HHS Public Access

Author manuscript

Proc SPIE Int Soc Opt Eng. Author manuscript; available in PMC 2021 March 10.

Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2021 February ; 11597: . doi:10.1117/12.2580358.

Use of biplane quantitative angiographic imaging with ensemble neural networks to assess reperfusion status during mechanical thrombectomy

Mohammad Mahdi Shiraz Bhurwani[1,2], **Kenneth V. Snyder**[2,3], **Muhammad Waqas**[2,3], **Maxim Mokin**[5], **Ryan A. Rava**[1,2], **Alexander R. Podgorsak**[1,2], **Kelsey N. Sommer**[1,2], **Jason M. Davies**[2,3,4], **Elad I. Levy**[2,3], **Adnan H. Siddiqui**[2,3], **Ciprian N. Ionita**[1,2,3]

[1]Department of Biomedical Engineering, University at Buffalo, Buffalo, NY 14260

[2]Canon Stroke and Vascular Research Center, Buffalo, NY 14203

[3]University Dept. of Neurosurgery, University at Buffalo, Buffalo, NY 14203

[4]University Dept. of Biomedical Informatics, University at Buffalo, Buffalo, NY 14214,

[5]Department of Neurosurgery and Brain Repair, University of South Florida, Tampa, FL 33606

Abstract

Digital subtraction angiography (DSA) is the main imaging modality used to assess reperfusion during mechanical thrombectomy (MT) when treating large vessel occlusion (LVO) ischemic strokes. To improve this visual and subjective assessment, hybrid models combining angiographic parametric imaging (API) with deep learning tools have been proposed. These models use convolutional neural networks (CNN) with single view individual API maps, thus restricting use of complementary information from multiple views and maps resulting in loss of relevant clinical information. This study investigates use of ensemble networks to combine hemodynamic information from multiple bi-plane API maps to assess level of reperfusion. Three-hundred-eighty-three anteroposterior (AP) and lateral view DSAs were retrospectively collected from patients who underwent MTs of anterior circulation LVOs. API peak height (PH) and area under time density curve (AUC) maps were generated. CNNs were developed to classify maps as adequate/inadequate reperfusion as labeled by two neuro-interventionalists. Outputs from individual networks were combined by weighting each output, using a grid search algorithm. Ensembled, AP-AUC, AP-PH, lateral-AUC, and lateral-PH networks achieved accuracies of 83.0% (95% confidence-interval: 81.2%–84.8%), 74.4% (72.0%–76.7%), 74.2% (72.8%–75.7%), 74.9% (72.2%–77.7%), and 76.9% (74.4%–79.5%); area under receiver operating characteristic curves of 0.86 (0.84–0.88), 0.81 (0.79–0.83), 0.83 (0.81–0.84), 0.82 (0.8–0.84), and 0.84 (0.82–0.87); and Matthews correlation coefficients of 0.66 (0.63–0.70), 0.48 (0.43–0.53), 0.49 (0.46–0.52), 0.51 (0.45–0.56), and 0.54 (0.49–0.59) respectively. Ensembled network performance was significantly better than individual networks (McNemar's p-value<0.05). This study proved feasibility of using ensemble networks to combine hemodynamic information from multiple bi-plane API maps to assess level of reperfusion during MTs.

Keywords

Ensemble network; machine learning; angiographic parametric imaging; acute ischemic stroke; large vessel occlusion; mechanical thrombectomy; thrombolysis in cerebral infarction (TICI)

1. INTRODUCTION

Angiographic parametric imaging (API) is a quantitative angiographic tool that uses a digital subtraction angiograph (DSA) to semi-quantitatively analyze blood flow through vessels.[1–6] Time density curves (TDC) are obtained by tracking image intensity at a pixel across the sequence (Figure 1). Parametrization of a TDC results in the generation of several parameters such as mean transit time, time to peak, peak height (PH) and area under the TDC (AUC). Parametrization of TDCs at every pixel in a DSA enables the generation of multiple API maps where each map encodes one specific hemodynamic property (Figure 2) [7]. Moderate correlation between the imaging biomarkers and blood flow conditions makes the interpretation of the individual API maps a challenging task. This challenge may be solved by using a hybrid approach, where hemodynamic information encoded in API maps is combined with deep learning tools such as convolutional neural networks (CNN). Deep learning tools have already been used in the evaluation of neurovascular pathologies, such as stroke assessment with Alberta stroke program early computed tomography score (ASPECT) [8], identification of neurovascular pathologies [9] and using API to make predictions regarding aneurysm treatment success [3, 10]. A challenge to using traditional CNNs with API maps is that they cannot fully utilize the connectivity between the multiple views and maps to aid in decision making. However, intra-procedurally, clinicians use multiple views to make their decision. In addition, different API maps contain complementary information regarding hemodynamic properties. Thus, there is a need for a tool that enables the utilization of complementary information encoded in different API maps from multiple views.

Ensemble learning is a deep learning technique that allows for the utilization of information encoded in multiple sources to provide a decision. Each source is used to train a separate neural network. A weight is calculated for each network based on the performance of the network on the testing set. This weight is indicative of the contribution of the specific network to the final decision and is used to combine predictions from each individual network to give a final comprehensive prediction. An ensemble of networks performs better than single networks as it improves the generalization ability [11]. Such a tool allows for the use of information from multiple sources such as different modalities [12]. In this study, we aim to test the feasibility of using ensemble networks to combine hemodynamic information obtained from two API maps derived from bi-plane angiographic systems (two views), and compare the performance with single view individual API map CNNs. One space where such a tool may be implemented is the intra-procedural evaluation of reperfusion in acute ischemic stroke (AIS) patients.

Large vessel occlusions (LVO) cause AISs and account for up to 46% of all AIS cases [13–15]. LVO AISs are the second leading cause of death after coronary artery disease killing

more than 130,000 Americans each year [16]. Treatment of such AISs involve recanalization of the blockage using treatments like mechanical thrombectomy (MT) or intravenous thrombolysis with recombinant tissue-type plasminogen activators [17]. Currently, intra-procedural assessment of each level of reperfusion during MT is achieved through subjective visual assessments of cerebral DSAs, commonly by providing a score on the modified thrombolysis in cerebral infarction (mTICI) scale [18, 19]. This leads to internal inconsistencies [20] as well as inclusion of bias by operators [21].

In previous studies, we have shown feasibility of using ensemble networks to assess level of reperfusion using either multiple view single API maps [22] or multiple API maps from a single view [23]. In this study, we aim to use ensemble networks to combine hemodynamic information obtained from both views of a bi-plane angiographic system as well as from multiple API maps to provide a robust tool that automatically assesses the level of reperfusion in AIS patients undergoing a MT.

2. MATERIALS AND METHODS

Patient data was collected and analyzed at a single center, and was within the scope of a research protocol approved by the institution's Institutional Review Board. DSAs were retrospectively collected from 223 patients with anterior circulation LVOs undergoing MT to treat an AIS. From each patient, multiple DSAs were collected pre-, post- and intra-procedurally, resulting in a total of 525 DSAs. Presence of motion artifacts in DSAs, and occlusions in the posterior circulation resulted in 142 cases being excluded, thus 383 cases were used for analysis. For each case considered, anteroposterior (AP) and lateral view DSAs were collected.

Two experienced neuro-interventionalists analyzed the AP and lateral view DSAs for each case and assigned a label. Any disagreements in labels were resolved by consensus decision. Operators labeled the data independently of each other, were not involved in the procedure, and were blinded to the API maps and clinical outcomes. Each case was labeled on the mTICI scale which includes the following 6 categories: no perfusion (grade 0), partial perfusion beyond initial occlusion but not in distal arteries (grade 1), partial perfusion less than 50% (grade 2a), partial perfusion more than 50% but less than 100% (grade 2b), complete but delayed perfusion (grade 2c) and complete perfusion (grade 3) [24]. These measures were taken to prevent inclusion of bias in the labels used to train the network.

Assessment of reperfusion during MT is generally conducted based on the level of tissue perfusion during the capillary phase of the DSA [19, 25]. In order to limit the overlapping structures from arterial and venous phases of the DSA, each DSA was temporally cropped to only include the capillary phase and exclude arterial and venous phases. TDCs were created at each pixel and then the peak height and area under each TDC were calculated to obtain PH and AUC maps, respectively. Four datasets were created, AP-AUC, AP-PH, lateral-AUC, and lateral-PH maps. The overall workflow of the study is displayed in Figure 3.

Development of CNNs were done using Keras [26]. The architecture development was an iterative process based on optimizing the accuracy of classifying API maps as having

insufficient reperfusion (mTICI 0,1,2a) or sufficient reperfusion (mTICI 2b,2c,3). The final architectures used are displayed in Figure 3. The network architecture was kept the same for each of the 4 datasets. Following the guidelines proposed by Radiology [27], and in order to prevent overfitting of the model, each dataset was split into a training set (70%, 268 cases), a validation set for hyperparameter tuning during training (10%, 39 cases), and a testing set (20%, 77 cases). In addition, a 20-fold Monte Carlo Cross Validation (MCCV) [28] was used to ensure network robustness and that the results obtained were not due to a specific training testing split.

The networks are ensembled by combining outputs from each contributing network into one final output, Figure 3. The combination can be done using different techniques, such as averaging, which assumes equal contribution to the task at hand from each source, thus the outputs from each network are averaged to provide a final output [29]. However, in most cases, each source does not have an equal contribution to the task at hand, and thus each network output needs to be weighted differently. The weights are decimal values between 0 and 1, and are treated like a percentage, such that the sum of the weights for each network equals one. The weights were calculated using two methods; grid search [29], and differential evolution directed optimization [30]. Grid search is a simple, but exhaustive method to finding the optimal weights where a course grid of values from 0 to 1 with a step size of 0.1 is generated, a cartesian product is used to generate all possible combinations of weights for each network. These weights are forced to sum up to one by calculating the sum of the absolute weight values (L1 norm) and dividing each weight by that value. Each combination of weights is tested, and the best performing weight combination is chosen as the final weights to be used for the classification task. The differential evolution directed optimization process is a stochastic method to find the global minimum of the loss function. The optimization was implemented using SciPy [31] and the algorithm is from Storn and Price [32]. A total of 10000 iterations were used with a tolerance of 1E-7. The loss function used in the optimization process was 1 - Matthews correlation coefficient (MCC). MCC is used in machine learning models to evaluate quality of binary classifications [33]. It has proven to be advantageous, as it takes into account class imbalance and uses every factor in the confusion matrix (true positives, false positives, true negatives and false negatives) [34]. Once the weights were calculated, the output from the final dense layer of each network was multiplied by the weight for the network, and the sum of the weighted outputs from each network was the final level of perfusion prediction.

The ensemble network was evaluated using accuracy, receiver operating characteristic (ROC) curves, sensitivity, specificity, and MCC. Each metric was averaged over the 20-fold MCCV. Sub group analysis was done using each of the four networks (AP-AUC, AP-PH, lateral-AUC, and lateral-PH) independently and each of the three ensembling techniques (averaging, grid search, and differential evolution directed optimization). Two-tailed McNemar's p-test values were calculated in order to evaluate significance of any performance differences ($p<0.05$).

3. RESULTS

Each API map was created and normalized in 250 ms, each network took an average of 9 minutes to train, the ensembled weights were calculated in approximately 700 ms, and a single case can be classified in approximately 0.6 ms. Results including classification accuracy, area under the ROC curve (AUROC), MCC, sensitivity, and specificity from the four individual networks, as well as the ensembled network, are displayed in Table 1. The results indicate that best performance is achieved when using an ensembled network with weights obtained using either the grid search or differential evolution directed optimization methods.

Two-tailed McNemar's test were conducted between each of the subsets in order to evaluate significance of any performance differences, these are displayed in Table 2. Grid weighted and differential evolution optimization weighted networks perform significantly better than the averaged network or any of the individual networks (AP-AUC, AP-PH, lateral-AUC, and lateral-PH maps networks). There is no significant difference in performance between the grid weighted and differential evolution optimization weighted networks.

Results of analysis on one specific case from the testing set are displayed in Figure 4. Using just the independent networks we see that the correct classification was obtained by the AP-AUC, AP-PH, and lateral-AUC maps, but not the lateral-PH map. This shows misclassifications can occur by a single network independently, however, when information from both views and multiple maps are combined using an ensemble network, the tool is able to correctly classify the case into the appropriate group. In each CNN classification table, the green highlight indicates the network classification.

4. DISCUSSION

In this study, we established two main findings. First, we demonstrated that the performance of assessing reperfusion in patients undergoing MT to treat an AIS using ensemble networks, to incorporate hemodynamic information from multiple API maps and imaging, is significantly better than using just a single API map from a single view. Second, we established that each single view map has a different contribution to the final decision, and using either grid searched or differential evolution optimization weighting, significantly outperforms simply averaging the outputs from each network.

In order to utilize information encoded in multiple API maps from different views, we trained each single view API map (AP-AUC, AP-PH, lateral-AUC, and lateral-PH) on a separate network and ensembled their outputs to provide a single final classification using different techniques (averaging, grid searched weights, and differential evolution optimization weights). Since each of the input maps were the same size and the vasculature dictating the level of reperfusion within each map were the same size, we used the same network architecture for each of the four maps. A 20-fold MCCV was used to ensure the robustness of results. In each fold of the MCCV, the data was split into a training set (70%, 268 cases), validation set for hyperparameter tuning during training (10%, 39 cases), and a testing set (20%, 77 cases) to ensure none of the networks were overfit. The results in

assessing the level of reperfusion on the independent testing set for each of the four individual networks, as well as the ensembled networks are given in Table 1. Networks ensembled with grid searched or differential evolution optimization weights outperformed the other networks in every metric. MCC values were greater than 0.6, and this indicates strong positive relationship between the classifications and the ground truth labels [35, 36].

The two-tailed McNemar's test was conducted to evaluate significance of difference in performance between the subgroups. The p-values are displayed in Table 2. It is observed that the ensemble networks with grid searched and differential evolution optimization weights significantly outperform all of the individual map networks as well as the average ensemble network (p-value < 0.05). There is no significant difference in performance between grid searched and differential evolution optimization networks (p-value > 0.05).

Previous studies have been conducted to automatically assess the level of reperfusion using API maps, however, these either only used two API maps from a single view [23] or used single API maps from two views [22]. These studies reported accuracies of 79.6% and 81.0%, AUROCs of 0.85 and 0.86, and MCCs of 0.59 and 0.62, respectively. With an average accuracy of 83.0%, AUROC of 0.86, and MCC of 0.66, the method used in this study performed better than those in the previous studies. This method also incorporates additional information and thus is potentially more robust in assessing level of reperfusion in patients undergoing a MT to treat an AIS.

There were limitations to this study. First, we are limited to only demonstrating a technical feasibility of our method due to the size of our monocentric dataset. Second, we are only classifying cases as having sufficient (mTICI 2b,2c,3) or insufficient (mTICI 0,1,2a) reperfusion, instead of classifying each case into the specific mTICI class. This was done due to the low number of cases per mTICI grade (mTICI 0 – 82, mTICI 1 – 5, mTICI 2a – 82, mTICI 2b – 140, mTICI 2c – 48, mTICI 3 – 26) which leads to a drastic decrease in performance. These two limitations may be addressed by collecting additional cases from multiple centers. Lastly, we are currently only using PH and AUC maps, other maps such as time to peak may be used as they contain additional information that may aid in providing a finer classification instead of just two classes.

Such a tool that uses multiple sources of information to assess level of reperfusion can be a valuable resource for neuro-interventionalists. Instead of having to rely on the potentially subjective visual assessment of angiograms, this tool can intra-operatively assess the level of reperfusion and aid the clinician. In addition, this study proves that quantitative angiographic information can be used with such deep learning techniques to make decisions and aid clinicians. This opens the door to making decisions on other treatment outcome scales and for other endovascular interventions such as aneurysm treatment outcome predictions. This tool can also be used as a training tool for medical students instructing them how to score angiograms on the mTICI scale.

5. CONCLUSION

In this technical feasibility study, we developed an ensemble neural network that uses different quantitative angiographic information from multiple views to automatically assess the level of reperfusion in patients undergoing a MT to treat an AIS. Best performance was achieved when using the grid search technique to calculate weights of each contributing map, as it achieved an accuracy of 83.0%, AUROC of 0.86, MCC of 0.66, sensitivity of 0.90, and specificity of 0.74. This study has demonstrated the ability to combine hemodynamic information incorporated in multiple quantitative angiographic maps and views to assess the level of reperfusion during MT procedures. The improvement in performance was significant, and opens the door to using such tools for other endovascular interventions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work is supported by the James H. Cummings Foundation and NIH grant 1R01EB030092 and R21 NS109575-01.

REFERENCES

[1]. Ionita CN, Garcia VL, Bednarek DR et al., “Effect of injection technique on temporal parametric imaging derived from digital subtraction angiography in patient specific phantoms,” Proc SPIE Int Soc Opt Eng, 9038, 90380L (2014).

[2]. Balasubramoniam A, Bednarek DR, Rudin S et al., “Sensitivity evaluation of DSA-based parametric imaging using Doppler ultrasound in neurovascular phantoms,” Proc SPIE Int Soc Opt Eng, 9788, (2016).

[3]. Bhurwani MMS, Waqas M, Williams KA et al., “Predicting treatment outcome of intracranial aneurysms using angiographic parametric imaging and recurrent neural networks.” 11314, 113142O.

[4]. Rava RA, Mokin M, Snyder KV et al., “Performance of angiographic parametric imaging in locating infarct core in large vessel occlusion acute ischemic stroke patients,” J Med Imaging (Bellingham), 7(1), 016001 (2020). [PubMed: 32064301]

[5]. Ryan AR, Ariana BA, Stephen R et al., “Effect of truncated singular value decomposition on digital subtraction angiography derived angiographic parametric imaging maps.” 11312.

[6]. Shiraz Bhurwani MM, Waqas M, Podgorsak AR et al., “Feasibility study for use of angiographic parametric imaging and deep neural networks for intracranial aneurysm occlusion prediction,” J Neurointerv Surg, 12(7), 714–719 (2020). [PubMed: 31822594]

[7]. Balasubramoniam A, Bednarek DR, Rudin S et al., “Sensitivity evaluation of DSA-based parametric imaging using Doppler ultrasound in neurovascular phantoms.” 9788, 97882H.

[8]. Albers GW, Wald MJ, Mlynash M et al., “Automated Calculation of Alberta Stroke Program Early CT Score: Validation in Patients With Large Hemispheric Infarct,” Stroke, 50(11), 3277–3279 (2019). [PubMed: 31500555]

[9]. Podgorsak AR, Rava RA, Bhurwani MMS et al., “Automatic radiomic feature extraction using deep learning for angiographic parametric imaging of intracranial aneurysms,” Journal of NeuroInterventional Surgery, 12(4), 417–421 (2020). [PubMed: 31444288]

[10]. Bhurwani MMS, Waqas M, Podgorsak AR et al., “Feasibility study for use of angiographic parametric imaging and deep neural networks for intracranial aneurysm occlusion prediction,” Journal of NeuroInterventional Surgery, (2019).

[11]. Perrone MP, and Cooper LN, [When networks disagree: Ensemble methods for hybrid neural networks] BROWN UNIV PROVIDENCE RI INST FOR BRAIN AND NEURAL SYSTEMS, (1992).

[12]. Yin Z, Zhao M, Wang Y et al., "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," Computer methods and programs in biomedicine, 140, 93–110 (2017). [PubMed: 28254094]

[13]. Rennert RC, Wali AR, Steinberg JA et al., "Epidemiology, natural history, and clinical presentation of large vessel ischemic stroke," Neurosurgery, 85(suppl_1), S4–S8 (2019). [PubMed: 31197329]

[14]. Waqas M, Mokin M, Primiani CT et al., "Large Vessel Occlusion in Acute Ischemic Stroke Patients: A Dual-Center Estimate Based on a Broad Definition of Occlusion Site," Journal of Stroke and Cerebrovascular Diseases, 29(2), 104504 (2020). [PubMed: 31761735]

[15]. Waqas M, Rai AT, Vakharia K et al., "Effect of definition and methods on estimates of prevalence of large vessel occlusion in acute ischemic stroke: a systematic review and meta-analysis," Journal of neurointerventional surgery, 12(3), 260–265 (2020). [PubMed: 31444289]

[16]. French BR, Boddepalli RS, and Govindarajan R, "Acute ischemic stroke: current status and future directions," Missouri medicine, 113(6), 480 (2016). [PubMed: 30228538]

[17]. Evans MR, White P, Cowley P et al., "Revolution in acute ischaemic stroke care: a practical guide to mechanical thrombectomy," Practical neurology, 17(4), 252–265 (2017). [PubMed: 28647705]

[18]. Almekhlafi MA, Mishra S, Desai JA et al., "Not all "successful" angiographic reperfusion patients are an equal validation of a modified TICI scoring system," Interventional Neuroradiology, 20(1), 21–27 (2014). [PubMed: 24556296]

[19]. Higashida RT, and Furlan AJ, "Trial design and reporting standards for intra-arterial cerebral thrombolysis for acute ischemic stroke," Stroke, 34(8), e109–e137 (2003). [PubMed: 12869717]

[20]. Kallmes DF, "TICI: if you are not confused, then you are not paying attention," American Journal of Neuroradiology, 33(5), 975–976 (2012). [PubMed: 22241379]

[21]. Sattar A, Asif K, Teleb M et al., "E-011 TICI Quantified: Automated Cerebral Revascularization Grading in Acute Ischemic Stroke," Journal of NeuroInterventional Surgery, 6(Suppl 1), A43–A43 (2014).

[22]. Bhurwani MMS, Snyder KV, Waqas M et al., "Use of quantitative angiographic methods with a data-driven model to evaluate reperfusion status (mTICI) during thrombectomy," Neuroradiology, 1–11.

[23]. Bhurwani M, Snyder KV, Waqas M et al., "Improving Deep Learning Automated Intra-operative Assessment of Reperfusion During Mechanical Thrombectomy Using Complementary Quantitative Angiographic Maps," Neurosurgery, 67(Supplement_1), nyaa447_226 (2020).

[24]. Mokin M, Khalessi AA, Mocco J et al., "Endovascular treatment of acute ischemic stroke: the end or just the beginning?," Neurosurgical focus, 36(1), E5 (2014).

[25]. Zaidat OO, Yoo AJ, Khatri P et al., "Recommendations on angiographic revascularization grading standards for acute ischemic stroke: a consensus statement," Stroke, 44(9), 2650–2663 (2013). [PubMed: 23920012]

[26]. Chollet F, [Keras], (2015).

[27]. Bluemke DA, Moy L, Bredella MA et al., [Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the radiology editorial board] Radiological Society of North America, (2020).

[28]. Xu Q-S, and Liang Y-Z, "Monte Carlo cross validation," Chemometrics and Intelligent Laboratory Systems, 56(1), 1–11 (2001).

[29]. Salama ES, El-Khoribi RA, Shoman ME et al., "A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition," Egyptian Informatics Journal, (2020).

[30]. Coletta LF, Hruschka ER, Acharya A et al., "A differential evolution algorithm to optimise the combination of classifier and cluster ensembles," International Journal of Bio-Inspired Computation, 7(2), 111–124 (2015).

[31]. Virtanen P, Gommers R, Oliphant TE et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python," Nature methods, 17(3), 261–272 (2020). [PubMed: 32015543]

[32]. Storn R, and Price K, “Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces,” Journal of global optimization, 11(4), 341–359 (1997).

[33]. Matthews BW, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” Biochimica et Biophysica Acta (BBA)-Protein Structure, 405(2), 442–451 (1975).

[34]. Chicco D, and Jurman G, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” BMC genomics, 21(1), 6 (2020). [PubMed: 31898477]

[35]. Cramer H, [Mathematical Methods of Statistics.] Princeton: Princeton University Press, 282 (1946).

[36]. Powers DM, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” (2011).

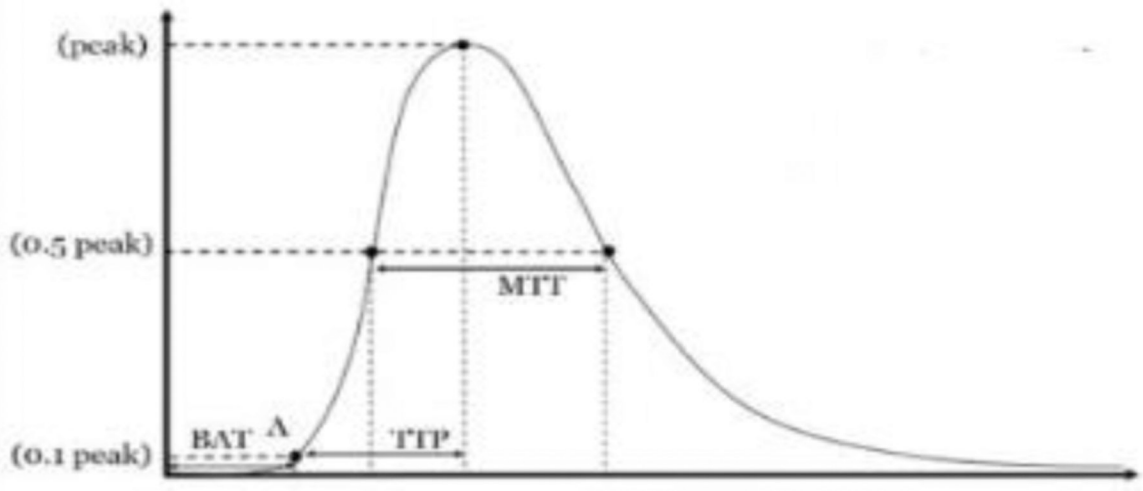

Figure 1:
Parametric Imaging time-density curve

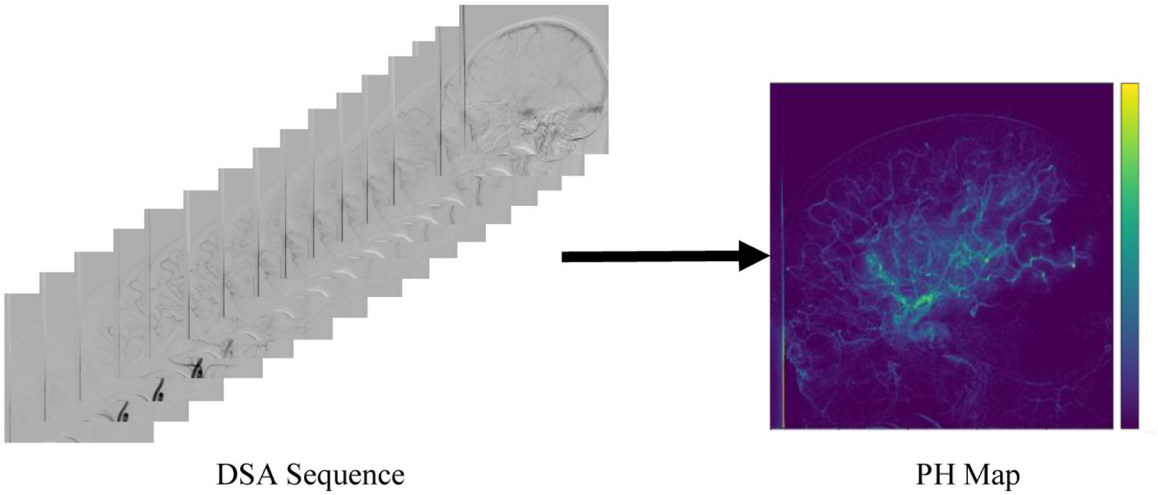

Figure 2:
Example of a peak height (PH) angiographic parametric imaging map created from a digital subtraction angiograph (DSA)

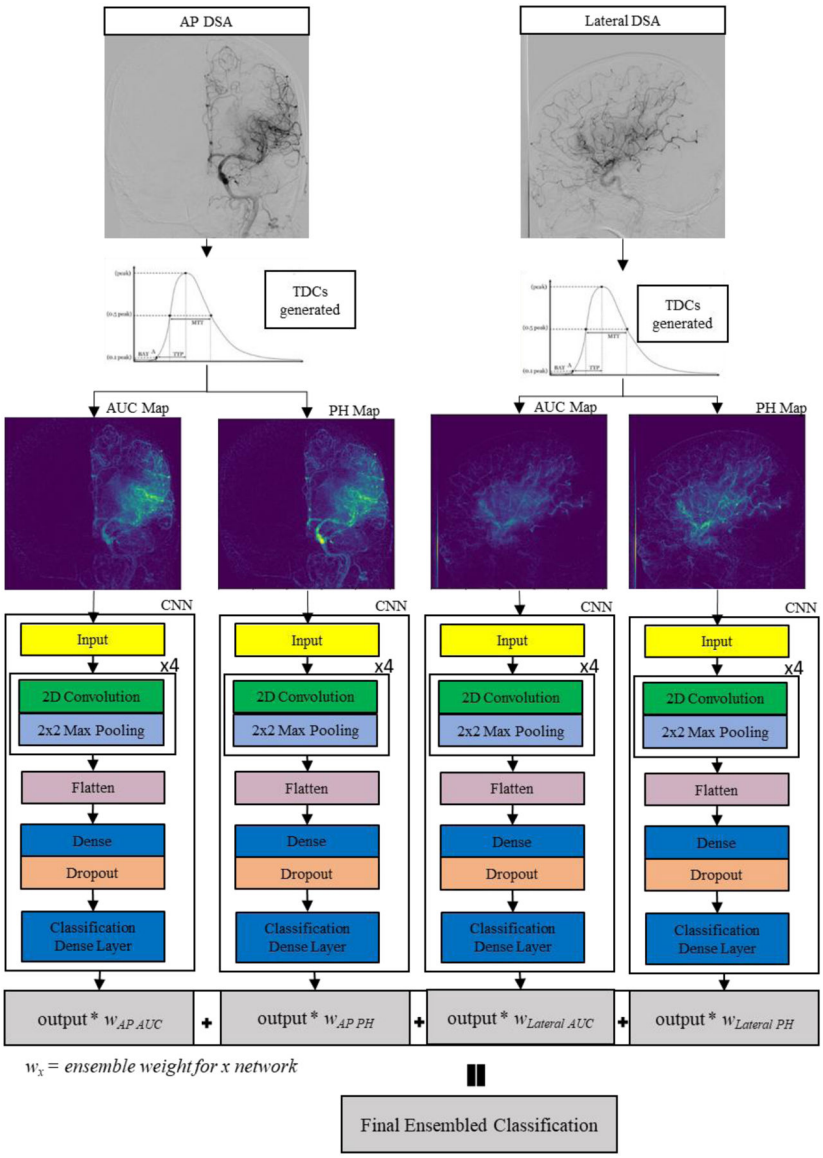

Figure 3:
Overall workflow of the study.

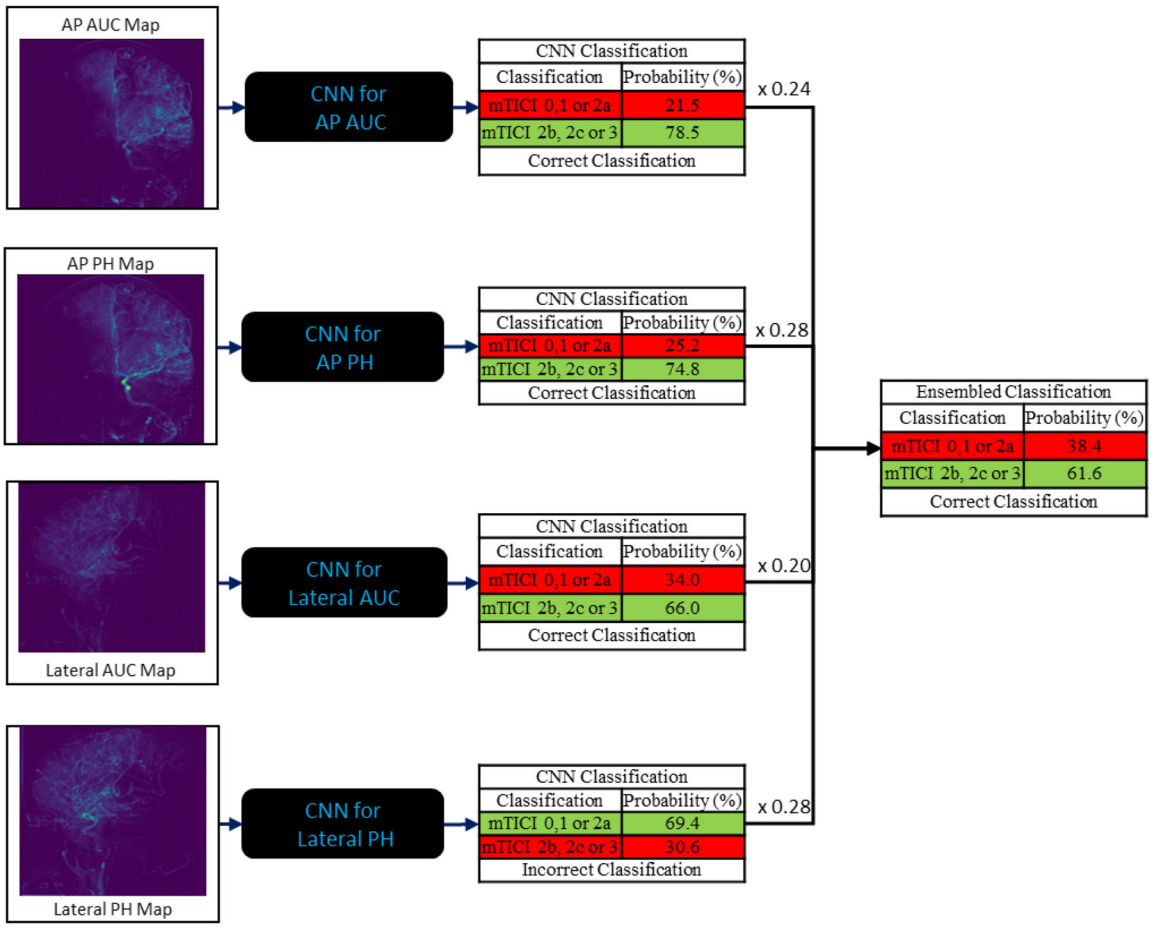

Figure 4:

An example case showing the anteroposterior (AP) and lateral view peak height (PH) and area under the time density curve (AUC) maps, the predictions from their respective networks, the weights given to each network, and the final ensembled classification after using the weights. The weights were obtained using the differential evolution optimization method.

Table 1:

Performance of each individual neural network (AP-AUC, AP-PH, lateral-AUC, and lateral-PH) and the ensembled networks (averaging, grid search, and differential evolution directed optimization) in classifying the respective case based on the level of reperfusion as defined by the mTICI scale. Performance is displayed in the form of average accuracies, AUROCs, MCCs, sensitivities, and specificities along with their standard deviations and 95% confidence intervals. The best results are in bold.

Metric	Reperfusion level assessment: mTICI 0,1,2a versus mTICI 2b,2c,3						
	AP AUC	**AP PH**	**Lateral AUC**	**Lateral PH**	**Averaged**	**Grid Weights**	**Differential Evolution Directed Optimization**
Accuracy	74.4 ± 5.4 (72.0, 76.7)	74.2 ± 3.3 (72.8, 75.7)	74.9 ± 6.3 (72.2, 77.7)	76.9 ± 5.9 (74.4, 79.5)	78.3 ± 5.1 (76.1, 80.5)	**83.0 ± 4.2 (81.2, 84.8)**	82.7 ± 4.5 (80.7, 84.7)
AUROC	0.81 ± 0.05 (0.79, 0.83)	0.83 ± 0.04 (0.81, 0.84)	0.82 ± 0.05 (0.8, 0.84)	0.84 ± 0.05 (0.82, 0.87)	0.86 ± 0.04 (0.84, 0.88)	**0.86 ± 0.05 (0.84, 0.88)**	**0.86 ± 0.05 (0.84, 0.88)**
MCC	0.48 ± 0.11 (0.43, 0.53)	0.49 ± 0.07 (0.46, 0.52)	0.51 ± 0.12 (0.45, 0.56)	0.54 ± 0.11 (0.49, 0.59)	0.56 ± 0.10 (0.5, 0.61)	**0.66 ± 0.08 (0.63, 0.70)**	**0.66 ± 0.09 (0.62, 0.70)**
Sensitivity	0.78 ± 0.08 (0.75, 0.81)	0.78 ± 0.1 (0.74, 0.83)	0.78 ± 0.13 (0.72, 0.84)	0.84 ± 0.1 (0.79, 0.88)	0.83 ± 0.10 (0.80, 0.87)	**0.90 ± 0.09 (0.87, 0.93)**	0.89 ± 0.08 (0.86, 0.93)
Specificity	0.70 ± 0.10 (0.65, 0.74)	0.69 ± 0.12 (0.64, 0.75)	0.71 ± 0.13 (0.65, 0.77)	0.68 ± 0.14 (0.62, 0.75)	0.72 ± 0.10 (0.67, 0.76)	**0.74 ± 0.09 (0.70, 0.78)**	**0.74 ± 0.09 (0.70, 0.78)**

Table 2:

Two-tailed McNemar's test p-values between each subgroup.

	AP AUC	AP PH	Lateral AUC	Lateral PH	Averaged	Grid Searched Weights	Differential Evolution Directed Optimization Weights
AP AUC	-	0.95	**0.03**	**0.03**	**4.66E-06**	**6.70E-18**	**4.15E-17**
AP PH	-	-	0.59	**0.02**	**9.42E-06**	**6.13E-20**	**4.28E-18**
Lateral AUC	-	-	-	**0.03**	**0.17E-05**	**4.44E-20**	**8.57E-19**
Lateral PH	-	-	-	-	0.13	**2.24E-14**	**2.34E-13**
Averaged	-	-	-	-	-	**3.19E-12**	**1.36E-11**
Grid Searched Weights	-	-	-	-	-	-	0.49

P-values less than 0.05 are indicated in bold and signify presence of significant difference in performance.