples of syndromes in which a gene mutation can lead to a clinically severe and life threatening syndrome. This study suggests that a ML/DL based approach could also be used for the diagnosis of these inherited cardiac syndromes.

### Limitations

This study is performed on a (relatively) small dataset. It is known that deep learning is a technique which is highly dependent of the amount of data it is trained on. Because PLN cardiomyopathy is a rare disease, it is difficult to bring together a much larger number of PLN patients. We augmented our data by using multiple beats from a single ECG as individual samples. Moreover, we decided to first only use the patients from our own center, to prove the concept that it is possible to predict carrier status of a specific mutation leading to heart disease using ML/DL based ECG analysis.

Also, for this analysis we chose to evaluate genetically proven carriers of the PLN p.Arg14del mutation, which were either symptomatic or asymptomatic. This was done to identify possible ECG features which are present in both these groups and not only in symptomatic patients.

The main goal of this study was to evaluate the predictive value of ECG for the diagnosis of PLN cardiomyopathy. Therefore, we did not included basic demographics to like age and gender, especially due to the risk of bias given these parameters could influence the ECG by itself.<sup>22</sup>

### Conclusion

In conclusion, this study has shown that ML and DL can improve diagnosis of PLN p.Arg14del cardiomyopathy and our results find regions of the surface ECG that are related to PLN p.Arg14del mutation and therefore suggest that the T-wave is of added importance to diagnosing PLN mutation caused heart disease, even before they become symptomatic.

### References

1. Van Der Zwaag PA, Van Rijsingen IAW, Asimaki A, et al.: Phospholamban R14del mutation in patients diagnosed with dilated cardiomyopathy or arrhythmogenic right ventricular cardiomyopathy: Evidence supporting the concept of arrhythmogenic cardiomyopathy. *Eur J Heart Fail* 2012; .
2. van Rijsingen IAW, van der Zwaag PA, Groeneweg JA, et al.: Outcome in Phospholamban R14del Carriers. *Circ Cardiovasc Genet* 2014; .
3. Bosman LP, Verstraelen TE, van Lint FHM, et al.: The Netherlands Arrhythmogenic Cardiomyopathy Registry: design and status update. *Netherlands Hear J* 2019; .
4. van der Zwaag PA, van Rijsingen IAW, de Ruiter R, et al.: Recurrent and founder mutations in the Netherlands-Phospholamban p.Arg14del mutation causes arrhythmogenic cardiomyopathy. *Netherlands Hear J* 2013; .
5. Posch MG, Perrot A, Geier C, et al.: Genetic deletion of arginine 14 in phospholamban causes dilated cardiomyopathy with attenuated electrocardiographic R amplitudes. *Hear Rhythm* 2009; .
6. Cheung CC, Healey JS, Hamilton R, et al.: Phospholamban cardiomyopathy: a Canadian perspective on a unique population. *Netherlands Hear J* 2019; .
7. Towbin JA, McKenna WJ, Abrams DJ, et al.: 2019 HRS expert consensus statement on evaluation, risk stratification, and management of arrhythmogenic cardiomyopathy. *Hear Rhythm* 2019; .
8. Hof IE, van der Heijden JF, Kranias EG, et al.: Prevalence and cardiac phenotype of patients with a phospholamban mutation. *Netherlands Hear J* 2019; .
9. Krittanawong C, Johnson KW, Rosenson RS, et al.: Deep learning for cardiovascularmedicine: A practical primer. *Eur. Heart J.* 2019,.
10. Lecun Y, Bengio Y, Hinton G: Deep learning. *Nature.* 2015,.
11. Hannun AY, Rajpurkar P, Haghpanahi M, et al.: Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 2019; .
12. Chattopadhyay A, Sarkar A, Howlader P: Grad-CAM ++ : Improved Visual Explanations for Deep Convolutional Networks. *IEEE Winter Conf Appl Comput Vis* 2018; .
13. Baalman SWE, Schroevers FE, Oakley AJ, et al.: A morphology based deep learning model for atrial fibrillation detection using single cycle electrocardiographic samples. *Int J Cardiol* 2020; .
14. Martis RJ, Acharya UR, Min LC: ECG beat classification using PCA, LDA, ICA and Discrete Wavelet Transform. *Biomed Signal Process Control* 2013; .
15. Kumar M, Pachori RB, Acharya UR: Characterization of coronary artery disease using flexible analytic wavelet transform applied on ECG signals. *Biomed Signal Process Control* 2017; .
16. McNemar Q: Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; .



## Computer versus cardiologist: Is a machine learning algorithm able to outperform...

17. Zeiler MD, Fergus R: Visualizing and understanding convolutional networks. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2014;.
18. Cohen J: A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 1960; .
19. Hermans BJM, Bennis FC, Vink AS, et al.: Improving long QT syndrome diagnosis by a polynomial-based T-wave morphology characterization. Hear Rhythm 2020; .
20. Ko WY, Siontis KC, Attia ZI, et al.: Detection of Hypertrophic Cardiomyopathy Using a Convolutional Neural Network-Enabled Electrocardiogram. J Am Coll Cardiol 2020; .
21. 2014 ESC Guidelines on diagnosis and management of hypertrophic cardiomyopathy. Eur Heart J 2014; .
22. Attia ZI, Friedman PA, Noseworthy PA, et al.: Age and Sex Estimation Using Artificial Intelligence from Standard 12-Lead ECGs. Circ Arrhythmia Electrophysiol 2019; .

# Supplemental material

## Supplemental Methods

### *I - Data pre-processing*

To increase the amount of training data, we extracted all beats from each 10-second ECG available and used them as individual samples during training. The number of beats per scan ranged from 3 to 13. To extract the beats from each ECG signal we used the Python BioSPPY, a toolbox developed to process biological signals.<sup>1</sup> We interpolated each beat to the standard size of 256 data points. Therefore, the final input to the models of size (8x256), eight leads of 256 data points for each patient (Supplemental Figure I).<sup>2,3</sup>

### *II - Wavelets*

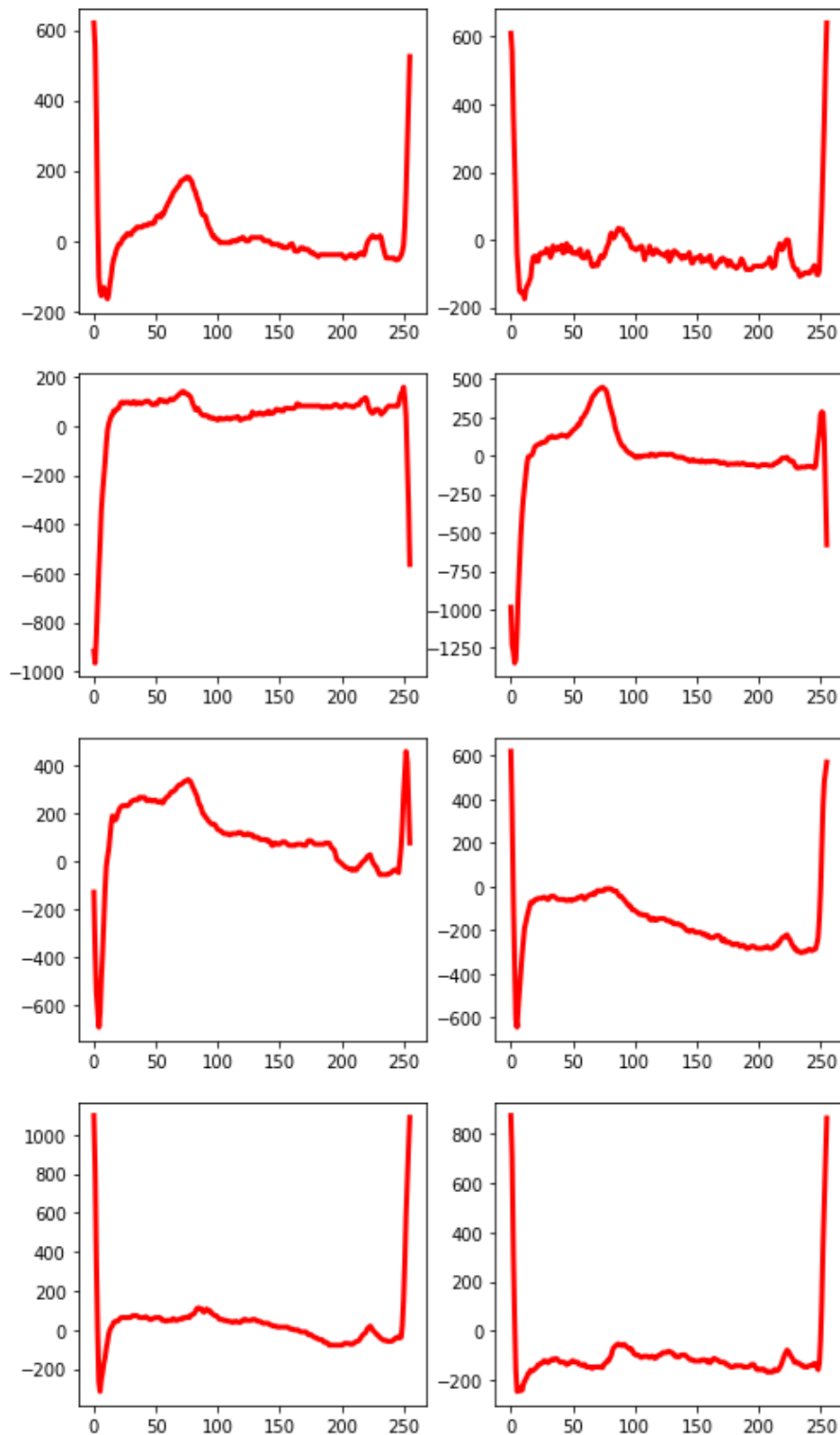
The wavelet transform is a technique used to extract information from signals since it provides information in both time and frequency domain. In our case we used the multilevel 1D discrete wavelet transform following the approach from Martis et al.<sup>4</sup> For this we used the PyWavelets library.<sup>5</sup>

### *III - Deep learning*

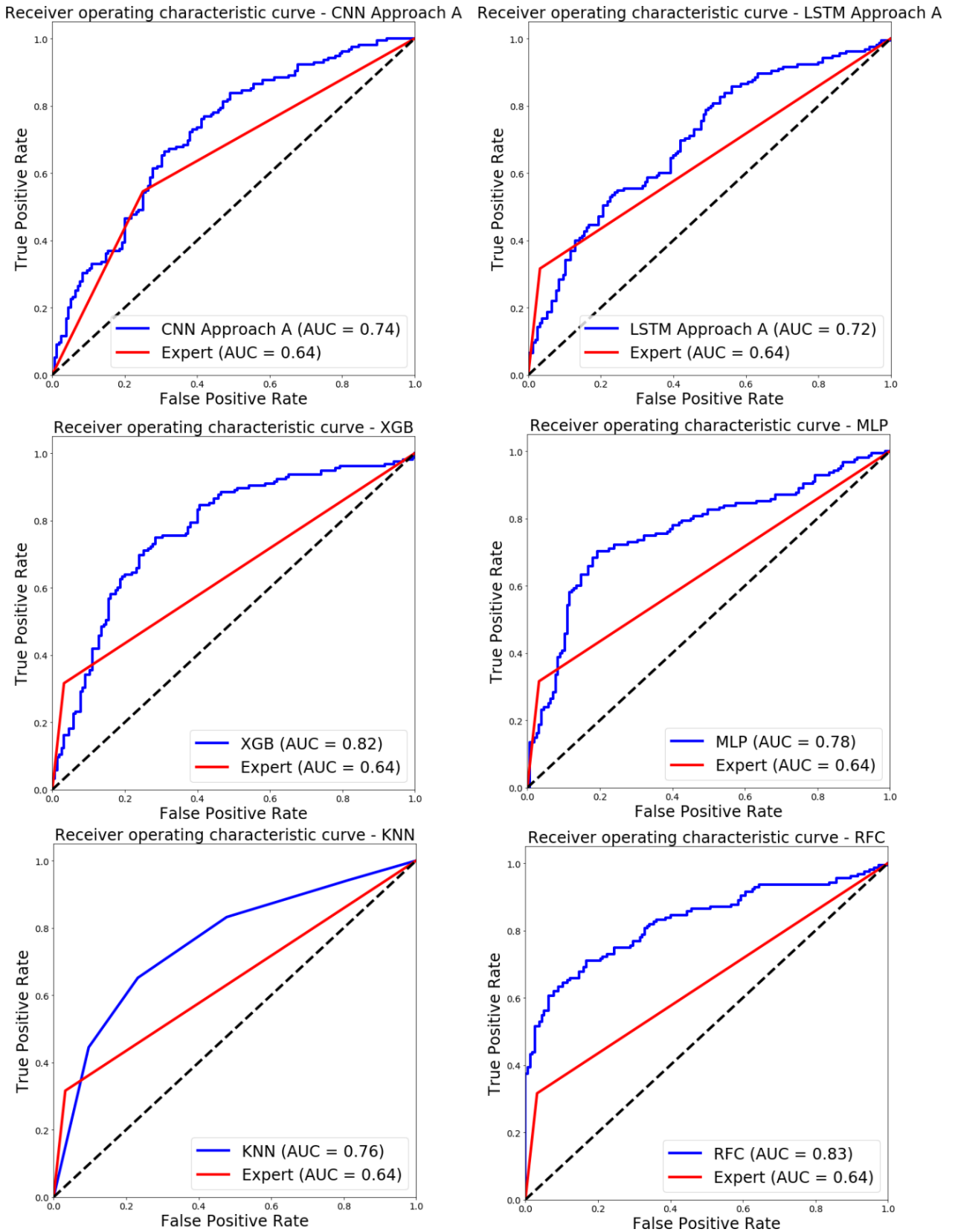
The main difference between approach A and B is the way the convolutional kernels are used. In Approach A, the CNN learns an individual kernel per lead, while in approach B, one kernel is learned and shared for all leads.

We explored two approaches, in the first one, named approach A, each lead was used as a channel in the following order (I, II, V1-V6), with an input of size 256x8 (256 data points for 8 different channels). The second approach, named approach B, the leads were placed one under the other, with an input of size (8x256x1). In approach B we used a 2D CNN, but set one of the dimensions of the kernels to one, to prevent the convolution of including more than one lead per operation. Finally, since the ECG signal is a time-series, we developed recurrent models using Long Short-Term Memory (LSTM) networks. The LSTM models were developed using approach A and B, using the libraries Keras and Pytorch.<sup>6</sup> We used three subsequent R to R beats as input to the LSTM models. We picked three since that was the minimum number of heartbeats found in some ECGs. Therefore, the final input size for the LSTM models was (3x256x8).

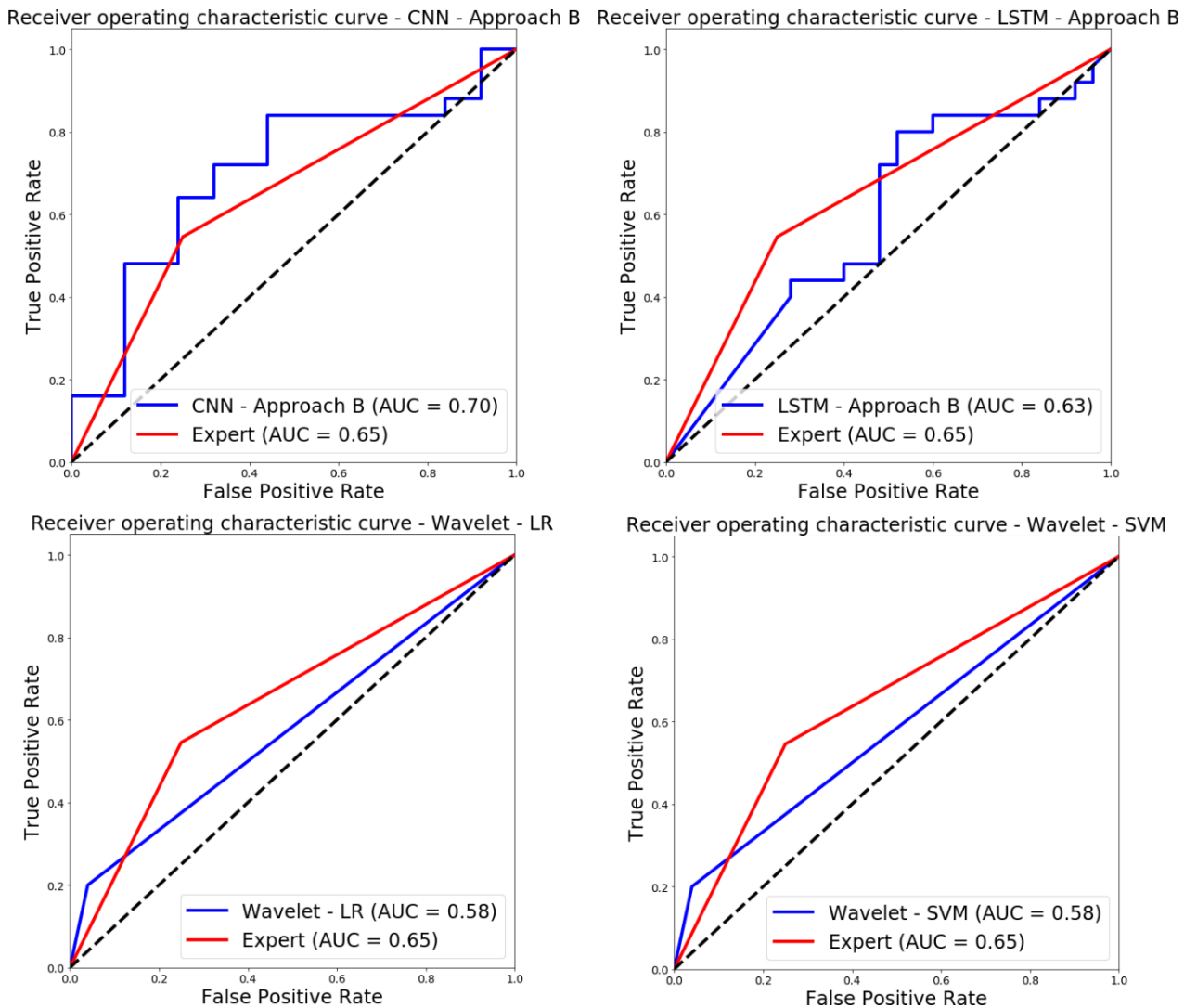
## Supplemental Figures and Tables



**Supplemental Figure I.** Example of the input (8x256) used to develop the models. Each plot corresponds to one of the 8 channels.



**Supplemental Figure II.** Repeater operating curves (ROC) for the best performing expert compared to other DL and ML approaches on the Amsterdam Data.



**Supplemental Figure III.** Repeater operating curves (ROC) for the best performing expert and the best performing models on the Murcia Data.

**Supplemental Table I.** Hyper-parameters used during model optimization. Values in bold resulted in the best validation results and were used to develop the final model.

Method	Hyper-parameter	Value range
CNN/LSTM	Number of convolutional layers followed by max pooling	(3,4)
	Number of filters	[ <b>16,32,64</b> ], [16,16,32], [16,16,32,64]
	Size of convolutional kernels	[1,32], [ <b>1,36</b> ]
	Size of pooling kernel	[0], [ <b>1,2</b> ]
	Number of dense layers	( <b>1</b> ,2)
	Activation	<b>ReLU</b>
	Dropout	[ <b>0.2,0.2,0.2,0.5</b> ], [0.3,0.3,0.3,0.6], [0.2,0.2,0.2,0.2]
	Nodes in dense layer	[128],[ <b>256</b> ],[128,256]
	Epochs	[50, <b>100</b> ,150]
	Learning rate	[ <b>0.00001</b> , 0.0005, 0.0001, 0.001]
	Mini batch size	[64, <b>128</b> ]
	RFC	Number of Trees
Max depth of trees		[2,5, <b>10</b> ]
Quality of split		<b>Gini</b> or Entropy
Minimum number of samples required to split an internal node		[2, <b>4</b> ,6,8]
Minimum number of samples required to be at a leaf node		[2,4,6,8,10]
SVM	Kernel type	Linear, <b>Radial basis function</b> , Polynomial
	Penalty parameter C	[0.001, 0.01, <b>0.1</b> , 1, 10, 100]
	Kernel coefficient $\gamma$ (gamma)	[1, 0.1, 0.01, 0.001, 0.0001]
	Degree of the Polynomial kernel	[2,3,4,5]
LR	Regularization	[0.001, 0.01, <b>0.1</b> , 1, 10, 100]
	Optimization algorithm	[newton-cg, lbfgs, <b>liblinear</b> , sag, saga]
NN	Hidden Layer sizes	[25], [25,25], [25,25,25], [50], [ <b>50,50</b> ], [50,50,50], [100], [100,100], [100,100,100],
	Activation	<b>ReLU</b>
	Regularization parameter	[0.1, 0.01, <b>0.001</b> , 0.0001]
	Batch size	[32, 64, 128]
	Learning rate	[0.01, 0.001, 0.005, <b>0.0001</b> ]
	Optimization algorithm	<b>Adam</b>
XGB	Learning rate	[0.1, <b>0.01</b> , 0.001, 0.005]
	Minimum sum of instance weight (hessian) needed in a child	[1, <b>5</b> , 10]
	Minimum loss reduction required to make a further partition on a leaf node of the tree	[0, 0.5, <b>1</b> , 1.5, 2, 5]

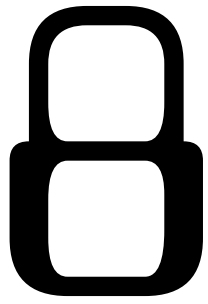
	Subsample ratio of the training instances	[0.7, <b>0.8</b> , 0.9, 1.0]
	Parameters for subsampling the columns	[ <b>0.7</b> ,0.9]
	Maximum depth of a tree	[3, <b>7</b> , 13]
KNN	Number of neighbors	[1, <b>2</b> , 3, 4, 5]
	Weight function	[ <b>uniform</b> , distance]
	Leaf size	[ <b>1</b> , 10, 20 ,30]

**Supplemental Table II.** ECG features visualized by occlusion maps for the True Positives (TP) and True Negatives (TN).

True positives			True negatives		
Feature	N	%	Feature	N	%
T-wave	67	63%	T-wave	3	3%
Whole signal	15	14%	Whole signal	59	53%
Other/random	24	23%	Other/random	49	44%
Total	106	100%	Total	111	100%

### Supplemental References

1. Carreiras C, Alves A, Lourenço A, Canento F, Silva H, Fred A: BioSPPy: Biosignal Processing in Python [Internet]. 2017,. Available from: <https://github.com/PIA-Group/BioSPPy/>
2. Baalman SWE, Schroevers FE, Oakley AJ, et al.: A morphology based deep learning model for atrial fibrillation detection using single cycle electrocardiographic samples. Int J Cardiol 2020; .
3. Baalman SWE, Oakley A, Schroevers FE, et al.: DEEP LEARNING ALGORITHM FOR ATRIAL FIBRILLATION DETECTION BASED ON SINGLE BEAT MORPHOLOGY. Hear Rhythm 2019; .
4. Martis RJ, Acharya UR, Min LC: ECG beat classification using PCA, LDA, ICA and Discrete Wavelet Transform. Biomed Signal Process Control 2013; .
5. Lee G, Gommers R, Waselewski F, Wohlfahrt K, O’Leary A: PyWavelets: A Python package for wavelet analysis. J Open Source Softw 2019; .
6. Keras [Internet]. Available from: <https://github.com/fchollet/keras>





# CHAPTER 8.

## Quantitative analysis of EEG reactivity for neurological prognostication after cardiac arrest

Admiraal MM, Ramos LA\*, Olabarriaga SD, Marquering HA, Horn J, van Rootselaar AF. Quantitative analysis of EEG reactivity for neurological prognostication after cardiac arrest. *Submitted to Clinical Neurophysiology.*

\*Shared first author

## Abstract

*Objective:* To test whether 1) quantitative analysis of EEG reactivity (EEG-R) using machine learning (ML) is superior to visual analysis, and 2) combining quantitative analyses of EEG-R and EEG background pattern increases prognostic value for prediction of poor outcome after cardiac arrest (CA).

*Methods:* ML models were trained with twelve quantitative features derived from EEG-R and EEG background data of 134 adult CA patients. Poor outcome was a Cerebral Performance Category score of 3-5 within 6 months.

*Results:* A random forest classifier trained on EEG-R data was most accurate and predicted poor outcome with 46% sensitivity (95%-CI 40-51) and 89% specificity (95%-CI 86-92). Visual analysis of EEG-R had 80% sensitivity and 65% specificity. A random forest classifier with EEG background data at 24h after CA showed 62% sensitivity (95%-CI 57-67) and 84% specificity (95%-CI 79-88). Combining both classifiers reduced the number of false positives.

*Conclusions:* Quantitative analysis using ML on EEG-R data predicts poor outcome with higher specificity, but lower sensitivity compared to visual analysis, and is of some additional value to ML on EEG background data.

*Significance:* Quantitative EEG-R using ML is a promising alternative to visual analysis and of some added value to ML on EEG background data.

## Introduction

Reliable prognostication is one of the major challenges in patients admitted to the intensive care unit after cardiac arrest (CA) (Rossetti et al. 2016). Hypoxic-ischemic brain injury is the main determinant of neurological outcome (Lemiale et al. 2013). Accurate outcome prediction can reduce unnecessary treatment in patients without a chance of recovery and avoid inappropriate withdrawal of life sustaining treatment. Recent studies have shown that electroencephalography (EEG) contains information that can be used to predict patient outcome with high accuracy (Hofmeijer et al. 2015; Sivaraju et al. 2015; Westhall et al. 2016; Rossetti et al. 2017; Ruijter et al. 2019). A continuous normal voltage EEG background in the first 12 h post CA predicts a good outcome (Hofmeijer et al. 2015; Ruijter et al. 2019). A low voltage EEG background or generalized periodic discharges on a suppressed background when present at 24 h after CA, or presence of identical bursts at any time after CA are strongly associated with a poor outcome. (Hofmeijer et al. 2015; Sivaraju et al. 2015; Westhall et al. 2016; Rossetti et al. 2017; Ruijter et al. 2019).

Another marker for prognostication of a poor outcome is the absence of EEG reactivity (Sandroni et al. 2014; Rossetti et al. 2017; Azabou et al. 2018). However, prognostic value of visual analysis of EEG-R varies widely with sensitivity reported between 60% and 96% and specificity between 67% and 100% (Rossetti et al. 2010; 2012; Alvarez et al. 2013; Noirhomme et al. 2014; Oddo and Rossetti 2014; Suys et al. 2014; Sivaraju et al. 2015; Amorim et al. 2016; Fantaneanu et al. 2016; Rossetti et al. 2017; Tsetsou et al. 2018; Benghanem et al. 2019).

Quantitative analysis utilizing a Machine Learning (ML) approach has the potential to improve the prognostic value of EEG-R by overcoming the subjective nature of the visual assessment. Machine Learning methods have been extensively used with EEG data in the clinical setting for diverse tasks, such as the prediction of epileptic seizures and outcome after postanoxic coma (Usman et al. 2017; Tjepkema-Cloostermans et al. 2019). Most studies regarding consciousness disorders focus on group level classification approaches (Noirhomme et al. 2017). Machine Learning methods are often able to use data more effectively, by modelling the interactions between

features, including non-linearities, and focus on individual level predictions (Noirhomme et al. 2017). Also, a fully automated approach would allow for faster and easier interpretation.

In this study, we aimed to investigate the prognostic value for prediction of poor outcome after CA of ML methods trained on quantitative features extracted from EEG-R data and compared it to the prognostic value of visual analysis of EEG-R. Since it is currently unknown whether ML using EEG-R data is superior to ML using EEG background data for prediction of poor outcome, we also aimed to compare and combine their prognostic values.. We hypothesized 1) that ML classifiers trained on quantitative features from EEG-R data can predict poor outcome more accurately than visual analysis of EEG-R, and 2) that combining ML classifiers of both EEG-R and EEG background increases prognostic value for prediction of poor outcome.

## Methods

### Participants

In this post-hoc analysis, cEEG registrations of 138 adult CA patients were available from a prospective cohort (Admiraal et al. 2019). In brief, between April 2015 and February 2018, consecutive comatose adult CA patients, admitted to the intensive cares of two university hospitals and one large teaching hospital in The Netherlands, where cEEG monitoring was started within 24 hours after CA, were included. Patients were treated according to the local hospital protocol including 24 hours of targeted temperature management with sedation. Decisions for withdrawal of life sustaining treatment were guided by the Dutch recommendations for prognostication after cardiac arrest, which included neurological examination after clearance of sedative drugs, subsequent SSEP testing, and EEG background after 72h post CA (Nederlandse Vereniging voor NeurologieNederlandse Vereniging voor Intensive Care 2011). Recommendations at the time did not include EEG-R in any form, nor EEG background patterns within 72 h after CA. The Medical Ethical Committee of the Amsterdam University Medical Centers, location AMC, waived the need for informed consent. The original trial was registered at [trialregister.nl](http://trialregister.nl) (identifier: NTR6231).

## EEG recordings

cEEG was recorded with nine electrodes, placed according to the international 10-20 system. Additionally, a ground and reference electrode were placed in the midline. EEGs were recorded with a Viasys Nicolet (CareFusion, Middleton, WI) or a BrainQuick ICU (Micromed, Mogliano Veneto, Italy).

During cEEG monitoring, EEG-R was tested twice a day according to a strict protocol, consisting of five different stimuli: clapping, calling out the patient's name, passive eye opening, nasal tickling, and sternal rub. Each stimulus was applied three times, with a duration of five seconds and inter-stimulus interval of 30 seconds. Synchronized with the start of each stimulus, an annotation was placed in the EEG recording to enable offline reviewing. For each patient, the EEG-R assessment of 15 stimuli closest to 24 hours after CA (limited to 12-36 hours, median 23 hours; IQR 19-25) was selected. We did not exclude patients based on the EEG background. Results of visual analysis of EEG-R and EEG background were available (see Table 8.1). Methods of the visual analysis have been described elsewhere (Admiraal et al. 2019) and are summarized in Table 8.2.

## Feature extraction

EEG data were resampled to 256 Hz to ensure equal sampling rate in all recordings. Epochs of five seconds immediately before and after each stimulation were exported. Also, a 5-minute artefact free EEG background clip at 24 hours after CA was extracted and divided into 5-second epochs. Epochs were preprocessed with a bipolar montage and band-pass filtered with a third order Butterworth filter between 0.5 and 30 Hz.

Quantitative features from three domains were extracted from the EEG data: time domain (standard deviation), frequency domain (total power (1-25 Hz), delta power (0.5-4 Hz), theta power (4-8 Hz), alpha power (8-13Hz), beta power (13-25 Hz), alpha/delta ratio, spectral edge frequency, and peak frequency), and information theory (approximate entropy, Shannon entropy (Shannon 1948) and Higuchi fractal dimension (Higuchi 1987)). Power spectra were estimated using a Welch periodogram with a Hamming window and 50% overlap. Features were calculated and averaged over all EEG channels. We selected these features since they have shown to be among the

most relevant features for neurologic prognostication (Bai et al. 2017; Amorim et al. 2019; Rizal and Estananto 2019).

From EEG-R data, features from the epochs before and after each stimulation were subtracted to highlight the EEG changes. Preprocessing and feature extraction were done using Matlab 2017a (Natick, MA, USA). Outliers were excluded based on the upper limit of boxplots (interquartile range) of the feature values. We excluded outliers per patient, to make the comparison between stimuli more fair and trustworthy and to guarantee that the differences between results were due to stimulus type and not differences in the population.

All code used to pre-process the data and develop the models can be found in: [https://github.com/L-Ramos/EEG\\_reactivity](https://github.com/L-Ramos/EEG_reactivity)

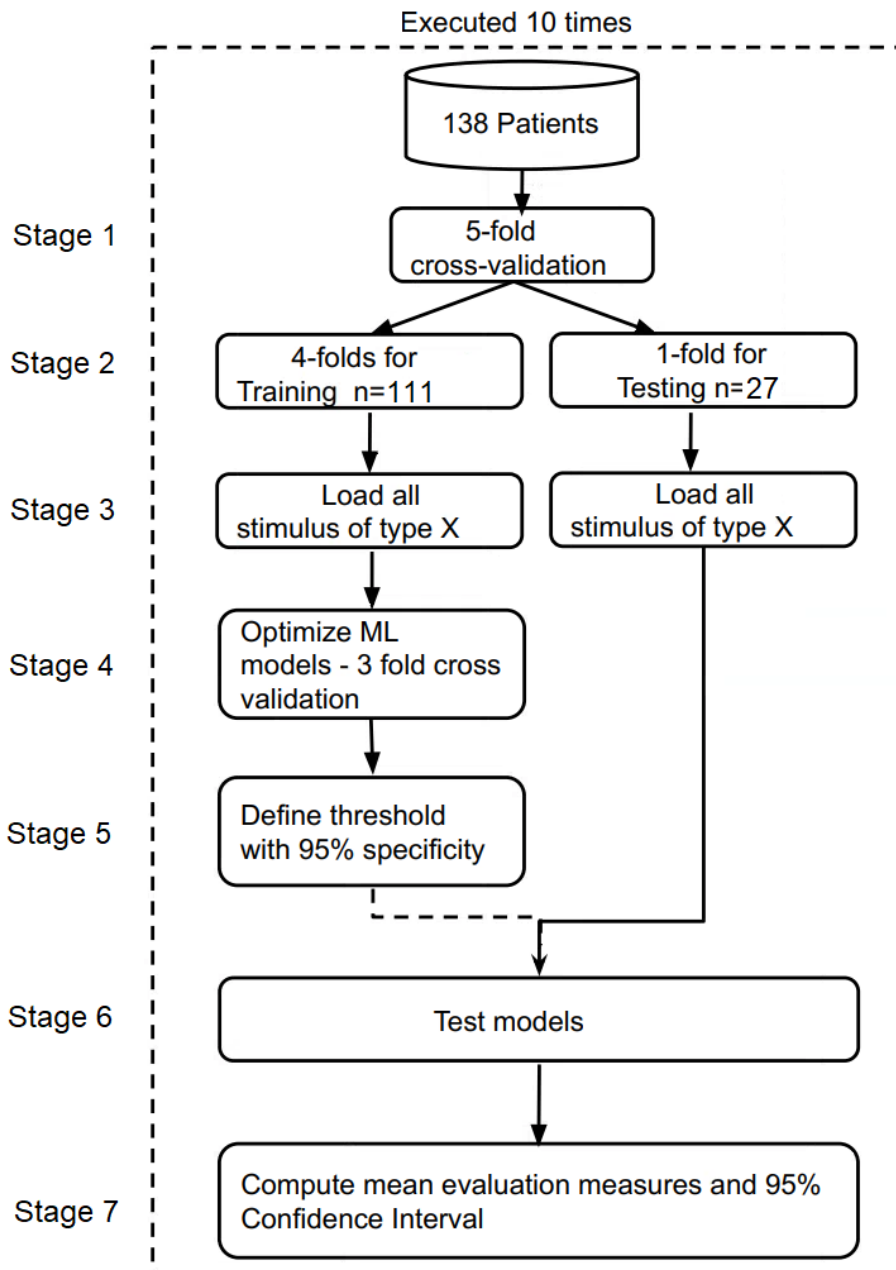
### Outcome assessment

Outcome was assessed as the best score on the Cerebral Performance Category (CPC) scale within six months after CA (Booth et al. 2004). Scores were dichotomized as poor (CPC 3-5: moderate cerebral impairment, vegetative stage, or death, N=54 (39%)) and good neurological outcome (CPC 1-2: nor or mild cerebral impairment, N=84 (61%)). See Table 8.1 for demographic details.

### Machine learning

We selected the following ML methods: Logistic Regression (LR), Support Vector Machine (SVM) (Cortes and Vapnik 1995), Random Forest (RF) (Breiman 2001), Neural Networks (NN) (Bishop 1995), and Gradient Tree Boosting (GTB) (Chen and Guestrin 2016). We included these classifiers since for each one of them, learning occurs in a very different way, with linear and non-linear interactions between variables and to explore the best possible learning setup (Fernández-Delgado et al.). For all classifiers we used implementations available in the Scikit-learn toolkit (Pedregosa et al. 2011). Parameters used for optimization are presented in the online supplement (Supplemental Tables I and II).

Figure 8.1 shows an overview of the workflow. Firstly, we split the dataset into training and testing sets using 5-fold cross validation, preserving the percentage of samples for each outcome class and complete patient data sets in each fold (Figure 8.1, stage 1 and 2). Then, the training set was split into training and validation using 3-fold cross-validation to optimize the hyper-parameters of all models (Figure 8.1, stage 3). The hyper-parameters that generated the best performing model over all iterations (highest mean area under of the receiver operator curve, AUC) were used to create the final model (Figure 8.1, stage 4). The final model was applied to the testing set, and the evaluation measures were computed (Figure 8.1, stage 5). This experimental setup was executed 10 times to obtain more reliable performance estimates (Lemm et al. 2011). To prevent inappropriate withdrawal of life sustaining care, cut-off for the probabilities resulting from the classifier was determined at  $\geq 95\%$  specificity in the training set. In total, 250 models using EEG-R data (50 for each stimulus type) and 50 models using EEG background data were generated.



**Figure 8.1.** Machine Learning Workflow. All five stimulus types were evaluated separately (the type is denoted X in the second stage). In total, 50 models from EEG reactivity data were generated, 10 for each stimulus type (using 5-fold cross validation). The same was done with EEG background data at 24 hours after cardiac arrest, generating 10 models.

To further assess whether the models created with features from EEG-R and EEG background data differ in their predictions, we compared the predictions in a patient-wise comparison, using the results from 5-fold cross-validation. Furthermore, we created an ensemble classifier to check whether by combining both a classifier training on EEG-R and EEG background, we



could reduce the number of miss-classified patients. For the combination, the probabilities of the best performing models using EEG-R and EEG background data were combined by computing the average of each patient-wise prediction.

### Performance assessment

To assess the performance of ML using quantitative features from EEG-R and EEG background data, sensitivity, specificity, positive prediction value (PPV), false positive rate (FPR) and AUC were obtained for each ML classifier on the test set. To estimate the uncertainty introduced by random variations in the data, the average and 95%-CI are computed for all measures (Noirhomme et al. 2017). The difference between models was considered statistically significant if the CI of the AUCs were non-overlapping. We also present the confusion matrices for the leave-one-out experiments.

## Results

### Demographics and visual analysis

Of 138 patients, an EEG-R assessment between 12 and 36 h after CA was available. Based on outlier inspection of the feature values, four more patients were excluded (two with good outcome). Visual inspection of the EEG recordings of those four patients showed muscle artifacts in two patients, burst-suppression EEG in one patient and no clear artifacts in the fourth excluded patient. Demographics and EEG characteristics of included patients can be found in Table 8.1.

**Table 8.1.** Demographics and EEG characteristics of included patients

	<b>Poor outcome n = 60</b>	<b>Good outcome n = 74</b>	<b>p- Value</b>
<b>Demographics</b>			
Age	66 (52-74)	62 (49-69)	0.19
Sex (male)	43/60 (72%)	61/74 (82%)	0.15
OHCA	52/60 (87%)	68/74 (92%)	0.40
Witnessed arrest	42/59 (71%)	62/72 (86%)	0.05
Time to ROSC (min)	22 (15-32)	12 (9-17)	<0.001
Initial rhythm shockable	32/58 (55%)	64/69 (93%)	<0.001
Cardiac etiology	32/53 (60%)	60/69 (87%)	0.001
<b>EEG characteristics</b>			
Time of EEG reactivity assessment (in h since CA)	24 (20-25)	23 (19-24)	0.25
EEG reactive on visual analysis	16/60 (27%)	48/74 (65%)	<0.001
EEG background pattern at 24h after CA			<0.001
Continuous normal voltage	25/57 (44%)	59/71 (83%)	
Discontinuous normal voltage	9/57 (16%)	10/71 (14%)	
Burst-suppression without identical bursts	4/57 (7%)	2/71 (3%)	
Burst-suppression with identical bursts	3/57 (5%)		
Low voltage/suppressed	16/57 (28%)		

Data presented as median (interquartile range) or n/total (%). OHCA: out of hospital cardiac arrest, ROSC: return of spontaneous circulation.

**Table 8.2.** Details of the visual analysis of EEG reactivity and EEG background

	<b>EEG-R</b>	<b>EEG background</b>
<b>Timepoint of assessment</b>	Closest to 24h after CA (limited to 12-36h, median 23h; IQR 19-25)	24h after CA
<b>Raters</b>	3 fully blinded raters	3 fully blinded raters
<b>Categories</b>	- Present - Absent (uncertain classified as present)	According to ACNS criteria (Hirsch et al. 2013): - Continuous normal voltage - Discontinuous normal voltage - Burst-suppression without identical bursts - Burst-suppression with identical bursts - Low voltage/suppressed
<b>Inter-rater reliability</b>	ICC 0.85 (95% CI 0.82-0.88) (Admiraal et al. 2019)	

EEG-R: EEG reactivity, CA: cardiac arrest, IQR: inter quartile range, ACNS: American Clinical Neurophysiology Society, ICC: Intra-class correlation coefficient.

Results of visual analysis of EEG-R were available from our previous study (Admiraal et al. 2019). In the currently analysed patient selection, absence of EEG-R showed 80% sensitivity (95%-CI 66-89) and 65% specificity (95%-CI 54-76) for poor outcome. Burst-suppression with identical bursts, low voltage, or suppressed EEG background at 24 hours after CA on visual analysis predicted poor outcome with 35% sensitivity (95%-CI 22-49) and 100% specificity (95%-CI 95-100).

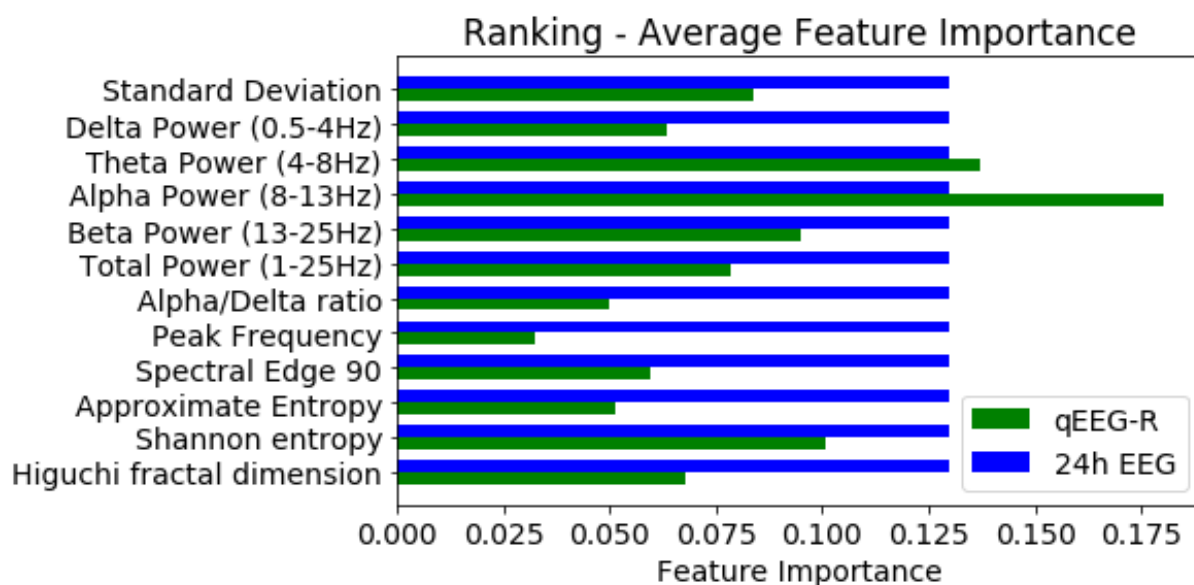
### Quantitative analysis - EEG reactivity data

The highest average AUC for ML using quantitative features from EEG-R data was obtained with the RF classifier and the passive eye opening stimulus AUC: 0.83 (95%-CI 0.80-0.86) with 46% sensitivity (95%-CI 40-51) and 89% specificity (95%-CI 86-92), the highest PPV: 86% (95%-CI 83-90), and lowest FPR 11 (95%-CI 8-14) see Table 8.3. The difference in accuracy between the various stimulus types was not statistically significant. The results obtained with the other classifiers are presented in Supplemental Table III and the standard deviation of the measures is shown in the Supplemental Table IV. To enable the interpretability of the RF model, the average feature importance for the best performing stimulus (the passive eye opening) was extracted and shown in Figure 8.2 (in green).

**Table 8.3.** Performance for prediction of poor outcome of the RF models based on features extracted from EEG-R data of each stimulus type, and EEG background data at 24 hours after CA (24 hours EEG)

	<b>AUC</b> <b>(95%-CI)</b>	<b>Sensitivity</b> <b>(95%-CI)</b>	<b>Specificity</b> <b>(95%-CI)</b>	<b>PPV</b> <b>(95%-CI)</b>	<b>FPR</b> <b>(95%-CI)</b>
<b>Clapping</b>	0.78 (0.76-0.81)	0.41 (0.35-0.48)	0.88 (0.84-0.92)	0.83 (0.77-0.88)	0.12 (0.08-0.16)
<b>Calling out patient's name</b>	0.78 (0.76-0.81)	0.46 (0.40-0.51)	0.86 (0.82-0.90)	0.81 (0.76-0.87)	0.14 (0.10-0.18)
<b>Passive eye opening</b>	0.83 (0.80-0.86)	0.46 (0.40-0.51)	0.89 (0.86-0.92)	0.86 (0.83-0.90)	0.11 (0.08-0.14)
<b>Nasal tickle</b>	0.81 (0.79-0.83)	0.45 (0.38-0.52)	0.87 (0.83-0.92)	0.80 (0.73-0.87)	0.13 (0.08-0.17)
<b>Sternal rub</b>	0.76 (0.73-0.79)	0.30 (0.23-0.36)	0.88 (0.84-0.92)	0.73 (0.64-0.81)	0.12 (0.08-0.16)
<b>24h EEG</b>	0.85 (0.83-0.88)	0.62 (0.57-0.67)	0.84 (0.79-0.88)	0.85 (0.81-0.89)	0.16 (0.12-0.21)

Sensitivity and specificity are optimized for prediction of poor outcome with the probability threshold shifted towards high specificity.



**Figure 8.2.** Average importance of the twelve features included in the random forest classifier extracted from EEG reactivity data and from EEG background data at 24 hours after cardiac arrest.

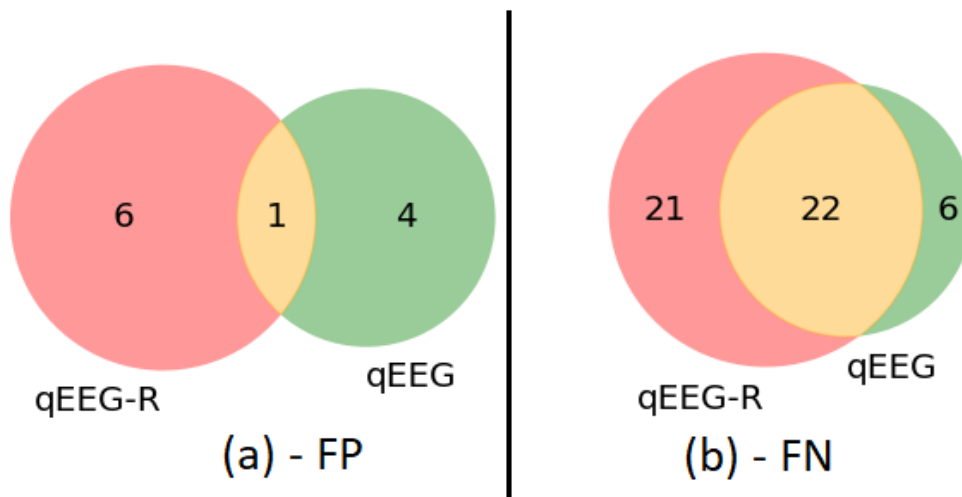
### Quantitative analysis - EEG background data

The highest average AUC for the EEG background data at 24 hours resulted from the RF classifier, with an AUC of 0.85 (95%-CI 0.83-0.88). Sensitivity was 62% (95%-CI 57-67) with a specificity of 84% (95%-CI 79-88). Despite

showing a high PPV value 85% (95%-CI 81 - 89), the EEG background had worse FPR 16% (95%-CI 12 - 21) when compared to EEG reactivity. The results obtained with the other classifiers are available in Supplemental Table III. The average feature importance is shown in Figure 8.2 (in blue).

### Quantitative analysis - Patient-wise comparison

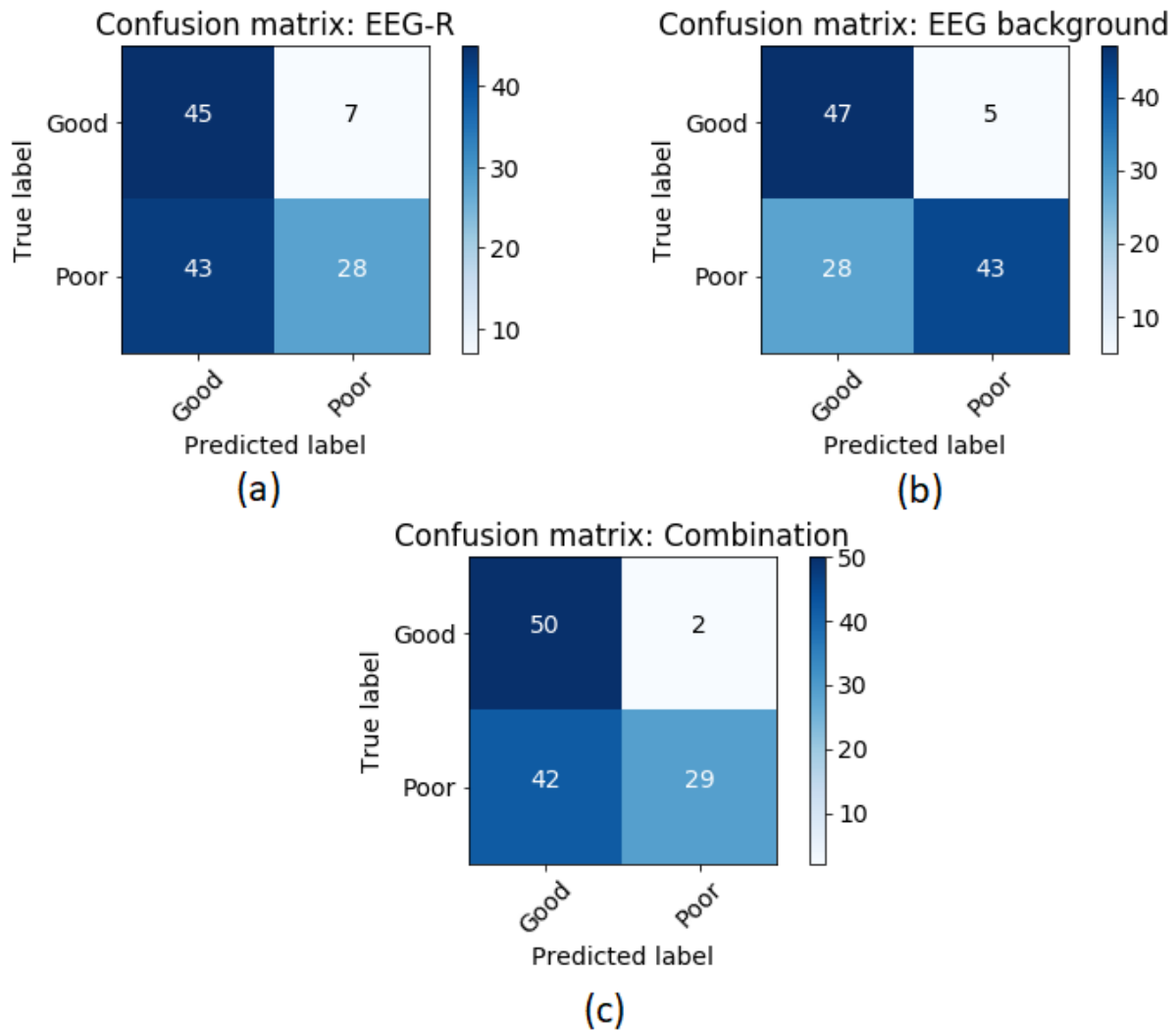
In the patient-wise comparison, FP and FN predictions of ML models of EEG-R data from the passive eye opening stimulus and EEG background data at 24 h after CA, were compared. (Figure 8.3). The FP predictions (Figure 8.3a) show an overlap of 11% (1 patient) and FN predictions (Figure 8.3b) 45% (22 patients) For both predictions there is no complete overlap, showing additional value.



**Figure 8.3.** Venn diagram for random forest classifier predictions built with quantitative features extracted from EEG reactivity data and EEG background data.

A) False positive predictions and B) False negative predictions.

In Figure 8.4, we show the confusion matrices of the RF models based on EEG-R data (a) from the passive eye opening stimulus, EEG background data at 24 hours and their combination using the averaged probabilities. The combination of probabilities of the RF models using EEG-R and EEG background data resulted in an AUC of 85%, 41% sensitivity and 96% specificity. The combination reduced the number of FP from 7 (EEG-R) and 5 (EEG background) to 2.



**Figure 8.4.** Confusion matrix of the averaged probabilities of the two random forest classifier models based on EEG-R data from the nasal tickle stimulus and EEG background data at 24 hours. Labels are good (CPC 1-2) and poor neurological outcome (CPC 3-5).

## Discussion

This post-hoc analysis of a prospective cohort study on the prognostic value of EEG-R after CA showed that ML models trained on quantitative features extracted from EEG-R data can predict poor neurological outcome with higher specificity than visual analysis of EEG-R data. ML on EEG background data at 24 hours after CA outperformed ML with EEG-R data in this cohort with promising prognostic values. We also found that combining probabilities resulting from ML models using EEG-R and EEG background data slightly improved the number of FP in outcome prediction.

EEG-R is widely used and visual evaluation of EEG-R is restricted to highly trained personnel and time consuming. Quantitative analysis of EEG-R has major advantages over visual analysis: It is faster, it is not subjected to inter-rater variability, and it can be fully automated. A few other studies have described quantitative analysis of EEG-R data. These studies indicate that this approach is at least as good as visual analysis, even though methodologies vary widely (Noirhomme et al. 2014; Hermans et al. 2015; Liu et al. 2016; Duez et al. 2018). Several studies calibrated the quantitative algorithm to obtain the highest concordance with EEG-R as determined by visual analysis, not to patient outcome (Noirhomme et al. 2014; Hermans et al. 2015). They found substantial agreement of quantitative analysis with visual analysis by EEG-certified neurologists. Prediction of poor outcome using quantitative analysis compared to outcome prediction using visual EEG-R analysis was reported to have 100% specificity for poor outcome (Duez et al. 2018). This study used only one feature as quantitative analysis and did not perform separate training and testing, nor performed cross-validation (Duez et al. 2018). A recent article by (Amorim et al. 2019) described a ML model similar to ours for prediction of good outcome after cardiac arrest with good results. All these results indicate that quantitative EEG-R analysis is a tool which could substitute visual analysis of EEG-R in hypoxic-ischemic encephalopathy.

Besides EEG-R, the already available EEG background pattern is known to predict neurological outcome in postanoxic coma with high accuracy, using both visual and quantitative analysis (Hofmeijer et al. 2015; Sivaraju et al. 2015; Spalletti et al. 2016; Westhall et al. 2016; Asgari 2018; Ruijter et al. 2019; Tjepkema-Cloostermans et al. 2019). A substantial number of studies on quantitative analysis of EEG background for prognostication after CA report similar prognostic value as visual analysis. Again, studies vary widely in methodology (Asgari 2018). A recent study by Nagaraj et al. (Nagaraj et al. 2018) with very similar methodology to our study reported 60% sensitivity and 100% specificity for prediction of poor outcome at 24 hours after CA, whereas we found 72% sensitivity and 84% specificity. Major differences from our study are a much larger number of investigated features (44 vs. 12) and a much larger sample size (551 vs. 138).

The limited number of samples in our study did not allow us to assess a large number of features, since the accuracy of ML methods may be reduced when the number of training samples is small compared to the number of features. This is also known as the curse of dimensionality (Duda et al. 2001). We accounted for this by extracting only well-known EEG features available in the literature (Bao et al. 2011). These features also largely overlapped with the features most often selected by the random forest model in the study on quantitative EEG-R by (Amorim et al. 2019). The limited sample size could also have made our study more prone to outliers. However, we expect that the effect would have been similar for quantitative analyses of both EEG-R and EEG background and not disproportionately at the expense of EEG-R.

Despite optimizing a probability threshold to  $\geq 95\%$  specificity, the specificity of some models was still lower than our threshold. Since our current dataset is relatively small, there is a possibility that the specificity threshold found with our validation set did not generalize well for the testing set. Nevertheless, estimating a threshold using the validation set is essential to prevent overoptimistic results and to keep our development setting as close to clinical practice as possible, where prior knowledge about the “testing” patients is not available.

In our study we opted for a multi-subject approach. We trained a model in a population of multiple patients which can later be applied to a single patient. One of the main advantages over the individualized approach is the robustness to bias, since features are combined in a multivariate and non-linear way, which can highlight previously undetected information (Noirhomme et al. 2017).

One of our aims was to investigate whether a certain stimulus could be more predictive than others. Therefore, we did not combine all stimulus types into one single dataset. Moreover, in case a certain stimulus type was less predictive than others, that could lead to negative effects in learning. Furthermore, most machine learning methods (like the SVM and NN) assume that the data is independently and identically distributed (IID) (Dundar et al. 2007), and combining all stimulus types could severely hurt this principle.



Our study is the first to compare and combine quantitative analysis of EEG-R and EEG background to predict poor outcome. Since EEG-R is an additional test during EEG registration, EEG background is always available, irrespectively whether EEG-R is tested or not. We found a small increase in prognostic accuracy when classifiers based on EEG-R and EEG background data were combined. This is in line with findings by Kustermann et al. where spectral analysis of the EEG background combined with visual analysis of EEG reactivity led to a higher prognostic value than spectral analysis of the EEG background alone (Kustermann et al. 2019). Although, in our study, EEG-R and EEG background were not tested exactly at the same time after CA (EEG-R between 12-36h vs EEG background at 24h), we do not assume that this was of major influence on the results at the expense of EEG-R. Most EEG-R assessments were in fact very close to 24 hours after CA as shown by the narrow inter-quartile ranges.

Major strengths of our study are: a prospective data set with standardized EEG-R testing, selection of multiple ML classifiers that have shown state-of-the-art results in many studies, follow-up data collected until six months to allow for patient recovery, leading to a more reliable label, and the number of cross-validation iterations, enabling a fair assessment of potential generality of the method. Since EEG reactivity and EEG background in any form were not part of decisions to withdraw life sustaining treatment, we assume our results are not biased by a self-fulfilling prophecy. As EEG-R was tested relatively early, most patients received sedative medication during the recording. In our previous publication we found no difference in sedation between reactive and unreactive patients (Admiraal et al. 2019). Also, EEG-R was more often seen in EEGs recorded during sedation. Therefore, we assume this is not a confounding factor in the current investigation.

## Conclusion

Our results indicate that ML using quantitative features extracted from EEG-R data predicts poor outcome after cardiac arrest with higher specificity than visual analysis of EEG-R and is therefore a promising alternative to visual analysis. ML using quantitative features from EEG background at 24 hours after CA outperforms ML using features from EEG-R data. Combined

probabilities of the models resulted in a decrease in false positives, showing some additional value of EEG-R for prediction of poor outcome. The value for prediction of good outcome, however, should be subject of future research.

### References

- Admiraal MM, van Rootselaar A-F, Hofmeijer J, Hoedemaekers CWE, van Kaam CR, Keijzer HM, et al. Electroencephalographic reactivity as predictor of neurological outcome in postanoxic coma: A multicenter prospective cohort study. *Ann Neurol*. 2019;86 (1):17–27.
- Alvarez V, Sierra-Marcos A, Oddo M, Rossetti AO. Yield of intermittent versus continuous EEG in comatose survivors of cardiac arrest treated with hypothermia. *Crit Care*. 2013;17 (5):R190.
- Amorim E, Rittenberger JC, Zheng JJ, Westover MB, Baldwin ME, Callaway CW, et al. Continuous EEG monitoring enhances multimodal outcome prediction in hypoxic-ischemic brain injury. *Resuscitation*. 2016;109:121–6.
- Amorim E, van der Stoel M, Nagaraj SB, Ghassemi MM, Jing J, O'Reilly U-M, et al. Quantitative EEG reactivity and machine learning for prognostication in hypoxic-ischemic brain injury. *Clin Neurophysiol*. 2019;130 (10):1908–16.
- Asgari S. Quantitative measures of EEG for prediction of outcome in cardiac arrest subjects treated with hypothermia: a literature review. *J Clin Monit Comput*. 2018;32 (6):977–92.
- Azabou E, Navarro V, Kubis N, Gavaret M, Heming N, Cariou A, et al. Value and mechanisms of EEG reactivity in the prognosis of patients with impaired consciousness: a systematic review. *Crit Care*. 2018;22 (1):184.
- Bai Y, Xia X, Li X. A Review of Resting-State Electroencephalography Analysis in Disorders of Consciousness. *Front Neurol*. 2017;8:471.
- Bao FS, Liu X, Zhang C. PyEEG: an open source Python module for EEG/MEG feature extraction. *Comput Intell Neurosci*. 2011;2011 (2):406391–7.
- Benghanem S, Paul M, Charpentier J, Rouhani S, Ben Hadj Salem O, Guillemet L, et al. Value of EEG reactivity for prediction of neurologic outcome after cardiac arrest: Insights from the Parisian registry. *Resuscitation*. 2019;142:168–74.
- Bishop CM. *Neural Networks for Pattern Recognition*. Oxford; 1995.
- Booth CM, Boone RH, Tomlinson G, Detsky AS. Is This Patient Dead, Vegetative, or Severely Neurologically Impaired?: Assessing Outcome for Comatose Survivors of Cardiac Arrest. *JAMA*. 2004;291 (7):870–9.
- Breiman L. Random Forests. *Machine Learning* 2001;45 (1):5–32.
- Chen T, Guestrin C. XGBoost. *KDD '16*. New York, New York, USA: ACM Press; 2016. pp. 785–94.
- Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995;20 (3):273–97.
- Duda RO, Hart PE, Stork DG. *Pattern Classification 2nd Edition*. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2001.

- Duez CHV, Ebbesen MQ, Benedek K, Fabricius M, Atkins MD, Beniczky S, et al. Large inter-rater variability on EEG-reactivity is improved by a novel quantitative method. *Clin Neurophysiol.* 2018;129 (4):724–30.
- Dundar M, Krishnapuram B, Bi J, Bharat Rao R. Learning Classifiers When the Training Data Is Not IID. *IJCAI'07: Proceedings of the 20th international joint conference on Artificial intelligence.* 2007;756–61.
- Fantaneanu TA, Tolchin B, Alvarez V, Friolet R, Avery K, Scirica BM, et al. Effect of stimulus type and temperature on EEG reactivity in cardiac arrest. *Clin Neurophysiol.* 2016;127 (11):3412–7.
- Fernández-Delgado M, Cernadas E, machine SBTJO, 2014. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* 2014;14:3133-81
- Hermans MC, Westover MB, van Putten MJAM, Hirsch LJ, Gaspard N. Quantification of EEG reactivity in comatose patients. *Clin Neurophysiol.* 2015;127 (1):571–80.
- Higuchi T. Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena.* 1987;31 (2):277–83.
- Hirsch LJ, LaRoche SM, Gaspard N, Gerard E, Svoronos A, Herman ST, et al. American Clinical Neurophysiology Society's Standardized Critical Care EEG Terminology. *J Clin Neurophysiol.* 2013;30 (1):1–27.
- Hofmeijer J, Beernink TMJ, Bosch FH, Beishuizen A, Tjepkema-Cloostermans MC, van Putten MJAM. Early EEG contributes to multimodal outcome prediction of postanoxic coma. *Neurology.* 2015;85:137–43.
- Kustermann T, Nguissi NAN, Pfeiffer C, Haenggi M, Kurmann R, Zubler F, et al. Electroencephalography-based power spectra allow coma outcome prediction within 24 h of cardiac arrest. *Resuscitation.* 2019;1–6.
- Lemiale V, Dumas F, Mongardon N, Giovanetti O, Charpentier J, Chiche J-D, et al. Intensive care unit mortality after cardiac arrest: the relative contribution of shock and brain injury in a large cohort. *Intensive Care Med.* 2013;39 (11):1972–80.
- Lemm S, Blankertz B, Dickhaus T, Müller K-R. Introduction to machine learning for brain imaging. *Neuroimage.* 2011;56 (2):387–99.
- Liu G, Su Y, Jiang M, Chen W, Zhang Y, Zhang Y, et al. Electroencephalography reactivity for prognostication of post-anoxic coma after cardiopulmonary resuscitation: A comparison of quantitative analysis and visual analysis. *Neurosci Lett.* 2016;626:74–8.
- Nagaraj SB, Tjepkema-Cloostermans MC, Ruijter BJ, Hofmeijer J, van Putten MJAM. The revised Cerebral Recovery Index improves predictions of neurological outcome after cardiac arrest. *Clin Neurophysiol.* 2018;129 (12):2557–66.
- Nederlandse Vereniging voor Neurologie, Nederlandse Vereniging voor Intensive Care. Richtlijn Prognose van Post-Anoxisch Coma. 2011. Available from: <https://nvic.nl/sites/nvic.nl/files/Richtlijnen%20aanmaken/Richtlijn%20Postanoxisch%20coma.pdf>
- Noirhomme Q, Brecheisen R, Lesenfants D, Antonopoulos G, Laureys S. “Look at my classifier's result”: Disentangling unresponsive from (minimally) conscious patients. *Neuroimage.* 2017;145 (Pt B):288–303.

Noirhomme Q, Lehembre R, Lugo ZDR, Lesenfants D, Luxen A, Laureys S, et al. Automated analysis of background EEG and reactivity during therapeutic hypothermia in comatose patients after cardiac arrest. *Clin EEG Neurosci.* 2014;45 (1):6–13.

Oddo M, Rossetti AO. Early multimodal outcome prediction after cardiac arrest in patients treated with hypothermia. *Crit Care Med.* 2014;42 (6):1340–7.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011;12 (Oct):2825–30.

Rizal A, Estananto R. Epileptic EEG Signal Classification Using Multiresolution Higuchi Fractal Dimension. *Int J Eng Res Technol.* 2019;12 (4):508–11.

Rossetti AO, Carrera E, Oddo M. Early EEG correlates of neuronal injury after brain anoxia. *Neurology.* 2012;78 (11):796–802.

Rossetti AO, Rabinstein AA, Oddo M. Neurological prognostication of outcome in patients in coma after cardiac arrest. *Lancet Neurol.* 2016;15 (5):597–609.

Rossetti AO, Tovar Quiroga DF, Juan E, Novy J, White RD, Ben-Hamouda N, et al. Electroencephalography Predicts Poor and Good Outcomes After Cardiac Arrest: A Two-Center Study. *Crit Care Med.* 2017;45 (7):e674–82.

Rossetti AO, Urbano LA, Delodder F, Kaplan PW, Oddo M. Prognostic value of continuous EEG monitoring during therapeutic hypothermia after cardiac arrest. *Crit Care.* 2010;14 (5):R173.

Ruijter BJ, Tjepkema-Cloostermans MC, Tromp SC, Bergh WM, Foudraïne NA, Kornips FHM, et al. Early EEG for outcome prediction of postanoxic coma: a prospective cohort study. *Ann Neurol.* 2019;86 (2):203–14.

Sandroni C, Cariou A, Cavallaro F, Cronberg T, Friberg H, Hoedemaekers C, et al. Prognostication in comatose survivors of cardiac arrest: An advisory statement from the European Resuscitation Council and the European Society of Intensive Care Medicine. *Intensive Care Med.* 2014;40 (12):1816–31.

Shannon CE. A Mathematical Theory of Communication. *Bell System Technical Journal.* 1948;27 (3):379–423.

Sivaraju A, Gilmore EJ, Wira CR, Stevens A, Rampal N, Moeller JJ, et al. Prognostication of post-cardiac arrest coma: early clinical and electroencephalographic predictors of outcome. *Intensive Care Med.* 2015;41 (7):1264–72.

Spalletti M, Carrai R, Scarpino M, Cossu C. Single electroencephalographic patterns as specific and time-dependent indicators of good and poor outcome after cardiac arrest. *Clin Neurophysiol.* 2016;127:2610–7.

Suys T, Bouzat P, Marques-Vidal P, Sala N, Payen J-F, Rossetti AO, et al. Automated Quantitative Pupillometry for the Prognostication of Coma After Cardiac Arrest. *Neurocrit Care.* 2014;21 (2):300–8.

Tjepkema-Cloostermans MC, da Silva Lourenço C, Ruijter BJ, Tromp SC, Drost G, Kornips FHM, et al. Outcome Prediction in Postanoxic Coma With Deep Learning. *Crit Care Med.* 2019;47 (10):1424–32.

Tsetsou S, Novy J, Pfeiffer C, Oddo M, Rossetti AO. Multimodal Outcome Prognostication After Cardiac Arrest and Targeted Temperature Management: Analysis at 36 °C. *Neurocrit Care.* 2018;28 (1):104–9.

Usman SM, Usman M, Fong S. Epileptic Seizures Prediction Using Machine Learning Methods. *Comput Math Methods Med.* 2017;2017 (7):1–10.

Westhall E, Rossetti AO, van Rootselaar A-F, Kjær TW, Horn J, Ullén S, et al. Standardized EEG interpretation accurately predicts prognosis after cardiac arrest. *Neurology.* 2016;86 (16):1482–90.

## Supplemental material

**Supplemental Table I.** Hyper-parameters used for SVM (Support Vector Machine)

Classifier	Kernel Type	Penalty parameter $C$	Kernel coefficient $\gamma$	Degree of the Polynomial kernel
SVM	Linear	[0.001, 0.01, 0.1, 1, 10, 100]	n.a.	n.a.
	Radial basis function	[0.001, 0.01, 0.1, 1, 10, 100]	[1, 0.1, 0.01, 0.001, 0.0001]	n.a.
	Polynomial	[0.001, 0.01, 0.1, 1, 10, 100]	[1, 0.1, 0.01, 0.001, 0.0001]	[1,2,3,4]

**Supplemental Table II.** Hyper-parameters used for optimizing RFC (Random Forest Classifier), NN (Neural Network), LR (Logistic Regression) and GTB (Gradient Tree Boosting)

Classifier	Parameter Name	Parameter Value
RFC	Number of Trees	[50,100,200,400,500,600,800,1000]
	Max features for split	auto, sqrt and log2,1,2,4,7
	Quality of split	Gini or Entropy
	Max Depth	10, 20, 30
	Minimum number of samples per leaf	2, 3, 4, 5, 6, 7
	Minimum number of samples required to split an internal node	2, 3, 4, 5, 6, 7
NN	Hidden Layer sizes	[12], [24], [36], [24,12], [24,36,12]
	Regularization	[0.1, 0.01, 0.001, 0.0001]
	Batch size	[32, 64]
	Learning rate	[0.01, 0.001, 0.0001]
	Optimizer Activation	SGD, Adam Relu, Logistic
LR	Regularization	0.001, 0.01, 0.1, 1, 10, 100
	Optimization Algorithm	newton-cg, lbfgs, liblinear, sag, saga
GTB	Learning rate	0.1, 0.01, 0.001
	Minimum sum of instance weight needed in a child	1, 5, 10
	Regularization	0, 0.5, 1, 1.5, 2, 5
	Subsample ratio of the training instances	0.7, 0.8, 0.9, 1.0
	Subsample ratio of columns when constructing each tree	0.3, 0.4, 0.5, 0.6, 0.7, 0.8
Maximum depth of a tree	3, 5, 7, 9, 10	

**Supplemental Table III.** Performance for prediction of poor outcome (average and 95% confidence interval) of the support vector machine (SVM), linear regression (LR), gradient tree boosting (GTB), and neuronal network (NN) models based on features extracted from EEG-R data of each stimulus type, and EEG background data at 24 h after CA (24 h EEG). Sensitivity and specificity are optimized for prediction of poor outcome with the probability threshold shifted towards high specificity.

	<b>AUC (95%-CI)</b>	<b>Sensitivity (95%-CI)</b>	<b>Specificity (95%-CI)</b>	<b>PPV (95% - CI)</b>	<b>FPR (95% -CI)</b>
<b>Clapping</b>					
	0.78	0.41	0.88	0.83	0.12
<b>RFC</b>	(0.76 - 0.81)	(0.35 - 0.48)	(0.84 - 0.92)	(0.77 - 0.88)	(0.08 - 0.16)
	0.75	0.55	0.66	0.73	0.34
<b>SVM</b>	(0.72 - 0.78)	(0.45 - 0.65)	(0.55 - 0.77)	(0.66 - 0.80)	(0.23 - 0.45)
	0.77	0.39	0.89	0.74	0.11
<b>LR</b>	(0.74 - 0.80)	(0.33 - 0.45)	(0.86 - 0.92)	(0.66 - 0.82)	(0.08 - 0.14)
	0.76	0.3	0.89	0.73	0.11
<b>GTB</b>	(0.74 - 0.79)	(0.28 - 0.43)	(0.85 - 0.92)	(0.64 - 0.82)	(0.08 - 0.15)
	0.75	0.49	0.77	0.79	0.23
<b>NN</b>	(0.72 - 0.77)	(0.41 - 0.57)	(0.67 - 0.86)	(0.73 - 0.84)	(0.14 - 0.33)
<b>Calling out patient's name</b>					
	0.78	0.46	0.86	0.81	0.14
<b>RFC</b>	(0.76 - 0.81)	(0.40 - 0.51)	(0.82 - 0.90)	(0.76 - 0.87)	(0.10 - 0.18)
	0.80	0.52	0.77	0.79	0.23
<b>SVM</b>	(0.78 - 0.83)	(0.44 - 0.60)	(0.68 - 0.85)	(0.74 - 0.85)	(0.15 - 0.32)
	0.78	0.45	0.85	0.81	0.15
<b>LR</b>	(0.75 - 0.80)	(0.39 - 0.50)	(0.81 - 0.89)	(0.75 - 0.86)	(0.11 - 0.19)
	0.77	0.42	0.86	0.80	0.14
<b>GTB</b>	(0.75 - 0.80)	(0.36 - 0.48)	(0.82 - 0.89)	(0.74 - 0.85)	(0.11 - 0.18)
	0.78	0.52	0.77	0.79	0.23
<b>NN</b>	(0.75 - 0.81)	(0.45 - 0.59)	(0.69 - 0.85)	(0.74 - 0.84)	(0.15 - 0.31)
<b>Passive eye opening</b>					
	0.83	0.46	0.89	0.86	0.11
<b>RFC</b>	(0.80 - 0.86)	(0.40 - 0.51)	(0.86 - 0.92)	(0.83 - 0.90)	(0.08 - 0.14)
	0.75	0.49	0.84	0.86	0.16
<b>SVM</b>	(0.72 - 0.78)	(0.41 - 0.56)	(0.76 - 0.92)	(0.81 - 0.90)	(0.08 - 0.24)
	0.76	0.30	0.88	0.68	0.12
<b>LR</b>	(0.72 - 0.79)	(0.23 - 0.37)	(0.83 - 0.93)	(0.58 - 0.78)	(0.07 - 0.17)
	0.80	0.47	0.87	0.82	0.13
<b>GTB</b>	(0.78 - 0.83)	(0.41 - 0.53)	(0.83 - 0.90)	(0.77 - 0.87)	(0.10 - 0.17)
	0.73	0.62	0.53	0.65	0.47
<b>NN</b>	(0.70 - 0.76)	(0.52 - 0.73)	(0.40 - 0.66)	(0.57 - 0.73)	(0.34 - 0.60)
<b>Nasal tickle</b>					
	0.81	0.45	0.87	0.80	0.13
<b>RFC</b>	(0.79 - 0.83)	(0.38 - 0.52)	(0.83 - 0.92)	(0.73 - 0.87)	(0.08 - 0.17)
	0.77	0.46	0.79	0.78	0.21
<b>SVM</b>	(0.75 - 0.80)	(0.39 - 0.54)	(0.72 - 0.86)	(0.73 - 0.83)	(0.14 - 0.28)
	0.75	0.35	0.84	0.59	0.16
<b>LR</b>	(0.72 - 0.78)	(0.27 - 0.43)	(0.78 - 0.90)	(0.49 - 0.69)	(0.10 - 0.22)


Supplemental Table III (continued)

	0.79	0.42	0.86	0.76	0.14
<b>GTB</b>	(0.76 - 0.81)	(0.35 - 0.49)	(0.83 - 0.90)	(0.68 - 0.83)	(0.10 - 0.17)
	0.74	0.61	0.54	0.66	0.46
<b>NN</b>	(0.71 - 0.77)	(0.51 - 0.70)	(0.42 - 0.67)	(0.59 - 0.72)	(0.33 - 0.58)
<b>Sternal rub</b>					
	0.76	0.30	0.88	0.73	0.12
<b>RFC</b>	(0.73 - 0.79)	(0.23 - 0.36)	(0.84 - 0.92)	(0.64 - 0.81)	(0.08 - 0.16)
	0.67	0.56	0.59	0.69	0.41
<b>SVM</b>	(0.64 - 0.70)	(0.45 - 0.66)	(0.48 - 0.70)	(0.62 - 0.76)	(0.30 - 0.52)
	0.71	0.28	0.80	0.54	0.20
<b>LR</b>	(0.67 - 0.74)	(0.20 - 0.37)	(0.72 - 0.87)	(0.44 - 0.64)	(0.13 - 0.28)
	0.74	0.21	0.90	0.46	0.10
<b>GTB</b>	(0.71 - 0.77)	(0.14 - 0.27)	(0.86 - 0.94)	(0.35 - 0.57)	(0.06 - 0.14)
	0.71	0.40	0.67	0.43	0.33
<b>NN</b>	(0.68 - 0.74)	(0.28 - 0.52)	(0.56 - 0.79)	(0.33 - 0.54)	(0.21 - 0.44)
<b>24 h EEG</b>					
	0.85	0.62	0.84	0.85	0.16
<b>RFC</b>	(0.83 - 0.88)	(0.57 - 0.67)	(0.79 - 0.88)	(0.81 - 0.89)	(0.12 - 0.21)
	0.85	0.72	0.72	0.82	0.28
<b>SVM</b>	(0.82 - 0.87)	(0.65 - 0.79)	(0.63 - 0.80)	(0.78 - 0.87)	(0.20 - 0.37)
	0.86	0.58	0.88	0.80	0.12
<b>LR</b>	(0.83 - 0.88)	(0.50 - 0.65)	(0.85 - 0.92)	(0.72 - 0.88)	(0.08 - 0.15)
	0.86	0.63	0.85	0.86	0.15
<b>GTB</b>	(0.83 - 0.88)	(0.57 - 0.69)	(0.81 - 0.89)	(0.82 - 0.89)	(0.11 - 0.19)
	0.86	0.66	0.77	0.78	0.23
<b>NN</b>	(0.84 - 0.88)	(0.58 - 0.73)	(0.69 - 0.85)	(0.71 - 0.85)	(0.15 - 0.31)



**Supplemental Table IV.** Performance for prediction of poor outcome (average and standard deviation) of the random forest (RFC), support vector machine (SVM), linear regression (LR), gradient tree boosting (GTB), and neuronal network (NN) models based on features extracted from EEG-R data of each stimulus type, and EEG background data at 24 h after CA (24 h EEG). Sensitivity and specificity are optimized for prediction of poor outcome with the probability threshold shifted towards high specificity.

	<b>AUC (95%-CI)</b>	<b>Sensitivity (95%-CI)</b>	<b>Specificity (95%-CI)</b>	<b>PPV (95% - CI)</b>	<b>FPR (95% -CI)</b>
<b>Clapping</b>					
<b>RFC</b>	0.78 ± (0.09)	0.41 ± (0.23)	0.88 ± (0.14)	0.84 ± (0.15)	0.12 ± (0.14)
<b>SVM</b>	0.75 ± (0.11)	0.55 ± (0.34)	0.66 ± (0.37)	0.73 ± (0.24)	0.34 ± (0.37)
<b>LR</b>	0.77 ± (0.10)	0.39 ± (0.21)	0.89 ± (0.10)	0.79 ± (0.21)	0.11 ± (0.10)
<b>GTB</b>	0.76 ± (0.09)	0.36 ± (0.26)	0.89 ± (0.13)	0.83 ± (0.17)	0.11 ± (0.13)
<b>NN</b>	0.75 ± (0.09)	0.49 ± (0.26)	0.77 ± (0.32)	0.79 ± (0.18)	0.23 ± (0.32)
<b>Calling out patient's name</b>					
<b>RFC</b>	0.78 ± (0.09)	0.46 ± (0.20)	0.86 ± (0.13)	0.83 ± (0.15)	0.14 ± (0.13)
<b>SVM</b>	0.80 ± (0.10)	0.52 ± (0.28)	0.77 ± (0.30)	0.79 ± (0.20)	0.23 ± (0.30)
<b>LR</b>	0.78 ± (0.10)	0.45 ± (0.20)	0.85 ± (0.14)	0.81 ± (0.20)	0.15 ± (0.14)
<b>GTB</b>	0.77 ± (0.09)	0.42 ± (0.21)	0.86 ± (0.12)	0.81 ± (0.15)	0.14 ± (0.12)
<b>NN</b>	0.78 ± (0.09)	0.52 ± (0.24)	0.77 ± (0.27)	0.79 ± (0.17)	0.23 ± (0.27)
<b>Passive eye opening</b>					
<b>RFC</b>	0.83 ± (0.09)	0.46 ± (0.19)	0.89 ± (0.11)	0.86 ± (0.13)	0.11 ± (0.11)
<b>SVM</b>	0.75 ± (0.11)	0.49 ± (0.26)	0.84 ± (0.27)	0.86 ± (0.15)	0.16 ± (0.27)
<b>LR</b>	0.76 ± (0.12)	0.30 ± (0.23)	0.88 ± (0.18)	0.79 ± (0.25)	0.12 ± (0.18)
<b>GTB</b>	0.80 ± (0.10)	0.47 ± (0.21)	0.87 ± (0.12)	0.84 ± (0.14)	0.13 ± (0.12)
<b>NN</b>	0.73 ± (0.10)	0.62 ± (0.36)	0.53 ± (0.44)	0.72 ± (0.19)	0.47 ± (0.44)
<b>Nasal tickle</b>					
<b>RFC</b>	0.81 ± (0.07)	0.45 ± (0.24)	0.87 ± (0.15)	0.86 ± (0.13)	0.13 ± (0.15)
<b>SVM</b>	0.77 ± (0.09)	0.46 ± (0.27)	0.79 ± (0.24)	0.78 ± (0.16)	0.21 ± (0.24)
<b>LR</b>	0.75 ± (0.11)	0.35 ± (0.27)	0.84 ± (0.20)	0.71 ± (0.25)	0.16 ± (0.20)
<b>GTB</b>	0.79 ± (0.08)	0.42 ± (0.24)	0.86 ± (0.14)	0.80 ± (0.19)	0.14 ± (0.14)
<b>NN</b>	0.74 ± (0.10)	0.61 ± (0.33)	0.54 ± (0.42)	0.70 ± (0.17)	0.46 ± (0.42)
<b>Sternal rub</b>					
<b>RFC</b>	0.76 ± (0.10)	0.30 ± (0.23)	0.88 ± (0.13)	0.77 ± (0.24)	0.12 ± (0.13)
<b>SVM</b>	0.67 ± (0.11)	0.56 ± (0.37)	0.59 ± (0.38)	0.69 ± (0.23)	0.41 ± (0.38)
<b>LR</b>	0.71 ± (0.12)	0.28 ± (0.28)	0.80 ± (0.27)	0.66 ± (0.28)	0.20 ± (0.27)
<b>GTB</b>	0.74 ± (0.11)	0.21 ± (0.22)	0.90 ± (0.13)	0.70 ± (0.27)	0.10 ± (0.13)
<b>NN</b>	0.71 ± (0.11)	0.40 ± (0.41)	0.67 ± (0.41)	0.63 ± (0.25)	0.33 ± (0.41)
<b>24 h EEG</b>					
<b>RFC</b>	0.85 ± (0.08)	0.62 ± (0.18)	0.84 ± (0.15)	0.85 ± (0.13)	0.16 ± (0.15)
<b>SVM</b>	0.85 ± (0.09)	0.72 ± (0.23)	0.72 ± (0.30)	0.82 ± (0.16)	0.28 ± (0.30)
<b>LR</b>	0.86 ± (0.09)	0.58 ± (0.26)	0.88 ± (0.13)	0.87 ± (0.17)	0.12 ± (0.13)
<b>GTB</b>	0.86 ± (0.08)	0.63 ± (0.21)	0.85 ± (0.13)	0.86 ± (0.12)	0.15 ± (0.13)
<b>NN</b>	0.86 ± (0.08)	0.66 ± (0.25)	0.77 ± (0.29)	0.83 ± (0.15)	0.23 ± (0.29)



9

# CHAPTER 9.

Discussion

## General Discussion

Artificial intelligence has been broadly applied in many clinical applications (1–8). These applications experienced a significant increase in performance when compared to classical computer vision and modelling approaches due to the learning capabilities offered by machine/deep learning models. Despite improved performance, machine/deep learning models are often hard to interpret, which can raise doubts about their trustworthiness for implementation in clinical practice (9). Moreover, given the large amount of data generated during patient care, it is challenging to develop models that can include all this information into account in a safe and trustworthy manner. This thesis focused on applications of machine/deep learning to cardiovascular diseases that, when applicable, combine multiple types of data and their role in the prediction task, while striving to make the models interpretable and transparent.

## Stroke

In this thesis, multiple models were developed and evaluated for various stroke-related tasks. The prediction of functional outcome at 3 months, represented by the modified Ranking Scale (mRS), was assessed in chapters 3 to 6. While prediction of functional outcome was often addressed in the literature (10–12), we explored a significantly larger number of variables using a large and heterogeneous dataset, the MR CLEAN registry, which contains data from patients from multiple hospitals all over the Netherlands (13). Moreover, we developed extensive and robust training and validation pipelines to prevent overoptimistic or biased results, and, when possible, externally validated the models. Despite the large number of variables included and the use of state-of-the-art machine learning models, the performance was similar to models developed using only a smaller number of known clinical predictors (10). We also showed that prediction models for good and poor mRS prediction have similar performance and can be optimized to greatly reduce the number of false positives, which can be useful in some clinical scenarios (14,15). Finally, by using various model visualization techniques we identified various variables that are associated with patient outcome, such as: age, collaterals, glucose level, baseline NIHSS, Glasgow Coma Scale, time from onset-to-groin, time from onset-to-first

hospital, pre-stroke mRS, among others (16,17). Despite many of these variables already had been identified in previous stroke research, this shows that our models were often using clinically relevant variables for prediction, increasing the trustworthiness of our results.

We also explored other tasks related to stroke, such as the prediction of brain tissue reperfusion, represented by the dichotomized modified Thrombolysis In Cerebral Infarction score (m-TICI) in patients with acute ischemic stroke, and Delayed Cerebral Ischemia (DCI) for patients with hemorrhagic stroke. For both tasks, although prediction accuracy from machine learning models was generally intermediate, our approaches involving deep learning often led to a significant improvement in prediction accuracy compared to baseline models, especially when automatically generated image features were taken into account during model development (chapters 2, 4 and 5). The increase in prediction accuracy compared to baseline models and traditional approaches was even more evident in the case of DCI, since the random forest models performed significantly better than logistic regression. Such increase in accuracy for DCI prediction suggests that, in some applications, the nonlinearities supported by machine learning methods can have a positive impact. Therefore, we suggest that future reperfusion and DCI prediction models should consider such machine learning approaches and automatic image features.

## Heterogeneous Data Combination

In this thesis, it was shown that the combination of multiple types of data, such as imaging, patient demographics and clinical scores, can lead to improvements in prediction accuracy and the discovery of new insights. We proposed multiple approaches for the combination of image and clinical data in chapters 2, 4 and 5. In chapter 2, we trained an auto-encoder on non-contrast CT images and used the output features from the encoding part as a feature generator for the scans. This approach greatly reduced the number of features in the images, reducing the risk of overfitting while keeping the most relevant features. Auto-encoder is an unsupervised method and does not require annotations, making it more feasible for application to large datasets.

Moreover, this unsupervised technique does not depend on any labels for training, making it more adaptable to new datasets and less biased.

We further studied the predictive value of image features in chapter 4, where we developed Residual Neural Networks (ResNets) and Structured Receptive Fields (RFNN) models, and compared the predictive value of models trained on CT angiography (CTA) scans with radiological scores (18,19). While models trained on CTA were as good as models trained on the manually/visual scores for the prediction of functional outcome, relevant image regions could be identified through model visualization using GRAD-CAM. For example, GRAD-CAM visualizations showed that the deep learning models focused mostly on the vessel occlusion location. Moreover, many brain regions that were considered relevant for outcome and reperfusion prediction in previous studies (20) were also deemed important in our machine learning approaches (chapter 5). In future work, other image modalities could be considered for combination, such as non-contrast CT and CT perfusion, since scores (such as ASPECTS) computed using these image modalities are associated with patient outcome (21).

Another promising field in deep learning is multi-task learning, where multiple tasks can be learned simultaneously to reduce the need for individual models per task and may subsequently significantly improve performance (22). The positive influence of multi-task learning has been shown by multiple studies. For example, the combination of four ICU-related prediction tasks using recurrent models was proposed in (23), and multi-task learning significantly outperformed the same models trained on the tasks individually. Mortality prediction at different moments using multi-task learning was addressed in (24), and again, significant improvements were found when compared to training of single tasks. Therefore, such approach could be extended to predict multiple stroke outcomes at once, incorporating multi image modalities in one model, and/or predicting genetic mutations such as phospholamban (PLN) and Long QT syndrome (LQTS) (25,26).

## Phospholamban

Regarding the identification of patient with the phospholamban (PLN) p.Arg14del gene mutation in ECG signals, we successfully developed

machine and deep learning models to classify PLN. Our models were able to identify patients even before they started showing symptoms, and the models performed better than experienced cardiologists in different PLN datasets. Furthermore, we assessed the robustness of our models by externally validating them in a dataset from Spain. Through this external validation, we found that a wavelet approach, despite performing well during internal validation, did not generalize well to the external dataset, which highlights the importance of external validation. Finally, through model visualization, we identified the T-wave to be the most important ECG region for PLN identification (26,27). In the future, provided that more data is available, models could be trained using the whole ECG signals instead of single heartbeats, which could lead to the discovery of new relevant ECG features.

## Cardiac Arrest

We compared the predictive value of EEG reactivity with EEG background for predicting poor outcome at 6 months after cardiac arrest and found that EEG background at 24 hours was the most predictive signal. Moreover, we evaluated multiple types of stimuli in EEG reactivity and found that passive eye opening was the most predictive stimulus. We showed that, when data is limited, feature engineering is a feasible option that can yield state-of-the-art results that are comparable to visual analysis by an expert. Future work should focus on a broader range of EEG features to be included in the model, as well as their combination with patient demographics.

## Limitations

There are some limitations shared among the studies in this thesis. Despite some models showing high AUC values in many chapters, currently there is no agreement about the minimum AUC value for a model to be considered for use in clinical practice, since different tasks may have different accuracy requirements (28). In our studies, we often included a large number of clinical variables in the prediction models, which often led to improvements in performance. A disadvantage of prediction models with so many variables is that all these variables have to be computed and available at baseline, which can be problematic in some cases because their assessment often requires specialized knowledge.

Since machine learning models are often deemed hard to interpret, we applied multiple model visualization techniques to make our models more transparent and trustworthy (16,17). Nevertheless, the results of such techniques are still not always easily interpretable and generalizable to the whole patient population, and further improvements are still necessary. Communication between data scientists and clinicians is of utmost importance to ensure that the models developed are robust and can offer interesting insights in daily use. Finally, despite the development of extensive training and validation pipelines, the applied internal validation approach does not replace the need of external validation. External validation was not possible in some studies due to the sensitive nature of data and patient privacy.

## Conclusion

In this thesis, we investigated the added value of machine/deep learning approaches for prediction modelling in multiple cardiovascular related conditions. It was shown that such approaches can lead to significant improvements in prognosis and diagnosis, provided that they are properly implemented and validated. Despite the controversy regarding interpretability, machine learning tools are slowly becoming more common in clinical research and practice thanks to model visualization techniques. Finally, the role of machine learning tools is often only of guidance and assistance, and the specialists should be the ones responsible for decision making.



---

## References

1. Nielsen A, Hansen MB, Tietze A, Mouridsen K. Prediction of Tissue Outcome and Assessment of Treatment Effect in Acute Ischemic Stroke Using Deep Learning. *Stroke*. 2018;49 (6):1394–401.
2. Tjepkema-Cloostermans MC, da Silva Lourenço C, Ruijter BJ, Tromp SC, Drost G, Kornips FHM, et al. Outcome Prediction in Postanoxic Coma With Deep Learning. *Crit Care Med*. 2019;47 (10):1424–32.
3. Salem M, Taheri S, Yuan JS. ECG Arrhythmia Classification Using Transfer Learning from 2- Dimensional Deep CNN Features. 2018 IEEE Biomed Circuits Syst Conf BioCAS 2018 - Proc. 2018;
4. Shen D, Wu G, Suk H-I. Deep Learning in Medical Image Analysis. *Annu Rev Biomed Eng*. 2017;19 (1):221–48.
5. Kuang H, Najm M, Chakraborty D, Maraj N, Sohn SI, Goyal M, et al. Automated aspects on noncontrast CT scans in patients with acute ischemic stroke using machine learning. *Am J Neuroradiol*. 2019;40 (1):33–8.
6. Alawieh A, Zaraket F, Alawieh MB, Chatterjee AR, Spiotta A. Using machine learning to optimize selection of elderly patients for endovascular thrombectomy. *J Neurointerv Surg*. 2019;1–6.
7. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: A 5-year multicentre prospective registry analysis. *Eur Heart J*. 2017;38 (7):500–7.
8. Giacobbe DR. Clinical interpretation of an interpreTable prognostic model for patients with COVID-19. *Nat Mach Intell* [Internet]. 2020;42256. Available from: <http://dx.doi.org/10.1038/s42256-020-0207-0>
9. Chandler C, Foltz PW, Elvevåg B. Using Machine Learning in Psychiatry: The Need to Establish a Framework That Nurtures Trustworthiness. *Schizophr Bull*. 2020;46 (1):11–4.
10. Venema E, Mulder MJHL, Roozenbeek B, Broderick JP, Yeatts SD, Khatri P, et al. Selection of patients for intra-arterial treatment for acute ischaemic stroke: Development and validation of a clinical decision tool in two randomised trials. *BMJ*. 2017;357.
11. Alaka SA, Menon BK, Brobbey A, Williamson T, Goyal M, Demchuk AM, et al. Functional Outcome Prediction in Ischemic Stroke: A Comparison of Machine Learning Algorithms and Regression Models. *Front Neurol*. 2020;
12. Forkert ND, Verleger T, Cheng B, Thomalla G, Hilgetag CC, Fiehler J. Multiclass support vector machine-based lesion mapping predicts functional outcome in ischemic stroke patients. *PLoS One*. 2015;10 (6):1–16.
13. Jansen IGH, Mulder MJHL, Goldhoorn RJB. Endovascular treatment for acute ischaemic stroke in routine clinical practice: Prospective, observational cohort study (MR CLEAN Registry). *BMJ*. 2018;360.
14. Goyal M, Almekhlafi MA, Cognard C, McTaggart R, Blackham K, Biondi A, et al. Which patients with acute stroke due to proximal occlusion should not be treated with endovascular thrombectomy? *Neuroradiology*. 2018;3–8.

15. Sarraj A, Albright K, Barreto AD, Boehme AK, Sitton CW, Choi J, et al. Optimizing prediction scores for poor outcome after intra-arterial therapy in anterior circulation acute ischemic stroke. *Stroke*. 2013;44 (12):3324–30.
16. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. Explainable AI for Trees: From Local Explanations to Global Understanding. 2019;1–72. Available from: <http://arxiv.org/abs/1905.04610>
17. Ribeiro MT, Singh S, Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. 2016;1135–44. Available from: <http://arxiv.org/abs/1602.04938>
18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. 2016;2016-Decem:770–8.
19. Jacobsen J-H, van Gemert J, Lou Z, Smeulders AWM. Structured Receptive Fields in CNNs. 2016; Available from: <http://arxiv.org/abs/1605.02971>
20. Ernst M, Boers AMM, Aigner A, Berkhemer OA, Yoo AJ, Roos YB, et al. Association of Computed Tomography Ischemic Lesion Location with Functional Outcome in Acute Large Vessel Occlusion Ischemic Stroke. *Stroke*. 2017;48 (9):2426–33.
21. Barber PA, Demchuk AM, Zhang J, Buchan AM. Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy. *Lancet*. 2000;355 (9216):1670–4.
22. Seraj RM. Multi-task Learning. *Learn to Learn*. 1997;75:95–133.
23. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. *Sci data [Internet]*. 2019;6 (1):96. Available from: <http://dx.doi.org/10.1038/s41597-019-0103-9>
24. Si Y, Roberts K. Deep Patient Representation of Clinical Notes via Multi-Task Learning for Mortality Prediction. *AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci [Internet]*. 2019;2019:779–88. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/31259035> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6568068>
25. Amin AS, Pinto YM, Wilde AAM. Long QT syndrome: Beyond the causal mutation. *Journal of Physiology*. 2013.
26. Van Der Zwaag PA, Van Rijsingen IAW, Asimaki A, Jongbloed JDH, Van Veldhuisen DJ, Wiesfeld ACP, et al. Phospholamban R14del mutation in patients diagnosed with dilated cardiomyopathy or arrhythmogenic right ventricular cardiomyopathy: Evidence supporting the concept of arrhythmogenic cardiomyopathy. *Eur J Heart Fail*. 2012;
27. Van Rijsingen IAW, Van Der Zwaag PA, Groeneweg JA, Nannenberg EA, Jongbloed JDH, Zwinderman AH, et al. Outcome in phospholamban R14del carriers results of a large multicentre cohort study. *Circ Cardiovasc Genet*. 2014;
28. English PA, Williams JA, Martini JF, Motzer RJ, Valota O, Buller RE. A case for the use of receiver operating characteristic analysis of potential clinical efficacy biomarkers in advanced renal cell carcinoma. *Futur Oncol*. 2016;12 (2):175–82.
29. Sapatinas T. Statistics for high-dimensional data. *J Appl Stat*. 2012.





# SUMMARY

Artificial intelligence in the prognostication and  
classification of cardiovascular diseases

In this thesis, I presented multiple applications of Artificial Intelligence in the field of cardiovascular-related diseases. Medical care for stroke patients involves multiple decisions that range from quick diagnosis and treatment selection, to proper monitoring for possible complications. A large amount of data is generated during stroke care such as signals, scores, patient characteristics, medical history, and images. Given the magnitude of medical data, it is possible that a lot of valuable information lies hidden in such data. I explored the predictive value of such data in chapters 2-6. I used various machine/deep learning methods to predict patient and treatment outcome as well as complications after ischemic and hemorrhagic stroke.

Delayed cerebral ischemia is a severe complication, which might occur after hemorrhagic stroke and is considered very difficult to predict. In **Chapter 2**, I explored the prediction of delayed cerebral ischemia using a combination of patient demographics, image scores and non-contrast CT features automatically extracted using deep/machine learning methods. I found that machine/deep learning can significantly improve prediction accuracy, especially when image features are taken into account.

I explored the prediction of functional outcome for patients who suffered a ischemic stroke in chapters 3-6.

**Chapter 3** is about the potential value of machine learning compared to other models from the literature. I developed models using a significantly larger number of variables than previous works, extensively validated multiple machine learning models and identified previously known and unknown predictive biomarkers.

**Chapter 4** describes a study about the predictive value of CT angiography for predicting functional outcome and reperfusion using deep learning approaches. I explored residual networks, structured receptive fields and unsupervised transfer learning using auto-encoders. I trained all models using the maximum intensity projection of the CT angiography scans. In many cases, I found a significant improvement in pre-training with auto-encoders and using structured receptive fields. I also identified relevant predictive regions in the brain using a state-of-the-art model visualization technique.

**Chapter 5** shows the predictive value of combining CT angiography with patient demographics and medical history to predict functional outcome and reperfusion. I focused on deep learning approaches that could take into account the 3D nature of the CT scans and explored the effect of attention (to direct focus during network training) using squeeze and excitation modules in our networks. I also investigated transfer learning from other medical related tasks and used Shapley additive explanations to visualize our models and increase transparency. I found a significant improvement in prediction for reperfusion when image data was combined with patient demographics, medical history and image derived scores. However, this improvement was not observed for predicting functional outcome. I also found that transfer learning can significantly improve accuracy and assist in training such large models when data is not abundant as it can be in other domains. Moreover, model visualization assisted in interpreting and validating our models.

**Chapter 6** focused on the prediction of poor functional outcome in. The accurate prediction of poor outcome could be used to prevent futile treatment, reducing the risks for complications and enabling a more efficient use of resources. I included all variables available at baseline and developed and validated multiple machine learning models. Despite achieving high areas under the curve and specificity, all models still presented a small number of false positives (non-poor outcome predicted as poor), which can hamper their use in clinical practice.

**Chapter 7** shifted focus to the identification of the cardiac genetic mutation Phospholamban. I developed deep/machine learning models to identify this mutation in asymptomatic patients and explored multiple approaches using 1D and 2D convolutional neural networks, recurrent models and wavelets transformations. The results were compared against multiple experts and externally validated in data from another center. I found that our models outperformed all specialists for most measures and that the approaches using convolutional neural networks generalized well in the external dataset.

Finally, in **Chapter 8** I investigated electroencephalograms for predicting outcome of patients in comatose after suffering of cardiac arrest. Multiple stimuli were applied to the patients to assess responsiveness. Therefore, I explored the predictive value of stimuli versus background signal using a

machine learning approach. I further compared our models to the visual analysis by experts. Feature engineering was applied to extract relevant information from the signals and to train multiple models. The results suggested that the background signal has more predictive value than the stimuli-related signals and that the models performed as good as visual analysis by experts.

To conclude, this thesis showed that machine/deep learning can be successfully applied to multiple prediction and classification tasks in the field of cardiovascular diseases and can lead to significant improvements in prognosis accuracy. In many studies, various sources of relevant information is not included in the prediction models due to the complex nature of data. If such information is properly pre-processed and included in a modelling pipeline, it can have a significant impact in model performance and lead to the discovery of new insights. Despite machine/deep learning models often being referred as “black box”, I also showed that various techniques can be applied to interpret their decision process and to increase transparency.







# NEDERLANDSE SAMENVATTING

Kunstmatige intelligentie voor de prognose en  
classificatie van hart- en vaatziekten

In deze thesis rapporteer ik over meerdere toepassingen van artificiële intelligentie die kunnen bijdragen aan een betere behandeling van beroerte. Tijdens de medische behandeling van patiënten met een beroerte worden er meerdere beslissingen gemaakt: van een snelle diagnose en behandelingskeuze tot aan de juiste monitoring van mogelijke complicaties na behandeling. Daarbij wordt er zeer veel data gecreëerd, waaronder fysiologische signalen, scores, patiënt karakteristieken, medische geschiedenis en medische beelden. Gegeven de hoeveelheid data, is het mogelijk dat er veel waardevolle informatie in deze data verborgen ligt. Daarom onderzoek ik de voorspellende waarde van deze data in hoofdstukken 2 t/m 6. Ik heb meerdere machine/deep learning methoden gebruikt voor het voorspellen van zowel de behandeluitkomst als de complicaties na herseninfarct of hersenbloeding.

Vertraagde cerebrale ischemie is een zware complicatie die mogelijk plaatsvindt na een hersenbloeding en moeilijk te voorspellen is. In **hoofdstuk 2** heb ik onderzocht of vertraagde cerebrale ischemie te voorspellen is door het combineren van patiënt demografie, scores van beeldvorming en kenmerken van CT-beelden. De resultaten laten zien dat machine/deep learning de nauwkeurigheid van het voorspellen van vertraagde cerebrale ischemie significant verbetert, vooral wanneer de beeldkenmerken zijn meegenomen.

In de hoofdstukken 3 t/m 6 heb ik onderzocht of ik de functionele uitkomst na de behandeling van patiënten met een herseninfarct kan voorspellen.

In **hoofdstuk 3** gaat over de potentiële waarde van machine learning vergeleken met die van andere modellen uit de literatuur. Ik heb modellen ontwikkeld die gebruik maken van een aanzienlijk groter aantal variabelen dan eerder beschreven in de literatuur. Ook heb ik al bekende, maar ook onbekende voorspellende biomarkers geïdentificeerd.

**Hoofdstuk 4** beschrijft een studie naar de waarde van CT-angiografie voor het voorspellen van functionele uitkomst en reperfusie, gebruik makend van deep learning methodes. Hiervoor heb ik ‘residual networks’, ‘structured receptive fields’ en ‘auto-encoders’ die getraind zijn met behulp van ‘unsupervised transfer learning’ gebruikt. Deze 4 verschillende modellen zijn

getraind op basis van de maximale intensiteit projectie van CT-scans. Voor de ‘structured receptive fields’ en de ‘auto-encoders’ vond ik in veel gevallen een significante verbetering in de voor-training. Ook heb ik met behulp van state-of-the-art visualisatie technieken relevante gebieden in de hersenen met voorspellende waarde geïdentificeerd.

**Hoofdstuk 5** behandelt de voorspellende waarde van het combineren van CT-angiografie met de demografie en medische geschiedenis van de patiënt voor het voorspellen van functionele uitkomst en reperfusie. Hierbij heb ik me gefocust op deep learning methodes die de 3D informatie van CT-scans kunnen meenemen. Binnen deze netwerken heb ik de ‘effects of attention’ onderzocht met behulp van ‘squeeze’ en ‘excitation’ modules, om hiermee het netwerk direct tijdens trainen te focussen op belangrijke features. Ook heb ik ‘transfer learning’ toegepast op basis van andere medisch gerelateerde taken en ‘Shapely additive explanations’ toegepast om de modellen te visualiseren en daarmee transparant te maken. De resultaten lieten een significante verbetering zien van het voorspellen van reperfusie wanneer de beelden werden gecombineerd met de demografie en medische geschiedenis van de patiënt en medische scores op basis van de beeldvorming. De resultaten lieten geen verbetering zien voor het voorspellen van functionele uitkomst. Ook zag ik dat ‘transfer learning’ de nauwkeurigheid van voorspellingen significant kan verbeteren en daarmee kan bijdragen in het trainen van grote modellen zoals in deze studie, wanneer data niet overvloedig is zoals in andere domeinen. Verder assisteerde de visualisatie van de modellen in het interpreteren en valideren van onze modellen.

**Hoofdstuk 6** gaat over het voorspellen van slechte functionele uitkomst. Het nauwkeurig voorspellen van slechte uitkomst kan gebruikt worden om overbodige behandeling te voorkomen, waarbij het risico van complicaties kan worden verminderd en medische middelen efficiënter kunnen worden gebruikt. Ik heb alle beschikbare baseline variabelen geïnccludeerd en meerdere machine learning modellen gevalideerd. Ondanks dat ik hoge ‘area under the curve’ en specificiteit van de resultaten heb bereikt, lieten de modellen nog steeds een klein aantal vals positieven zien (geen slechte uitkomst wordt als slechte uitkomst voorspeld), wat de medische toepassing kan belemmeren.

**Hoofdstuk 7** verplaatst de focus naar het hart, specifiek het identificeren van de genetische cardiale mutatie ‘Phospholamban’. Ik heb deep/machine learning modellen ontwikkeld die deze mutatie identificeert in asymptomatische patiënten en onderzochten meerdere 1D en 2D ‘convolutional neural networks’, ‘recurrent models’ en ‘wavelets transformations’. De resultaten werden vergeleken met meerdere experts en extern gevalideerd op data uit een ander centrum. De resultaten lieten zien dat onze modellen voor de meeste maten beter presteerden dan de experts en dat de ‘convolutional neural networks’ goed generaliseerden in de externe dataset.

Als laatste, in **hoofdstuk 8**, werden electroencefalogrammen onderzocht voor het voorspellen van de uitkomst van patiënten die in coma zijn geraakt na een hartstilstand. De patiënten ondergingen meerdere stimuli om hun responsiviteit te bepalen. Ik heb de voorspellende waarde van het signaal ten tijde van de stimuli vergeleken met het achtergrond signaal. ik heb de modellen vergeleken met de visuele beoordeling van experts. Ik past ‘feature engineering’ toe om relevante informatie van de signalen te extraheren en heb meerdere modellen getraind. De resultaten lieten zien dat het achtergrond signaal een hogere voorspellende waarde heeft vergeleken met de stimulus-gerelateerde signalen en dat de modellen net zo goed presteerden als de experts in de beoordeling.

Concluderend, het onderzoek in deze thesis laat zien dat machine/deep learning succesvol kan worden toegepast in verschillende voorspellingen en classificatie taken binnen cardiovasculaire ziektes en daarmee een significante verbetering kan geven in de prognostische nauwkeurigheid. In veel studies wordt veel relevante informatie niet meegenomen in predictie modellen door de complexe aard van de data. Als informatie op een juiste wijze wordt voorbereid en wordt meegenomen in een model, kan dit een significante impact hebben op de prestatie van een model en resulteren in nieuwe inzichten. Ondanks dat machine/deep learning modellen vaak als ‘black box’ wordt gezien, heb ik ook laten zien dat technieken toegepast kunnen worden om modellen te interpreteren en hun keuzeprocessen inzichtelijk te maken.







# ABBREVIATIONS

AE	Auto-Encoders
AI	Artificial intelligence
AOL	Arterial Occlusive Lesion
aSAH	Aneurysmal Subarachnoid Hemorrhage
ASPECTS	Alberta Stroke Program Early CT Score
AUC	Area Under the Curve
AUPRC	Area Under the Precision Recall Curve
CA	Cardiac Arrest
CBS	Clot Burden Score
cEEG	Continuous Electroencephalogram
CNN	Convolutional Neural Network
CNTK	The Microsoft Cognitive Toolkit
CRP	C-Reactive Protein
CT	Computed Tomography
CTA	Computed Tomography Angiography
DCI	Delayed Cerebral Ischemia
DL	Deep Learning
DSA	Digital Subtraction Angiography
ECG	Electrocardiogram
EEG	Electroencephalogram
EEG-R	Electroencephalogram Reactivity
EVT	Endovascular Treatment
GCS	Glasgow Coma Scale
GPUs	Graphical Processing Units
Grad-CAM	Gradient-weighted Class Activation Mapping
GTB	Gradient Tree Boosting
GWGBP	Gradient Weighted Guided Backpropagation
ICA	Internal Carotid Artery
ICA-T	Internal Carotid Artery Terminus
ICU	Intensive Care Unit
INR	International Normalized Ratio
IV	Intravenous Alteplase
LASSO	Least Absolute Shrinkage and Selection Operator
LIME	Local Interpretable Model-agnostic Explanations
LPBA40	Laboratory of Neuro Imaging Probabilistic Brain Atlas

LR	Logistic Regression
LVO	Large Vessel Occlusion
MICE	Multiple Imputation by Chained Equations
MLP	Multi-layer Perceptron
MR CLEAN	Registry Multicenter Randomized Clinical Trial of Endovascular Treatment for Acute Ischemic Stroke in the Netherlands
mRS	modified Rankin Scale
mTICI	Modified Thrombolysis in Cerebral Infarction
NIHSS	National Institutes of Health Stroke Scale
NN	Neural Network
PCA	Principal Component Analysis
PLN	Phospholamban
ReLU	Rectified Linear Unit
ResNet	Residual Neural Network
RFC	Random Forest Classifier
RFNN	Structured Receptive Field Network
ROC	Operating Characteristic Curve
SAH	Subarachnoid Hemorrhagic Stroke
SDCAE	Stacked Denoising Convolutional Auto-encoder
SE	Squeeze and Excitation
SVM	Support Vector Machine
TBV	Total Blood Volume
TIA	Transient Ischemic Attack
ULTRA	Ultra-Early Tranexamic Acid After Subarachnoid Hemorrhage
WFNS	World Federation of Neurosurgical Societies
XGB	Gradient boosting



# PHD PORTIFOLIO

**Name PhD student:** Lucas Alexandre Ramos

**PhD period:** 09/2016 – 09/2020

**Name PhD supervisor:** Dr. H.A. Marquering and Dr. S.D Olabarriaga

<b>PhD training</b>	<b>Year</b>	<b>ECTS</b>
<b>General courses</b>		
E-science	2016	0.6
Unix	2016	0.5
Entrepreneurship in Health and Life Sciences	2017	1.5
Practical Biostatistics	2017	1.1
Oral Presentation in English	2018	0.8
Scientific Writing in English for Publication	2018	1.5
<b>Specific courses</b>		
Machine Learning (Coursera)	2016	2.0
Advanced Patter Recognition (ASCI)	2017	1.5
Deep Learning Course UvA (Masters in Artificial Intelligence)	2017	1.5
NFBIA – Summer School on Medical Image Analysis	2017	1.5
	2018	1.5
MISS – Summer School - Medical Imaging meets Deep Learning	2018	1.0
IQ Winter School – Machine Learning Applied to Quantitative Analysis of Medical Images	2019	1.5
ASCI - Computer Vision by Learning		
<b>Seminars, workshops and master classes</b>		
Weekly Cardiovascular Engineering Meeting	2016-2020	4.0

Weekly Clinical Epidemiology, Biostatistics and Bioinformatics Seminar	2016-2020	4.0
Bi-weekly Machine Learning Meeting	2017-2020	1.0
Monthly Stroke Research Meeting	2016-2020	2.0
Journal Club - KEBB	2018	1.0
Journal Club – Machine Learning (BMEP)	2019	1.0
<b>Presentations</b>		
ESOC – European Stroke Organization Conference (2x)	2018	0.5
MEMTAB - Methods for Evaluation of medical prediction Models, Tests And Biomarkers	2018	0.5
ISAH – International Conference on Subarachnoid Haemorrhage	2019	1.0
World AI Summit	2020	0.5
<b>Conferences</b>		
Fall Meeting of the Dutch Society of Pattern Recognition and Image Processing	2016	0.25
Institute Quantivision Conference 2017	2017	0.25
MISP - Medical Imaging Symposium for PhD Students	2017	0.25
ESOC – European Stroke organization Conference	2018	0.75
MEMTAB - Methods for Evaluation of medical prediction Models, Tests And Biomarkers	2018	0.5
MIDL - Medical Imaging with Deep Learning	2018	0.75
MISP - Medical Imaging Symposium for PhD Students	2018	0.25

World AI Summit	2018	0.5
Symposium on Advances in Deep Learning	2018	0.25
ISAH	2019	0.75
MIDL - Medical Imaging with Deep Learning	2019	0.75
Amsterdam Public Health Annual Meeting	2019	0.25
PLN Meeting - Spain	2019	0.75
World AI Summit	2020	0.25

<b>Teaching</b>	<b>Year</b>	<b>ECTS</b>
<b>Lecturing and Tutoring</b>		
MIK – Advanced Medical Imaging Processing	2016	1.5
IQ Winter School – Machine Learning Applied to Quantitative Analysis of Medical Images	2018	0.5
MIK - Advanced Medical Imaging (2x)	2018-2019	1.5
<b>Supervising</b>		
Master Internship - Martin van Kuik	2017	0.5
Master Thesis – Dyantha van der Sluijs	2017	1.0
Master Thesis - Ton Koning	2018	1.0
Master Thesis - Thabiso Epema	2018	1.0
Master Thesis - Bouke Postma	2018	1.0
Master Thesis - Nicole Kappelhof	2018	1.0
Master Thesis - Jur Haartman	2019	1.0
Master Thesis - Andrei Pauliuc	2019	1.0
Master Internship - Bella Nicholson	2019	0.5



Other	Year	ECTS
Reviewer European Radiology, Physica Medical, Neurorehabilitation & Neural Repair, Future Generation Computer Systems, Applied Sciences MDPI, Conference: Recent trends in Image Processing & Pattern Recognition	2016-2020	1.0

## List of Publications

1. Van Os HJA, **Ramos LA**, Hilbert A, Van Leeuwen M, Van Walderveen MAA, Kruyt ND, et al. Predicting outcome of endovascular treatment for acute ischemic stroke: Potential value of machine learning algorithms. *Front. Neurol.* 2018;9:1–8.
2. **Ramos LA**, Van Der Steen WE, Sales Barros R, Majoie CBLM, Van Den Berg R, Verbaan D, et al. Machine learning improves prediction of delayed cerebral ischemia in patients with subarachnoid hemorrhage. *J. Neurointerv. Surg.* 2018;1–7.
3. Lopes RR, van Mourik MS, Schaft E V., **Ramos LA**, Baan J, Vendrik J, et al. Value of machine learning in predicting TAVI outcomes. *Netherlands Hear. J.* 2019;27:443–450.
4. van der Steen WE, Marquering HA, Boers AMM, **Ramos LA**, van den Berg R, Vergouwen MDI, et al. Predicting Delayed Cerebral Ischemia with Quantified Aneurysmal Subarachnoid Blood Volume. *World Neurosurg.* 2019;130:e613–e619.
5. Hilbert A, **Ramos LA**, van Os HJA, Olabarriaga SD, Tolhuisen ML, Wermer MJH, et al. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. *Comput. Biol. Med.* 2019;115:103516.
6. Ottenhoff MC, **Ramos LA**, Potters W, Janssen MLF, Hubers D, Piñafuentes D, et al. Predicting Mortality Of Individual Covid-19 Patients : A Multicenter Dutch Cohort. 2020;1–27.

7. Prasetya H, **Ramos LA**, Epema T, Treurniet KM, Emmer BJ, van den Wijngaard IR, et al. qTICI: Quantitative assessment of brain tissue reperfusion on digital subtraction angiograms of acute ischemic stroke patients. *Int. J. Stroke*. 2020;0.
8. Baalman SWE, Schroevers FE, Oakley AJ, Brouwer TF, van der Stuijt W, Bleijendaal H, et al. A morphology based deep learning model for atrial fibrillation detection using single cycle electrocardiographic samples. *Int. J. Cardiol*. 2020;316:130–136.
9. Van Der Steen WE, Marquering HA, **Ramos LA**, Van Den Berg R, Coert BA, Boers AMM, et al. Prediction of outcome using quantified blood volume in aneurysmal SAH. *Am. J. Neuroradiol*. 2020;41:1015–1021.
10. Bleijendaal H, **Ramos LA**, Lopes RR, Verstraelen TE, Baalman SWE, Oudkerk Pool MD, et al. Computer versus cardiologist: Is a machine learning algorithm able to outperform an expert in diagnosing a phospholamban p.Arg14del mutation on the electrocardiogram? *Hear. Rhythm*. 2021;18:79–87.
11. **Ramos LA**, Kappelhof M, van Os HJA, Chalos V, Van Kranendonk K, Kruyt ND, et al. Predicting Poor Outcome Before Endovascular Treatment in Patients With Acute Ischemic Stroke. *Front. Neurol*. 2020;11:1–12.

## Under Review

12. Ramos LA, Van Os H, Hilbert A, Ernst M, Olabarriaga S, Wermer M, et al. Combination of Radiological and Clinical Baseline Data for Outcome Prediction of Patients with an Acute Ischemic Stroke. *Computers in Biology and Medicine*.
13. Lopes RR, Bleijendaal H, Ramos LA, Verstraelen TE; Amin AS; Wilde AAM, et al. Improving electrocardiogram-based detection of rare genetic heart disease using transfer learning. An application to Phospholamban p.Arg14del mutation carriers. *Computers in Biology and Medicine*.

14. Admiraal MM, Ramos LA, Olabarriaga SD, Marquering HA, Horn J, van Rootselaar AF. Quantitative analysis of EEG reactivity for neurological prognostication after cardiac arrest. *Clinical Neurophysiology*.
15. Kappelhof N, Ramos LA, Kappelhof M, van Os HJA, Chalos V, van Kranendonk KR. Evolutionary algorithms and decision trees for predicting poor outcome after endovascular treatment for acute ischemic stroke. *Computers in Biology and Medicine*.



DANKWOORD

After all this time, it feels weird to finally reach the end of my PhD. I faced many challenges since the start, from moving to a new country and adapting to a new lifestyle, to working in a different environment. I must say that despite the many challenges, I enjoyed every minute of my journey.

I would not have come this far without the help and influence of many people and I would like to mention some of them.

I would like to start by thanking my parents, Marcio and Karina. Without their support from the very beginning, I would never be where I am. I'm aware of the many sacrifices they had to make to give me a better future and for that I'm eternally grateful.

My supervisors, promotors and co-promotors, who assisted me throughout all the work that led to this thesis.

Henk, thank you for the guidance throughout all these years. I'm grateful for having a supervisor who always made me feel comfortable in expressing my opinions, even if you did not agree with them, and made it clear by adding 400 comments to my manuscripts. I'm also especially grateful for the assistance with new projects and collaborations, I learned a lot more than I would have if I had worked alone. I'm glad we will continue working together and I'm happy to inform you that I'm now very confident in riding bikes with pedal brakes.

Silvia, I remember very clearly the first day we met since it was my first day of work. I was very nervous while you showed me around the AMC, but your friendly attitude put me at ease very quickly. I was happy to have a Brazillian face around during my PhD and I'm grateful for everything you taught me about research and The Netherlands, but I'm especially grateful for you always pushing me further.

Koos, I just realized I have no idea why we call you Koos as I see no K's in Aeilko Having Zwinderman. That's maybe a question I forgot to ask among the many questions I had to ask you during my PhD. Something that always impressed me was that you were always eager to go to the whiteboard and explain everything from scratch, even if it was something I was supposed to know. Thank you especially for checking my modeling pipelines, appropriate statistical tests, and everything else. I believe you saved me many rejections and rebuttal iterations with reviewers.

Gustav, I especially enjoyed our monthly updates. Your feedback was always on point and I lost count of how many times I entered your office clueless and left with a perfect solution for my problems. You were often very positive about my work, which I think is something rare in academia, where everyone is so focused on critizing and findings flaws. You highlight the positive aspects of our work and was responsible for keeping me motivated throughout these years.

I must also thank all my colleagues for their support in multiple aspects of my life and work.

Manon Tolhuisen, you pushed me out of my comfort zone right after we met, introduced me to other people, and made me feel at home (quite literally by sharing an apartment with me) when I felt completely out of place. I'm glad we share so many memories together, and I cherish our friendship a lot. Lauren, I can't wait for us to be able to speak broken dutch together.

Little Manon, as I sit now to think about what I would like to write, I see how we evolved from colleagues to good friends very fast. From sharing an office and rarely talking to each other, to working in many projects together, and even bouldering and eating out. We always have a good time together. You taught me many things, work and non-work related, like how to perform high-quality dance moves like "doing the dishes".

Bruce, the first time I met you I thought we did not have anything in common. You seemed like such a serious guy, barely talked and so focused on his own work. I have no idea what the turning point was, but I'm glad I was wrong. We share more in common than I can write here, from childhood experiences to most of our hobbies. I'm happy I had you on this journey and I hope we have many gaming days ahead of us. Plus Ultra!

Eva, jij bent een trouwe vriend die altijd bereid is om te helpen. Jij helpt mij toen ik dakloos was, toen werkloos was en nog veel meer. Ik heb geen familie in Nederland, maar ik heb het gevoel dat je een lid van my familie bent. Bedankt voor de leuk spelletjesavonds en de ik wil weer pizza eten.

Praneeta, thanks for the long talks, for the fun trips and outings, and for the countless pictures. Google photos recognizes you in many of them and has automatically created a special folder for you.

Ricardo, Riaan, and Henk, thanks for the many hours of discussion we had in the office, I think your feedback was relevant in many decisions I had to make.

Marit, Merel, Raquel, Nerea, Haryadi, Bart, Wessel, Marcela, and all the others, thank you for the fun trips and parties together.

Matthan, I enjoyed very much working and teaching with you, and hope we can work together again in the future.

I would also like to thank all my co-authors for their contribution and constant feedback. In special, Wessel, Adam, Ricardo, Marjolein, Manon, Hine and Hidde, were so closely involved in many of the chapters included in this thesis.

Tim, ik moet dit echt in Nederlands schrijven. Jij was de eerste die zei “ik spreek geen Engels meer met je” en dat was verschriklijk. Maar, ik heb zoveel geleerd. Nu kan ik met mensen praten en voel ik me niet altijd als een buitenlander meer. Ik heb ook geleerd dat jij moet altijd opletten met leeuwen in de nederlandse bossen. Jij hulp mij mijn doel om Nederlander te worden te bereiken. Je bent beter in Nederlands, maar ik ben tenminste beter met boulderen.

Renan, your experience was very useful during my PhD, and saved me a lot of time and stress. But most of all, your friendship helped me the most. I always feel energized and motivated after our talks.

Thank you to my Brazillian friends (Tonho, Firmo and Eduardo) that despite the distance, still pay an effort to remain in contact with me.

I would also like to thank my colleagues from Pacmed, where I had the opportunity to do a 4-month internship. I learned a lot from you guys and completely changed my way of working and programming.

And last but not least I would like to thank the person who supported me throughout this whole PhD. Marcia, leaving everything behind us to start a new life in a new country was scary for both of us. Together, we faced our fears and insecurities and achieved more than we thought we would ever do. I’m glad I have you to share this experience and the many that are to come. Te amo!







# ABOUT THE AUTHOR

## Lucas Alexandre Ramos

Lucas Alexandre Ramos was born in Botucatu on 5 of August of 1991. He discovered his passion for programming around the age of 13, which led him to pursue a bachelor's degree in Computer Science at the Sao Paulo State University (UNESP). During one of his courses, he discovered an interest in biometrics and image analysis and decided to study this topic further in a master's of Pattern Recognition and Computer Vision. During his masters, he explored multi-scale approaches for image retrieval and classification. After working extensively on the development of new classification methods, Lucas decided to study the applications of Artificial Intelligence in the clinical field. He started as a PhD candidate at the Amsterdam University Medical Centers at the department of Biomedical Engineering and Physics, where he applied Artificial Intelligence to cardiovascular disease. He works currently as a post-doc at the departments of Psychiatry and Biomedical Engineering and Physics (Amsterdam UMC).

