



OPEN ACCESS

EDITED BY

Daniel Donoho,
Children's National Hospital, United States

REVIEWED BY

Matthias Wilms,
University of Calgary, Canada
Jesse Klostranec,
McGill University Health Centre, Canada
Alex Buioite Stella,
University of Trieste, Italy

*CORRESPONDENCE

Freda Werdiger
✉ freda.werdiger@unimelb.edu.au

SPECIALTY SECTION

This article was submitted to
Stroke,
a section of the journal
Frontiers in Neurology

RECEIVED 20 November 2022

ACCEPTED 02 February 2023

PUBLISHED 23 February 2023

CITATION

Werdiger F, Parsons MW, Visser M, Levi C,
Spratt N, Kleinig T, Lin L and Bivard A (2023)
Machine learning segmentation of core and
penumbra from acute stroke CT perfusion data.
Front. Neurol. 14:1098562.
doi: 10.3389/fneur.2023.1098562

COPYRIGHT

© 2023 Werdiger, Parsons, Visser, Levi, Spratt,
Kleinig, Lin and Bivard. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).
The use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in this
journal is cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Machine learning segmentation of core and penumbra from acute stroke CT perfusion data

Freda Werdiger^{1,2*}, Mark W. Parsons^{3,4,5}, Milanka Visser^{1,2},
Christopher Levi^{6,7}, Neil Spratt^{6,7}, Tim Kleinig⁸, Longting Lin^{6,7} and
Andrew Bivard^{1,2}

¹Melbourne Brain Centre, Department of Neurology, The Royal Melbourne Hospital, Melbourne, VIC, Australia, ²Department of Medicine, University of Melbourne, Melbourne, VIC, Australia, ³Southwestern Sydney Clinical School, University of New South Wales, Sydney, NSW, Australia, ⁴Department of Neurology, Liverpool Hospital, Liverpool, NSW, Australia, ⁵Ingham Institute for Applied Medical Research, Liverpool, NSW, Australia, ⁶Hunter Medical Research Institution, University of Newcastle, Newcastle, NSW, Australia, ⁷Department of Neurology, John Hunter Hospital, University of Newcastle, Newcastle, NSW, Australia, ⁸Department of Neurology, Royal Adelaide Hospital, Adelaide, SA, Australia

Introduction: Computed tomography perfusion (CTP) imaging is widely used in cases of suspected acute ischemic stroke to positively identify ischemia and assess suitability for treatment through identification of reversible and irreversible tissue injury. Traditionally, this has been done *via* setting single perfusion thresholds on two or four CTP parameter maps. We present an alternative model for the estimation of tissue fate using multiple perfusion measures simultaneously.

Methods: We used machine learning (ML) models based on four different algorithms, combining four CTP measures (cerebral blood flow, cerebral blood volume, mean transit time and delay time) plus 3D-neighborhood (patch) analysis to predict the acute ischemic core and perfusion lesion volumes. The model was developed using 86 patient images, and then tested further on 22 images.

Results: XGBoost was the highest-performing algorithm. With standard threshold-based core and penumbra measures as the reference, the model demonstrated moderate agreement in segmenting core and penumbra on test images. Dice similarity coefficients for core and penumbra were 0.38 ± 0.26 and 0.50 ± 0.21 , respectively, demonstrating moderate agreement. Skull-related image artefacts contributed to lower accuracy.

Discussion: Further development may enable us to move beyond the current overly simplistic core and penumbra definitions using single thresholds where a single error or artefact may lead to substantial error.

KEYWORDS

acute ischemic stroke, CT perfusion imaging, machine learning, ischemic core, penumbra

1. Introduction

Rapid diagnosis of acute ischemic stroke is of vital importance and is confirmed by computed tomography (CT) or magnetic resonance (MR) imaging. Historically improved patient outcomes were obtained by early reperfusion treatment, with significant effort and resources being provided to improve both stroke detection and clinical workflows to facilitate faster treatment (1–3). Recently, clinical trials have demonstrated that patients with a favorable perfusion imaging profile benefit from treatment up to 9 h from symptom onset/mid-point of wake-up with thrombolysis and up to 24 h with thrombectomy (4–7). Perfusion imaging allows estimation of salvageable brain tissue (penumbra) and tissue

already infarcted or destined for infarction irrespective of reperfusion (ischemic core) (4, 7–11). Patient outcomes have been shown to be strongly related to the estimated volume of ischemic core at baseline (12, 13). As a result, CT perfusion (CTP) is increasingly being used in clinical practice around the world, with several software providing automated estimates of salvageable and ischemic core derived through various mathematical models (hemodynamic maps) (14, 15).

The hemodynamic maps generated by CTP are obtained by tracking a contrast medium as it flows into and out of the brain. The data is then processed using one of several different algorithms (14, 15). The estimation of salvageable tissue and ischemic core is then performed by applying a single threshold to one or two maps (9, 16, 17). However, there is significant variation between algorithms used when estimating tissue perfusion, and single-value thresholds have been shown to both under and overestimate the size of the infarct core and penumbra (18, 19). This may be partly due to the misclassification of image voxels as core or penumbra that results from single-value thresholding of core and penumbra. More sophisticated methods of processing CTP maps are required that can, for example, delineate artifactual signals from those caused by perfusion deficit.

The currently used perfusion thresholds have been validated to some degree and have shown success in selecting patients for treatment through clinical trials (6). However, a predictive model that uses all available perfusion data and spatial context of voxels may provide a more nuanced representation of the pathophysiology of evolving ischemic stroke, improving the accuracy of the images and the robustness of the output. Furthermore, shifting from a rigid single threshold model to a trained Machine Learning (ML) model is highly advantageous as the ML model may continue to improve performance with the addition of data.

There are many studies that develop and test ML and Deep Learning (DL) models for lesion segmentation and there have been great advances in developing applications of ML and DL to healthcare in general [e.g., (20, 21)]. However, there are challenges in widespread deployment such as lack of standardized methods to evaluate performance. Furthermore, the inner mathematical processes of ML and DL are often difficult to understand, and their outputs difficult to interpret. These issues of “explainability” and “interpretability” lead to ML being approached as a “black box” problem, without understanding of internal mechanisms. This has hampered implementation into medical practice. It is therefore essential to integrate ML in small, explainable steps rather than large, black-box overhauls that will result in issues of reliability (22). In this study we investigate if single-value thresholds for measurement of ischemic core and penumbra can be replaced with a ML-based method. We also outline challenges that must be addressed for successful integration into acute stroke assessment protocols.

2. Materials and methods

We developed an early ML model that is trained to delineate both ischemic core and penumbra from surrounding tissue using acute CTP data. We used retrospective data from an acute ischemic stroke patient cohort to develop models based on four ML algorithms (Logistic regression, Random Forest, XGBoost and

Support Vector Machine). We tested performance of the model on an additional set of new, unseen patient data.

2.1. Data acquisition

We analyzed CTP images from the International Stroke Perfusion Imaging Registry (INSPIRE), which is a database of acute stroke perfusion imaging and associated clinical information. For this study we used consecutive patients presenting with acute ischemic stroke who had whole brain CTP and who were recruited into INSPIRE between 2010 and 2017 at the John Hunter Hospital, Newcastle, Australia. For standardization, only one site was used at this stage. As is routine in INSPIRE, patients all underwent baseline multimodal CT imaging with non-contrast CT, CTA, and CTP. Written informed consent was obtained from all participants, and the INSPIRE study was approved by the site’s ethics committee (23).

To obtain the perfusion images, a total of 19 acquisitions occurred over 60 s. The CTP data were processed by commercial software MISTar (Apollo Medical Imaging Technology, Melbourne, VIC, Australia). CTP parameters were generated by applying the mathematical algorithm of singular value decomposition with delay and dispersion correction (24). The following four CTP parameters were generated: cerebral blood flow (CBF), cerebral blood volume (CBV), mean transit time (MTT), and delay time (DT). The penumbra and core volumes were defined with dual thresholds: DT at the threshold of 3 s for total ischemic lesion volume and CBF at the threshold setting of 30% for acute core volume (8, 16, 25). After single-value thresholding, core/penumbra areas were limited to a single lesion and artifactual or erroneous regions were removed. The resulting map was used as the ground truth (GT). Core/penumbra were reviewed by experts to ensure they were accurate.

To develop the model, we used 86 acute ischemic stroke patients with a large vessel occlusion (LVO): M1 segment of the middle cerebral artery (MCA) or internal carotid artery (ICA). To provide additional testing and external validation, 25 patients were used, with both LVO and non-LVO occlusions. This was done to observe whether a model trained only on lesions resulting from an occlusion of large vessel will perform as well when testing on a variety of occlusion sites. Each patient in the test set underwent follow-up MR diffusion-weighted imaging (DWI) between 24 and 72 h after onset. The volume (mL) of the infarct core, as estimated by MR-DWI, was recorded and used for external validation. On follow-up imaging, all patients had a thrombolysis in cerebral infarction (TICI) score of at least 2b, indicating relatively complete reperfusion of initially hypoperfused regions. In these cases, the volume of the acute CTP core should more closely match that of the follow-up infarct core and could therefore be used to validate the predictions.

2.2. Creating labeled data

2.2.1. Class labels

The four hemodynamic maps (hereafter referred to as *features*) and core-penumbra segmentation maps (hereafter referred to as *lesion map*) were used in the development of the algorithm. The

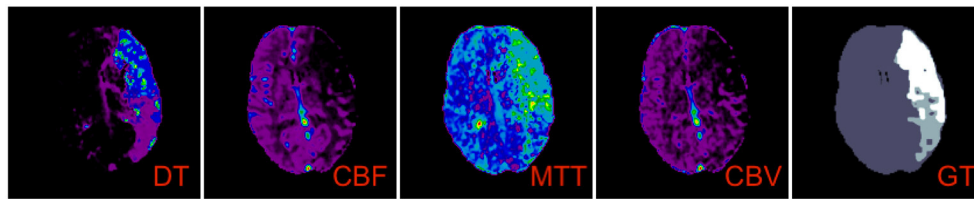


FIGURE 1 Feature maps and lesion map corresponding to the M1 test image. A single axial slice is shown with corresponding perfusion data for delay time (DT), cerebral blood flow (CBF), mean transit time (MTT) and cerebral blood volume (CBV). The corresponding class labels which make up the lesion map, used as ground truth (GT) in the algorithm, is shown on the far right.

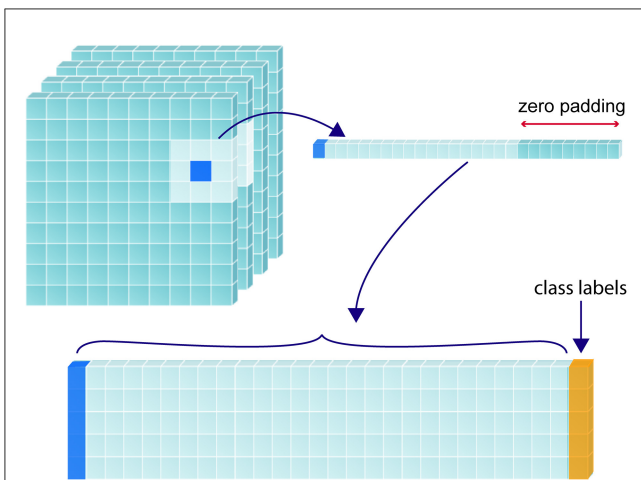


FIGURE 2 Construction of training matrix through sampling and patch extraction. For a given randomly selected sample (shown in dark blue), its corresponding perfusion map value/s and the values corresponding to its 26 immediate neighbors are collapsed into a 1-dimensional array, with the corresponding class label (yellow) added at the furthestmost right position. If multiple perfusion maps are used, the 27 values from each map are recursively added to extend the 1-D array to the left of the class label. The 1-D array for each label are stacked to form a 2-D training matrix.

lesion map, together with the spatial coordinates of the mean baseline image from the CTP acquisition, was used to create a 3-D array of tissue class labels, where each voxel was one of four values: 0—background; 1—non-ischemic brain tissue; 2—penumbra; 3—core). Figure 1 shows the features alongside their class label array for a single patient.

2.2.2. Under-sampling

For this early model, we avoided the issue of class imbalance by sampling the same number of voxels from each class in each image. We processed all lesion maps in the training data, counting the number of voxels belonging to each class. The smallest core volume contained 708 voxels and the smallest penumbra volume contained 8,436 voxels, and two images in the group had a penumbra but no core. We then randomly sampled 300 voxels from each class in each image. For the two images with no core, 300 extra healthy

tissue samples were randomly taken from the image, ensuring 1,200 voxels were sampled from each feature channel.

2.2.3. Patch analysis

To predict the tissue status of a single sample (i.e., voxel of interest), we included the feature values associated with the coordinates of that voxel as well as the values associated with every direct neighboring voxel (26 in total), creating a patch-wise analysis. This was done to include spatial context in the determination of sample tissue status. Zero padding was used for samples that lay around the edges of the image. Figure 2 demonstrates this process for a single voxel of interest, where a 1-D array is created from the sample and its neighbors. Each sample resides in a single row of the training matrix, alongside its class label. All feature channels are concatenated along the same row.

2.3. Machine learning models

The sampled training data was further split into training (60%) and validation (40%) cohorts. Optimization and training were performed on the training data and evaluation was performed on the validation data. All data was standardized to $[-1, 1]$ using the Standard Scalar function in Scikit-Learn in Python (v 0.0) (26).

We used Scikit-learn to optimize four models, based on logistic regression (LR), random forest (RF), XGBoost (XGB) and support vector machine (SVM), respectively. Except for SVM, a randomized search was initially performed, to estimate the best hyperparameters for each algorithm, after which a grid search was performed to narrow down the best hyperparameters. The chosen range for each hyperparameter was determined based on recommendations in Scikit-learn documentation. For each unique parameter combination, three-fold cross validation was performed.

2.4. Impact of added features

For this early model, we wished to determine whether performance was enhanced by including all four CTP maps vs. CBF and Delay Time alone. In particular, we wish to learn whether using four maps reduced the presence of artifactual perfusion lesions. Therefore, each model was trained twice; first with data only from

CBF and Delay Time and then on data from CBF, Delay Time, MTT and CBV.

2.5. Performance evaluation

All the data used to train and optimize the model comprised random samples from images. However, the model will ultimately be used to process whole patient images and provide a prediction that can be displayed as an image. Therefore, we used an additional 25 whole brain patient images to further test the model's performance as it would be applied in a clinical scenario, and to provide a visualization of the model's accuracy. The images were processed as follows: from each voxel in the image, a 3D neighborhood patch was extracted and added to a matrix as in [Figure 2](#). Each 3D patch from the image was forwarded through the model, and the resulting predictions were accumulated in a common space, preserving their spatial location and allowing the image to be reconstructed.

2.5.1. Quantitative performance evaluation

The predictive model was trained using random samples, evenly distributed among the classes. For the test images, however, classes were severely imbalanced. Using receiver-operating characteristics (ROC) or average accuracy would favor the majority class and it is the minority classes that are of interest in this case. Furthermore, the area under the ROC curve (AUC) metric rewards positively predicted background pixels. Therefore, it is not a fair representation of the accuracy of a brain lesion segmentation, whereby background pixels constitute much of the image. For this reason, it was more appropriate to choose a metric more in line with perceptual quality, which reflects both size and localization agreement.

The Dice similarity coefficient (DSC) is a measure of spatial overlap for two regions (A, B), and is given by $DSC(A, B) = 2(A \cap B) / (A + B)$, where \cap is the intersection. It can be seen as the percentage overlap between A and B. A perfect intersection between A and B will give a DSC of 1, and if there is no intersection between the two regions, the score is 0. DSC is sensitive to both size and location differences and is a highly intuitive manner of expressing similarity between two regions. We calculated the DSC between the ground truth and predicted images for the core and penumbra regions separately. After (27), DSC can be separated in a similar manner to the Kappa coefficient for agreement, into the following six categories (28, 29): 0, "No Agreement"; 0–0.2, "Slight agreement"; 0.2–0.4, "Fair agreement"; 0.4–0.6, "Moderate agreement"; 0.6–0.8, "Substantial agreement"; "0.8–1"; "Almost perfect agreement".

The Jaccard Index (JI), also known as the Intersection of Union (IoU), like the DSC, ranges from 0 (no agreement) to 1 (perfect agreement). The JI is mathematically represented by $IoU(A, B) = A \cap B / A \cup B$, where \cup is the union. The relationship between JI and DSC can therefore be described as $JI = DSC / (2 - DSC)$. The DSC tends to be higher as it counts the true positive classifications twice in both the numerator and denominator of its equation, while the JI gives a greater penalty for bad classifications. Therefore,

providing an average score over a set of classification will lead the average DSC and average JI to diverge from one another. The two metrics will always be positively correlated, however, we found it worthwhile to analyse the distinction as both are used throughout literature to evaluate segmentation tasks. The DSC and JI values for each the core and penumbra were calculated for all 25 images, and the differences between them were evaluated using paired *t*-tests.

Finally, lesion volume, one of the most important predictors of outcome after ischemic stroke, was calculated for the additional test images. The volumes of the core and penumbra were calculated for each of the ground truth and the predicted lesion by counting the number of voxels assigned to each area (30). Using pixel information encoded in the image, the absolute volume in milliliters could be calculated. As an external validation, the predicted core volume was compared with the follow-up (24–72h) infarct core derived from MR-DWI imaging and reviewed by the expert stroke neurologist (MP).

2.5.2. Qualitative performance evaluation

We identified eleven images within the cohort affected by artifacts relating to the skull. In brain CT imaging, beam hardening from the dense skull region or, to a lesser degree, contrast-enhanced arteries, may result in a characteristic "streaking" artifact (31). When the skull, a highly attenuating region is adjacent to less attenuating tissue, such as soft tissue, and there is limited CT resolution, partial volume averaging may also occur. Here, the image intensity of affected voxels is a mixture, or an average, or the intensity of both these regions (32). [Figure 3](#) shows an example of the partial volume artifact in Subject 3. Upon CTP processing, such voxels near the edge of the brain shows increased Delay Time. However, these artifacts are common and, if the image is otherwise of good quality, artifactual perfusion lesions are easy to identify to the trained eye. Therefore, we did not exclude these cases from the study and instead prefer to investigate the impact of artifact on model performance. We qualitatively compared the ability of the algorithms to make a correct prediction around those areas, based on both the inclusion of all four CTP maps and the additional spatial information provided by the 3D patches.

3. Results

For the training set, 55 patients had an occlusion of the M1 segment of the middle cerebral artery (MCA), and 31 had an occlusion of the internal carotid artery (ICA). Forty-three patients were female (50%), and the median onset age was 74 (IQR 63–82). The median baseline NIHSS (National Institutes of Health Stroke Scale) was 17 (IQR 14–20). Of these, 70 patients had a known time of onset; the median time between onset and CT imaging was 121 min (IQR 95–157). One patient had a wake-up stroke, and 15 patients had an unknown time of onset. Seventy-six patients received intravenous (IV) thrombolysis, one received intraarterial (IA) thrombectomy, two received both, five received no treatment and two patients did not have any treatment documented.

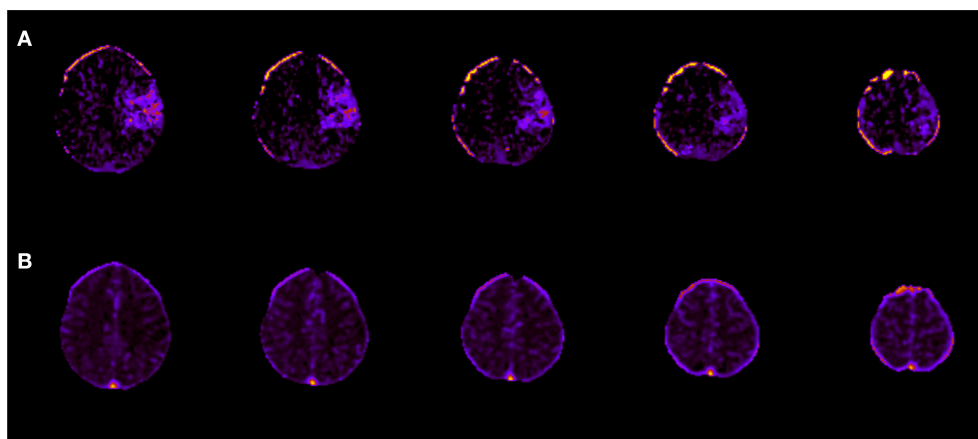


FIGURE 3

Skull artifacts. For subject 3, skull artifacts can be seen in their (A) Delay time and (B) CBF maps near the top of the skull.

TABLE 1 Class representations across the training and validation cohorts.

	Background	Non-ischemic brain	Core	Penumbra
Train	15,485	15,870	15,141	15,424
Validation	10,315	10,530	10,059	10,376
Total	25,800	26,400	25,200	25,800

Samples were split into these categories using Scikit-Learn (26).

This was done to ensure the model did not bias any class.

Three patients were discarded from the test set due to considerable infarct growth. For the remaining patients in the test set, 16 patients had an M2-MCA occlusion, four had an M3-MCA occlusion, and one each with an occlusion of the anterior cerebral artery (ACA) and ICA. Thirteen patients (59%) were female, and the median onset age was 79 (IQR 74–83). The median baseline NIHSS was 11 (IQR 6–16). In total, 20 patients had a known time of onset; the median time between onset and CT imaging for these patients was 110 min (IQR 96–168). The remaining patients had an unknown time of onset. Of all the patients in the test set, 20 received IV treatment, one received IA treatment and one received no treatment. Sixteen were given a TIC1 3 score, and 6 were given a TIC1 2b score. The median day of DWI image after stroke onset was 1 (IQR 1–2, min-max 0–12). The median size of the follow-up DWI core was as 10 mL (IQR 6–33). A Pearson correlation test shown a strong correlation ($p < 0.005$, two tailed) between the data used for the ground truth core measurement and the expert assessed MR-DWI measurements for core volume.

For model development, a total of 103,200 patch samples was used. Table 1 shows the class instances for the train and validation groups used to develop the model.

Table 2 shows details of optimizing each model. Each model was trained using six computer processing units (CPU) in parallel. For SVM, only a random search for the two-map model was carried due to the excessive training times (>22 h), and only

polynomial and linear kernels were tested, with the polynomial kernel outperforming the linear kernel. Table 3 shows results for each model on the under-sampled data. XGBoost was the highest performing algorithm, and there was an improvement in performance when all four CTP maps were included.

The performance of the best performing model (shown in bold in Table 3) was tested on the remaining 22 images in the test set. The results are shown in Supplementary Table A1. Figure 4 shows axial slices of lesion predictions (overlayed on non-contrast CT image slices) using the model based on all four CTP maps for a selection of datasets (subjects 7, 8, and 1 with reference to Supplementary Table A1).

For all 22 patients, the mean DSC values for core and penumbra were 0.39 (SD 0.26) and 0.50 (SD 0.22), respectively, and the mean JI values for core and penumbra were 0.28 (SD 0.23) and 0.36 (SD 0.20), respectively. For both core and penumbra, JI and DSC were significantly different across the dataset (core: paired t -test, $p < 0.0001$; penumbra: paired t -test, $p < 0.0001$).

To explore the difference between performance on core and penumbra, a volume analysis was performed. Each similarity measured varied significantly with volume: A Pearson's correlation for DSC variation with volume showed ($r = 0.56$, $p = 0.0065$) for penumbra and ($r = 0.71$, $p = 0.0002$) for core. For JI a Pearson's correlation calculation showed ($r = 0.61$, $p = 0.0028$) for penumbra and ($r = 0.72$, $p < 0.0002$) for core.

Out of the 22 testing images, 16 lesions were due to an occlusion of the M2 segment of the MCA. The DSC scores for core and penumbra averaged to 0.34 (SD 0.23) and 0.50 ± 0.20 , respectively. The mean volume of core and penumbra for M2 lesions was 9.89 mL (SD 8.17) and 38.34 mL (SD 22.5), respectively, lower as compared with the entire testing set.

There was no significant correlation between the XGB-predicted core and the 24 h DWI infarct core (Pearson's r ; $r = 0.18$, $p = 0.41$). However, visual inspection confirmed that artifacts due to the skull were present in half the cases ($n = 11$) and led to overestimation of perfusion regions. When considering the test cases with no obvious skull artifacts, there was a significant correlation between the predicted core and the follow-up DWI

TABLE 2 Details of model training.

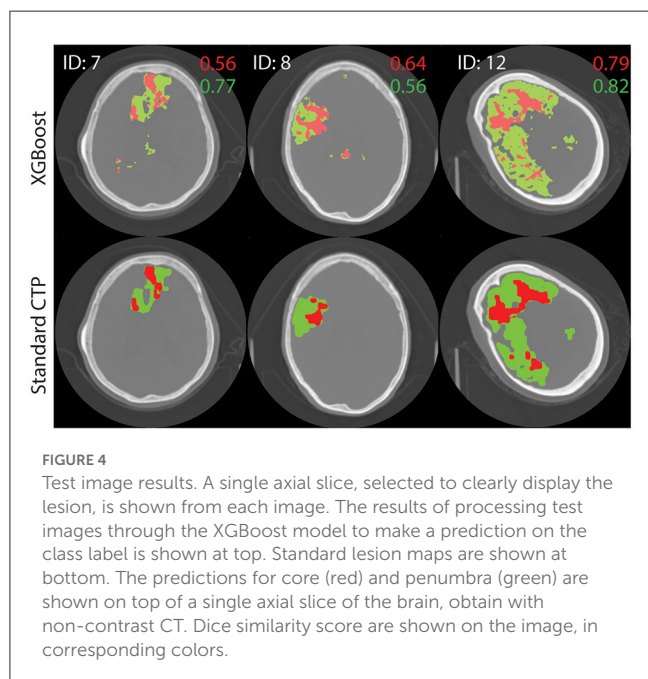
Algorithm	# Parameters optimized	# Candidates in random search	Time taken (2 map, 4 map)	# Candidates in grid search	Time taken (2 map, 4 map)
LR	6	28	5 min, 33 min	3	20 min, 53 min
RF	7	80	1 h 53 min, 3 h 1min	81	4 h 26 min, 6 58 min
XGB	5	10	30 min, 57 min	27	1 h 38 min, 2 h 53 min
SVM	3	30	22 h 33 min, N/A	N/A	N/A

Four different algorithms were used to train models: Logistic Regression (LR), Random Forest (RF), XGBoost (XGB) and Support Vector Machine (SVM). Each algorithm has different hyperparameters, and the number of different hyperparameters that were optimized here is shown (“#Parameters optimized”). Except for SVM, the parameters were optimized by first running a random search, training and testing models with a number of different random hyperparameter models (“#Candidates in random search”). The best performing combination was used to create the range for a more refined grid search. The number of candidates tested in the grid search was determined by the number of parameters that were optimized and the possible values for each parameter (e.g., whether values were discrete or continuous). The time that was taken to run all the different combinations is included. Six CPUs were used in parallel.

TABLE 3 Results of models on validation data.

	ROC-AUC	DSC (core)	DSC (pen)	Jl (core)	Jl (pen)
LR	0.9757	0.8438	0.7874	0.7298	0.6494
	0.9776	0.848	0.7907	0.736	0.6538
RF	0.9825	0.8553	0.8172	0.7471	0.6908
	0.9841	0.8611	0.8269	0.7561	0.7048
XGB	0.983	0.8552	0.8185	0.7470	0.6927
	0.9844	0.8610	0.8275	0.7559	0.7057
SVM	0.9799	0.8467	0.8081	0.7341	0.678

Eight models in total were optimized. Three algorithms (Logistic Regression/LR, Random Forest/RF, XGBoost/XGB) were trained twice, one on Cerebral Blood Flow (CBF) and Delay Time (DT) data (top), and once on data from CBF, DT, Cerebral Blood Flow and Mean Transit Time (bottom). Support Vector Machine (SVM) was only trained for two maps due to excessive training times. Results on the validation data are shown for each model. The highest performance across all categories was obtained for XGB, trained on all four CTP maps.



core (Pearson’s r ; $r = 0.82$, $p = 0.0018$). Figure 5 shows a comparison of results from each algorithm for the subject shown in Figure 3. This subject had significant CTP artifacts due to the skull. While LR could not distinguish actual from artifactual

perfusion lesions, in this case all the other algorithms were able to.

4. Discussion

This study proposes a machine learning algorithm using the entire perfusion map datasets as an alternative to measuring the penumbra and ischemic core using binary thresholds with CTP data. Models based on four different well-known ML algorithms were tested. Accuracy was tested both quantitatively, using similarity measurements, and qualitatively, by using visual inspected to determine which algorithm was better at prediction on artifactual CTP hyperintensities. Simple neighborhood analysis was used to make a prediction on a single voxel; all surrounding voxels were considered. Our model may easily be expanded to include additional input channels, such as non-contrast CT, or relevant clinical information such as time-from-onset, blood pressure, clinical severity measurements and age.

Out of the four algorithms tested, XGBoost performed best in the quantitatively analysis, achieving good accuracy in mimicking the CTP perfusion lesions derived by the clinically used software MIStar. There was an improvement in performance when all four CTP maps were used compared to only CBF and DT for this early model. Future versions of the model will continue to use all four CTP maps to make a prediction.

Ideally, an automated CTP algorithm should differentiate between genuine and artifactual hypoperfusion patterns, just as an

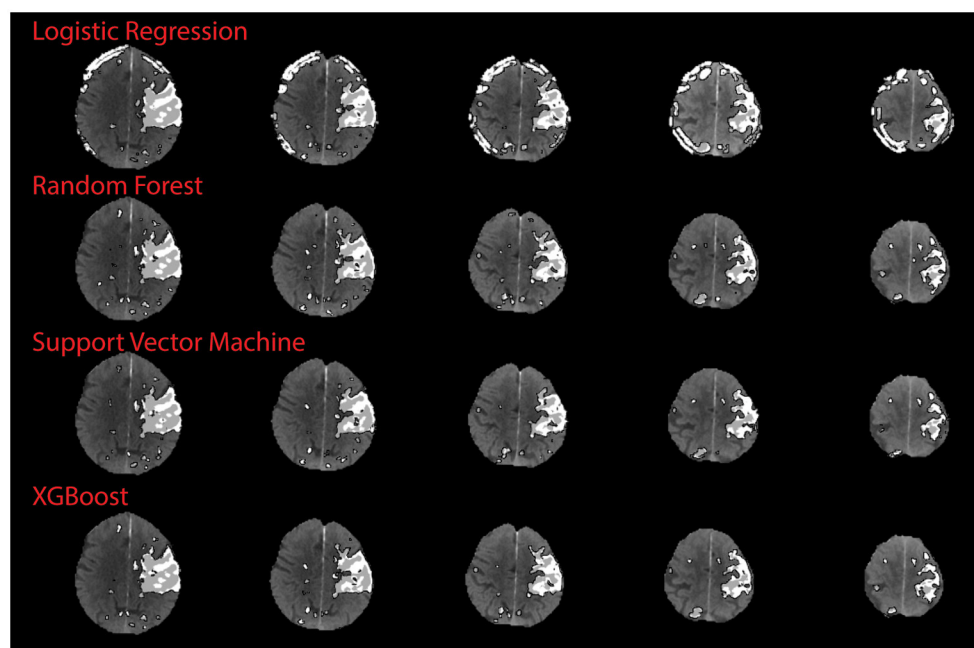


FIGURE 5
Qualitative results on skull artifact. Predictions made on subject 3 (shown in Figure 3) for each algorithm. Linear regression performed the poorest in terms of identifying artifact as areas of perfusion, as demonstration by subject 3 axial slices.

experienced stroke physician should be able to determine whether the pattern is topographically consistent with stroke phenotype (33, 34). For the qualitative study, SVM, XGB and RF improved on the ability of the LR algorithm to distinguish real from artifactual CTP hyperintensities. This is because LR is the only algorithm based on linear first-order interactions between variables, whereas the other three are more sophisticated, and able to model non-linear and higher interactions. This is shown clearly in Figure 5, although in other cases the ML model still derived artifact in making a prediction, leading to an overall worse correlation with the DWI infarct core for images with obvious artifact. As CT artifacts are difficult to avoid altogether in a clinical setting, this is a useful insight. Further development is required to ensure future versions of this model to not derive artifactual perfusion lesions.

For the testing images, DSC and JI scores were shown to vary significantly even though they are both commonly used similarity metrics. In addition, both metrics varied significantly with volume. Therefore, the DSC or JI score for a large lesion may not represent the same accuracy as for a small lesion, even though (27) has proposed otherwise (35, 36). For example, the large core in Figure 4 (ID = 12, M1) receives an almost perfect DSC value, while the smaller cores received lower DSC scores; these differences may have resulted merely from size differences. The same behavior was seen with JI. An average DSC or JI score that is a result of the summing over results from lesions of different sizes will not be an accurate representation of the overall performance of a model. We propose a weighted mean DSC/JI to account for size variation before these scores can be fully interpretable. Further studies will explore the application of a weighted mean. In lieu of a robust and subjective

model performance metric, benchmark data [ISLES 2018 (37)] will be used in future studies to report performance.

The most significant limitation to this study is that, as a first step, we have used the CTP core and penumbra estimations derived by MISTar as the ground truth, even though these only approximate the ground truth. The gold standard in the determination of acute ischemic tissue is an expertly segmented MR-DWI lesion, either with (core) or without reperfusion (penumbra) (9, 38). Without a perfect ground truth, it remains difficult to interpret model performance in an objective fashion. For example, as MISTar maps are based on a simple thresholding method, a meaningful comparison of this ML method to a thresholding method against MISTar maps is challenging. Although MISTar and other software (39) CTP core estimates have been shown to be a fair approximation of DWI lesions previously, there are certainly ongoing issues (18), one of which is that the reference standard for core is imperfect (16). Nonetheless, future studies will adopt manually segmented DWI images as ground truth and therefore be able to provide performance metrics that are more robust and interpretable. In addition, this model uses derived perfusion parameters rather than raw CTP time series images, risking a loss of valuable information contained in the raw images which may be lost in the derivation process. The model uses a simple approach over more advanced approaches that have been tested in the literature, such as those based on Deep Learning. With DL, features may be automatically extracted from images, both locally and globally, to make predictions with efficiency (40, 41). Future models will adopt DL, however, the current analysis using more explainable algorithms, was a necessary first step.

With this study we have shown that a Machine Learning method is capable of mimicking common use perfusion lesion measurements to a high accuracy. With the increasing prevalence of CTP assessments for treatment selection of ischemic stroke patients, particularly in the extended time window, it is vital that the measurements be accurate and representative of the underlying pathophysiology. There is significant scope for the current single threshold methods to overestimate the ischemic core depending on onset to reperfusion speed, and other factors. The proposed model may prove more accurate with further development than the currently used single threshold maps and can consider physiologically relevant information such as blood pressure, cardiac output and fluid status which would influence contrast flow and hence perfusion measures on the CTP. Imaging metadata such as time may also influence accuracy, as “ghost cores” have been noted in the hyperacute phase (42).

While the model is simple in its current form, we were able to demonstrate salient points about CTP-based predictions of stroke infarct. We have demonstrated that similarity indices such as DSC and JI have some difficulty in interpretations and further development of performance metrics is required. We have also demonstrated that non-linear algorithms are more adept at making predictions on common CT artifacts that linear model such as logistic regression. Further studies will use manually segmented DWI volumes as ground truth, as well as digest raw CTP data rather than post-processed CTP maps for Deep Learning predictions. Benchmark datasets will be used to measure performance. In addition, the role of clinical data and imaging metadata will be explored in making predictions.

5. Conclusion

We have described a Machine Learning model for the delineation of ischemic tissue from CTP data which is based on the XGBoost algorithm combined with 3D neighborhood analysis. The model is trained on lesion segmentations derived by clinically used software and can derive perfusion lesions to high accuracy. The model improves on clinically available software in that it is able to use multiple input channels but is currently limited by the lack of validation against gold standard lesion segmentations. Nonetheless, the model allowed us to demonstrate useful insight

into CTP-based prediction of stroke infarct which will be used to make future developments.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: De-identified data from the current study are available for qualified investigators upon reasonable request to the corresponding author. Requests to access these datasets should be directed to freda.werdiger@unimelb.edu.au.

Author contributions

FW performed conception of the work, data analysis and interpretation, and drafting of the article. AB, MP, MV, TK, CL, NS, and LL provided critical revision of the article and final approval of the version to be published. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fneur.2023.1098562/full#supplementary-material>

References

- Goyal M, Menon BK, van Zwam WH, Dippel DWJ, Mitchell PJ, Demchuk AM, et al. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. *Lancet*. (2016) 387:1723–31. doi: 10.1016/S0140-6736(16)00163-X
- Jovin TG, Chamorro A, Cobo E, de Miquel MA, Molina CA, Rovira A, et al. Thrombectomy within 8 hours after symptom onset in ischemic stroke. *N Engl J Med*. (2015) 372:2296–306. doi: 10.1056/NEJMoa1503780
- Saver JL, Goyal M, van der Lugt A, Menon BK, Majoie CBLM, Dippel DW, et al. Time to treatment with endovascular thrombectomy and outcomes from ischemic stroke: a meta-analysis. *JAMA*. (2016) 316:1279. doi: 10.1001/jama.2016.13647
- Albers GW, Marks MP, Kemp S, Christensen S, Tsai JP, Ortega-Gutierrez S, et al. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. *N Engl J Med*. (2018) 378:708–18. doi: 10.1056/NEJMoa1713973
- Campbell BCV, Ma H, Ringleb PA, Parsons MW, Churilov L, Bendzus M, et al. Extending thrombolysis to 4.5–9 h and wake-up stroke using perfusion imaging: a systematic review and meta-analysis of individual patient data. *Lancet*. (2019) 394:139–47. doi: 10.1016/S0140-6736(19)31053-0
- Ma H, Campbell BCV, Parsons MW, Churilov L, Levi CR, Hsu C, et al. Thrombolysis guided by perfusion imaging up to 9 hours after onset of stroke. *N Engl J Med*. (2019) 380:1795–803. doi: 10.1056/NEJMoa1813046

7. Nogueira RG, Jadhav AP, Haussen DC, Bonafe A, Budzik RF, Bhuva P, et al. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. *N Engl J Med.* (2018) 378:11–21. doi: 10.1056/NEJMoa1706442
8. Campbell BCV, Christensen S, Levi CR, Desmond PM, Donnan GA, Davis SM, et al. Cerebral blood flow is the optimal CT perfusion parameter for assessing infarct core. *Stroke.* (2011) 42:3435–40. doi: 10.1161/STROKEAHA.111.618355
9. Bivard A, Levi C, Spratt N, Parsons M. Perfusion CT in acute stroke: a comprehensive analysis of infarct and penumbra. *Radiology.* (2013) 267:543–50. doi: 10.1148/radiol.12120971
10. Campbell BCV, Mitchell PJ, Kleinig TJ, Dewey HM, Churilov L, Yassi N, et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. *N Engl J Med.* (2015) 372:1009–18. doi: 10.1056/NEJMoa1414792
11. Bivard A, Spratt N, Levi C, Parsons M. Perfusion computer tomography: imaging and clinical validation in acute ischaemic stroke. *Brain.* (2011) 134:3408–16. doi: 10.1093/brain/awr257
12. Albers GW, Goyal M, Jahan R, Bonafe A, Diener H-C, Levy EI, et al. Ischemic core and hypoperfusion volumes predict infarct size in SWIFT PRIME: CT perfusion volumes. *Ann Neurol.* (2016) 79:76–89. doi: 10.1002/ana.24543
13. Lansberg MG, Christensen S, Kemp S, Mlynash M, Mishra N, Federau C, et al. Computed tomographic perfusion to Predict Response to Recanalization in ischemic stroke: results of the CRISP study. *Ann Neurol.* (2017) 81:849–56. doi: 10.1002/ana.24953
14. Kudo K, Sasaki M, Yamada K, Momoshima S, Utsunomiya H, Shirato H, et al. Differences in CT perfusion maps generated by different commercial software: quantitative analysis by using identical source data of acute stroke patients. *Radiology.* (2010) 254:200–9. doi: 10.1148/radiol.254082000
15. Zussman BM, Boghosian G, Gorniak RJ, Olszewski ME, Read KM, Siddiqui KM, et al. The relative effect of vendor variability in CT perfusion results: a method comparison study. *Am J Roentgenol.* (2011) 197:468–73. doi: 10.2214/AJR.10.6058
16. Lin L, Bivard A, Krishnamurthy V, Levi CR, Parsons MW. Whole-brain CT perfusion to quantify acute ischemic penumbra and core. *Radiology.* (2016) 279:876–87. doi: 10.1148/radiol.2015150319
17. McVerry F, Dani KA, MacDougall NJJ, MacLeod MJ, Wardlaw J, Muir KW. Derivation and evaluation of thresholds for core and tissue at risk of infarction using CT perfusion: CT perfusion thresholds for core and tissue at risk. *J Neuroimaging.* (2014) 24:562–8. doi: 10.1111/jon.12134
18. Parsons MW. Automated measurement of computed tomography acute ischemic core in stroke: does the emperor have no clothes? *Stroke.* (2021) 52:642–4. doi: 10.1161/STROKEAHA.120.032998
19. Bivard A, Churilov L, Ma H, Levi C, Campbell B, Yassi N, et al. Does variability in automated perfusion software outputs for acute ischemic stroke matter? Reanalysis of EXTEND perfusion imaging CNS. *Neurosci Ther.* (2022) 28:139–44. doi: 10.1111/cns.13756
20. Anand VK, Khened M, Alex V, Krishnamurthy G. Fully automatic segmentation for ischemic stroke using CT perfusion maps. In: Crimi A, Bakas S, Kuijff H, Keyvan F, Reyes M, van Walsum T, editors. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Lecture Notes in Computer Science. Cham: Springer International Publishing. (2019). p. 328–334. doi: 10.1007/978-3-030-11723-8_33
21. Zhang J, Shi F, Chen L, Xue Z, Zhang L, Qian D. Ischemic Stroke Segmentation from CT Perfusion Scans Using Cluster-Representation Learning. In: Kia SM, Mohy-ud-Din H, Abdulkadir A, Bass C, Habes M, Rondina JM, Tax C, Wang H, Wolfers T, Rathore S, et al., editors. *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*. Cham: Springer International Publishing (2020). p. 67–76
22. Gupta R, Krishnam SP, Schaefer PW, Lev MH, Gonzalez RG. An east coast perspective on artificial intelligence and machine learning. *Neuroimaging Clin N Am.* (2020) 30:467–78. doi: 10.1016/j.nic.2020.08.002
23. Gao L, Tan E, Moodie M, Parsons M, Spratt NJ, Levi C, et al. Reduced impact of endovascular thrombectomy on disability in real-world practice, relative to randomized controlled trial evidence in Australia. *Front Neurol.* (2020) 11:593238. doi: 10.3389/fneur.2020.593238
24. Lin L, Bivard A, Kleinig T, Spratt NJ, Levi CR, Yang Q, et al. Correction for delay and dispersion results in more accurate cerebral blood flow ischemic core measurement in acute stroke. *Stroke.* (2018) 49:924–30. doi: 10.1161/STROKEAHA.117.019562
25. Bivard A, Levi C, Krishnamurthy V, McElduff P, Miteff F, Spratt NJ, et al. Perfusion computed tomography to assist decision making for stroke thrombolysis. *Brain.* (2015) 138:1919–31. doi: 10.1093/brain/awv071
26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *J Mach Learn Res.* (2011) 12:2825–30. doi: 10.48550/arXiv.1201.0490
27. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Trans Med Imaging.* (1994) 13:716–24. doi: 10.1109/42.363096
28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* (1977) 33:159. doi: 10.2307/2529310
29. Pajula J, Kauppi J-P, Tohka J. Inter-subject correlation in fMRI: method validation against stimulus-model based analysis. *PLoS ONE.* (2012) 8:e41196. doi: 10.1371/journal.pone.0041196
30. Hope TMH, Seghier ML, Leff AP, Price CJ. Predicting outcome and recovery after stroke with lesions extracted from MRI images. *NeuroImage Clin.* (2013) 2:424–33. doi: 10.1016/j.nicl.2013.03.005
31. Barrett JF, Keat N. Artifacts in CT: Recognition and avoidance. *RadioGraphics.* (2004) 24:1679–91. doi: 10.1148/rg.246045065
32. Coolens C, Mohseni H, Dhodi S, Ma S, Keller H, Jaffray DA. Quantification accuracy for dynamic contrast enhanced (DCE) CT imaging: phantom and quality assurance framework. *Eur J Radiol.* (2018) 106:192–8. doi: 10.1016/j.ejrad.2018.08.003
33. Kaw F, Heit JJ, Martin BW, van Ommen F, Kappelle LJ, Velthuis BK, et al. Computed tomography perfusion data for acute ischemic stroke evaluation using rapid software: pitfalls of automated postprocessing. *J Comput Assist Tomogr.* (2020) 44:75–7. doi: 10.1097/RCT.0000000000000946
34. Siegler JE, Olsen A, Pulst-Korenberg J, Cristancho D, Rosenberg J, Raab L, et al. Multicenter volumetric assessment of artifactual hypoperfusion patterns using automated CT perfusion imaging. *J Neuroimaging.* (2019) 29:573–9. doi: 10.1111/jon.12641
35. Cheng D-C, Chi J-H, Yang S-N, Liu S-H. Organ contouring for lung cancer patients with a seed generation scheme and random walks. *Sensors.* (2020) 20:4823. doi: 10.3390/s20174823
36. Ghaffari M, Sanchez L, Xu G, Alaraj A, Zhou XJ, Charbel FT, et al. Validation of parametric mesh generation for subject-specific cerebroarterial trees using modified Hausdorff distance metrics. *Comput Biol Med.* (2018) 100:209–20. doi: 10.1016/j.combiomed.2018.07.004
37. Hakim A, Christensen S, Winzeck S, Lansberg MG, Parsons MW, Lucas C, et al. Predicting infarct core from computed tomography perfusion in acute ischemia with machine learning: lessons from the ISLES challenge. *Stroke.* (2021) 52:2328–37. doi: 10.1161/STROKEAHA.120.030696
38. Bivard A, Kleinig T, Miteff F, Butcher K, Lin L, Levi C, et al. Ischemic core thresholds change with time to reperfusion: a case control study. *Ann Neurol.* (2017) 82:995–1003. doi: 10.1002/ana.25109
39. Cereda CW, Christensen S, Campbell BC, Mishra NK, Mlynash M, Levi C, et al. A benchmarking tool to evaluate computer tomography perfusion infarct core predictions against a DWI standard. *J Cereb Blood Flow Metab.* (2016) 36:1780–9. doi: 10.1177/0271678X15610586
40. Robben D, Boers AMM, Marquering HA, Langezaal LLCM, Roos YBWEM, van Oostenbrugge RJ, et al. Prediction of final infarct volume from native CT perfusion and treatment parameters using deep learning. *Med Image Anal.* (2020) 59:101589. doi: 10.1016/j.media.2019.101589
41. Amador K, Wilms M, Winder A, Fiehler J, Forkert ND. Predicting treatment-specific lesion outcomes in acute ischemic stroke from 4D CT perfusion imaging using spatio-temporal convolutional neural networks. *Med Image Anal.* (2022) 82:102610. doi: 10.1016/j.media.2022.102610
42. Boned S, Padroni M, Rubiera M, Tomasello A, Coscojuela P, Romero N, et al. Admission CT perfusion may overestimate initial infarct core: the ghost infarct core concept. *J NeuroInterventional Surg.* (2017) 9:66–9. doi: 10.1136/neurintsurg-2016-012494