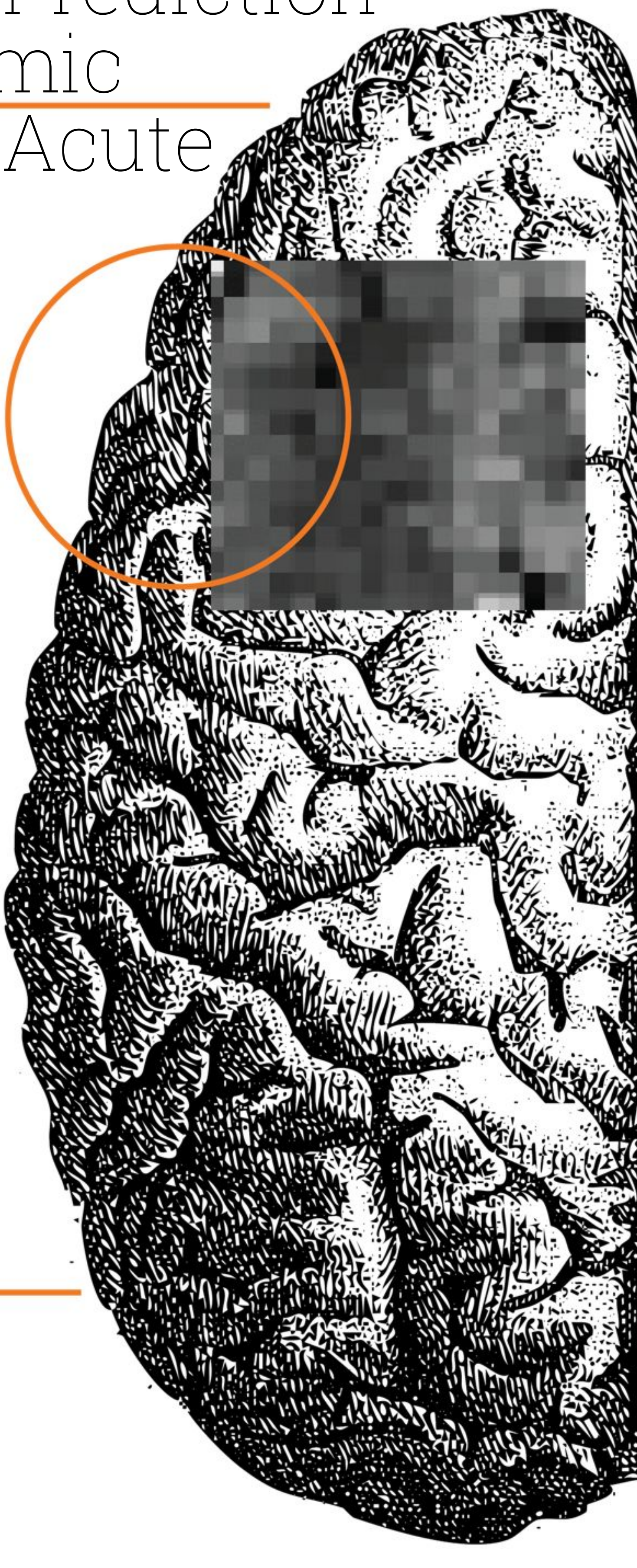
Simplified Annotations for Radiomics-Based Prediction of Subacute Ischemic Lesion Volume in Acute Ischemic Stroke

BM51035: Graduation Project

Emma A.G. Borggreven



Delft University of Technology

TU Delft

Simplified Annotations for Radiomics-Based Prediction of Subacute Ischemic Lesion Volume in Acute Ischemic Stroke

by

Emma A.G. Borggreven

in partial fulfilment of th requirements of Master of Science Biomedical Engineering Track Medical Physics

at the Delft University of Technology,

to be defended publicly on Thursday July 17, 2025 at 12:30.

Student number:	4861965	
Project duration:	November 14, 2024 – July 17, 2025	
Thesis advisors:	W. (Wiktor) Olszewski,	Amsterdam UMC, daily supervisor
	O. (Odysseas) Papakyriakou,	Amsterdam UMC
	Prof. dr. H.A. (Henk) Marquering,	Amsterdam UMC
	Dr. ir. F.J.H. (Frank) Gijsen,	TU Delft
Thesis committee:	Dr. ir. F.J.H. (Frank) Gijsen,	TU Delft, chair
	W. (Wiktor) Olszewski,	Amsterdam UMC
	Dr. ir. M.W.A. (Matthan) Caan,	Amsterdam UMC
	Dr. ir. Q. (Qian) Tao,	TU Delft

Simplified Annotations for Radiomics-Based Prediction of Subacute Ischemic Lesion Volume in Acute Ischemic Stroke

Emma Borggreven[1]

June 2025

Supervisors: Wiktor Olszewski[2], Odysseas Papakyriakou[2], Henk Marquering[2], and Frank Gijsen[1]

[1]Delft University of Technology, The Netherlands
[2]Amsterdam UMC, The Netherlands

Background. Accurate prediction of ischemic lesion volume (ILV) in the subacute phase is essential to estimate functional outcome, as the two are positively associated. Ischemic lesions can continue to evolve between 24 hours and 1 week after stroke onset, even after successful treatment. Radiomics offers a promising approach for ILV prediction using non-contrast computed tomography (NCCT), the first-line imaging modality in AIS. However, applying CT radiomics in AIS remains challenging, as ischemic lesion segmentation is time consuming and challenging, due to its low contrast. **Objective.** This study aims to investigate whether radiomic features extracted from post-treatment NCCT scans, acquired at 24 hours after stroke onset, can be used to predict the subacute ischemic lesion volume at 1 week. In addition, it explores whether simplified annotations are feasible for radiomic feature extraction. As a secondary analysis, this study explores whether incorporating clinical data has an added value for this prediction task. **Methods.** Patients from the MR CLEAN-NOIV trial, with 24-hour and 1-week follow-up NCCT scans available, were included. The included patients were randomly divided into a pre-training set (80%) and a test set (20%). Radiomic features were extracted from the 24-hour NCCT scan using three annotation types: (1) the original segmentation, (2) a bounding box annotation, and (3) a circle annotation. Feature selection included reproducibility filtering, low-variance filtering, correlation-based clustering, and Least Absolute Shrinkage and Selection Operator (LASSO) regression. Three XGBoost radiomics regression models were trained, using five-fold cross-validation. Additionally, three combined models, using a combination of clinical and radiomic features, and two clinical models, using only clinical features, were constructed. The performance of the models was evaluated on the test set using the coefficient of determination (R^2), concordance correlation coefficient (CCC), mean absolute error (MAE), and root mean squared error (RMSE). Feature importance was assessed using SHapley Additive exPlanations (SHAP). **Results.** The radiomics model based on the original segmentation achieved a high predictive performance (R^2 = 0.89, CCC = 0.95, MAE = 24 mL, RMSE = 31 mL). The radiomics model based on the bounding box achieved comparable performance, and the model based on the circle annotation yielded significantly lower performance. Incorporating clinical features did not significantly improve the predictive performance of the radiomics models. Across all well-performing models including radiomic features, the Run Length Non-Uniformity radiomic feature was a strong predictor of the 1-week ILV. **Conclusion.** Radiomic features extracted from 24-hour NCCT scans can accurately predict the subacute ILV at 1-week. A simplified bounding box annotation is a simpler and effective alternative to the detailed lesion segmentation, whereas the circle annotation showed poor performance and is not a good alternative for radiomic feature extraction in this context. These findings demonstrate the potential of radiomics and the use of simplified annotations for feature extraction to predict patient prognosis and guide personalized stroke care. However, further research is required before these models can be considered for clinical use.

Keywords: acute ischemic stroke, radiomics, non-contrast computed tomography, ischemic lesion volume, simplified annotation, prediction model, machine learning, regression, XGBoost

1 Introduction

Stroke is a major global health challenge, ranking as the second-leading cause of death worldwide, with approximately 12 million new cases reported annually [1]. Acute ischemic stroke (AIS), caused by the occlusion of an intracranial artery, accounts for the majority of these cases [1]. This occlusion severely reduces cerebral blood flow, resulting in brain tissue that either infarcts and becomes irreversibly damaged (ischemic core), or becomes functionally impaired but remains salvageable (ischemic penumbra) [2, 3, 4]. Brain cells within the penumbra can potentially be saved with timely treatment due to residual perfusion and the presence of collaterals [2]. Collateral vessels act as alternative pathways that provide just enough oxygen and nutrients to delay permanent tissue damage [5]. In some cases, AIS may also lead to the development of cerebral edema, a secondary injury that occurs as a response to the initial infarct and can emerge hours to days after stroke onset. Edema progression is typically developed in three phases. Cytotoxic edema occurs within minutes after stroke onset and is characterized by swelling of brain cells due to energy failure [6]. This is followed by ionic edema, which arises in the early hours as a result of disrupted ion gradients caused by cytotoxic edema [6]. Vasogenic edema develops later and results from the blood-brain barrier (BBB) breakdown, allowing fluid to accumulate in the extracellular space, which causes tissue swelling [6].

Non-contrast computed tomography (NCCT) is the first-line imaging modality for stroke assessment due to its rapid acquisition, low cost, and wide availability [7]. It is primarily used to differentiate between ischemic and hemorrhagic stroke [7, 8]. Ischemic tissue appears hypodense on NCCT due to water uptake in the affected region, while hemorrhagic lesions appear hyperdense as a result of blood accumulation [9]. The ischemic lesion assessed on NCCT scans consists of both infarcted and edematous volume [9].

AIS is primarily treated with intravenous thrombolysis (IVT) and/or endovascular thrombectomy (EVT), with the choice of treatment depending on the time since symptom onset, the location of the occlusion, and the size of the ischemic core [10, 11, 12, 13]. IVT involves the administration of a thrombolytic agent, such as alteplase, to chemically dissolve the occluding thrombus and is limited by a therapeutic time window of up to 9 hours after stroke onset [14]. EVT is a minimally invasive procedure in which the thrombus is mechanically removed and can be performed within an extended time window of up to 24 hours after stroke onset [11]. Both treatments aim to rapidly restore blood flow to the ischemic brain tissue and prevent further lesion expansion [15, 16]. Early treatment is crucial for both IVT and EVT, as the ischemic core expands over time if blood flow is not restored [7].

Despite these treatments, not all patients achieve favorable outcomes. Treatment effectiveness depends on multiple factors, such as time from stroke onset to treatment and the initial ischemic core volume [11, 13]. In recent years, numerous studies have used prediction models to estimate the outcome after AIS treatment, with the primary aim of identifying patients most likely to benefit from treatment [17, 18, 19]. While some studies aim to estimate tissue outcome [17], which refers to predicting, at voxel level, which brain tissue will eventually infarct, the majority focuses on functional outcome prediction [18, 19], as this directly reflects a patient's post-stroke quality of life and independence. Extensive research has shown a significant positive association between follow-up ischemic lesion volume (ILV) and 90-day functional outcome [20, 21, 22]. This relationship has increased the interest in predicting final ILV, as it, unlike functional outcome, directly quantifies the extent of brain tissue damage, making it informative for understanding treatment effect and underlying physiological processes.

These prediction models typically define the final ILV as the volume measured at 24 hours post stroke onset. However, research has shown that the ischemic lesion can continue to evolve during the subacute phase (24 hours to 1 week after stroke onset), even after successful reperfusion treatment [21, 22]. This continued lesion evolution suggests that the volume assessed on the 24-hour NCCT may underestimate the final ILV. Subacute lesion growth is thought to result from reperfusion injury, a process in which the sudden restoration of blood flow causes the tissue to be further damaged through physiological processes like excitotoxicity, oxidative stress, inflammation, microvascular injury, and BBB breakdown [22, 23]. These mechanisms can lead to delayed infarction and vasogenic edema. As the ILV continues to evolve in the days following treatment, the 1-week ILV provides a more accurate representation of the final ILV, and therefore long-term functional outcome and patient prognosis. This is supported by Krongold et al., who reported that final infarct volume can be approximated as early as 7 days post-stroke onset [24]. Since the lesion progression is thought to result from post-treatment physiological responses and treatment related variables, such as type of treatment and completeness of reperfusion, accurate prediction of the 1-week ILV likely requires treatment information and information from post-treatment imaging. Predicting the 1-week volume from the 24-hour NCCT scan allows for early estimation of the patient's prognosis, which is valuable when informing the patient and their relatives about the expected functional outcome, and for personalizing the rehabilitation plan. In addition, the continued growth of the lesion beyond 24 hours suggests that some patients may benefit from secondary treatments, such as neuroprotective agents, even after the acute phase. Although no such therapies have yet proven effective, ongoing research is investigating various neuroprotective therapies [7, 25]. In this context, accurate prediction of the 1-week ILV could serve as a useful reference for evaluating treatment response in clinical trials. By comparing the predicted ILV to the actual volume after secondary treatment, it can be used to determined whether the additional treatment is effective.

Prediction models for ILV mainly rely on magnetic resonance imaging (MRI) and perfusion CT (CTP), as these modalities offer high sensitivity for detecting ischemic regions and provide detailed information about tissue perfusion, respectively. However, the clinical utility of MRI is often limited by factors such as high cost, restricted availability, longer acquisition times, and contraindications like metallic implants [8, 26, 27]. Given its broader availability and rapid acquisition, computed tomography (CT) is more commonly used in clinical practice, making it a more feasible imaging modality for developing prediction models. Since CTP requires multiple time-resolved scans, it is highly sensitive to motion artifacts, which can compromise image quality and limit its use for prediction. As an alternative to CTP-based models, some studies have investigated prediction based on CT angiography (CTA) [28, 29], and despite NCCT being the first-line imaging modality in AIS, only one study has explored prediction based solely on NCCT images [30]. This study indicates that radiomics, an advanced imaging analysis technique, can extract quantitative features from NCCT scans that enhace their utility for prediction in AIS research. Radiomics has been widely applied in various fields such as oncology, neuroscience, and cardiology, demonstrating promising results [31, 32]. It enables the extraction of quantitative features, such as shape, intensity, and texture, that are mostly not visible to the human eye. These features can be used independently or combined with clinical data, to develop prediction models for diagnosis, treatment response and prognosis [33, 34, 35]. However, radiomics requires a detailed delineation of the volume of interest (VOI) from which features are extracted. On NCCT, ischemic lesion segmentations are particularly challenging due to its low contrast, making it difficult to distinguish ischemic tissue from normal brain tissue. In radiomics, this poses a major limitation, as the low contrast of NCCT makes accurate lesion delineation in AIS time consuming and subject to inter-observer variability.

Prior research in oncology has attempted to overcome the challenges of segmentations required for radiomics by using an alternative annotation, the smallest bounding box enclosing the lesion. These studies showed that bounding boxes could perform as well as, or even outperform, detailed tumor segmentation in classification prediction tasks, where the goal was to predict binary outcomes such as whether early tumor growth would occur or whether cancer has reached a particular stage [36, 37, 38]. Similarly, a recent study in AIS demonstrated that radiomic features extracted from a fixed 1 cm diameter spherical volume were able to predict hemorrhagic transformation [39]. These findings suggest that for certain prediction tasks, radiomic features extracted from simplified annotations can provide comparable or even superior predictive performance while reducing annotation effort and inter-observer variability.

This study aims to investigate whether radiomics features extracted from post-treatment NCCT scans acquired at 24 hours after stroke onset can be used to predict the subacute ischemic lesion volume at 1 week. Additionally, it explores whether simplified annotations could serve as feasible alternatives to detailed segmentations for radiomic feature extraction in this context. As a secondary analysis, this study explores whether there is an added value in using clinical data in combination with radiomics for predicting ILV at 1 week.

2 Materials and Methods

2.1 Patients and Imaging Acquisition

This study used data from patients enrolled in the Multicenter Randomized Clinical Trial of Endovascular Treatment of Acute Ischemic Stroke in the Netherlands - No Intravenous Thrombolysis (MR CLEAN-NOIV) trial [40]. The trial included adults ($\geq$ 18 years) with an AIS due to an intracranial proximal occlusion of the anterior circulation, eligible for EVT and IVT with alteplase within 4.5 hours. Patients were randomized to receive either EVT alone or IVT followed by EVT. Depending on the specific center capabilities and patient contraindications, follow-up imaging was performed with either MRI at 24 hours (±12 hours) after reperfusion or NCCT after both 24 hours (±12 hours) and 1 week (5-7 days). More details on the inclusion and exclusion criteria of the trial are described in the study protocol [41]. For this study, a pre-selected subset of 116 patients was used. This subset was defined in a previous study by Konduri et al. [42] that included only patients with NCCT imaging available at baseline, and at both 24-hour and 1-week follow-up. Their exclusion criteria were clear evidence of extensive contrast extravasation, poor quality scans that were incomplete or included movement artifacts, beam-hardening effects, and other technical errors [42]. For this study, patients with an ILV smaller than 5 mL on the 24-hour NCCT scan were also excluded, as lesions this small were not compatible with the annotation protocol (described in detail in section *2.2 Segmentations*). The eligible patients were randomly split into a pre-training set (80%) and a test set (20%), while ensuring that each patient was included in only one of the two groups. This split was fixed and maintained throughout all subsequent analyses.

The NCCT scans used in this study were acquired at multiple participating centers. However, the MR CLEAN-NOIV study did not report specific scanner models or acquisition protocols. The known acquisition parameters were a matrix size of 512×512 and a slice thickness of 5 mm. The scans covered the entire head, ranging from the base of the skull to the vertex.

2.2 Segmentations

Ischemic Lesion Segmentation. Ischemic lesions were semi-automatically segmented on baseline, 24-hour, and 1-week NCCT scans using a deep-learning based software developed by Nico.lab [43]. Two trained observers (a trained neurologist with >5 years of experience and an experienced neuroradiologist with >15 years of experience) manually corrected the segmentations using a fixed window width of 40 Hounsfield Units (HU) and a center level of 40 HU when needed [42]. The observers were blinded to all clinical information, except the occlusion location [42]. The ischemic lesion on NCCT was defined as the intra-axial hypodense areas in the affected hemisphere, and included brain and edema swelling extending into the contralateral hemisphere or resulting in sulcal or ventricular effacement [42]. Hemorrhages, defined as hyperdense regions, were excluded from the ischemic lesion segmentation. The ILVs were calculated by multiplying the number of voxels in the segmentation with the voxel size.

Simplified Annotations. In addition to the original segmentation, two simplified annotations were manually drawn on the 24-hour NCCT scan using 3D Slicer [44]. These annotations were based on the original 24-hour ischemic lesion segmentation and using a fixed window width of 40 HU and a center level of 40 HU. The first annotation, referred to as the bounding box, includes both the ischemic lesion and adjacent healthy brain tissue. In each axial slice the smallest possible rectangle was drawn that fully encloses the original lesion segmentation. Although the bounding box may initially include components other than brain tissue (e.g., cerebrospinal fluid (CSF), hemorrhage, or regions outside the brain), these were later excluded during pre-processing steps by intensity thresholding, described in section *2.3 Pre-processing and Radiomic Feature Extraction.* The second annotation, referred to as the circle, aimed to capture only the most severely affected tissue. A circle with a diameter of 1 cm was placed in the most hypodense location of the ischemic lesion, as this region is assumed to represent the most extensive tissue damage.

2.3 Pre-processing and Radiomic Feature Extraction

Prior to radiomic feature extraction, all NCCT images underwent skull stripping using a custom Python script (Python version 3.8.0). Additional pre-processing steps were performed using the PyRadiomics Python package (version 3.1.0) [45]. These steps included resampling of both the NCCT scans and corresponding segmentations to an isotropic voxel space of $1 \times 1 \times 1$ mm (using the default interpolators: linear interpolation for the image and nearest neighbor interpolation for the segmentation), intensity thresholding to exclude voxels with intensities outside the range of 20 to 80 HU as these represent components other than brain tissue (e.g., CSF, hemorrhage, or regions outside the brain) [46], and gray-evel discretization using a fixed bin width of 1 HU. For all other PyRadiomics settings, the default option was used. A complete overview of these settings is provided in Supplementary Table S1.

PyRadiomics [45] was used to extract radiomic features separately from three different VOI shapes (the original segmentation, bounding box, and circle) per patient. For the original segmentation VOIs, all default features from the following feature classes were extracted (n = 107): shape, first-order, gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), gray level dependence matrix (GLDM), and neighborhood gray tone difference matrix (NGTDM). For the bounding box and circle VOIs, shape features were excluded. These simplified annotations do not follow the actual boundaries of the ischemic lesion, meaning that the resulting VOIs do not represent the true shape of the infarct. As a result, shape features that describe the geometric properties of the VOI, are not meaningful for these annotation types. The radiomic features were extracted only from the original, unfiltered NCCT images. A description of each radiomic feature class is provided in Supplementary Note S2, and a full list of all the extracted radiomic features is included in Supplementary Table S3.

2.4 Clinical Features

In addition to radiomic features, the added value of clinical features, when combined with radiomic features, was also investigated. Clinical features considered as candidate predictors were limited to baseline characteristics and follow-up data collected within 24 hours after stroke onset. ILVs at baseline and 24 hours were also included as part of the clinical feature set. However, the 24-hour volume was calculated only for the original segmentation, as it depends on the voxel count within the segmentation. This voxel count is then multiplied by the voxel size to compute the ILV. Since, the bounding box and circle annotations do not represent the actual ischemic lesion shape, it is not possible to derive the 24-hour ILV from these annotations. Accordingly, this volume feature was not computed for the bounding box and circle annotation, reflecting the clinical scenario where simplified annotation would not yield the ILV.

Missing values in the clinical dataset were imputed using Multiple Imputation by Chained Equations (MICE) [47] in R version 4.5.0 [48]. MICE imputation was performed separately for the pre-training and test sets. All available clinical data from the MR CLEAN-NOIV study was used for the MICE imputation. In addition, the ILVs at the three time points, were also included in the imputation model. Variables were excluded from the imputation model if they were random and independent of other variables, binary versions of more informative variables, or categorical variables with only one observed class and missing values. The details of the MICE imputation are provided in Supplementary Note S4.

An overview and description of all clinical variables, is provided in Supplementary Table S5. The table indicates which variables were excluded from the MICE imputation and which variables were excluded from the clinical features used in this study.

2.5 Feature Selection

Radiomic feature selection was performed to reduce the risk of overfitting and consisted of four steps: (1) reproducibility filtering based on the intraclass correlation coefficient (ICC), (2) low-variance filtering, (3) correlation-based clustering, and (4) feature selection using LASSO regression. Feature selection was performed on the pre-training set to prevent information from the test set influencing the feature selection. To address the small yet commonly occurring interobserver variability introduced by manual or semiautomatic segmentations [49], ICC filtering was applied first. Since all segmentations were generated by a single observer, simulated segmentations were created for a randomly selected subset of 50% of the pre-training set. Radiomic features were extracted from the simulated and original segmentations, and the ICC was calculated for each feature using the Pingouin Python package (version 0.5.5) [50]. Radiomic features with an ICC ≥ 0.80 were retained, as these features are robust to small segmentation variations. Details on generating the simulated segmentations are provided in Supplementary Note S6. In the second step, retained features were excluded if their variance between patients was lower than 0.0001, as such low-variance features are unlikely to contribute to meaningful predictions. Third, to remove redundant features, pairwise Spearman correlation coefficients were calculated between all remaining features. Features with a correlation coefficient $\rho \geq 0.9$ were grouped into clusters. From each cluster, one representative feature was retained. This feature was selected based on having the highest average correlation with the other features in the same cluster, as it best represents the shared information. The final subset of features was selected using Least Absolute Shrinkage and Selection Operator (LASSO) regression, implemented with the scikit-learn Python package (version 1.3.2) [51]. The remaining features were standardized using Z-score normalization before LASSO regression to ensure comparability of scale. Normalization was applied at this stage, as ICC, variance, and correlation calculations are scale-invariant and would not benefit from standardization. LASSO was applied to the standardized feature set to identify the most relevant features for the prediction task. The regularization parameter (α) was tuned to ensure that the number of selected features was in accordance with the commonly used guideline of maintaining at least 10 outcome events per predictor variable [52, 53]. If the number of features after correlation filtering was already within this limit, LASSO regression was not applied. Details of the LASSO regression are provided in Supplementary Note S7.

Combined feature sets, included the radiomic features that were retained after the ICC selection step, and all clinical features. The combined feature sets were then subjected to the remaining selection steps: low-variance filtering, correlation-based filtering, Z-score normalization, and LASSO regression. For the feature sets only containing clinical features, only the final three selection steps (low-variance filtering, correlation-based filtering, and LASSO regression) were applied, as ICC filtering was not applicable for the clinical variables.

2.6 Model Development

A total of eight gradient boosting regression models were developed to predict the ILV at 1-week follow-up. The ground truth for the prediction task was defined as the volume derived from the ischemic lesion segmentation on the 1-week follow-up NCCT scan. The models were implemented using the eXtreme Gradient Boosting (XGBoost) algorithm in Python via the xgboost package (version 2.1.4) [54]. XGBoost was selected for its suitability for tabular data, and strong performance in regression task [55]. It uses an ensemble of decision trees that sequentially reduces the prediction error, enabling the model to capture non-linear relationships and achieve high predictive accuracy [55]. Additionally, XGBoost is computational efficient and supports feature importance analysis, which enhances the interpretability of the model [56, 55].

Three models were developed using only radiomic features (R models), extracted from the different VOI annotations (original ischemic lesion segmentation, bounding box, and circle annotation), three other models used a combination of radiomic and clinical features (RC models) for each annotation type, and the last two models used only clinical features (C models), one including the 24-hour ILV (C-ILV model) and the other not including it (C-noILV model). As such, each model was trained on a distinct set of features.

In this study, a standardized pipeline was used for all models to ensure consistency and reproducibility. This pipeline consisted of feature selection, hyperparameter tuning, model training with cross-validation, and final evaluation on an independent test set.

Hyperparameter tuning was performed using the RandomizedSearchCV function from the scikit-learn library (version 1.3.2) [51], applying 5 fold cross-validation on the pre-training set. Fifty parameter combinations were randomly sampled, and the best configuration was selected based on the lowest mean squared error (MSE) in the validation set. The settings for the hyperparameter tuning are listed in Supplementary Table S8. Following parameter optimization, each model was trained across 5 folds using the optimal parameter configuration. Within each fold, 80% of the pre-training set was used for training and 20% for validation. Within each fold, the root mean squared error (RMSE) of the validation set was monitored. To prevent overfitting, early stopping was used to stop the training if the validation RMSE did not improve for 10 boosting rounds. All models were implemented using the XGBoost Regressor with the squared error as loss function. A schematic overview of the pipeline used for the development of the radiomics models is shown in Figure 1.

2.7 Model Performance and Feature Importance

Model performance was evaluated on an independent test set, which was not used during feature selection, hyperparameter tuning, or model training. Each of the five models trained during cross-validation was used to predict on the test set, and the resulting predictions were averaged to obtain the final prediction. Performance metrics included the coefficient of determination (R^2), concordance correlation coefficient (CCC), mean absolute error (MAE), and RMSE. The 95% confidence intervals (CIs) of the performance metrics were estimated using bootstrapping with 1000 resamples [57]. To assess agreement between predicted and reference values, Bland–Altman plots were constructed. These plots visualize the mean difference (bias) and the ±95% limits of agreement between predicted and true volumes.

To gain insight into the contribution of individual features to the model prediction, feature importance was evaluated with a SHapley Additive exPlanations (SHAP) analysis using the SHAP Python package (version 0.44.1) [58]. For each model, SHAP values were computed across the five training folds with TreeExplainer, and the average values were used to estimate global feature importance. This procedure was conducted separately for each of the eight models.

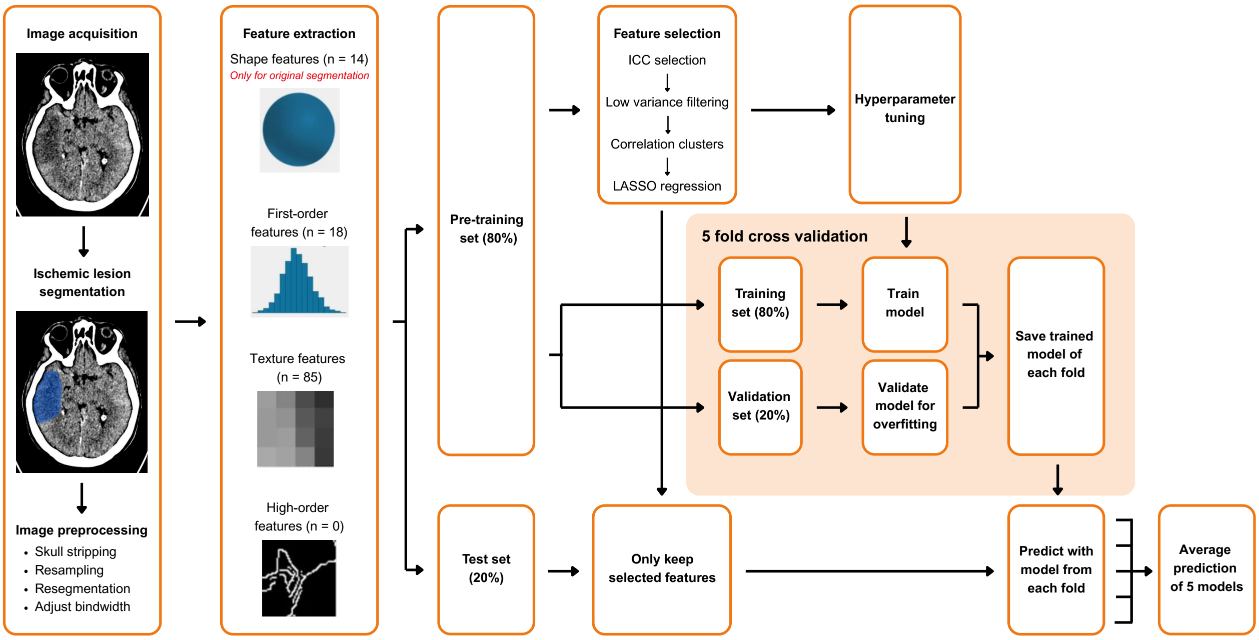


Figure 1: Systematic overview of the pipeline used for the development of the radiomics models.

2.8 Statistical Analysis

To assess whether the differences in prediction errors were statistically significant, the absolute and squared errors were compared between models. Pairwise comparisons were performed across models for each test subject. The radiomics models based on the bounding box and circle annotation were each compared separately to the radiomics model using the original segmentation. Additionally, for each annotation type, the radiomics model and the clinical model were individually compared to the combined model. The normality of the paired differences in errors was evaluated using the Shapiro–Wilk test. If the differences were normally distributed, a paired *t*-test was used to test for significance. Otherwise, the Wilcoxon signed-rank test was applied. All hypothesis tests were two-sided, and a p-value ≤ 0.05 was considered statistically significant. All statistical analyses were conducted in Python using the statistical functions from the SciPy package (version 1.10.1) [59].

3 Results

3.1 Patient Characteristics

From the 116 patients in the MR CLEAN-NOIV subset defined by Konduri et al. [42], 16 were excluded due to missing imaging data (n=1) or an ILV below 5 mL at 24-hour follow-up (n=15). As a result, 100 patients were included in this study (median age, 71 years; interquartile range (IQR), 59-76 years; 37 female). The pre-training set included 80 patients and the test set included 20 patients. The clinical characteristics of the patients included in this study are provided in Table 1.

3.2 Lesion Characteristics

An example of the original segmentation and simplified annotations is shown in Figure 2. Among the 100 included patients, the median ILV was 13 (IQR: 4.7 – 34) mL at baseline, 28 (IQR: 16 – 87) mL at 24-hour follow-up, and 45 (IQR: 20 – 125) mL at 1-week follow-up. Subacute lesion growth, defined as an increase in ILV between the 24-hour and 1-week follow-up scans, was observed in 71 (71%) patients. The median ILV difference over this interval was 9.3 (IQR: -0.51 – 30) mL.

3.3 Feature Selection

For the radiomic features sets, 107 features were extracted from the original segmentation, and 93 features were extracted from both the bounding box and the circle annotation. After applying the ICC selection step, 101 features remained for the original segmentation, 92 for the bounding box annotation, and 26 for the circle annotation. Variance filtering excluded 4 features for the original lesion segmentation and 8 for the bounding box annotation, while no features were excluded for the circle annotation. After correlation-based selection, the number of features was further reduced to 27, 22, and 7 features for the original lesion segmentation, bounding box annotation, and circle annotation, respectively. Finally, LASSO regression was used to select the 10 most relevant features for the original segmentation and bounding box annotation. For the circle annotation, LASSO regression was not applied, as the number of features was already below the selection threshold following correlation filtering.

Prior to feature selection, the combined feature sets, consisted of 134 features for the original segmentation, 124 features for the bounding box annotation, and 58 features for the circle annotation. The clinical feature sets contained 33 for the C-ILV model and

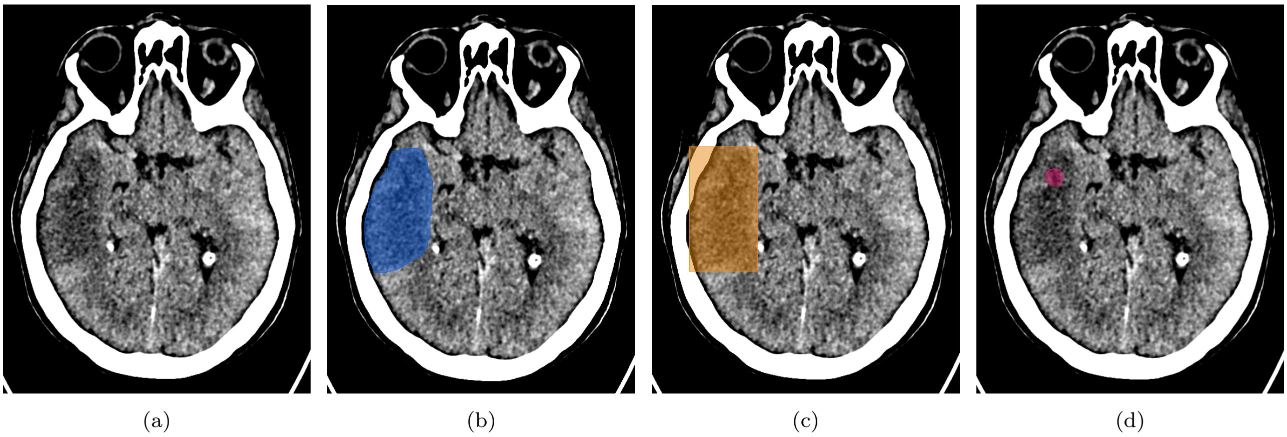


Figure 2: Example of a 24-hour NCCT scan (a) with the three segmentation approaches: (b) the original ischemic lesion segmentation, (c) the bounding box annotation, and (d) the circle annotation.

Table 1. Clinical characteristics

	Included patients (n=100)
Age (years)	71 (59 – 76)
Sex (female)	37 (37%)
Medical history	
Ischemic stroke	12 (12%)
Atrial fibrillation	16 (16%)
Diabetes mellitus	17 (17%)
Hypertension	42 (42%)
Pre-stroke mRS > 2	3 (3%)
Clinical parameters	
Systolic blood pressure (mmHg)	151.0 (132.5 – 170.0)
Blood glucose (mmol/L)	6.7 (5.9 – 8.3)
Left affected hemisphere side	51 (51%)
Baseline NIHSS	16 (11 – 19)
Radiological parameters	
ASPECTS score	9 (8 – 10)
Occlusion location	
Intracranial ICA	0 (0%)
Terminal ICA	19 (19%)
M1	64 (64%)
Proximal M2	15 (15%)
None	2 (2%)
Tandem lesion	20 (20%)
Collateral score	
0	6 (6%)
1	29 (29%)
2	45 (45%)
3	18 (18%)
Treatment parameters	
IVT administered	58 (58%)
Type of treatment (randomization)	
IVT followed by EVT	54 (54%)
EVT alone	46 (46%)
Median duration (min)	
From stroke onset to randomization	93.5 (70.0 – 144.3)
From stroke onset to needle	92.0 (78.0 – 141.0)
From stroke onset to reperfusion	180.0 (149.0 – 247.8)
From stroke onset to groin puncture	147.5 (112.5 – 210.5)
From door to needle	32.0 (25.0 – 40.0)
From door to groin puncture	69.0 (53.5 – 92.3)
Post treatment parameters (24 hours)	
Successful reperfusion	78 (78%)
Hemorrhage	32 (32%)
mAOL score	3 (3 – 3)
NIHSS	8 (4 – 15)
Sich	4 (4%)

Values are represented as a number (%) or median (interquartile range).

Abbreviations: ASPECTS: Alberta Stroke Program Early Computed Tomography Score, EVT: endovascular thrombectomy, ICA: intracranial cartoid artery, IVT: intravenous thrombolysis, mAOL: modified Arterial Occlusion Lesion, mRS: modified Rankin scale, NIHSS: National Institutes of Health Stroke Scale, sich: symptomatic intracranial hemorrhage.

32 features for the C-noILV model. After variance filtering, correlation-based selection, and LASSO regression, 10 features were retained for each feature set.

Supplementary Figures S9 and S10 show the correlation matrices and LASSO regression plots for each feature set, respectively. An overview of the number of features retained after each selection step is provided in Supplementary Figure S11, and Supplementary Table S12 lists the selected features per model.

3.4 Model Performance and Feature Importance

Following hyperparameter tuning, the optimized parameters for each XGBoost regression model are provided in Supplementary Table S13.

Radiomics Models. The predictive performance of the three radiomics models is summarized in Table 2. The model based on radiomic features extracted from the original segmentation achieved an R^2 of 0.89 (95% CI: 0.74 – 0.96), CCC of 0.95 (95% CI: 0.89 – 0.98), MAE of 24 mL (95% CI: 16 – 33), and RMSE of 31 mL (95% CI: 19 – 42). The model based on the bounding box annotation yielded a higher R^2 and CCC, and lower MAE and RMSE, compared to the original segmentation model. The differences in absolute errors ($p = 0.29$) and squared errors ($p = 0.43$) were not statistically significant. The model based on the circle annotation showed a lower R^2 and CCC, and higher error values compared to the model based on the original segmentation. The absolute errors ($p < 0.01$) and squared errors ($p < 0.01$) were significantly higher than those of the original segmentation model. Scatter plots comparing the predicted versus actual ILVs at 1-week follow-up for each radiomics model are shown in Figure 3a-c, with a combined overview in Figure 3d. Bland–Altman plots for each radiomics model are presented in Figure 4. The mean differences between predicted and actual ILVs were 16 mL, 3 mL, and −9 mL for the original segmentation, bounding box, and circle, respectively.

SHAP analysis was used to assess the contribution of each feature to the model outcome. For the model based on the original segmentation, the three features with the largest impact on the model output were the *Run Length Non-Uniformity*, a texture feature derived from the GLRLM, the *Minor Axis Length* shape feature, and the *Maximum 2D Diameter Column* shape feature. For the bounding box model these were the *Run Length Non-Uniformity*, the *Maximum*, a first-order feature, and the *Busyness* derived from the NTGDM. For the circle model they were the*Small Area Low Gray Level Emphasis* and *Large Area Emphasis* derived from the GLSZM, and the *Contrast* feature derived from the GLCM. The SHAP summary plots, visualizing the impact of all included features on the model output, are shown in Figure 5.

Combined and Clinical Models. The predictive performance of the combined (RC) and clinical (C) models for each annotation type is summarized in Table 3, alongside the previous reported results of the radiomics (R) models for comparison. For the original segmentation, the RC model achieved the highest R^2 (0.91; 95% CI: 0.79 – 0.95) and CCC (0.95; 95% CI: 0.89 – 0.97), and lowest error values (MAE: 24 mL; 95% CI: 17 – 31, and RMSE: 29 mL; 95% CI: 22 – 34). Compared to the clinical model, the RC model showed statistically significantly lower absolute errors ($p = 0.02$) and squared errors ($p = 0.03$). Among the bounding box based models, the R model yielded the highest R^2 (0.92; 95% CI: 0.81 – 0.97) and CCC (0.95; 95% CI: 0.89 – 0.98), and lowest error values (MAE: 20 mL; 95% CI: 13 – 28 and RMSE: 27 mL; 95% CI: 17 – 36). Additionally, the bounding box RC model demonstrated significantly lower absolute

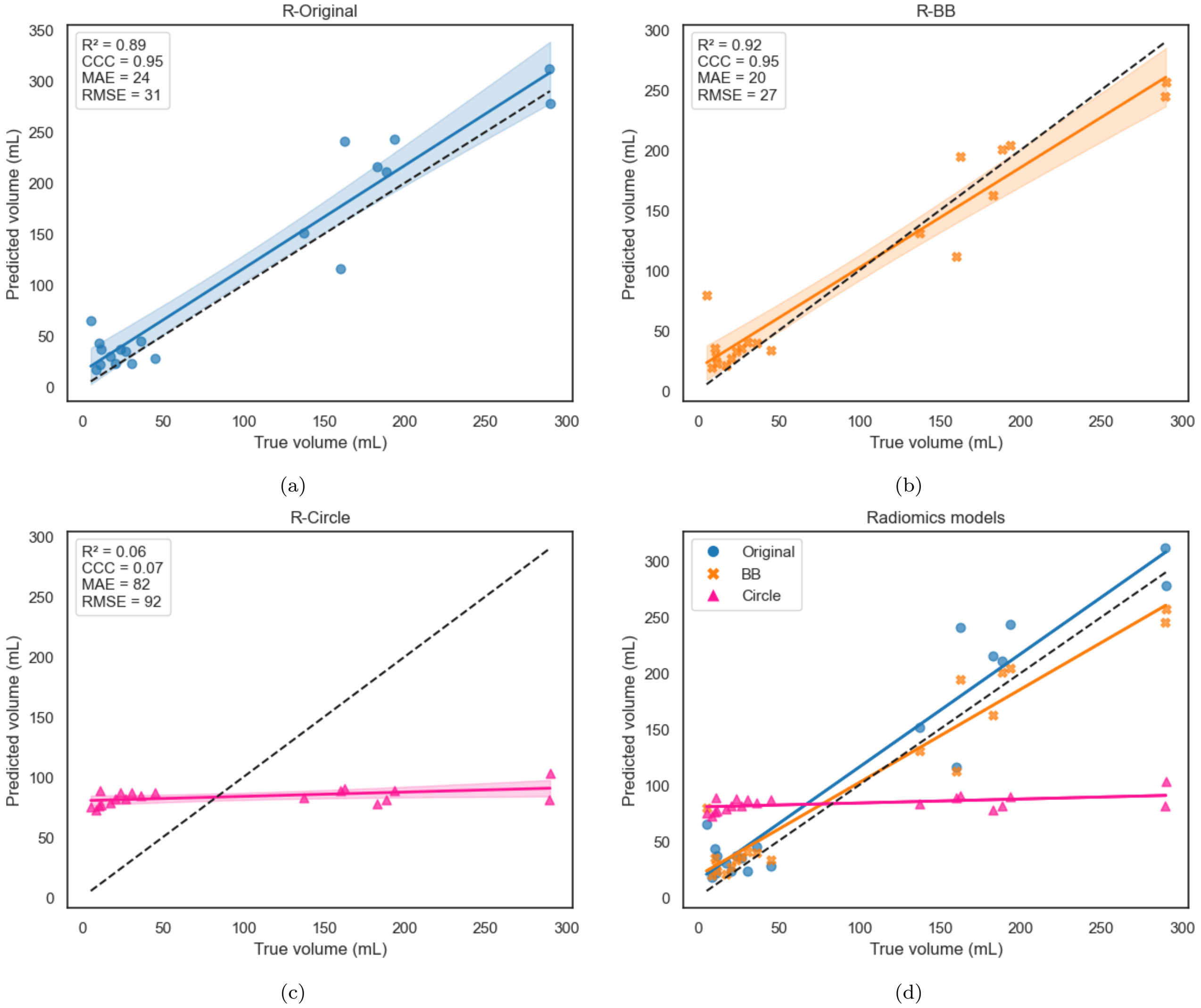
Table 2. Comparing the predictive performance metrics of the radiomics regression models on the test set

	R-Original	**R-BB**	***p*-value**	**R-Circle**	***p*-value**
R^2	0.89 (0.74 – 0.96)	0.92 (0.81 – 0.97)	–	0.06 (-0.25 – 0.12)	–
CCC	0.95 (0.89 – 0.98)	0.95 (0.89 – 0.98)	–	0.07 (-0.01 – 0.14)	–
MAE (mL)	24 (16 – 33)	20 (13 – 28)	0.29	82 (66 – 103)	$< 0.01^*$
RMSE (mL)	31 (19 – 42)	27 (17 – 36)	0.43	92 (67 – 118)	$< 0.01^*$

*: p-value ≤ 0.05

Values are presented as the mean (95%CIs). Two-sided paired t-test or two-sided Wilcoxon signed-rank test was performed to compare the prediction errors between the radiomics models based on the simplified annotations with the radiomics model based on the original segmentation.

Abbreviations: CCC: concordance correlation coefficient, MAE: mean absolute error, R^2: coefficient of determination, R-BB: model based on radiomic features extracted from the bounding box annotation, R-Circle: model based on radiomic features extracted from the circle annotation, RMSE: root mean squared error, R-Original: model based on radiomic features extracted from the original segmentation.



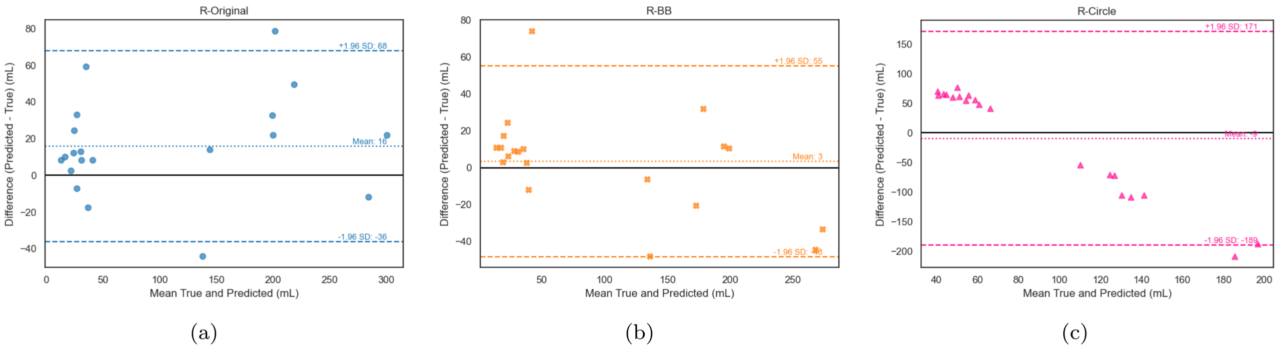
(a)

(b)

(c)

(d)

Figure 3: Scatter plots showing the agreement between the ground-truth and predicted values and the 95% confidence intervals of the radiomics models based on (a) the original segmentation, (b) the bounding box annotation, and (c) the circle annotation. Plot (d) compares the predictions of all three models in a single plot.

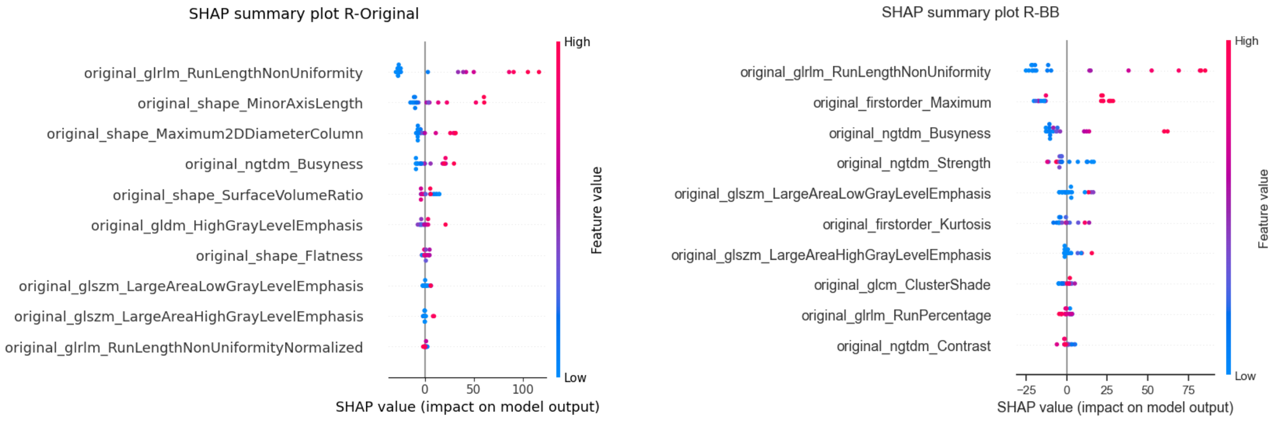


(a)

(b)

(c)

Figure 4: Bland-Altman plots indicating the bias and ±95% limits of agreement for the radiomics models based on the (a) original segmentation, (b) the bounding box annotation, and (c) the circle annotation.



(a)

(b)

Figure 5: The feature importance of the radiomics models based on (a) the original segmentation and (b) the bounding box annotation, using the SHapley Additive exPlanations (SHAP) analysis.

errors ($p < 0.01$) and squared errors ($p < 0.01$) compared to the C model. For the circle annotation, the RC model showed the highest R^2 (0.31; 95% CI: -0.18 – 0.51) and CCC (0.41; 95% CI: 0.15 – 0.60), and lowest error values (MAE: 69 mL; 95% CI: 53–86, and RMSE: 79 mL; 95% CI: 61 – 95). No statistically significant differences in absolute errors or squared errors were observed between the RC model and either the R or C models. Scatter plots comparing the radiomics, combined, and clinical models per annotation type are shown in Figure 6. Both the individual scatter plots and Bland-Altman plots for the combined models and clinical models are shown in Supplementary Figure S14.

SHAP analysis showed that in both the RC model based on the original segmentation and the RC model based on the bounding box, the *Run Length Non-Uniformity* feature derived from the GLRLM was the feature with the largest impact on the model output. In the C-ILV model, the 24-hour ILV had the largest impact on the model output. The SHAP summary plots of these models are provided in Figure 7, and the SHAP summary plots of the other models are provided in Supplementary Figure S15.

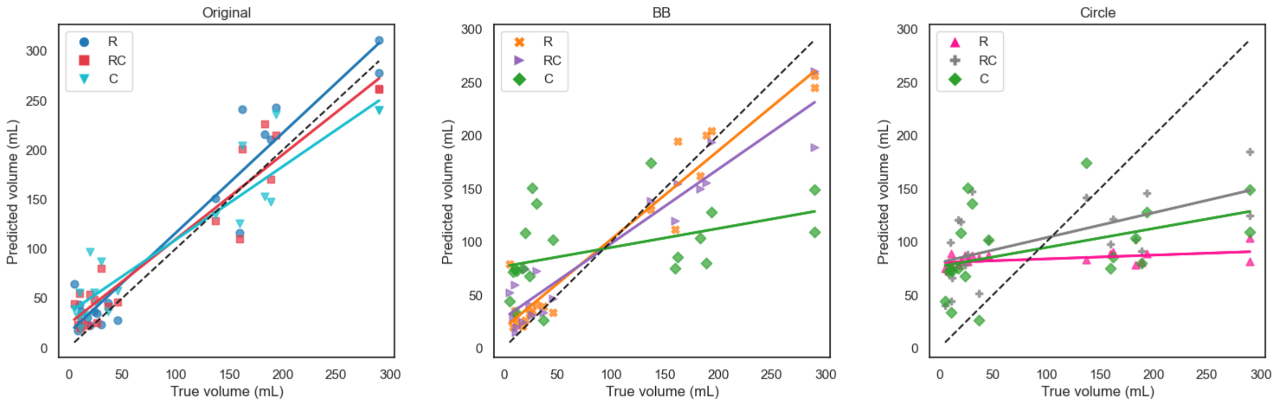
Table 3. Comparing the predictive performance measures of the radiomics (R) and clinical (C) models with the combined (RC) models on the test set

		RC	**R**	*p*-value	**C**	*p*-value
Original	$\mathbf{R^2}$	0.91 (0.79 – 0.95)	0.89 (0.74 – 0.96)	–	0.84 (0.69 – 0.91)	–
	CCC	0.95 (0.89 – 0.97)	0.95 (0.89 – 0.98)	–	0.91 (0.82 – 0.94)	–
	MAE (mL)	24 (17 – 31)	24 (16 – 33)	0.99	32 (23 – 40)	0.02*
	RMSE (mL)	29 (22 – 34)	31 (19 – 42)	0.73	37 (28 – 44)	0.03*
BB	$\mathbf{R^2}$	0.86 (0.75 – 0.94)	0.92 (0.81 – 0.97)	–	0.19 (-0.27 – 0.38)	–
	CCC	0.91 (0.83 – 0.96)	0.95 (0.89 – 0.98)	–	0.31 (0.07 – 0.50)	–
	MAE (mL)	25 (15 – 37)	20 (13 – 28)	0.30	75 (58 – 94)	< 0.01*
	RMSE (mL)	35 (22 – 49)	27 (17 – 36)	0.67	85 (66 – 104)	< 0.01*
Circle	$\mathbf{R^2}$	0.31 (-0.18 – 0.51)	0.06 (-0.25 – 0.12)	–	0.19 (-0.27 – 0.38)	–
	CCC	0.41 (0.15 – 0.60)	0.07 (-0.01 – 0.14)	–	0.31 (0.07 – 0.50)	–
	MAE (mL)	69 (53 – 86)	82 (66 – 103)	0.16	75 (58 – 94)	0.29
	RMSE (mL)	79 (61 – 95)	92 (67 – 118)	0.20	85 (66 – 104)	0.26

*: p-value ≤ 0.05

Values are presented as the mean (95%CIs). Two-sided paired t-test or two-sided Wilcoxon signed-rank test was performed to compare the prediction errors between the radiomics model and the combined model, and the clinical model and the combined model, for each annotation type.

Abbreviations: BB: bounding box, C: model based on clinical features, CCC: concordance correlation coefficient, MAE: mean absolute error, R^2: coefficient of determination, R: model based on radiomic features, RC: model based on both radiomic and clinical features, RMSE: root mean squared error.

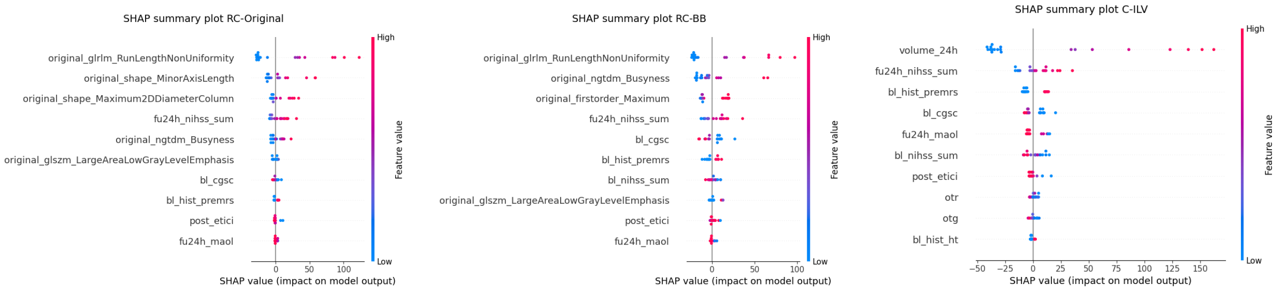


(a)

(b)

(c)

Figure 6: Scatter plots comparing the predictions of three models: radiomic features only (R), radiomics combined with clinical features (RC), and clinical features only (C). The models are based on (a) the original segmentation, (b) the bounding box annotation, and (c) the circle annotation. Each subfigure shows the predicted versus ground-truth 1-week ischemic lesion volumes, illustrating the performance of the three models for each segmentation strategy.



(a)

(b)

(c)

Figure 7: The feature importance of the combined models based on (a) the original segmentation and (b) the bounding box annotation, and of (c) the clinical model including the 24-hour ischemic lesion volume. The feature importance is assessed using the SHapley Additive exPlanations (SHAP) analysis.

4 Discussion

This study aimed to investigate whether radiomic features extracted from post-treatment NCCT scans acquired at 24 hours after stroke onset can be used to predict the subacute ischemic lesion volume at 1-week follow-up in patients with AIS, and whether simplified annotations could serve as feasible alternatives to detailed segmentations for radiomic feature extraction. To address this, three regression models were developed, each trained on radiomics features extracted from another VOI annotation (original segmentation, bounding box annotation, and circle annotation). The radiomics model based on the original segmentation demonstrated high predictive performance, indicating that radiomic features extracted from the 24-hour NCCT scan can be used to accurately predict ILV at 1-week follow-up. The model based on the bounding box annotation showed comparable predictive performance. Utilizing this bounding box annotation can significantly reduce the time needed for defining the VOI without compromising the predictive performance of the model, offering a feasible alternative to detailed segmentations. In contrast, the model based on the circle annotation resulted in significantly lower predictive performance compared to the model based on the original segmentation, indicating that this simplified annotation method is not a suitable alternative for radiomic feature extraction in this context.

This appears to be the first study to predict the 1-week ILV using radiomic features extracted from 24-hour post-treatment NCCT scans in patients with AIS. The high predictive performance observed in the radiomics model based on the original segmentation is confirmed by prior work by Wang et al. [30], who also demonstrated the utility of NCCT-based radiomics for the prediction of ILV. However, their study focused on predicting ILV at 24-hour follow-up from baseline NCCT and stratified patients by treatment type, whereas the current study predicts ILV at 1 week and the model was developed for all including patients.

Furthermore, their approach relied on deep learning-based lesion segmentation, while this study evaluated both semi-automatic segmentation and simplified annotation approaches. The finding that a bounding box-based model achieved comparable performance to the model based on the detailed segmentation is in line with studies in oncology, where bounding box annotations have been shown to yield equal or even better performance than detailed segmentations [36, 37, 38]. Although prior work has examined variations in box size and dimensionality, these studies were limited to classification tasks in oncology. This study extends these findings by demonstrating that bounding box annotations are also feasible alternatives for continuous outcome prediction in AIS. Conversely, the poor performance of the model based on the circle annotation is in contrast with a study by Heo et al. [39], who reported accurate classification of hemorrhagic transformation using a circle annotation. While both studies employed the same annotation strategy, their research focused on a classification task, whereas this study predicted a continuous outcome. The small size of the circle VOI may not provide enough spatial or contextual information for accurate regression based prediction. Overall, these results confirm existing observations on the value of radiomics in NCCT, and introduce a novel approach by demonstrating that bounding box annotations offer a practical, reproducible alternative to detailed segmentation for volumetric outcome prediction in AIS.

SHAP analysis of the two best performing radiomics models revealed that the feature with the greatest impact in both the original segmentation and bounding box models was *Run Length Non-Uniformity (RLNU)*, a texture feature derived from the GLRLM. This feature quantifies the variability in the lengths of consecutive voxels with the same gray level, where higher values indicate more heterogeneity among these run lengths. In both models, high RLNU values were associated with higher SHAP values, meaning they pushed the prediction toward a larger predicted ILV, while low RLNU values were linked to negative SHAP values, reducing the predicted ILV. This observed heterogeneity may reflect underlying tissue damage caused by reperfusion injury, which is thought to contribute to subacute ischemic lesion growth [22]. Reperfusion injury encompasses a range of processes that further damage brain tissue, including microvascular injury, BBB breakdown, and inflammation [23]. While these processes are not directly visible on NCCT, they could potentially influence voxel intensities and spatial patterns, leading to increased textural heterogeneity in the ischemic lesion. The specific impact of these processes on NCCT voxel intensities is unknown. However, the positive association between higher RLNU and larger predicted volumes across all high performing models including radiomic features suggests that it has an effect on the heterogeneity of the ischemic lesion. As such, it is possible that RLNU is a marker for ongoing secondary injury.

There are several factors that may explain why the model based on the bounding box annotation performs comparable to the model based on the original segmentation, despite not including shape features and encompassing adjacent normal brain tissue. Firstly, although the bounding box includes adjacent normal brain tissue, the ischemic lesion still dominates in terms of voxel count and intensity distribution. As a result, it seems possible that first-order and texture features extracted from the bounding box are still mainly influenced by the ischemic lesion. Therefore, these features could possibly remain representative of the underlying tissue characteristics of the ischemic lesion, although some adjacent normal brain tissue is included. Secondly, the ICC feature selection step in this study was performed using simulated second-observer segmentations that involved geometric transformations (e.g., translations, dilation, and erosion). These operations preserve the overall shape of the VOI and therefore do not completely reflect the actual variability introduced by inter-observer delineations. As a result, certain non-reproducible features may not have been excluded during feature selection. Since the bounding box annotation has a more robust shape, this may have contributed to the strong predictive performance observed for the model based on this annotation compared to the model based on the original segmentation. Thirdly, segmentation of ischemic lesions on NCCT is difficult due to low contrast, which can lead to under segmentation. The bounding box, includes a broader region and may therefore be more likely to capture the entire ischemic lesion, especially in cases where the boundaries are unclear. Lastly, in oncology it has been suggested that bounding boxes perform well despite missing shape features because they include the microscopic peritumor environment, which is clinically relevant for tumor growth but not directly visible on imaging [36, 37, 38]. While speculative, a similar phenomenon may occur in ischemic stroke, where adjacent tissue could contain early or subtle changes, such as microvascular damage, cytotoxic edema, or BBB breakdown, which are not visible on images, but could be detected through radiomic analysis. This could provide additional prognostic information to predict ILV at 1 week.

Due to the poor predictive performance of the circle based model, this simplified annotation cannot be considered a good alternative. One likely factor contributing to the poor predictive performance of the model based on the circle annotation is the result of the ICC filtering step. Due to the small size of the VOI (1 cm circle), even minor variations in the placement of the circle led to substantial differences in the voxels included in the VOI. This placement variability resulted in low reproducibility for many of the extracted features, which were excluded during the ICC selection step. As a result, a considerable number of potentially informative radiomic features were excluded from the feature set used in prediction. Automating the circle annotation step could reduce the variability of the placement and preserve a greater number of potentially informative features, which may improve predictive performance. Another possible explanation of the model's poor performance is the voxel intensity thresholding applied during preprocessing. As described in the methods section (*2.3 Pre-processing and Radiomics Feature Extraction*), only voxels with HU values between 20 and 80 were retained to exclude non-brain tissue components. To ensure a fair comparison, the same pre-processing pipeline was applied for each annotation type. However, since the circle was specifically placed in the most hypodense region of the ischemic lesion, it often contained voxels with HU values below 20. Although this threshold was applied to exclude CSF and ensure consistency across VOIs, it may have unintentionally excluded severely infarcted tissue, which in the subacute phase can present with HU values below this threshold [60]. This thresholding step further reduced the size of the VOI and may have removed relevant image information. It is plausible that these effects, in combination with the already limited spatial context of the circle annotation, reduced the predictive value of the extracted features for this regression-based prediction task. This study also explored whether the incorporation of clinical data can improve the predictive performance of the radiomics models. To investigate this, three combined models were trained using both radiomic and clinical features, and additionally two clinical models were trained using only clinical features. The combined models did not outperform any of the radiomics models, although they significantly outperformed the clinical models for the original segmentation and bounding box annotation. This suggests that, for predicting the subacute ILV at 1 week from 24-hour NCCT scans, clinical variables do not provide additional prognostic value beyond what is already captured by radiomic features. Conversely, adding radiomic features to clinical models significantly improved predictive performance. That combined models perform better than clinical models is in line with previous studies [61, 62]. However, the findings that combined models do not outperform radiomics models differ from these studies. Zhou et al. and Zhang et al. used radiomic features derived from baseline imaging (MRI and NCCT, respectively) to predict functional outcomes and reported that clinical variables significantly improved the predictions [61, 62]. A key difference is that the current study used NCCT scans acquired at 24 hours, instead of at baseline. These post-treatment images may contain information about reperfusion injury, which is believed to contribute to subacute ischemic lesion growth [22]. These findings could suggest that, at this stage, clinical variables may no longer add significant predictive value, since radiomic features already capture critical prognostic information.

Notable is the high predictive performance of the C-ILV model. Although this model was significantly improved by adding radiomic features from the original segmentation, the high predictive performance of the model still makes it interesting to look at its feature importance. SHAP analysis indicated that 24-hour ILV was the feature with the highest impact on its predictions. Given the wide range of ILVs in the dataset, from 5 to 400 mL, the 24-hour ILV may serve as a useful starting point. The model then uses the other features to estimate the likelihood and extent of further lesion progression. However, the radiomics models achieved even higher predictive performance without explicitly including the 24-hour ILV as input feature. A study by Volpe et al. reported that certain radiomic features, particularly texture features, are strongly correlated with the volume of the VOI [63]. This suggests that radiomics models may indirectly incorporate information about the ischemic lesion size through these features. The significant improvement observed when adding radiomic features to the C-ILV model indicates that radiomic features capture additional information about the subacute evolution of ischemic lesions.

4.1 Limitations

This study has several limitations. Firstly, the dataset was relatively small, as inclusion required the availability of NCCT scans at both 24 hours and 1 week after stroke onset. This requirement excluded a substantial number of patients from the MR CLEAN-NOIV trial. A larger training set may improve the accuracy and robustness of the predictive models. Secondly, the generalizability of the model is limited by both the composition of the study cohort and the absence of external validation. All patients included in this study received EVT and were eligible for IVT, meaning they were treated within 4.5 hours after symptom onset. As a result, the model was trained exclusively on early presenters, who typically have smaller baseline ILVs and potentially less complex EVT procedures. This narrow treatment profile may limit the applicability of the model to patients treated at later time points, received only IVT, or received no treatment. While 5 fold cross-validation was used and the final predictions were obtained by averaging the outputs of five models trained on different patient splits, it cannot compensate for the low diversity in the dataset. Also, the data used in this study was acquired across multiple centers, but all patients came from a single clinical trial. Testing the model on an external dataset is necessary to assess its robustness before it can be applied in clinical practice. Thirdly, the subacute ILV at 1 week was assessed on NCCT, which is less sensitive than MRI for detecting ischemic tissue. While NCCT was used because MRI scans at 1-week follow-up were not available in the MR CLEAN-NOIV trial, MRI can provide a more accurate representation of the true ILV and could serve as a more reliable ground truth. Another limitation is that the high-performing models consistently overestimated small ILVs at 1 week, as observed in the Bland-Altman plots (Figure 4 and Supplementary Figure S14. This indicates a systematic positive bias for lower volume predictions. A possible explanation could be the choice of the RMSE as the loss function during model training. RMSE is sensitive to large errors because it squares the difference between predicted and actual values before averaging. This means that the model places greater emphasis on minimizing larger errors, often at the expense of smaller ones. As a result, it may slightly overestimate smaller ILVs to balance overall error, leading to a systematic positive bias for small values, particularly in datasets with a wide range of target values, as in this study. Finally, EVT is a technically complex procedure that varies between patients and can cause treatment complications like vessel rupture [64]. Treatment specific variables, such as the number of attempts or procedure duration, could be associated with ischemic lesion evolution after treatment, however these were not considered in this study. This is also supported by a study from Wang et al. [30], which reported a lower prediction accuracy for ILV in patients who received EVT. Their findings suggest that the complexity of the procedure may have an influence on the ischemic lesion development, and when unaccounted for, can limit the predictive performance of a model.

4.2 Future research

Future research should aim to increase the size of the dataset, include patients who receive EVT at later time points, IVT alone, or no treatment at all, and validate on an external test set, such as the Imaging Repository of Ischemic Stroke (IRIS) cohort. The IRIS cohort is particularly suitable for validation as it includes NCCT scans acquired at both 24 hours and 1 week after stroke onset, which is in line with the imaging time points used in the current study. Treatment-specific variables related to EVT, such as the number of retrieval attempts, should also be considered, as they may influence lesion evolution. In addition, future research should apply strategies to mitigate volume-related bias. One approach is to explore custom loss functions that weight errors based on the magnitude of the target variable, thereby balancing the influence of small and large ILVs during model training. An alternative solution could be to stratify model development based on the 24-hour ILV, training separate models for small and large lesions. Although this does not narrow the prediction range, it reduces the influence of large-volume cases on the loss function. This may help prevent systematic overestimation of small lesions by allowing the model to better minimize errors within each subgroup.

Future research could also further investigate the simplified annotations. Assessing how the size of bounding box and circle annotations influences predictive performance may help determine the best alternative to detailed segmentations for radiomics based modeling. This could offer insight into when and how simplified annotations are most effective. In addition, given the strong performance and reduced annotation time shown by the bounding box approach in this study, it could be explored as a alternative for other radiomics based prediction tasks requiring a time-consuming manual segmentation step.

Future research could also explore more specific predictive objectives. Rather than predicting ILV, it may be valuable to predict infarct and edema volumes separately. Previous research has shown that both components continue to evolve after 24 hours and are associated with functional outcome [9]. Separately predicting infarct and edema volumes could clarify which component dominates total ILV and inform future treatment strategies if secondary therapies are developed. Additionally, future research could also focus on predicting ischemic lesion growth, such as the absolute or relative increase in volume between 24 hours and 1 week, or by classifying patients with or without lesion progression. Shifting the focus from volume to growth could help identify radiomic features that specifically associated with ongoing physiological processes related to ischemic lesion progression in the subacute phase. These features could provide valuable biological insights and potentially support future research into secondary treatment strategies.

5 Conclusion

This study showed that radiomic features extracted from post-treatment NCCT scans acquired at 24 hours after stroke onset can accurately predict the subacute ischemic lesion volume at 1 week. A bounding box offers a feasible and faster alternative to the detailed ischemic lesion segmentation for radiomic feature extraction, without compromising predictive performance. Conversely, the circle annotation showed poor performance and does not appear to be a suitable alternative for radiomic feature extraction in this context. Adding clinical data to the radiomic features did not improve model performance, suggesting that clinical variables do not provide additional prognostic value beyond what is already captured by radiomic features at this point. Across all well-performing models, the ischemic lesion heterogeneity consistently emerged as a strong predictor of the 1-week volume, possibly containing information about ongoing secondary injury. Despite these promising results, challenges still remain, particularly the systematic overestimation of small ILVs and the limited generalizability due to the composition of the dataset. Future research should address these limitations before such radiomics based prediction models can be considered for clinical use.

Disclaimer

This thesis was written with the help of ChatGPT Plus [65]. The AI tools was used to structure sentences, refine word choices, and improve overall flow of paragraphs. The AI-assisted suggestions were carefully reviewed, and the content, interpretation of data, and conclusions were formed without the use of AI. Example prompts used for this theses include: *"Give other examples for this sentence.", "Replace this word with something else.", "Write this paragraph in the same style as this section.", "Add this information to this paragraph."*.

Acknowledgement

I would like to thank Henk Marquering for introducing me to this amazing topic, giving me the opportunity to work on it in such a great team, and for his supervision throughout my thesis. I am also very grateful to Frank Gijsen for his enthusiasm and supervision from Delft and for believing in a good collaboration with the people from Amsterdam UMC. Odysseas Papakyriakou, thank you for helping me get familiar with machine learning. It is cool to see how much I have learned as I started with no background in AI. I am especially thankful to my daily supervisor Wiktor Olszewski, whose enthusiasm, support, and helped me through challenges and encouraged my ideas. I truly could not have wished for a better supervisors. I would also like to thank the assessment committee for their time and effort in evaluating my thesis. To the DoMiBo, thank you for the countless coffee breaks. Even as a non-coffee drinker, I needed every one of them. I want to thank my roommate Fee, for her amazing design skills and for creating the figure on my title page, which turned out even better than I could have imagined. Last but not least, I want to thank my family and friends for always believing in me and supporting me throughout this journey. I could not have done it without them.

References

1. Feigin VL, Brainin M, Norrving B, Marting SO, Pandian J, Lindsay P, Grupper MF, and Rautalin I. World Stroke Orginization: Global Stroke Fact Sheet 2025. International Journal of Stroke 2024
2. Campbell BC, De Silva DA, Macleod MR, Coutts SB, Schwamm LH, Davis SM, and Donnan GA. Ischaemic stroke. Nature reviews Disease primers 2019; 5:70
3. Jung S, Gilgen M, Slotboom J, El-Koussy M, Zubler C, Kiefer C, Luedi R, Mono ML, Heldner MR, Weck A, et al. Factors that determine penumbral tissue loss in acute ischaemic stroke. Brain 2013; 136:3554–60
4. Zhang H, Prabhakar P, Sealock R, and Faber JE. Wide genetic variation in the native pial collateral circulation is a major determinant of variation in severity of stroke. Journal of Cerebral Blood Flow & Metabolism 2010; 30:923–34
5. Liebeskind DS. Collateral circulation. Stroke 2003; 34:2279–84
6. Chen S, Shao L, and Ma L. Cerebral edema formation after stroke: emphasis on blood–brain barrier and the lymphatic drainage system of the brain. Frontiers in cellular neuroscience 2021; 15:716825
7. Dammavalam V, Lin S, Nessa S, Daksla N, Stefanowski K, Costa A, and Bergese S. Neuroprotection during thrombectomy for acute ischemic stroke: a review of future therapies. International Journal of Molecular Sciences 2024; 25:891
8. Birenbaum D, Bancroft LW, and Felsberg GJ. Imaging in acute stroke. Western Journal of Emergency Medicine 2011; 12:67
9. Konduri P, Kranendonk K van, Boers A, Treurniet K, Berkhemer O, Yoo AJ, Zwam W van, Oostenbrugge Rv, Lugt A van der, Dippel D, et al. The role of edema in subacute lesion progression after treatment of acute ischemic stroke. Frontiers in neurology 2021; 12:705221

10. Hilkens NA, Casolla B, Leung TW, and Leeuw FE de. Stroke. The Lancet 2024; 403:2820–36

11. Olthuis SG, Pirson FAV, Pinckaers FM, Hinsenveld WH, Nieboer D, Ceulemans A, Knapen RR, Robbe MQ, Berkhemer OA, Walderveen MA van, et al. Endovascular treatment versus no endovascular treatment after 6–24 h in patients with ischaemic stroke and collateral flow on CT angiography (MR CLEAN-LATE) in the Netherlands: a multicentre, open-label, blinded-endpoint, randomised, controlled, phase 3 trial. The Lancet 2023; 401:1371–80

12. Psychogios M, Brehm A, Ribo M, Rizzo F, Strbian D, Räty S, Arenillas JF, Martínez-Galdámez M, Hajdu SD, Michel P, et al. Endovascular Treatment for Stroke Due to Occlusion of Medium or Distal Vessels. New England Journal of Medicine 2025

13. Yoshimura S, Sakai N, Yamagami H, Uchida K, Beppu M, Toyoda K, Matsumaru Y, Matsumoto Y, Kimura K, Takeuchi M, et al. Endovascular therapy for acute stroke with a large ischemic region. New England Journal of Medicine 2022; 386:1303–13

14. Ma H, Campbell BC, Parsons MW, Churilov L, Levi CR, Hsu C, Kleinig TJ, Wijeratne T, Curtze S, Dewey HM, et al. Thrombolysis guided by perfusion imaging up to 9 hours after onset of stroke. New England Journal of Medicine 2019; 380:1795–803

15. Emberson J, Lees KR, Lyden P, Blackwell L, Albers G, Bluhmki E, Brott T, Cohen G, Davis S, Donnan G, et al. Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. The Lancet 2014; 384:1929–35

16. Goyal M, Menon BK, Van Zwam WH, Dippel DW, Mitchell PJ, Demchuk AM, Dávalos A, Majoie CB, Der Lugt A van, De Miquel MA, et al. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. The Lancet 2016; 387:1723–31

17. Benzakoun J, Charron S, Turc G, Hassen WB, Legrand L, Boulouis G, Naggara O, Baron JC, Thirion B, and Oppenheim C. Tissue outcome prediction in hyperacute ischemic stroke: Comparison of machine learning models. Journal of Cerebral Blood Flow & Metabolism 2021; 41:3085–96

18. Zihni E, Madai VI, Livne M, Galinovic I, Khalil AA, Fiebach JB, and Frey D. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. Plos one 2020; 15:e0231166

19. Liu Y, Yu Y, Ouyang J, Jiang B, Yang G, Ostmeier S, Wintermark M, Michel P, Liebeskind DS, Lansberg MG, et al. Functional outcome prediction in acute ischemic stroke using a fused imaging and clinical deep learning model. Stroke 2023; 54:2316–27

20. Meng X and Ji J. Infarct volume and outcome of cerebral ischaemia, a systematic review and meta-analysis. International Journal of Clinical Practice 2021; 75:e14773

21. Bucker A, Boers AM, Bot JC, Berkhemer OA, Lingsma HF, Yoo AJ, Van Zwam WH, Van Oostenbrugge RJ, Van der Lugt A, Dippel DW, et al. Associations of ischemic lesion volume with functional outcome in patients with acute ischemic stroke: 24-hour versus 1-week imaging. Stroke 2017; 48:1233–40

22. Konduri P, Voorst H van, Bucker A, Kranendonk K van, Boers A, Treurniet K, Berkhemer O, Yoo AJ, Zwam W van, Oostenbrugge R van, et al. Posttreatment ischemic lesion evolution is associated with reduced favorable functional outcome in patients with stroke. Stroke 2021; 52:3523–31

23. Fisher M and Savitz SI. Pharmacological brain cytoprotection in acute ischaemic stroke—renewed hope in the reperfusion era. Nature Reviews Neurology 2022; 18:193–202

24. Krongold M, Almekhlafi MA, Demchuk AM, Coutts SB, Frayne R, and Eilaghi A. Final infarct volume estimation on 1-week follow-up MR imaging is feasible and is dependent on recanalization status. NeuroImage: Clinical 2015; 7:1–6

25. Vos E, Geraedts V, Van Der Lugt A, Dippel D, Wermer M, Hofmeijer J, Es A van, Roos Y, Peeters-Scholte C, and Wijngaard I van den. Systematic review-combining neuroprotection with reperfusion in acute ischemic stroke. Frontiers in neurology 2022; 13:840892

26. Vilela P and Rowley HA. Brain ischemia: CT and MRI techniques in acute ischemic stroke. European journal of radiology 2017; 96:162–72

27. Baird AE and Warach S. Magnetic resonance imaging of acute stroke. Journal of Cerebral Blood Flow & Metabolism 1998; 18:583–609

28. Hokkinen L, Mäkelä T, Savolainen S, and Kangasniemi M. Evaluation of a CTA-based convolutional neural network for infarct volume prediction in anterior cerebral circulation ischaemic stroke. European radiology experimental 2021; 5:1–11

29. Qiu W, Kuang H, Ospel JM, Hill MD, Demchuk AM, Goyal M, and Menon BK. Automated prediction of ischemic brain tissue fate from multiphase computed tomographic angiography in patients with acute ischemic stroke using machine learning. Journal of stroke 2021; 23:234–43

30. Wang X, Meng Y, Dong Z, Cao Z, He Y, Sun T, Zhou Q, Niu G, Ding Z, Shi F, et al. Segmentation of infarct lesions and prognosis prediction for acute ischemic stroke using non-contrast CT scans. Computer Methods and Programs in Biomedicine 2025; 258:108488

31. Liu Z, Wang S, Dong D, Wei J, Fang C, Zhou X, Sun K, Li L, Li B, Wang M, et al. The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. Theranostics 2019; 9:1303

32. Chen Q, Xia T, Zhang M, Xia N, Liu J, and Yang Y. Radiomics in stroke neuroimaging: techniques, applications, and challenges. Aging and disease 2021; 12:143

33. Harding-Theobald E, Louissaint J, Maraj B, Cuaresma E, Townsend W, Mendiratta-Lala M, Singal AG, Su GL, Lok AS, and Parikh ND. Systematic review: radiomics for the diagnosis and prognosis of hepatocellular carcinoma. Alimentary pharmacology & therapeutics 2021; 54:890–901

34. Chen R, Fu Y, Yi X, Pei Q, Zai H, and Chen BT. Application of radiomics in predicting treatment response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: strategies and challenges. Journal of Oncology 2022; 2022:1590620

35. Tang Fh, Xue C, Law MY, Wong Cy, Cho Th, and Lai Ck. Prognostic prediction of cancer based on radiomics features of diagnostic imaging: the performance of machine learning strategies. Journal of Digital Imaging 2023; 36:1081–90

36. Zhu Y, Wei Y, Chen Z, Li X, Zhang S, Wen C, Cao G, Zhou J, and Wang M. Different radiomics annotation methods comparison in rectal cancer characterisation and prognosis prediction: a two-centre study. Insights into Imaging 2024; 15:211

37. Liu D, Zhang W, Hu F, Yu P, Zhang X, Yin H, Yang L, Fang X, Song B, Wu B, et al. A bounding box-based radiomics model for detecting occult peritoneal metastasis in advanced gastric cancer: a multicenter study. Frontiers in Oncology 2021; 11:777760

38. Zhou J, Zhang Y, Chang KT, Lee KE, Wang O, Li J, Lin Y, Pan Z, Chang P, Chow D, et al. Diagnosis of benign and malignant breast lesions on DCE-MRI by using radiomics and deep learning with consideration of peritumor tissue. Journal of Magnetic Resonance Imaging 2020; 51:798–809

39. Heo J, Sim Y, Kim BM, Kim DJ, Kim YD, Nam HS, Choi YS, Lee SK, Kim EY, and Sohn B. Radiomics using non-contrast CT to predict hemorrhagic transformation risk in stroke patients undergoing revascularization. European Radiology 2024; 34:6005–15

40. LeCouffe NE, Kappelhof M, Treurniet KM, Rinkel LA, Bruggeman AE, Berkhemer OA, Wolff L, Voorst H van, Tolhuisen ML, Dippel DW, et al. A randomized trial of intravenous alteplase before endovascular treatment for stroke. New England journal of medicine 2021; 385:1833–44

41. Treurniet KM, LeCouffe NE, Kappelhof M, Emmer BJ, Es AC van, Boiten J, Lycklama GJ, Keizer K, Yo LS, Lingsma HF, et al. MR CLEAN-NO IV: intravenous treatment followed by endovascular treatment versus direct endovascular treatment for acute ischemic stroke caused by a proximal intracranial occlusion—study protocol for a randomized clinical trial. Trials 2021; 22:1–15

42. Konduri P, Cavalcante F, Voorst H van, Rinkel L, Kappelhof M, Kranendonk K van, Treurniet K, Emmer B, Coutinho J, Wolff L, et al. Role of intravenous alteplase on late lesion growth and clinical outcome after stroke treatment. Journal of Cerebral Blood Flow & Metabolism 2023; 43:116–25

43. Barros RS, Tolhuisen ML, Boers AM, Jansen I, Ponomareva E, Dippel DW, Lugt A van der, Oostenbrugge RJ van, Zwam WH van, Berkhemer OA, et al. Automatic segmentation of cerebral infarcts in follow-up computed tomography images with convolutional neural networks. Journal of neurointerventional surgery 2020; 12:848–52

44. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magnetic resonance imaging 2012; 30:1323–41

45. Van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RG, Fillion-Robin JC, Pieper S, and Aerts HJ. Computational radiomics system to decode the radiographic phenotype. Cancer research 2017; 77:e104–e107

46. Broocks G, Flottmann F, Scheibel A, Aigner A, Faizy TD, Hanning U, Leischner H, Broocks SI, Fiehler J, Gellissen S, et al. Quantitative lesion water uptake in acute stroke computed tomography is a predictor of malignant infarction. Stroke 2018; 49:1906–12

47. Buuren S van and Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 2011; 45:1–67

48. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2021. Available from: `https://www.R-project.org/`

49. Cimflova P, Ospel JM, Marko M, Menon BK, and Qiu W. Variability assessment of manual segmentations of ischemic lesion volume on 24-h non-contrast CT. Neuroradiology 2022 :1–9

50. Vallat R. Pingouin: statistics in Python. Journal of Open Source Software 2018 Nov; 3:1026. DOI: `10.21105/joss.01026`

51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, and Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011; 12:2825–30

52. Peduzzi P, Concato J, Feinstein AR, and Holford TR. Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. Journal of clinical epidemiology 1995; 48:1503–10

53. Peduzzi P, Concato J, Kemper E, Holford TR, and Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology 1996; 49:1373–9

54. Chen T and Guestrin C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 2016 :785–94

55. Wiens M, Verone-Boyle A, Henscheid N, Podichetty JT, and Burton J. A tutorial and use case example of the eXtreme gradient boosting (XGBoost) artificial intelligence algorithm for drug development applications. Clinical and Translational Science 2025; 18:e70172

56. Liang D, Wang L, Zhong P, Lin J, Chen L, Chen Q, Liu S, Luo Z, Ke C, and Lai Y. Global Burden of Iodine Deficiency: Insights and Projections to 2050 Using XGBoost and SHAP. Advances in Nutrition 2025 :100384

57. Efron B and Tibshirani RJ. An Introduction to the Bootstrap. Chapman and Hall/CRC, 1993

58. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, and Lee SI. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence 2020; 2:2522–5839

59. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 2020; 17:261–72. DOI: `10.1038/s41592-019-0686-2`

60. Ali B, Amir A, Ago JL, and Khalid MI. Investigation of the Association between Computed Tomography Hounsfield Units and Cerebral Infarction Phases: A Cross-sectional Study. Journal of Neonatal Surgey 2025

61. Zhou Y, Wu D, Yan S, Xie Y, Zhang S, Lv W, Qin Y, Liu Y, Liu C, Lu J, et al. Feasibility of a clinical-radiomics model to predict the outcomes of acute ischemic stroke. Korean Journal of Radiology 2022; 23:811

62. Zhang L, Wu J, Yu R, Xu R, Yang J, Fan Q, Wang D, and Zhang W. Non-contrast CT radiomics and machine learning for outcomes prediction of patients with acute ischemic stroke receiving conventional treatment. European Journal of Radiology 2023; 165:110959

63. Volpe S, Isaksson LJ, Zaffaroni M, Pepa M, Raimondi S, Botta F, Presti GL, Vincini MG, Rampinelli C, Cremonesi M, et al. Impact of image filtering and assessment of volume-confounding effects on CT radiomic features and derived survival models in non-small cell lung cancer. Translational lung cancer research 2022; 11:2452

64. Pilgram-Pastor SM, Piechowiak EI, Dobrocky T, Kaesmacher J, Den Hollander J, Gralla J, and Mordasini P. Stroke thrombectomy complication management. Journal of neurointerventional surgery 2021; 13:912–7

65. OpenAI. ChatGPT Plus (Version 4-Turbo). 2025. Available from: `https://openai.com/chatgpt`

6 Supplement

Supplementary Table S1. Parameters used for PyRadiomics Feature Extraction

Feature Extractor Level	
Image Normalization	
normalize	False
normalizeScale	1
removeOutliers	None
Resampling the image/mask	
resampledPixelSpacing	[1, 1, 1]
interpolator	'sitkLinear'
padDistance	5
Pre-Cropping	
preCrop	False
Resegmentation	
resegmentRange	[20, 80]
resegmentMode	'absolute'
resegmentShape	False
Mask validation	
minimumROIDimensions	2
minimumROISize	None
geometryTolerance	None
correctMask	False
Miscellaneous	
additionalInfo	True

Feature Class Level	
Label	1
Image discretization	
binWidth	1
binCount	None
Forced 2D Extraction	
force2D	False
force2Ddimension	0
Texture matrix weighting	
weightingNorm	None
Distance to neighbour	
distances	[1]

Feature Class Specific Settings	
First Order	
voxelArrayShift	0
GLCM	
symmetricalGLCM	True
GLDM	
gldm_a	0

Non default parameter settings are marked grey.

Supplementary Note S2: Description of Radiomics Feature Classes

- Shape-based features describe the geometric properties of the volume of interest (VOIs), such as volume and maximum surface area. These features are derived from the 3D mesh representation of the VOI shape and are independent of gray level values [1].
- First-order features are derived from the gray level histogram of the image and characterise the distribution of individual voxel grey level values within the VOI [1]. Each voxel is analysed independently of its neighbors (single-voxel analysis), which defines these features as first-order. Examples include mean, median, and standard deviation.
- Second-order features, also referred to as "texture" features, evaluate the spatial relationships between voxel gray levels within the VOI [2]. PyRadiomics offers a set of matrix-based texture feature classes, each capturing different spatial dependencies. These include:
  - Gray Level Co-occurrence Matrix (GLCM): Features that describe how often voxel pairs with specific gray levels occur, at predefined distances and across 13 directions in 3D [2].
  - Gray Level Run Length Matrix (GLRLM): Features that measure the length and distribution of consecutive voxels with the same gray level in specified directions [3].
  - Gray Level Size Zone Matrix (GLSZM): Features that quantify the size of homogeneous zones, which are formed by connected voxels with the same gray level, regardless of direction [4].
  - Gray Level Dependence Matrix (GLDM): Features that describe the number of connected voxels within a predefined distance that depend on the gray level of the center voxel [5].
  - Neighborhood Gray Tone Difference Matrix (NGTDM): Features that quantify the difference between the gray level of a voxel and the average gray level of its neighboring voxels within a predefined distance [6].

- High-order features are extracted after applying filters or transformations to the original image, enabling the detection of complex patterns that are not visible in the native gray level domain [7]. Commonly used methods include Fourier transforms, Wavelet decomposition, and Gaussian filters. This approach significantly increases the number of extracted features, since features are computed from each transformed image representation.

Supplementary Table S3. Extracted Radiomic Features

Class	Features
Shape 3D* (n=14)	Elongation, Flatness, Least Axis Length, Major Axis Length, Maximum 2D Diameter Column, Maximum 2D Diameter Row, Maximum 2D Diameter Slice, Maximum 3D Diameter, Mesh Volume, Minor Axis Length, Sphericity, Surface Area, Surface Volume Ratio, Voxel Volume
First-order (n=18)	10th Percentile, 90th Percentile, Energy, Entropy, Interquartile Range, Kurtosis, Maximum, Mean Absolute Deviation, Mean, Median, Minimum, Range, Robust Mean Absolute Deviation, Root Mean Squared, Skewness, Total Energy, Uniformity, Variance
GLCM (n=24)	Autocorrelation, Cluster Prominence, Cluster Shade, Cluster Tendency, Contrast, Correlation, Difference Average, Difference Entropy, Difference Variance, Inverse Difference, Inverse Difference Moment, Inverse Difference Moment Normalized, Inverse Difference Normalized, Informational Measure of Correlation 1, Informational Measure of Correlation 2, Inverse Variance, Joint Average, Joint Energy, Joint Entropy, Maximal Correlation Coefficient, Maximum Probability, Sum Average, Sum Entropy, Sum of Squares
GLRLM (n=16)	Gray Level Non-Uniformity, Gray Level Non-Uniformity Normalized, Gray Level Variance, High Gray Level Run Emphasis, Long Run Emphasis, Long Run High Gray Level Emphasis, Long Run Low Gray Level Emphasis, Low Gray Level Run Emphasis, Run Entropy, Run Length Non-Uniformity, Run Length Non-Uniformity Normalized, Run Percentage, Run Variance, Short Run Emphasis, Short Run High Gray Level Emphasis, Short Run Low Gray Level Emphasis
GLSZM (n=16)	Gray Level Non-Uniformity, Gray Level Non-Uniformity Normalized, Gray Level Variance, High Gray Level Zone Emphasis, Large Area Emphasis, Large Area High Gray Level Emphasis, Large Area Low Gray Level Emphasis, Low Gray Level Zone Emphasis, Size-Zone Non-Uniformity, Size-Zone Non-Uniformity Normalized, Small Area Emphasis, Small Area High Gray Level Emphasis, Small Area Low Gray Level Emphasis, Zone Entropy, Zone Percentage, Zone Variance
GLDM (n=14)	Dependence Entropy, Dependence Non-Uniformity, Dependence Non-Uniformity Normalized, Dependence Variance, Gray Level Non-Uniformity, Gray Level Variance, High Gray Level Emphasis, Large Dependence Emphasis, Large Dependence High Gray Level Emphasis, Large Dependence Low Gray Level Emphasis, Low Gray Level Emphasis, Small Dependence Emphasis, Small Dependence High Gray Level Emphasis, Small Dependence Low Gray Level Emphasis
NGTDM (n=5)	Busyness, Coarseness, Complexity, Contrast, Strength

* Shape features were only extracted for the original segmentation, and not for the bounding box or circle annotation.

Abbreviations: GLCM:, Gray level Co-occurence Matrix, GLRLM: Gray Level Run Length Matrix, GLSZM: Gray Level Size Zone Matrix, GLDM: Gray Level Dependence Matrix, NGTDM: Neighborhood Gray tone Difference Matrix.

Supplementary Note S4: Description of MICE Imputation

Missing values in the clinical dataset were imputed using Multiple Imputation by Chained Equations (MICE) [8], in R version 4.5.0 [9]. MICE is an iterative method that uses prediction models to predict the missing values as a function of the other variables in the dataset [8]. The imputation method used depends on the variable class. For this imputation the default methods are used: predictive mean matching (PMM) for numerical variables, a proportional odds model (POLR) for ordinal variables, and logistic regression for categorical variables. Supplementary Table S5 provides a list of all the clinical variables and their class. The quickpred() function was used to automatically identify the most relevant predictors for each variable with missing data. This function computes pairwise correlations between variables and selects the variables with a correlation above the threshold [8]. The correlation threshold is set at 0.1 and a maximum of 7 predictors is set per variable that requires imputation. Additional MICE settings include a maximum of 15 iterations and a generation of five imputed datasets (m=5). The convergence was visually assessed using trace plots, and only the first imputed dataset was used as the final imputed dataset.

Supplementary Table S5. Overview and Description of Clinical Variables

Feature	**Class**	**Description**
volume_bl	numeric	Ischemic lesion volume at baseline.
volume_24h	numeric	Ischemic lesion volume at 24 hours after stroke onset.
volume_1wk[*]	numeric	Ischemic lesion volume at 1 week after stroke onset.
hem_24h	category	Hemorrhage present on 24-hour NCCT scan.
r_treatmentall[†]	category	Allocated type of treatment.
r_age	numeric	Age of the patient.
r_sex	category	Sex of the patient.
r_sidestroke	category	Side of the stroke (left or right hemisphere).
bl_nihss_sum	ordinal	NIHSS[a] score at baseline.
bl_hist_is	category	History of ischemic stroke.
bl_hist_af	category	History of atrial fibrillation.
bl_hist_dm	category	History of diabetes mellitus.
bl_hist_ht	category	History of hypertension.
bl_hist_premrs[‡]	ordinal	The mRS[b] score before stroke.
bl_hist_premrs_over2[†,‡]	category	mRS[b] before stroke is higher than 2.
bl_phy_rrsyst	numeric	Systolic blood pressure at baseline (mmHg).
bl_lab_glucose[‡]	numeric	Blood glucose at baseline (mmol/L).
bl_aspects	ordinal	ASPECTS score at baseline. The ASPECTS quantifies early ischemic changes in the brain on CT. A score of 10 indicates no changes, and 1 point is subtracted for each affected brain region.
bl_occloc	category	Location of the occlusion at baseline assessed on CTA.
bl_cgsc	ordinal	Collateral score at baseline. This score quantifies the extent of collateral flow visible on CTA. The score ranges from 0, indicating no collaterals, to 3, indicating collateral flow to 100% of the affected territory.
post_etici[‡]	ordinal	The eTICI[c] score assessed after reperfusion treatment.
tici2b3[†,‡]	category	eTICI[c] score after reperfusion treatment is 2B or 3.
fu24h_maol[‡]	ordinal	mAOL[d] score based on 24-hour follow-up CTA (default) or MRI.
fu24h_recan[‡]	category	mAOL[d] score is 3 at 24-hour follow-up.
otorep[‡]	numeric	Time from stroke onset to reperfusion.
otr	numeric	Time from stroke onset to randomization.
otn[‡]	numeric	Time from stroke onset to needle.
otg[‡]	numeric	Time from stroke onset to groin puncture.
dtg[‡]	numeric	Time from arriving at the hospital to groin puncture.
dtn[‡]	numeric	Time from arriving at the hospital to needle injection.
mrs_def[*]	ordinal	mRS[b] at 90 days after stroke onset.
mrs_rev[*,†]	ordinal	mRS[b] at 90 days after stroke onset with score levels reversed.
fu24h_nihss_sum[‡]	ordinal	NIHSS[a] score at 24 hours after stroke onset.
fu1wk_nihss_sum[*,‡]	ordinal	NIHSS[a] score at 1 week after stroke onset.

(Continued)

Supplementary Table S5. Continued

sich	category	Symptomatic intracranial hemorrhage according to Heidelberg criteria.
ivt_given[*,†,‡]	category	IVT given prior to EVT regardless of randomization.
tandemlesion[‡]	category	Ipsilateral extracranial carotid tandem lesion. Tandem lesion was defined as an intracranial target occlusion with ipsilateral extracranial carotid dissection, clinically significant atherosclerotic stenosis, or atherosclerotic occlusion.
ivt_admin	category	Any type of IVT admistered (this includes escape IVT when given).

[*] Excluded as candidate predictor (input for feature selection).

[†] Excluded from MICE imputation.

[‡] Contains missing values.

[a] Scores on the NIHSS range from 0 to 42, with higher scores indicating a more severe neurological deficit.

[b] Scores on the mRS range from 0 (no functional limitations) to 6 (death), with higher scores indicating more severe functional disability. A score of 2 or less indicates functional independence.

[c] Scores on the eTICI scale range from 0 (no perfusion) to 3 (complete perfusion), with higher scores indicating more successful reperfusion. A score of 2B or higher is generally considered successful recanalization.

[d] Scores on the mAOL scale range from 0 (no recanalization) to 3 (complete recanalization), with higher scores indicating better arterial recanalization. A score of 3 reflects complete recanalization.

Abbreviations: ASPECTS: Alberta Stroke Program Early Computed Tomography Score, CTA: computed tomography angiography, eTICI: expanded treatment in cerebral infarction, EVT: endovascular thrombectomy, IVT: intravenous thrombolysis, mAOL: modified Arterial Occlusive Lesion, NCCT: non-contrast computed tomography, NIHSS: National Institutes of Health Stroke Scale, MRI: magnetic resonance imaging, mRS: modified Rankin scale .

Supplementary Note S6: Simulated Segmentation Generation and ICC Calculation

Manual and semi-automatic segmentations of the ischemic lesion on a follow-up NCCT scan introduce inter-observer variability [10]. Research has shown that radiomics features can be sensitive to these differences, as variability in lesion delineation can affect the stability and reproducibility of extracted features [11]. To retain only features that are robust to segmentation variation, an ICC-based filtering step was applied. In the absence of segmentations from multiple observers, one simulated segmentation was generated per annotation type (original, bounding box, and circle) for a randomly selected 50% subset of the pre-training set. These simulations aimed to reflect realistic differences between observers in segmenting the ischemic lesion on NCCT.

For the original and bounding box segmentations, the simulated masks were generated by randomly applying a in-plane spatial shifts of up to 11 voxels, a binary dilation of up to 10 voxels, or a binary erosion of up to 6 voxels. These transformation parameters were derived from a study that assessed the inter-observer variability of ischemic lesion segmentations on NCCT [10]. This study reported a mean Dice Similarity Coefficient (DSC) of $72.8 \pm 23.0\%$. Given a median ischemic lesion volume of 28 mL at 24 hours and an average voxel spacing of $0.4 \times 0.4 \times 5$ mm (voxel volume $= 0.8\,\mathrm{mm}^3$), the transformation parameters that produce a segmentation that results in the DSC reported in the study of Cimflova et al. [10] were estimated. The full derivation of these values is provided below (*Calculation of Simulated Mask Transformations*). In addition, for simulated masks of the original segmentation, voxels outside the 20–80 HU range were excluded to remove non-brain tissue and ensure that the simulated segmentation contained only brain tissue.

For the circle annotation, a different approach was used because this annotation type is artificially defined and geometrically constrained. Unlike manual or bounding box segmentations, the circle does not reflect observer-drawn lesion boundaries. Inter-observer variability can only be caused by the placement of the circle. As a result, the inter-observer Dice Similarity Coefficient (DSC) values reported by Cimflova et al. [10] is not applicable. The simulated mask of the circle annotation was created by applying a random shift of up to 12 voxels in the x- and y-directions and 1 slice in the z-direction. To maintain anatomical plausibility, the shifted circle was always constrained to remain entirely within the original ischemic lesion segmentation.

Radiomic features were extracted from all simulated segmentations. The ICC(2,1) was then calculated for each feature using the Pingouin Python package (version 0.5.5) [12]. This model (two-way random effects, absolute agreement, single measurement) assumes that both patients and raters (in this case, segmentation conditions) are random samples. ICC was calculated over 40 patients (i.e., $k = 80$ observations per feature). Features with an ICC value ≥ 0.80 were considered reproducible and retained for further selection steps.

Calculation of Simulated Mask Transformations

Dice Similarity Coefficient (DSC)

The Dice Similarity Coefficient (DSC) is used to quantify the spatial overlap between two segmentations. It is defined as:

$$\text{DSC} = \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

Where:

- A and B are the voxel sets of the original and simulated segmentations,
- $|A|$ and $|B|$ are the number of voxels in each set,
- $|A \cap B|$ is the number of overlapping voxels.

Given parameters:

- Average voxel size: $0.4 \times 0.4 \times 5\,\text{mm}$,
- Average voxel volume: $0.8\,\text{mm}^3 = 0.0008\,\text{mL}$,
- Average DSC [10]: 72.8%.

Bounding Box and Segmentation Characteristics

Based on the dataset:

- Median 24-hour ischemic lesion volume: $28\,\text{mL}$,
- Average bounding box dimensions for a 28 mL lesion volume: $9 \times 50 \times 75\,\text{mm}$.

The number of voxels corresponding to a 28 mL lesion is calculated as:

$$\frac{28}{0.0008} = 35{,}000\,\text{voxels}$$

Assumption

To estimate the effects of spatial transformations, all calculations are performed using the bounding box dimensions. It is assumed that when a given transformation (e.g., shift, dilation, erosion) is applied to the bounding box and results in a specific DSC, the same transformation applied to the actual lesion segmentation yields an identical DSC. This assumption is based on the bounding box fully enclosing the lesion and undergoing the same spatial transformation.

Shift-Based Analysis

When only a spatial shift is applied, the shape and volume of the segmentation remain unchanged. Therefore, the number of voxels in both the original and simulated masks are the same:

$$|A| = |B| = 35{,}000\,\text{voxels}$$

Substituting in Eq. (1):

$$0.728 = \frac{2x}{35{,}000 + 35{,}000} \Rightarrow x = 25{,}480\,\text{voxels}$$

This implies that approximately 25,480 voxels overlap between the original and shifted mask. Assuming the overlapping region results from a symmetric shift in both the x- and y-dimensions within the bounding box (depth $= 9\,\text{mm}$):

$$25{,}480 = 9 \cdot (50 - a) \cdot (75 - a) \Rightarrow a \approx 11\,\text{voxels}$$

This corresponds to a maximum shift of approximately 11 voxels in both spatial directions.

Dilation-Based Analysis

For dilation, the shape is enlarged. Eq. (1) is used to calculate the number of voxels in the dilated mask B:

$$0.728 = \frac{2 \cdot 35{,}000}{35{,}000 + B} \Rightarrow B \approx 61{,}154\,\text{voxels}$$

Assuming the dilated segmentation has the same shape:

$$61{,}154 = 9 \cdot (50 + 2a) \cdot (75 + 2a) \Rightarrow a \approx 10\,\text{voxels}$$

Thus, the dilation corresponds to an expansion of approximately 10 voxels in each in-plane direction.

Erosion-Based Analysis

For erosion, the simulated segmentation is a smaller subset of the original. Eq. (1) is used to calculate the number of voxel in the eroded segmentation B:

$$0.728 = \frac{2 \cdot B}{35{,}000 + B} \Rightarrow B \approx 20{,}031\,\text{voxels}$$

Assuming the eroded segmentation has the same shape:

$$20{,}031 = 9 \cdot (50 - 2a) \cdot (75 - 2a) \Rightarrow a \approx 6\,\text{voxels}$$

This corresponds to a contraction of approximately 6 voxels in each in-plane direction.

Supplementary Note S7: Description of LASSO Regression Selection Step

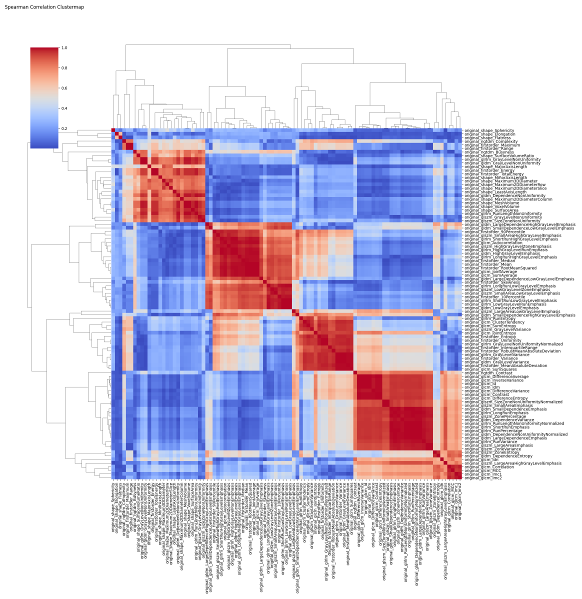
To select the most relevant features for the prediction task, a Least Absolute Shrinkage ans Selection Operator (LASSO) regression selection step was implemented. LassoCV from the scikit-learn package (version 1.3.2) [13] was used to perform a 5-fold cross-validation on the pre-training set and identify the optimal regularization parameter (α) that minimized the mean squared error (MSE). After determining the cross-validated α, the regularization strength was iteratively increased or decreased with a step size, which was reduced by 10% after each iteration. The loop was terminated once the number of features with non-zero coefficients met the predefined feature limit, or if (α) left the valid range ($\alpha < 10^{-6}\,\text{or}\,\alpha > 100$). The LASSO regression step was skipped if the previous selection steps had already reduced the number of features to the allowed maximum or fewer.

Implementation parameters:

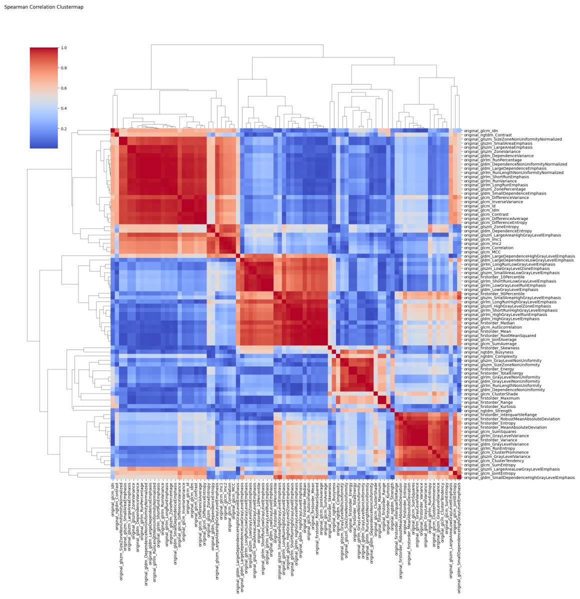
- Cross-validation: 5-fold,
- Maximum iterations: 10,000,
- Random state: 42,
- Initial stepsize: 1.1.

Supplementary Table S8. Parameter Ranges for Hyperparameter Tuning

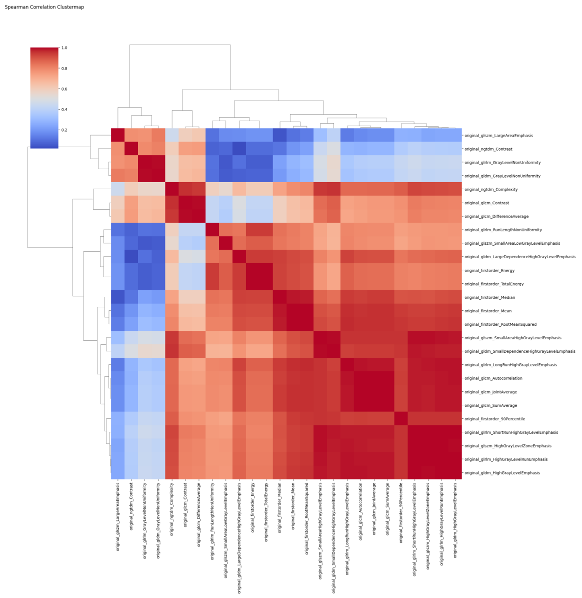
Parameter	Range
n_estimators	[50, 250]
learning_rate	[0.01, 0.20]
gamma	[0, 0.2]
max_depth	[4, 6]
min_child_weight	[1, 4]
subsample	[0.6, 1.0]
colsample_bytree	[0.6, 1.0]
reg_lambda	[0, 2]
reg_alpha	[0, 2]



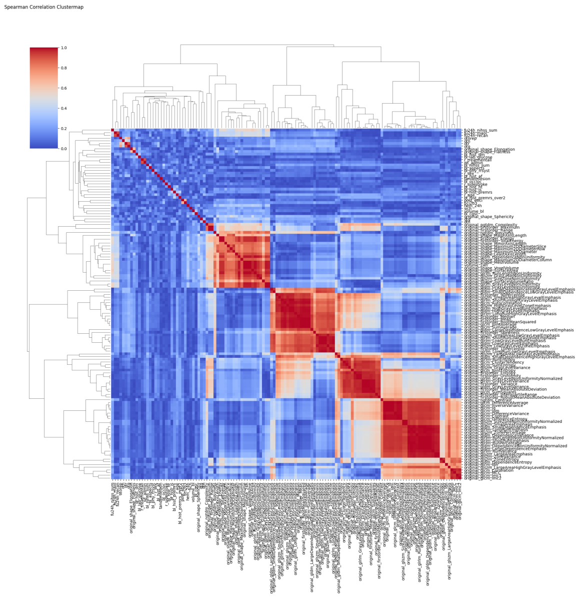
(a)



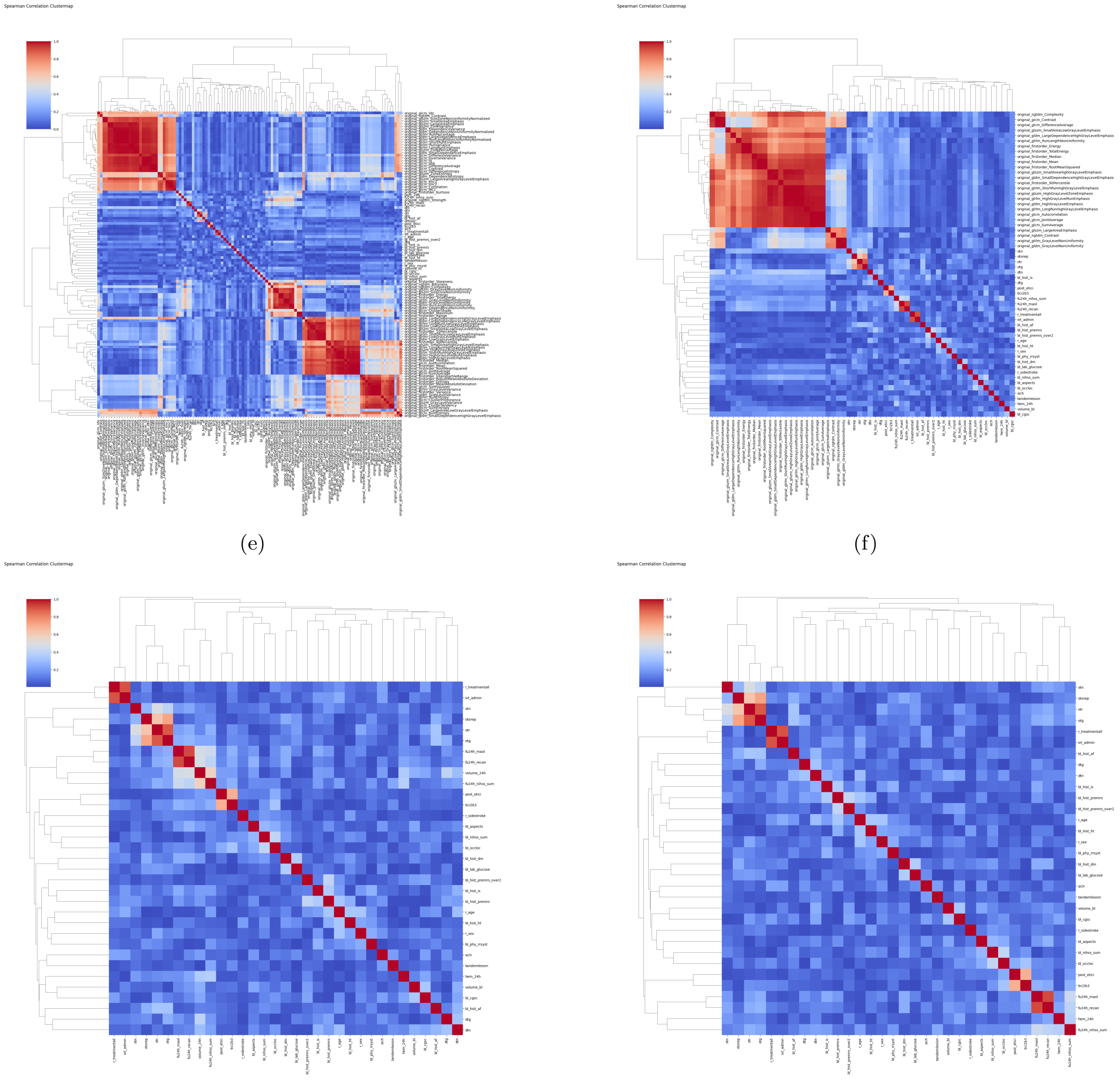
(b)



(c)



(d)



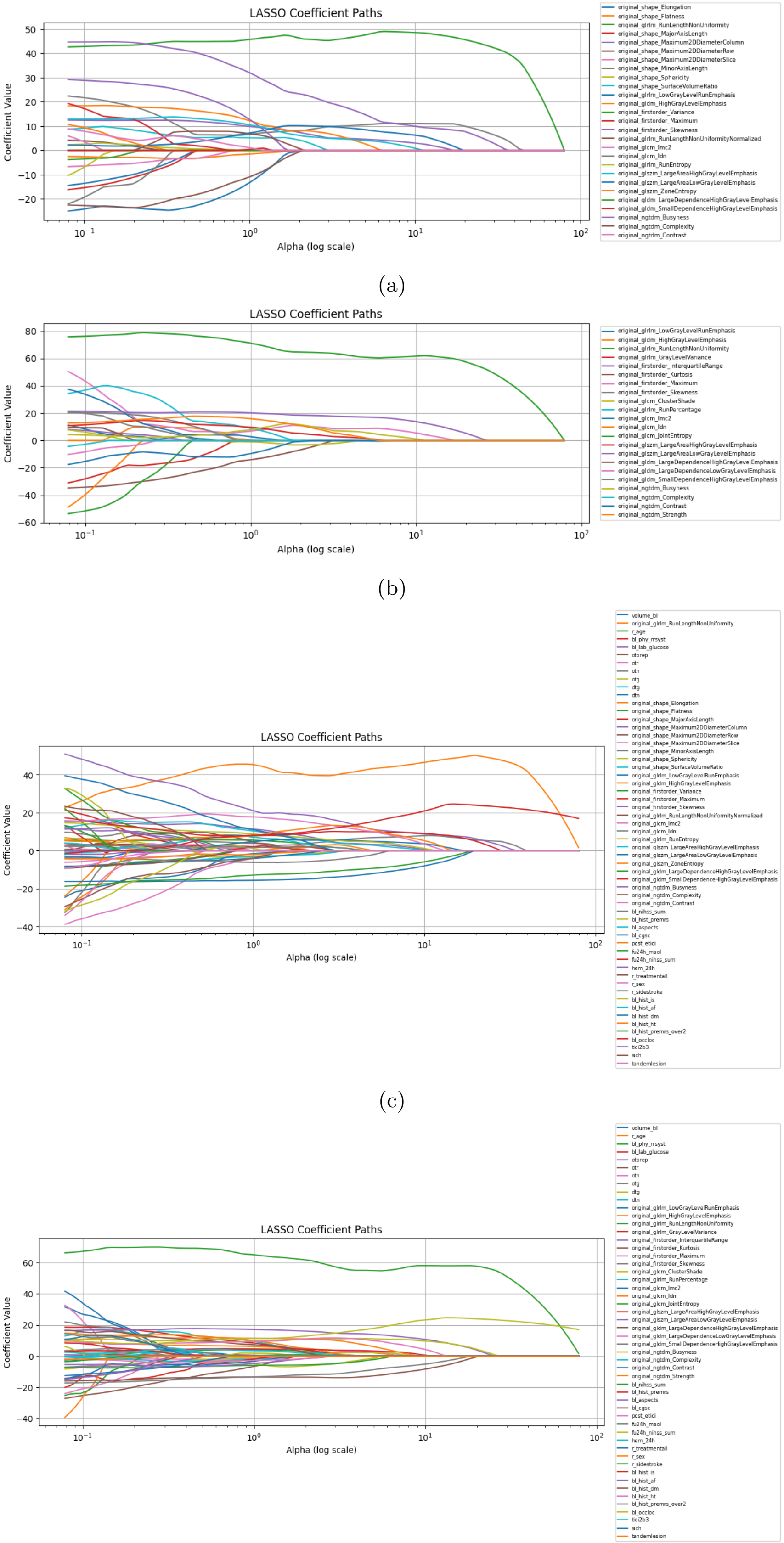
(e)

(f)

(g)

(h)

Supplementary Figure S9. Correlation Matrices for the (a) Original Segmentation Radiomic Feature Set, (b) Bounding Box Radiomic Feature Set, (c) Circle Radiomic Feature Set, (d) Original Segmentation Combined Feature set, (e) Bounding Box Combined Feature Set, (f) Circle Combined Feature Set, (g) Clinical Feature Set including 24-hour Ischemic Lesion Volume, and (h) Clinical Feature Set not including 24-hour Ischemic Lesion Volume.

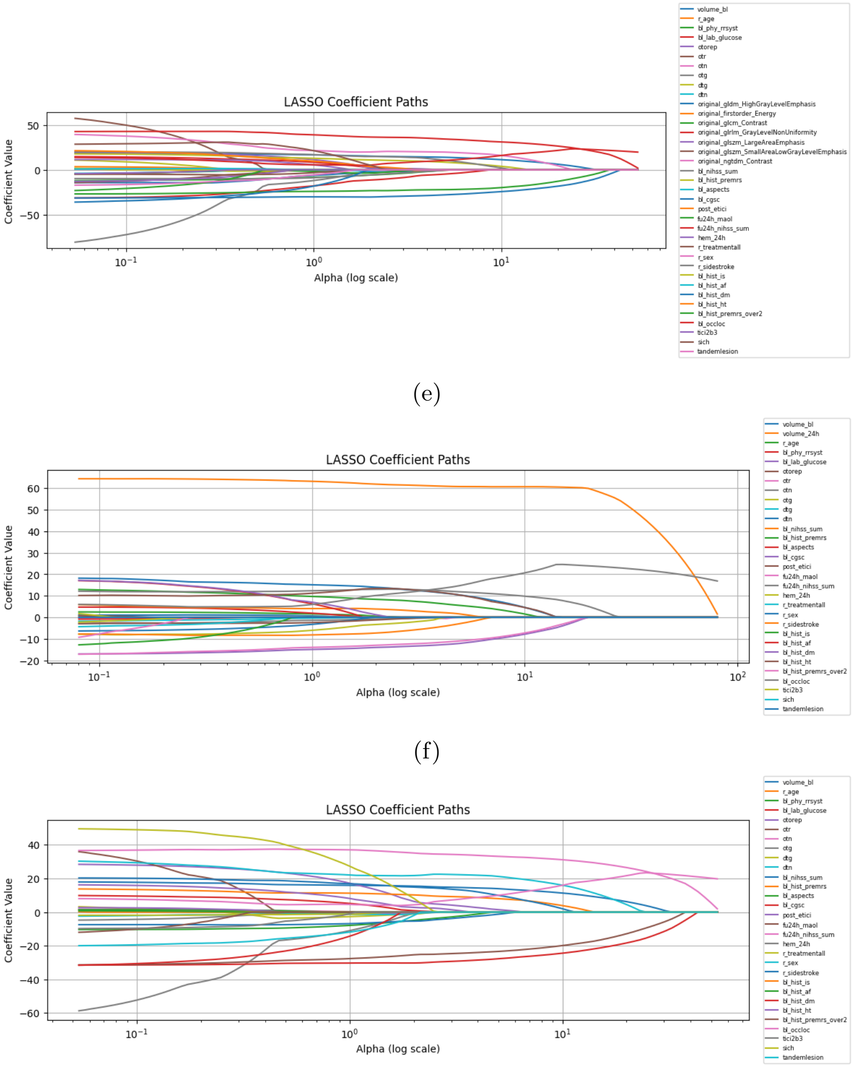

LASSO Coefficient Paths
Coefficient Value
50
40
30
20
10
0
−10
−20
10^{-1}
10^{0}
10^{1}
10^{2}
Alpha (log scale)
original_shape_Elongation
original_shape_Flatness
original_glrlm_RunLengthNonUniformity
original_shape_MajorAxisLength
original_shape_Maximum2DDiameterColumn
original_shape_Maximum2DDiameterRow
original_shape_Maximum2DDiameterSlice
original_shape_MinorAxisLength
original_shape_Sphericity
original_shape_SurfaceVolumeRatio
original_glrlm_LowGrayLevelRunEmphasis
original_gldm_HighGrayLevelEmphasis
original_firstorder_Variance
original_firstorder_Maximum
original_firstorder_Skewness
original_glrlm_RunLengthNonUniformityNormalized
original_glcm_Imc2
original_glcm_Idn
original_glrlm_RunEntropy
original_glszm_LargeAreaHighGrayLevelEmphasis
original_glszm_LargeAreaLowGrayLevelEmphasis
original_glszm_ZoneEntropy
original_gldm_LargeDependenceHighGrayLevelEmphasis
original_gldm_SmallDependenceHighGrayLevelEmphasis
original_ngtdm_Busyness
original_ngtdm_Complexity
original_ngtdm_Contrast
LASSO Coefficient Paths
Coefficient Value
80
60
40
20
0
−20
−40
−60
10^{-1}
10^{0}
10^{1}
10^{2}
Alpha (log scale)
original_glrlm_LowGrayLevelRunEmphasis
original_gldm_HighGrayLevelEmphasis
original_glrlm_RunLengthNonUniformity
original_glrlm_GrayLevelVariance
original_firstorder_InterquartileRange
original_firstorder_Kurtosis
original_firstorder_Maximum
original_firstorder_Skewness
original_glcm_ClusterShade
original_glrlm_RunPercentage
original_glcm_Imc2
original_glcm_Idn
original_glcm_JointEntropy
original_glszm_LargeAreaHighGrayLevelEmphasis
original_glszm_LargeAreaLowGrayLevelEmphasis
original_gldm_LargeDependenceHighGrayLevelEmphasis
original_gldm_LargeDependenceLowGrayLevelEmphasis
original_gldm_SmallDependenceHighGrayLevelEmphasis
original_ngtdm_Busyness
original_ngtdm_Complexity
original_ngtdm_Contrast
original_ngtdm_Strength
LASSO Coefficient Paths
Coefficient Value
40
20
0
−20
−40
10^{-1}
10^{0}
10^{1}
10^{2}
Alpha (log scale)
volume_bl
original_glrlm_RunLengthNonUniformity
r_age
bl_phy_rrsyst
bl_lab_glucose
otorep
otr
otn
otg
dtg
dtn
original_shape_Elongation
original_shape_Flatness
original_shape_MajorAxisLength
original_shape_Maximum2DDiameterColumn
original_shape_Maximum2DDiameterRow
original_shape_Maximum2DDiameterSlice
original_shape_MinorAxisLength
original_shape_Sphericity
original_shape_SurfaceVolumeRatio
original_glrlm_LowGrayLevelRunEmphasis
original_gldm_HighGrayLevelEmphasis
original_firstorder_Variance
original_firstorder_Maximum
original_firstorder_Skewness
original_glrlm_RunLengthNonUniformityNormalized
original_glcm_Imc2
original_glcm_Idn
original_glrlm_RunEntropy
original_glszm_LargeAreaHighGrayLevelEmphasis
original_glszm_LargeAreaLowGrayLevelEmphasis
original_glszm_ZoneEntropy
original_gldm_LargeDependenceHighGrayLevelEmphasis
original_gldm_SmallDependenceHighGrayLevelEmphasis
original_ngtdm_Busyness
original_ngtdm_Complexity
original_ngtdm_Contrast
bl_nihss_sum
bl_hist_premrs
bl_aspects
bl_cgsc
post_etici
fu24h_maol
fu24h_nihss_sum
hem_24h
r_treatmentall
r_sex
r_sidestroke
bl_hist_is
bl_hist_af
bl_hist_dm
bl_hist_ht
bl_hist_premrs_over2
bl_occloc
tici2b3
sich
tandemlesion
LASSO Coefficient Paths
Coefficient Value
60
40
20
0
−20
−40
10^{-1}
10^{0}
10^{1}
10^{2}
Alpha (log scale)
volume_bl
r_age
bl_phy_rrsyst
bl_lab_glucose
otorep
otr
otn
otg
dtg
dtn
original_glrlm_LowGrayLevelRunEmphasis
original_gldm_HighGrayLevelEmphasis
original_glrlm_RunLengthNonUniformity
original_glrlm_GrayLevelVariance
original_firstorder_InterquartileRange
original_firstorder_Kurtosis
original_firstorder_Maximum
original_firstorder_Skewness
original_glcm_ClusterShade
original_glrlm_RunPercentage
original_glcm_Imc2
original_glcm_Idn
original_glcm_JointEntropy
original_glszm_LargeAreaHighGrayLevelEmphasis
original_glszm_LargeAreaLowGrayLevelEmphasis
original_gldm_LargeDependenceHighGrayLevelEmphasis
original_gldm_LargeDependenceLowGrayLevelEmphasis
original_gldm_SmallDependenceHighGrayLevelEmphasis
original_ngtdm_Busyness
original_ngtdm_Complexity
original_ngtdm_Contrast
original_ngtdm_Strength
bl_nihss_sum
bl_hist_premrs
bl_aspects
bl_cgsc
post_etici
fu24h_maol
fu24h_nihss_sum
hem_24h
r_treatmentall
r_sex
r_sidestroke
bl_hist_is
bl_hist_af
bl_hist_dm
bl_hist_ht
bl_hist_premrs_over2
bl_occloc
tici2b3
sich
tandemlesion

(a)

(b)

(c)

(d)

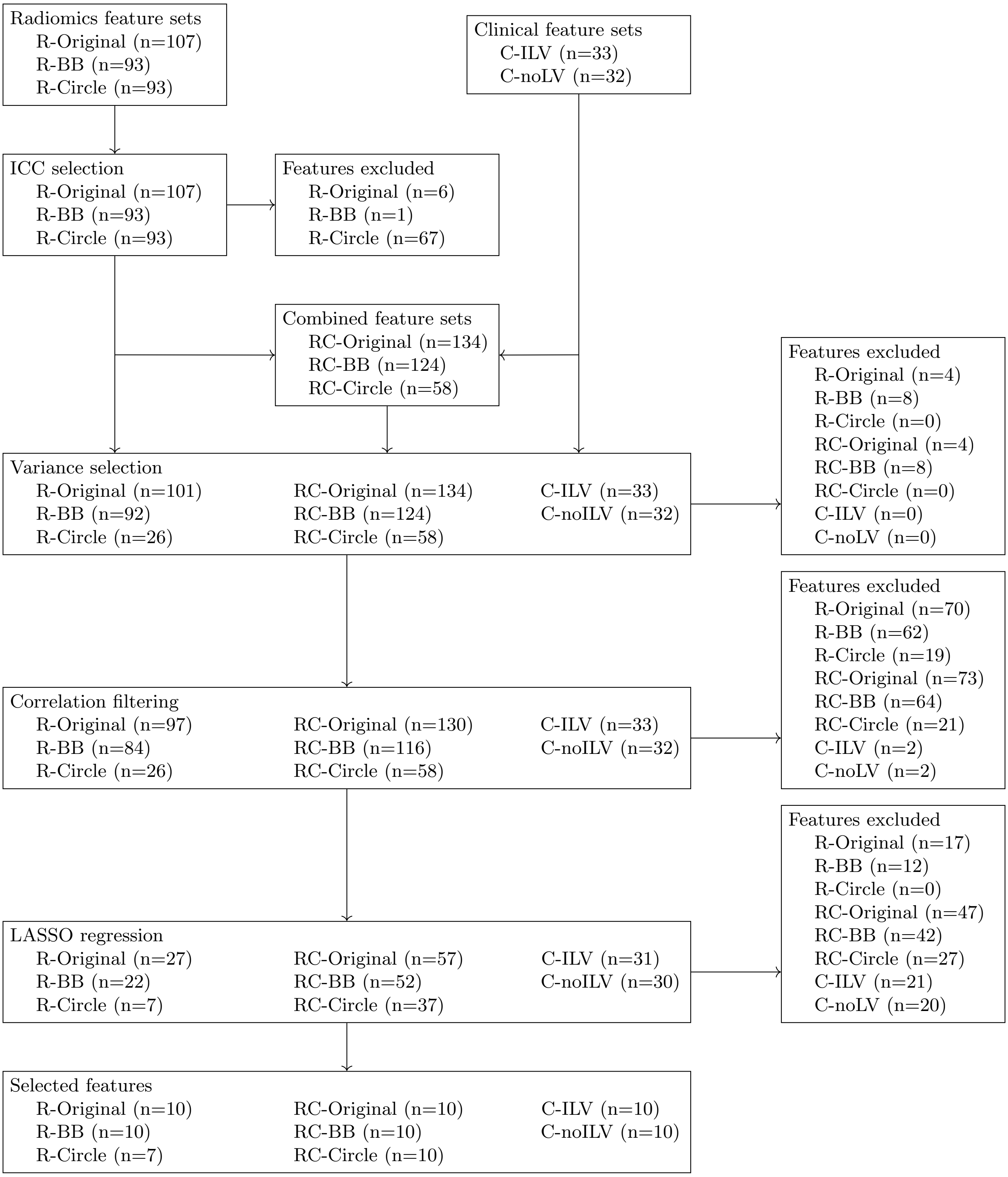


(e)

(f)

(g)

Supplementary Figure S10. Least Absolute Shrinkage and Selection Operator (LASSO) Regression Plots for the (a) Original Segmentation Radiomic Feature Set, (b) Bounding Box Radiomic Feature Set, (c) Original Segmentation Combined Feature set, (d) Bounding Box Combined Feature Set, (e) Circle Combined Feature Set, (g) Clinical Feature Set including 24-hour Ischemic Lesion Volume, and (h) Clinical Feature Set not including 24-hour Ischemic Lesion Volume.



Supplementary Figure S11. Features Excluded and Retained for each Model after each Feature Selection Step.

Supplementary Table S12. Features Selected per Model

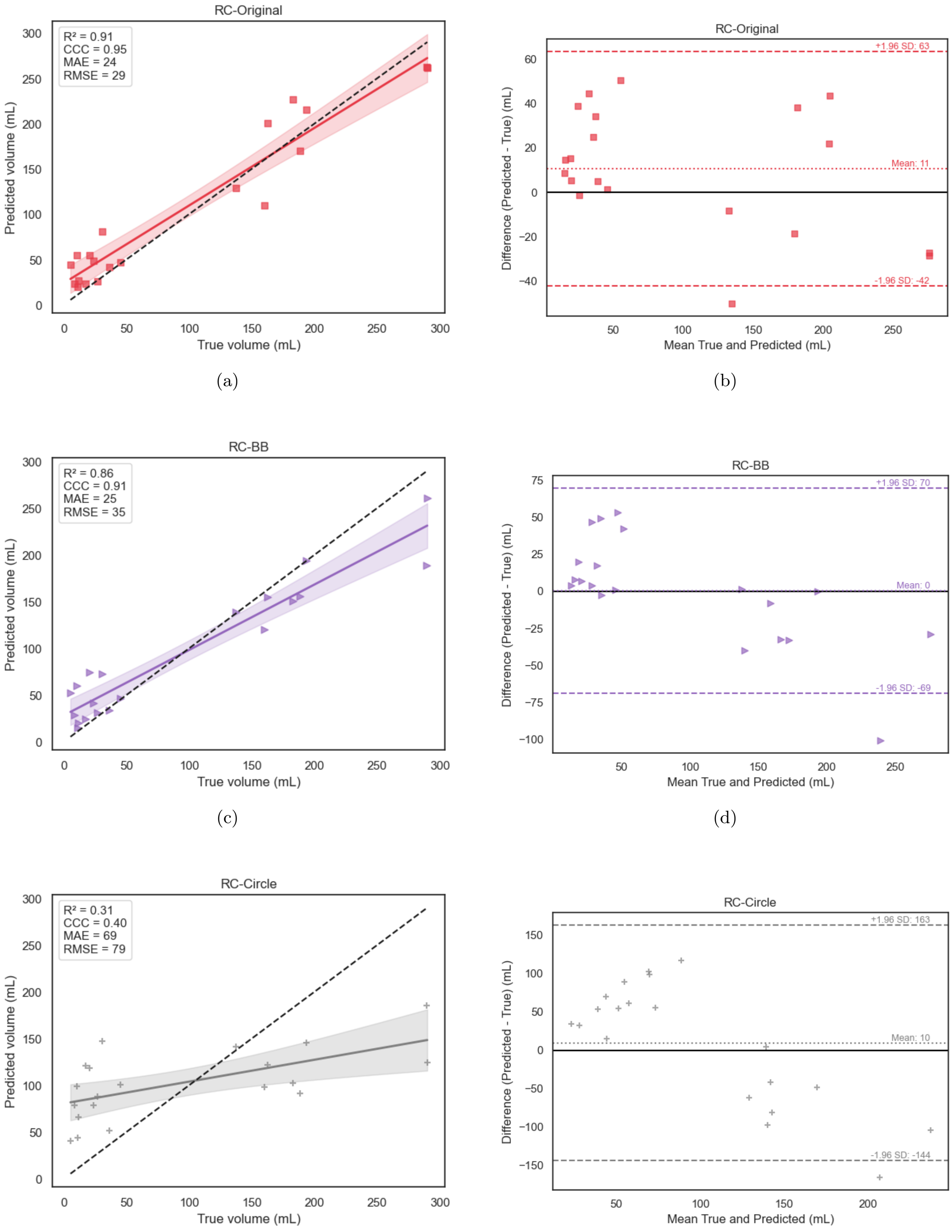
Model	Selected features
R-Original (n=10)	original_shape_Flatness, original_glrlm_RunLengthNonUniformity, original_shape_Maximum2DDiameterColumn, original_shape_MinorAxisLength, original_shape_SurfaceVolumeRatio, original_gldm_HighGrayLevelEmphasis, original_glrlm_RunLengthNonUniformityNormalized, original_glszm_LargeAreaHighGrayLevelEmphasis, original_glszm_LargeAreaLowGrayLevelEmphasis, original_ngtdm_Busyness
R-BB (n=10)	original_glrlm_RunLengthNonUniformity, original_firstorder_Kurtosis, original_firstorder_Maximum, original_glcm_ClusterShade, original_glrlm_RunPercentage, original_glszm_LargeAreaHighGrayLevelEmphasis, original_glszm_LargeAreaLowGrayLevelEmphasis, original_ngtdm_Busyness, original_ngtdm_Contrast, original_ngtdm_Strength
R-Circle(n=7)	original_gldm_HighGrayLevelEmphasis, original_firstorder_Energy, original_glcm_Contrast, original_glrlm_GrayLevelNonUniformity, original_glszm_LargeAreaEmphasis, original_glszm_SmallAreaLowGrayLevelEmphasis, original_ngtdm_Contrast
RC-Original (n=10)	original_glrlm_RunLengthNonUniformity, original_shape_Maximum2DDiameterColumn, original_shape_MinorAxisLength, original_glszm_LargeAreaLowGrayLevelEmphasis, original_ngtdm_Busyness, bl_hist_premrs, bl_cgsc, post_etici, fu24h_maol, fu24h_nihss_sum
RC-BB (n=16)	original_glrlm_RunLengthNonUniformity, original_firstorder_Maximum, original_glszm_LargeAreaLowGrayLevelEmphasis, original_ngtdm_Busyness, bl_nihss_sum, bl_hist_premrs, bl_cgsc, post_etici, fu24h_maol, fu24h_nihss_sum
RC-Circle (n=10)	volume_bl, original_firstorder_Energy, original_glrlm_GrayLevelNonUniformity, original_glszm_SmallAreaLowGrayLevelEmphasis, bl_nihss_sum, bl_hist_premrs, bl_cgsc, fu24h_maol, fu24h_nihss_sum, r_sex
C-ILV (n=10)	volume_24h, otr, otg, bl_nihss_sum, bl_hist_premrs, bl_cgsc, post_etici, fu24h_maol, fu24h_nihss_sum, bl_hist_ht
C-noILV (n=10)	volume_bl, bl_phy_rrsyst, otorep, bl_nihss_sum, bl_hist_premrs, bl_cgsc, fu24h_maol, fu24h_nihss_sum, r_sex, r_sidestroke

Abbreviations: BB: bounding box, C: clinical model, C-ILV: clinical model including the ischemic lesion volume at 24 hours after stroke onset, C-noILV: clinical model not including the ischemic lesion volume at 24 hours after stroke onset, R: radiomics model, RC: combined model.

Supplementary Table S13. Optimal Parameters after Hyperparameter Tuning for each Model

	Radiomics models			Combined models			Clinical models	
Parameter	**Original**	**BB**	**Circle**	**Original**	**BB**	**Circle**	**ILV**	**noILV**
n_estimators	200	196	57	90	196	157	130	186
learning_rate	0.11	0.19	0.14	0.06	0.19	0.03	0.07	0.19
gamma	0.06	0.03	0.05	0.07	0.03	0.05	0.08	0.01
max_depth	5	5	5	5	5	5	4	4
min_child_weight	1	2	1	1	2	3	3	3
subsample	0.68	0.78	0.70	0.61	0.78	0.84	0.97	
colsample_bytree	0.85	0.66	0.69	0.93	0.66	0.99	0.77	0.63
reg_lambda	1.08	0.54	1.82	0.33	0.54	0.62	1.58	0.12
reg_alpha	1.21	1.59	0.07	0.59	1.59	1.18	1.42	1.38

Abbreviations: BB: bounding box, ILV: including 24-hour ischemic lesion volume, noILV: not including 24-hour ischemic lesion volume.


RC-Original
R² = 0.91
CCC = 0.95
MAE = 24
RMSE = 29
Predicted volume (mL)
True volume (mL)
0
50
100
150
200
250
300
RC-Original
+1.96 SD: 63
Mean: 11
-1.96 SD: -42
Difference (Predicted - True) (mL)
Mean True and Predicted (mL)
60
40
20
0
−20
−40
50
100
150
200
250
RC-BB
R² = 0.86
CCC = 0.91
MAE = 25
RMSE = 35
Predicted volume (mL)
True volume (mL)
RC-BB
+1.96 SD: 70
Mean: 0
-1.96 SD: -69
Difference (Predicted - True) (mL)
Mean True and Predicted (mL)
75
50
25
0
−25
−50
−75
−100
RC-Circle
R² = 0.31
CCC = 0.40
MAE = 69
RMSE = 79
Predicted volume (mL)
True volume (mL)
RC-Circle
+1.96 SD: 163
Mean: 10
-1.96 SD: -144
Difference (Predicted - True) (mL)
Mean True and Predicted (mL)
150
100
50
0
−50
−100
−150

(a)

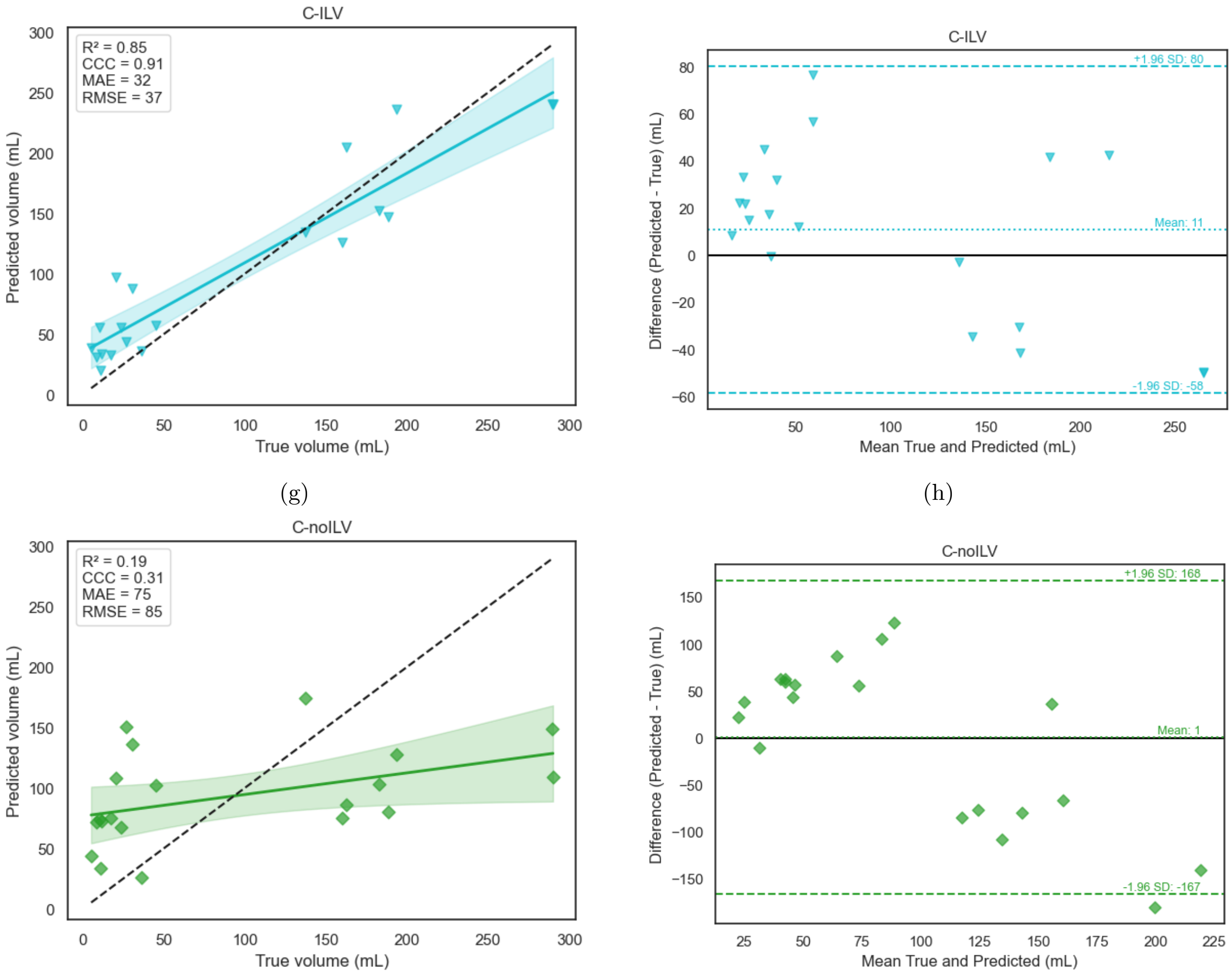
(b)

(c)

(d)

(e)

(f)



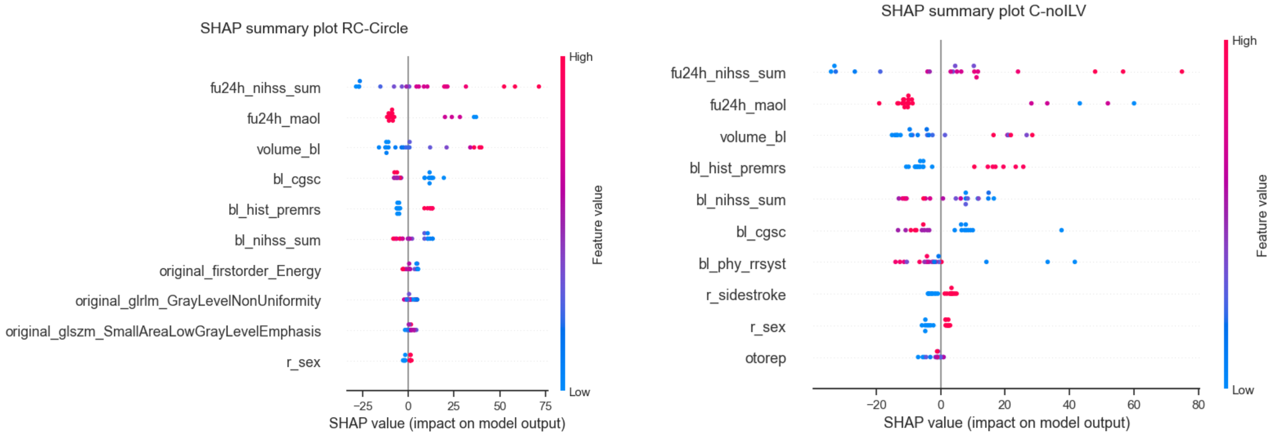
(g)

(h)

(i)

(j)

Supplementary Figure S14. Scatter Plots and Bland-Altman Plots of the Combined Models based on (a,b) the Original Segmentation, (c,d) the Bounding Box Annotation, and (e,f) the Circle Annotation, and the Clinical Models (g,h) with the 24-hour Ischemic Lesion Volume, and (i,j) without.



(a)

(b)

Supplementary Figure S15. The Feature Importance of (a) the Combined Model based on the Circle Annotation, and (b) the Clinical Model without the 24-hour Ischemic Lesion Volume.

References

1. Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, and Cook G. Introduction to radiomics. Journal of Nuclear Medicine 2020; 61:488–95
2. Haralick RM, Shanmugam K, and Dinstein IH. Textural features for image classification. IEEE Transactions on systems, man, and cybernetics 2007 :610–21
3. Galloway MM. Texture analysis using gray level run lengths. Computer graphics and image processing 1975; 4:172–9
4. Thibault G, Fertil B, Navarro C, Pereira S, Cau P, Levy N, Sequeira J, and Mari JL. Shape and texture indexes application to cell nuclei classification. International Journal of Pattern Recognition and Artificial Intelligence 2013; 27:1357002
5. Sun C and Wee WG. Neighboring gray level dependence matrix for texture classification. Computer Vision, Graphics, and Image Processing 1983; 23:341–52
6. Amadasun M and King R. Textural features corresponding to textural properties. IEEE Transactions on systems, man, and Cybernetics 1989; 19:1264–74
7. Ge G, Zhang JZ, and Zhang J. The impact of high-order features on performance of radiomics studies in CT non-small cell lung cancer. Clinical Imaging 2024; 113:110244
8. Buuren S van and Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 2011; 45:1–67
9. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2021. Available from: https://www.R-project.org/
10. Cimflova P, Ospel JM, Marko M, Menon BK, and Qiu W. Variability assessment of manual segmentations of ischemic lesion volume on 24-h non-contrast CT. Neuroradiology 2022 :1–9
11. Haarburger C, Müller-Franzes G, Weninger L, Kuhl C, Truhn D, and Merhof D. Radiomics feature reproducibility under inter-rater variability in segmentations of CT images. Scientific reports 2020; 10:12688
12. Vallat R. Pingouin: statistics in Python. Journal of Open Source Software 2018 Nov; 3:1026. DOI: 10.21105/joss.01026
13. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, and Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011; 12:2825–30