Validating the Benefit of Combining Imaging and Clinical Data for Ischemic Stroke Outcome Prediction

Zeyad Abouyoussef[1,2] ZEYAD.ABOUYOUSSEF@UCALGARY.CA
Johanna Ospel[2,3] JOHANNA.OSPEL@UCALGARY.CA
Roberto Souza[2,4] ROBERTO.SOUZA2@UCALGARY.CA

[1] *Biomedical Engineering, University of Calgary, Calgary, Alberta, Canada*

[2] *Hotchkiss Brain Institute, University of Calgary, Calgary, Alberta, Canada*

[3] *Department of Diagnostic Imaging and Clinical Neurosciences, University of Calgary, Calgary, Alberta, Canada*

[4] *Electrical and Software Engineering, University of Calgary, Calgary, Alberta, Canada*

Editors: Under Review for MIDL 2026

Abstract

Endovascular treatment (EVT) has been proven to be a successful treatment for some cases of acute ischemic stroke. However, neuro-radiologists rely on a small set of clinical features to select patients for treatment. This leads to the exclusion of patients who would have benefited from treatment and the inclusion of others who would not have benefited from it. Deep learning has been used to predict stroke outcome from baseline imaging and clinical data, with most studies reporting that combining clinical and imaging data slightly outperforms traditional methods (e.g., logistic regression) trained on clinical data only. However, it is not clear how much of this improvement is attributed to the imaging data and whether it is robust to larger and more diverse test sets. We use one of the largest multi-center acute ischemic stroke datasets to determine whether combining imaging and clinical data outperforms traditional methods. We show that combining imaging and clinical data matches the performance of logistic regression (0.72 Area Under Receiver Operating Characteristic Curve (AUROC)) when evaluated on a multi-center test set of over 600 samples. We examine the models' predictions and weights and find that 1) both methods match each other's prediction for 78% of the samples, and 2) the weights associated with the imaging features are small compared to the clinical ones. This suggests that imaging features extracted from the deep learning model do not contribute to the prediction as much as the clinical ones.

Keywords: ischemic stroke, deep learning, outcome prediction, multimodal learning.

1. Introduction

Acute stroke is a leading cause of disability and death, with more than three million deaths worldwide (Martin et al., 2024). Acute ischemic stroke develops when a blood clot blocks one of the major arteries, causing brain tissue death (Martin et al., 2024). For acute ischemic stroke patients, time is of the essence. For every minute the patient is left untreated, approximately 2 million neurons die, leading to irreversible brain damage and a poor long-term prognosis (Saver, 2006). This risk requires fast decision-making to begin treatment as soon as possible and prevent long-term disabilities.

Endovascular treatment (EVT) is considered the most effective treatment for acute ischemic stroke, with multiple studies reporting improvement in patient outcomes (Berkhemer

et al., 2015; Goyal et al., 2015; Jovin et al., 2015). Clinicians rely on features like the Alberta Stroke Program Early Computed Tomography Score (ASPECTS), clot location, and stroke onset time to select patients for EVT (Powers et al., 2019). However, it is challenging to select patients who will benefit from the procedure using such a small subset of clinical features. Furthermore, radiological features like ASPECTS do not fully capture the entire imaging information, and fail to fully describe the clinical outcome described by the 90-day modified Rankin Scale (mRS) (Barber et al., 2000; Gupta et al., 2012; Farzin et al., 2016).

Multiple studies have used traditional machine learning models (e.g., logistic regression) to estimate the mRS from clinical variables (Nishi et al., 2019; Venema et al., 2021; Rajashekar et al., 2021). However, only a small number of imaging-derived features (e.g., ASPECTS and clot location) were used to train the models. While these features are useful, they lack the full imaging information present in the scans. On the other hand, multiple studies have used deep learning to automatically learn useful features from the scans and estimate mRS (Hilbert et al., 2019; Nishi et al., 2020; Samak et al., 2020; Bacchi et al., 2020; Samak et al., 2022; Jo et al., 2023; Samak et al., 2023; Amador et al., 2025; Liu et al., 2025; Diprose et al., 2025). When trained solely on medical images, deep learning models underperformed when compared to traditional models. Only when jointly trained on clinical and imaging data do the models start to marginally outperform logistic regression, albeit on small test sets. So far, it is unclear why deep learning models only start performing once clinical data is added, and whether the improvement over traditional methods is robust when tested on a relatively large multi-center test set.

Ramos et al. conducted several experiments on a cohort of $\sim$ 3000 patients, and their cross-validation results showed that combining imaging and clinical data underperformed logistic regression by 4% (Ramos et al., 2022). While they only trained a small ResNet model and only had access to CT Angiography (CTA) scans, their findings imply that adding clinical data does not help the model learn useful imaging features. Instead, we believe it encourages the model to partially ignore imaging features and focus more on clinical features.

In this study, we explore the effect of combining images with clinical data for stroke outcome prediction. We start by training models that either use clinical or imaging data. Then, we explore the effect of combining them and show that models trained with both perform similarly to traditional models. We examine model weights and show that clinical data have a bigger impact on the model's predictions than imaging data. Finally, we briefly investigate whether imaging-derived features are important to traditional models.

2. Methods

2.1. Dataset

We used the ESCAPE-NA1 trial (Hill et al., 2020) dataset to train and evaluate our models. It contains pre-treatment NCCT, CTA, and clinical data for acute ischemic stroke patients undergoing EVT, as well as the 90-day mRS. Data from 1105 patients was collected from 48 acute care centers spanning eight countries. The trial investigators did not specify a specific acquisition protocol for each scan, and it was left to each center to use its standard protocol. This provides us with a diverse imaging dataset that contains a variety of scanners and acquisition parameters. The dataset contained 1051 samples with full baseline clinical

and imaging data. We dichotomized the 90-day mRS such that a good outcome was defined as mRS ≤ 2, and a bad outcome was defined as mRS ≥ 3. Table 1 shows a summary of the variables in the dataset.

We split the data into training and testing splits. Unlike other studies (Samak et al., 2020, 2022, 2023) that pooled all samples from all care centers and randomly split, we grouped samples by care center and assigned all samples from each center to either training or testing. While assigning centers to the training set, we ensured that the final training set was balanced. This eliminates biases introduced from training on imbalanced data and testing on care centers already used during training. In total, our training set contained 400 samples from 18 care centers, while the test set contained 651 samples from 30 care centers. Finally, we performed 5-fold cross-validation on the training split for hyperparameter tuning. While this splitting scheme uses only about 40% of the available data for training, we generate a large test split that is unbiased to any specific center or acquisition protocol used in training. Furthermore, the amount of data used for training is similar to other studies (Samak et al., 2020, 2022, 2023).

2.2. Preprocessing

2.2.1. Imaging Data

We used TotalSegmentator (Wasserthal et al., 2023; Isensee et al., 2020) to automatically skull-strip the NCCT and CTA scans. Then, we performed foreground cropping to remove background regions. The NCCT scans were adjusted to a window level of 40 Hounsfield Units (HU) and a window width of 80 HU to highlight infarcted tissue, while the CTA scans were windowed on a 250 HU window level and a 400 HU window width to highlight blood vessels. To account for variation in the training set, all NCCT scans were resampled to the training set's median voxel spacing of $0.45 \times 0.45 \times 2.5\ mm^3$ and resized to a median image shape of $291 \times 356 \times 54$. As for the CTA scans, the median voxel spacing was $0.47 \times 0.47 \times 0.63\ mm^3$, and the median image shape after resampling was $278 \times 337 \times 221\ mm^3$. Voxel intensity was normalized based on the global mean ($\mu_{NCCT} = 29.81, \mu_{CTA} = 43.53$) and standard deviation ($\sigma_{NCCT} = 12.06, \sigma_{CTA} = 17.64$) of the foreground voxels in the training set. Finally, we performed random flipping on the fly during training time to reduce overfitting.

2.2.2. Model Training

We start by training logistic regression, random forest, and neural network models to estimate the 90-day mRS from 20 baseline clinical features. Features included were: age, sex, baseline mRS, baseline National Institutes of Health Stroke Scale (NIHSS), baseline ASPECTS, time between stroke onset and treatment, location of vessel occlusion, affected brain hemisphere, and medical history (i.e., smoking, diabetes, hypertension, atrial fibrillation, peripheral vascular disease, chronic renal failure, ischemic heart disease, congestive heart failure, high cholestrol, past stroke, recent stroke, major surgery). The location of vessel occlusion variable was one-hot encoded as it had 5 different values (i.e., distal M1 MCA, proximal M1 MCA, mid M1 MCA, M2/M3 MCA, ICA). The logistic regression model was trained for up to 5000 iterations with $l2$ regularization. The random forest model had 1000 estimators, with a maximum tree depth of 10, a minimum of 2 samples at leaf nodes, a minimum of 5 samples to split an internal node, and a maximum of 4 features considered

Table 1: Summary of the dataset variables with respect to the 90-day mRS. Continuous variables are represented by the median and interquartile range, while categorical variables are represented by percentages. ASPECTS, Alberta Stroke Program Early Computed Tomography Score; NIHSS, National Institutes of Health Stroke Scale; MCA, Middle Cerebral Artery.

Variable	All samples ($n = 1051$)	mRS ≤ 2 ($n = 640$)	mRS ≥ 3 ($n = 411$)
Age	70.8 (60.7-79.7)	66.4 (57.98-75.83)	76.65 (67.53-83.57)
Sex, Male	50.38%	53.08%	46.91%
Time between stroke onset and treatment	201 (135-326.75)	185 (127-286)	230 (148.5-369)
ASPECTS	8 (7-9)	8 (7-9)	8 (7-9)
NIHSS	17 (12-21)	16 (11-20)	18 (14-22)
Hemishphere Stroke, Left	47.26%	45.20%	50.47%
Occlusion Site, MCA	80.23%	84.11%	74.15%
Hypertension	69.87%	65.43%	76.81%
Smoker	49.15%	52.31%	47.69%
Diabetes	19.68%	14.66%	27.53%
Atrial Fibrillation	35.03%	29.63%	43.48%
Peripheral Vascular Disease	5.37%	4.32%	7.00%
High Cholestrol	46.67%	43.05%	52.42%
Chronic Renal Failure	5.46%	4.17%	7.49%
Chronic Heart Failure	12.24%	11.57%	13.29%
Recent Major Surgery	3.58%	3.70%	3.38%
Past Stroke	13.84%	12.96%	15.22%
Recent Stroke	4.50%	4.70%	4.20%
Ischemic Heart Disease	22.88%	20.37%	26.81%

for the best split. The neural network consisted of a linear layer with 512 output features, a batch normalization layer, a linear layer with 512 features, a ReLU, and a final classification layer.

We chose a 3D ResNeXt (Xie et al., 2017) with Squeeze-and-Excitation (SE) (Hu et al., 2018) modules as the backbone for extracting features from the scans. ResNeXt extends ResNet by using grouped convolutions to make the network more efficient. This reduces the high computational cost associated with training on 3D data, leading to faster convergence. Meanwhile, SE acts as a channel attention mechanism that helps the network capture the most important channels in a feature map. In total, the backbone had 29,401,840 parameters with the final layer outputting 512 features. The 512 features are batch normalized, then fed to a classification head consisting of a linear layer, a ReLU, and a final linear layer that predicts the mRS.

When combining NCCT with CTA, we use separate backbones for each scan type and concatenate the features after batch normalization. The concatenated features are then passed to the classification head for mRS prediction. Combining clinical data with one scan type follows a similar procedure. We feed the clinical data into a linear layer with 512 output features and concatenate them with the imaging ones. When combining all data, we feed each input to its own backbone, concatenate, and then proceed to the classification head. Figure 1 shows how we combine features in different training setups.

To compare our combined NCCT + Clinical model with the current state of the art, we retrain their TranSOP model on our data using the same hyperparameters they used (Samak et al., 2023). Their model had a similar setup as ours, but instead of using a ResNeXt backbone, they used a Swin transformer backbone (Liu et al., 2021).

In all experiments, we pick the fold with the highest F1 score and use it for evaluation on the test set. Models were developed using MONAI (Cardoso et al., 2022) and scikit-learn (Pedregosa et al., 2011) and trained on 80 GB A100 and H100 NVIDIA GPUs. Code is publicly available at https://github.com/zeyad-kay/multimodal_mrs_prediction

3. Results

In total, we trained 10 models to investigate how combining clinical and imaging data affects model performance. Table 2 summarizes the performance of each model along with uncertainty estimates for each metric. Logistic regression was the best clinical-only model with 0.72 Area Under Receiver Operating Characteristic Curve (AUROC). Figure 2 shows the SHapley Additive exPlanations (SHAP) values (Lundberg and Lee, 2017) for the logistic regression model. Features like age, ASPECTS, NIHSS, and time between stroke onset and treatment contributed the most to the model predictions.

Training with only the scans resulted in lower AUROC than the clinical-only models. The NCCT model was the best imaging-only model, followed by the NCCT + CTA combination. Combining imaging and clinical data outperformed the imaging-only models, with the NCCT + Clinical model matching the AUROC of logistic regression and achieving the best Area Under Precision Recall Curve (AUPRC) across all experiments. Meanwhile, the TranSOP model achieved the best F1 score and accuracy. While the CTA-only and NCCT + CTA experiments had high F1 scores, they had 96% and 85% false positive rates, respectively.

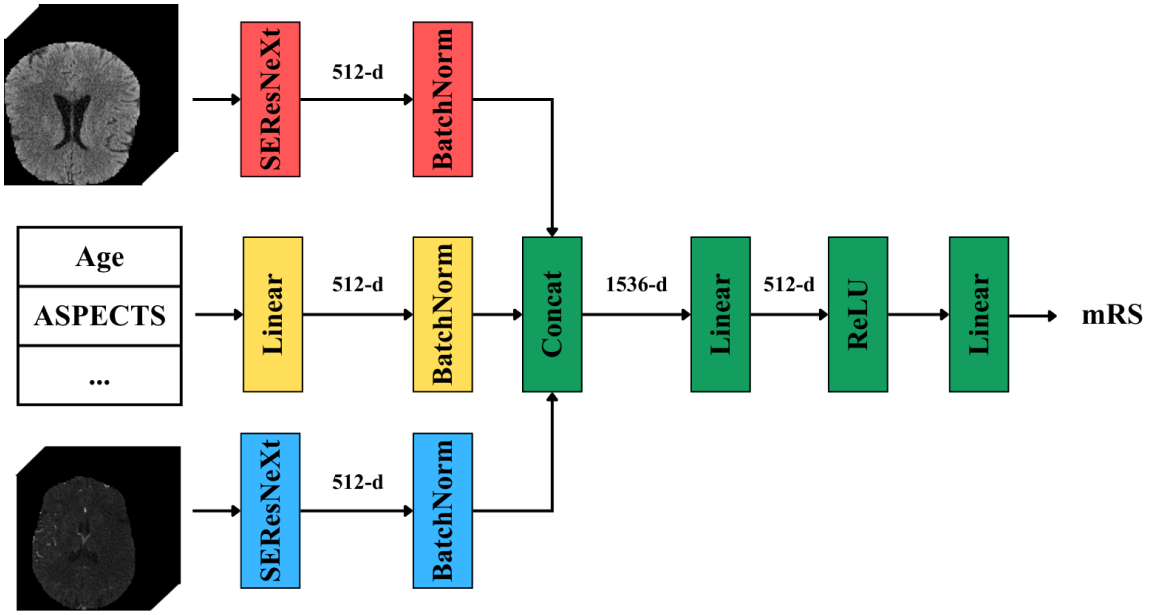

Figure 1: Network architecture used when training the models. Each input is fed into a separate backbone (red, yellow, blue) that extracts a 512-dimensional feature vector. The features are concatenated and fed to a linear and a final classification layer (green). Backbones are enabled/disabled depending on which input data we are using. For example, we only use the red backbone when training on NCCT, the yellow and blue when training on clinical data and CTA, and all backbones when training on NCCT, CTA, and clinical data. SEResNeXt, Squeeze-and-Excitation ResNeXt.

4. Discussion

The reduced performance of the imaging-only models compared to the ones using clinical data is consistent with current literature (Samak et al., 2020, 2022, 2023; Liu et al., 2025). For instance, Samak et al. reported that training with only clinical data outperformed training with NCCT by 6% AUROC (0.73 vs 0.67) (Samak et al., 2023). While Hilbert et al. reported better AUROC when training with 2D maximum intensity projections of CTA scans (0.71 vs 0.68), their results were based on internal cross-validation data (Hilbert et al., 2019). Interestingly, our results show that training with CTA performed worse than with NCCT, even when combined with the clinical data. This might be attributed to the HU window highlighting blood vessels as opposed to the one highlighting infarcts in NCCT. This assumption is based on several factors: 1) SHapley Additive exPlanations values (Lundberg and Lee, 2017) in Figure 2 show that baseline ASPECTS tends to have a larger impact on model predictions than occlusion location, 2) Diprose et al. (Diprose et al., 2025) reported that combining both imaging modalities yielded better performance, but their analysis showed that they are using a similar window for both modalities, one

Table 2: Test set performance across different models. Best model highlighted in bold. A random model has 0.67 AUPRC and 0.5 AUROC. Model labeled "All" is the NCCT + CTA + Clinical. CI, Confidence Interval; AUROC, Area Under Receiver Operating Characteristic Curve; AUPRC Area Under Precision Recall Curve; LR, Logistic Regression; RF, Random Forest; NN, Neural Network;

Models	AUROC (95% CI)	AUPRC (95% CI)	F1 (95% CI)	Accuracy (95% CI)
Clinical				
LR	**0.72 (0.68-0.77)**	0.83 (0.79-0.87)	0.75 (0.71-0.78)	0.68 (0.64-0.72)
RF	0.7 (0.66-0.74)	0.81 (0.77-0.85)	0.75 (0.71-0.78)	0.68 (0.64-0.71)
NN	0.65 (0.6-0.69)	0.77 (0.72-0.81)	0.68 (0.65-0.72)	0.61 (0.58-0.65)
Imaging				
NCCT	0.57 (0.52-0.62)	0.72 (0.66-0.77)	0.51 (0.47-0.56)	0.49 (0.45-0.53)
CTA	0.5 (0.45-0.55)	0.66 (0.61-0.71)	0.78 (0.76-0.81)	0.65 (0.61-0.69)
NCCT+CTA	0.55 (0.5-0.59)	0.71 (0.66-0.76)	0.77 (0.74-0.8)	0.65 (0.61-0.68)
Imaging+Clinical				
NCCT+Clinical	**0.72 (0.68-0.76)**	**0.84 (0.8-0.87)**	0.75 (0.72-0.79)	0.68 (0.65-0.72)
TranSOP	0.71 (0.66-0.75)	0.82 (0.77-0.86)	**0.79 (0.76-0.82)**	**0.7 (0.67-0.74)**
CTA+Clinical	0.62 (0.58-0.67)	0.76 (0.72-0.8)	0.77 (0.75-0.8)	0.66 (0.62-0.69)
All	0.66 (0.61-0.7)	0.8 (0.75-0.83)	0.67 (0.63-0.71)	0.6 (0.56-0.64)

that highlights infarcts. To validate this assumption, we retrained the CTA-only and CTA + Clinical models with the same window used in NCCT. The retrained CTA-only model improved AUROC to 0.56 and AUPRC to 0.72, while the combined model improved AUROC to 0.69 and AUPRC 0.79. This confirms that the wide window highlighting blood vessels is not as helpful to the model as the one highlighting infarcts.

The superior performance of the models combining imaging and clinical data over the imaging-only models is indicative of the importance of clinical data to the models. Again, several studies highlight this pattern (Samak et al., 2020, 2022; Ramos et al., 2022; Samak et al., 2023; Liu et al., 2025). Samak et al. combined baseline NCCT and clinical data to achieve a state-of-the-art performance of 0.85 AUROC (Samak et al., 2023). While their model showed a 20% improvement over training with only NCCT, they only tested on a cohort of 75 patients. To investigate how changing the backbone affects performance, we retrained their model. On our test set, the model's AUROC dropped to 0.71. This indicates that even when changing the imaging backbone to a more advanced architecture, it did not outperform logistic regression.

The near identical performance between logistic regression and NCCT + Clinical is interesting. When looking at the predictions from both models, we find that the predictions are identical for 78% of the test samples. This might imply that the combined model is just focusing on the clinical features. While it is challenging to fully determine how much the network learns from the imaging data, we try to partially answer this question by

Figure 2: SHapley Additive exPlanations (SHAP) values for the logistic regression model. Positive SHAP values are associated with good outcomes, while negative ones are associated with bad outcomes. Features are ordered based on their importance, with the most important ones being at the top. High feature values are depicted in red while low ones are in blue. For example, older patients (red) are associated with bad outcomes (negative SHAP value).

examining the weights of the linear layer following feature concatenation in Figure 3. In the NCCT + Clinical model, each neuron has a 1024-dimensional weight vector w with $w_{1,\ldots,512}$ being associated with imaging features and $w_{513,\ldots,1024}$ with the clinical features. If we clip

small weights with absolute value < 0.04 to zero, we see that almost all imaging weights across all neurons are clipped, while the clinical weights have larger absolute values and were not clipped. To some extent, these observations explain why the multimodal model's performance is nearly identical to logistic regression.

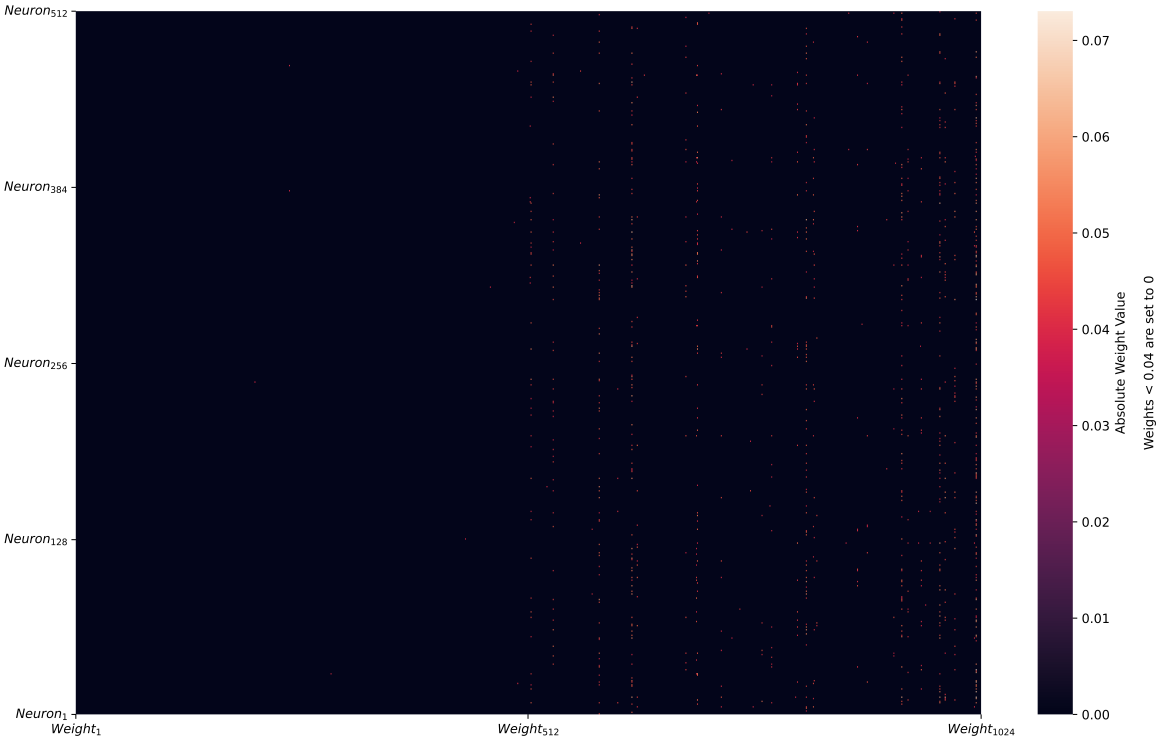

Figure 3: Heatmap of the weights of the linear layer following feature concatenation in the NCCT + Clinical model. The layer has 512 neurons (y-axis) and 1024 weights for each neuron (x-axis). The first 512 weights (left half) are the ones associated with the imaging features extracted from SEResNeXt, and the second 512 weights (right half) are the ones associated with the clinical features. Weights with absolute value < 0.04 are set to zero for better visualization.

It is surprising that the multimodal model ignores imaging features, especially since Figure 2 shows that ASPECTS, an imaging-derived feature, is important for logistic regression. To further investigate this, we retrain the logistic regression model without any imaging-derived information (i.e., ASPECTS and occlusion location) and find that the AUROC and accuracy stay the same, AUPRC increases by 1%, and $F1$ decreases by 1%. This indicates that even when training with logistic regression, imaging-derived features like ASPECTS do not affect performance. This is surprising since ASPECTS is considered a proxy to the total infarct volume, a variable shown by Ospel et al. to have a strong association with good outcomes (Ospel et al., 2024).

5. Conclusion

In this study, we investigated how different inputs affect stroke outcome prediction model performance. We found that training with only imaging data drastically reduces performance, while combining imaging and clinical data results in identical performance to logistic regression. Finally, we examined the multimodal model weights and found that imaging features do not contribute to the prediction as much as the clinical features.

Acknowledgments

The study was funded by an NSERC Discovery Grant awarded to Roberto Souza (#RGPIN-2021-02867) and is supported by the Hotchkiss Brain Institute. The authors also thank the Digital Research Alliance of Canada and the University of Calgary Research Computing Services for providing access to GPU clusters.

References

Kimberly Amador, Noah Pinel, Anthony J. Winder, Jens Fiehler, Matthias Wilms, and Nils D. Forkert. A cross-attention-based deep learning approach for predicting functional stroke outcomes using 4d ctp imaging and clinical metadata. *Medical Image Analysis*, 99: 103381, 1 2025. ISSN 1361-8415. doi: 10.1016/J.MEDIA.2024.103381.

Stephen Bacchi, Toby Zerner, Luke Oakden-Rayner, Timothy Kleinig, Sandy Patel, and Jim Jannes. Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes: A pilot study. *Academic Radiology*, 27:e19–e23, 2 2020. ISSN 1076-6332. doi: 10.1016/J.ACRA.2019.03.015. URL https://www.sciencedirect.com/science/article/pii/S1076633219301746?via%3Dihub.

Philip A. Barber, Andrew M. Demchuk, Jinjin Zhang, and Alastair M. Buchan. Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy. *Lancet*, 355:1670–1674, 5 2000. ISSN 01406736. doi: 10.1016/S0140-6736(00)02237-6. URL https://www.thelancet.com/action/showFullText?pii=S0140673600022376.

Olvert A. Berkhemer, Puck S.S. Fransen, Debbie Beumer, Lucie A. van den Berg, Hester F. Lingsma, Albert J. Yoo, et al. A randomized trial of intraarterial treatment for acute ischemic stroke. *New England Journal of Medicine*, 372:11–20, 1 2015. ISSN 0028-4793. doi: 10.1056/NEJMOA1411587/SUPPL_FILE/NEJMOA1411587_DISCLOSURES.PDF. URL https://www.nejm.org/doi/pdf/10.1056/NEJMoa1411587.

M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, et al. Monai: An open-source framework for deep learning in healthcare, 2022. URL https://arxiv.org/abs/2211.02701.

James P Diprose, William K Diprose, Tuan-Yow Chien, Michael T M Wang, Andrew McFetridge, Gregory P Tarr, et al. Deep learning on pre-procedural computed tomography and clinical data predicts outcome following stroke thrombectomy. *Journal of NeuroInterventional Surgery*, 17(3):266–271, 2025. ISSN 1759-8478. doi: 10.1136/jnis-2023-021154. URL https://jnis.bmj.com/content/17/3/266.

Behzad Farzin, Robert Fahed, Francois Guilbert, Alexandre Y. Poppe, Nicole Daneault, André P. Durocher, et al. Early ct changes in patients admitted for thrombectomy. *Neurology*, 87:249–256, 7 2016. ISSN 1526632X. doi: 10.1212/WNL.0000000000002860.

Mayank Goyal, Andrew M. Demchuk, Bijoy K. Menon, Muneer Eesa, Jeremy L. Rempel, John Thornton, et al. Randomized assessment of rapid endovascular treatment of ischemic

stroke. *New England Journal of Medicine*, 372:1019–1030, 3 2015. ISSN 0028-4793. doi: 10.1056/NEJMOA1414905/SUPPL_FILE/NEJMOA1414905_DISCLOSURES.PDF. URL https://www.nejm.org/doi/pdf/10.1056/NEJMoa1414905.

A. C. Gupta, P. W. Schaefer, Z. A. Chaudhry, T. M. Leslie-Mazwi, R. V. Chandra, R. G. González, et al. Interobserver reliability of baseline noncontrast ct alberta stroke program early ct score for intra-arterial stroke treatment selection. *American Journal of Neuroradiology*, 33:1046–1049, 6 2012. ISSN 0195-6108. doi: 10.3174/AJNR.A2942. URL https://www.ajnr.org/content/33/6/1046https://www.ajnr.org/content/33/6/1046.abstract.

A. Hilbert, L. A. Ramos, H. J.A. van Os, S. D. Olabarriaga, M. L. Tolhuisen, M. J.H. Wermer, et al. 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*, 115, 12 2019. ISSN 18790534. doi: 10.1016/j.compbiomed.2019.103516.

Michael D Hill, Mayank Goyal, Bijoy K Menon, Raul G Nogueira, Ryan A McTaggart, Andrew M Demchuk, et al. Efficacy and safety of nerinetide for the treatment of acute ischaemic stroke (escape-na1): a multicentre, double-blind, randomised controlled trial. *The Lancet*, 395:878–887, 3 2020. ISSN 01406736. doi: 10.1016/S0140-6736(20)30258-0. URL https://linkinghub.elsevier.com/retrieve/pii/S0140673620302580.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

Fabian Isensee, Paul F. Jaeger, Simon A.A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods 2020 18:2*, 18:203–211, 12 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-01008-z. URL https://www.nature.com/articles/s41592-020-01008-z.

Hongju Jo, Changi Kim, Dowan Gwon, Jaeho Lee, Joonwon Lee, Kang Min Park, and Seongho Park. Combining clinical and imaging data for predicting functional outcomes after acute ischemic stroke: an automated machine learning approach. *Scientific Reports 2023 13:1*, 13:1–8, 10 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-44201-8. URL https://www.nature.com/articles/s41598-023-44201-8.

Tudor G. Jovin, Angel Chamorro, Erik Cobo, María A. de Miquel, Carlos A. Molina, Alex Rovira, et al. Thrombectomy within 8 hours after symptom onset in ischemic stroke. *New England Journal of Medicine*, 372(24):2296–2306, 2015. doi: 10.1056/NEJMoa1503780. URL https://www.nejm.org/doi/full/10.1056/NEJMoa1503780.

Mingtian Liu, Nima Hatami, Laura Mechtouff, Tae-Hee Cho, Carole Lartizien, and Carole Frindel. Lesion-centered vision transformer for stroke outcome prediction from image and clinical data . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15974. Springer Nature Switzerland, September 2025.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Seth S. Martin, Aaron W. Aday, Zaid I. Almarzooq, Cheryl A.M. Anderson, Pankaj Arora, Christy L. Avery, et al. 2024 heart disease and stroke statistics: A report of us and global data from the american heart association. *Circulation*, 149:E347–E913, 2 2024. ISSN 15244539. doi: 10.1161/CIR.0000000000001209/SUPPL_FILE/SUB-SAHARAN. URL https://www.ahajournals.org/doi/10.1161/CIR.0000000000001209.

Hidehisa Nishi, Naoya Oishi, Akira Ishii, Isao Ono, Takenori Ogura, Tadashi Sunohara, et al. Predicting clinical outcomes of large vessel occlusion before mechanical thrombectomy using machine learning. *Stroke*, 50:2379–2388, 9 2019. ISSN 1524-4628. doi: 10.1161/STROKEAHA.119.025411. URL https://pubmed.ncbi.nlm.nih.gov/31409267/.

Hidehisa Nishi, Naoya Oishi, Akira Ishii, Isao Ono, Takenori Ogura, Tadashi Sunohara, et al. Deep learning-derived high-level neuroimaging features predict clinical outcomes for large vessel occlusion. *Stroke*, 51:1484–1492, 5 2020. ISSN 15244628. doi: 10.1161/STROKEAHA.119.028101/SUPPL_FILE/STR_STROKE-2019-028101_SUPP1.PDF. URL https://www.ahajournals.org/doi/10.1161/STROKEAHA.119.028101.

Johanna M. Ospel, Leon Rinkel, Aravind Ganesh, Andrew Demchuk, Manraj Heran, Eric Sauvageau, et al. How do quantitative tissue imaging outcomes in acute ischemic stroke relate to clinical outcomes? *Journal of Stroke*, 26:252–259, 5 2024. ISSN 2287-6391. doi: 10.5853/JOS.2023.02180. URL http://j-stroke.org/journal/view.php?doi=10.5853/jos.2023.02180.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

William J. Powers, Alejandro A. Rabinstein, Teri Ackerson, Opeolu M. Adeoye, Nicholas C. Bambakidis, Kyra Becker, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke a guideline for healthcare professionals from the american heart association/american stroke association. *Stroke*, 50:E344–E418, 12 2019. ISSN 15244628. doi: 10.1161/STR.0000000000000211/SUPPL_FILE/DATA. URL /doi/pdf/10.1161/STR.0000000000000211?download=true.

Deepthi Rajashekar, Michael D. Hill, Andrew M. Demchuk, Mayank Goyal, Jens Fiehler, and Nils D. Forkert. Prediction of clinical outcomes in acute ischaemic stroke patients: A comparative study. *Frontiers in Neurology*, 12:663899, 5 2021. ISSN 16642295. doi: 10.3389/FNEUR.2021.663899/BIBTEX. URL www.frontiersin.org.

Lucas A. Ramos, Hendrikus van Os, Adam Hilbert, Silvia D. Olabarriaga, Aad van der Lugt, Yvo B.W.E.M. Roos, et al. Combination of radiological and clinical baseline data for outcome prediction of patients with an acute ischemic stroke. *Frontiers in neurology*, 13, 4 2022. ISSN 1664-2295. doi: 10.3389/FNEUR.2022.809343. URL https://pubmed.ncbi.nlm.nih.gov/35432171/.

Zeynel A. Samak, Philip Clatworthy, and Majid Mirmehdi. Prediction of thrombectomy functional outcomes using multimodal data. *Communications in Computer and Information Science*, 1248 CCIS:267–279, 2020. ISSN 18650937. doi: 10.1007/978-3-030-52791-4_21/TABLES/3. URL https://link.springer.com/chapter/10.1007/978-3-030-52791-4_21.

Zeynel A. Samak, Philip Clatworthy, and Majid Mirmehdi. Fema: Feature matching auto-encoder for predicting ischaemic stroke evolution and treatment outcome. *Computerized Medical Imaging and Graphics*, 99:102089, 7 2022. ISSN 0895-6111. doi: 10.1016/J.COMPMEDIMAG.2022.102089.

Zeynel A. Samak, Philip Clatworthy, and Majid Mirmehdi. Transop: Transformer-based multimodal classification for stroke treatment outcome prediction. *Proceedings - International Symposium on Biomedical Imaging*, 2023-April, 2023. ISSN 19458452. doi: 10.1109/ISBI53787.2023.10230576.

Jeffrey L. Saver. Time is brain—quantified. *Stroke*, 37:263–266, 1 2006. ISSN 00392499. doi: 10.1161/01.STR.0000196957.55928.AB. URL https://www.ahajournals.org/doi/10.1161/01.STR.0000196957.55928.ab.

Esmee Venema, Bob Roozenbeek, Maxim J.H.L. Mulder, Scott Brown, Charles B.L.M. Majoie, Ewout W. Steyerberg, et al. Prediction of outcome and endovascular treatment benefit: Validation and update of the mr predicts decision tool. *Stroke*, 52:2764–2772, 9 2021. ISSN 15244628. doi: 10.1161/STROKEAHA.120.032935/SUPPL_FILE/STR_STROKE-2020-032935_SUPP1.PDF. URL https://www.ahajournals.org/doi/10.1161/STROKEAHA.120.032935.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, et al. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), September 2023. ISSN 2638-6100. doi: 10.1148/ryai.230024. URL http://dx.doi.org/10.1148/ryai.230024.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017. doi: 10.1109/CVPR.2017.634.