Radiomics machine learning study with small sample size: single random training-test set split may result in unreliable results

Chansik An

National Health Insurance Service Ilsan Hospital <https://orcid.org/0000-0002-0484-6658>

Yae Won Park

Yonsei University College of Medicine

Sung Soo Ahn (✉ SUNGSOO@yuhs.ac)

Yonsei University College of Medicine

Kyunghwa Han

Yonsei University College of Medicine

Hwiyoung Kim

Yonsei University College of Medicine

Seung-Koo Lee

Yonsei University College of Medicine

Research Article

Keywords: Artificial intelligence, Machine learning, Magnetic resonance imaging, Radiomics

Posted Date: December 4th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-105766/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Objective: This study aims to determine how randomly splitting a dataset into training and test sets affects the estimated performance of a machine learning model under different conditions, using real-world brain tumor radiomics data.

Materials and Methods: We conducted two classification tasks of different difficulty levels with magnetic resonance imaging (MRI) radiomics features: (1) “Simple” task, glioblastomas [n=109] vs. brain metastasis [n=58] and (2) “difficult” task, low- [n=163] vs. high-grade [n=95] meningiomas. Additionally, two undersampled datasets were created by randomly sampling 50% from these datasets. We performed random training-test set splitting for each dataset repeatedly to create 1,000 different training and test set pairs. For each dataset pair, the least absolute shrinkage and selection operator model was trained by five-fold cross-validation (CV) or nested CV with or without repetitions in the training set and tested with the test set, using the area under the curve (AUC) as an evaluation metric.

Results: The AUCs in CV and testing varied widely based on data composition, especially with the undersampled datasets and the difficult task. The mean ($\pm$ standard deviation) AUC difference between CV and testing was 0.029 ($\pm$ 0.022) for the simple task without undersampling and 0.108 ($\pm$ 0.079) for the difficult task with undersampling. In a training-test set pair, the AUC was high in CV but much lower in testing (0.840 and 0.650, respectively); in another dataset pair with the same task, however, the AUC was low in CV but much higher in testing (0.702 and 0.836, respectively). None of the CV methods helped overcome this issue.

Conclusions: Machine learning after a single random training-test set split may lead to unreliable results in radiomics studies, especially when the sample size is small.

Introduction

Since the advent of precision and personalized medicine, machine learning (ML) has received great interest as a promising tool for identifying the best diagnosis and treatment for an individual patient. ML research has expanded rapidly in various fields, including radiomics—a method to uncover disease characteristics using a large number of features extracted from medical images (1). According to the PubMed database with the search term “(machine learning OR deep learning) AND radiomics,” the number of published papers per year was 2 in 2015 and increased to 308 in 2019 (2).

Millions of observations are often required for an ML model to be robust and reach acceptable performance levels (3). However, radiomics-based ML studies are often conducted with limited samples, especially when rare diseases are involved. Nonetheless, many published papers have reported fair or good diagnostic performances. However, reproducibility is a big concern in ML, an intensely debated topic for the past few decades. ML algorithms have many parameters to train or set manually, and their training often involves randomness; all these pose unique challenges to reproducibility (4–6). Therefore,

caution should be taken when interpreting a reported model performance, especially when there is a lack of external validation, as it may be overly optimistic (7, 8).

The random split of a dataset into training and test sets may arguably be the most commonly used method in the absence of independent external validation data. However, we have observed that results often vary significantly based on how a single source dataset is split into training and test sets, especially with small sample size and difficult tasks. Previous studies have also suggested that the ML performance based on single random training-test set splitting may be over-optimistic (9–11). However, the impact of randomly splitting data on the reliability of model performance estimation in radiomics ML studies has not been investigated.

This study aims to determine how splitting a dataset randomly into training and test sets affects the estimated performance of an ML model using real-world brain tumor radiomics data under different conditions—the number of input features, sample size, and task difficulty.

Materials And Methods

Subjects

This retrospective study was approved by the Institutional Review Board, and informed consent from the patients was waived.

Two classification tasks of different difficulty levels were performed using radiomics features extracted from brain magnetic resonance imaging (MRI). The first task was a “simple” task of differentiating between glioblastoma (GBM) and single metastasis; the accuracy of radiomics-based ML for this task has been reported to be up to 90% (12, 13). In the current study, the first task dataset consisted of 167 adult patients with pathologically confirmed single GBM (n=109) or brain metastasis (n=58) following brain MRI between January 2014 and December 2017. The second task was a “difficult” task of differentiating between low- and high-grade meningioma, for which reported accuracies are <76% with conventional MRI (14, 15). The second task dataset consisted of 258 adult patients with a low- (n=163) or high-grade (n=95) meningiomas diagnosed between February 2008 and September 2018. Both the datasets were taken from the same tertiary academic hospital; some subsets of these patients were used in previous studies (12, 16). Additionally, two undersampled datasets were created by randomly sampling 50% of the datasets to determine the effect of sample size. Therefore, our final study subjects consisted of the four datasets: 1) simple task (GBM vs. metastasis) without undersampling, 2) simple task with undersampling, 3) difficult task (low- vs. high-grade meningioma) without undersampling, and 4) difficult task with undersampling.

Figure 1 shows the flow of this study. MRI acquisition, image preprocessing, and radiomics feature extractions are described in Supplementary Appendices 1 and 2, and Supplementary Table 1.

Random training-test set split

Each dataset was randomly split into training and test sets with a 7:3 ratio while maintaining the proportions of the outcome classes. This random split was repeated for each dataset while changing the random seed number from 0 to 999 to create 1,000 training-test set pairs with different data compositions.

Feature selection

Radiomics features were repeatedly selected from the 1,000 different training sets based on the coefficient or feature importance employing four ML models: Least absolute shrinkage and selection operator (LASSO), linear support vector machine (SVM), adaptive boosting, and random forest. We measured each feature's selection frequencies out of 1,000 trials, calculated the average across the four ML models, and selected the top k features in descending order of their average frequency, where k is a hyperparameter.

Model performance and stability

We used LASSO for ML—one of the least flexible algorithms—to minimize the effect of model selection on the results. For each training-test dataset pair, a model was trained by five-fold CV and grid search in the training set and tested with the test set. The mean cross-validated area under the curve (AUC) and the test AUC were used as an evaluation metric for model performance in the training and test sets, respectively. A model with a smaller difference between the mean CV AUC and the test AUC was considered more stable than that with a larger difference. Additionally, when model stability was compared using a training set alone, the relative standard deviation (RSD) was used as a metric, which is calculated by dividing the standard deviation (SD) of a group of values by their average (17).

To investigate the effects of the number of input features and CV method on model performance and stability, the whole process mentioned above was repeated while increasing the number of radiomics features (i.e., k in the top k features) from 1 to 50 and changing the CV method: five-fold CV, five-fold CV with 100 repetitions, nested CV, and nested CV with 100 repetitions. In CV with n repetitions, CV is repeated n times by shuffling data for each combination of hyperparameters. The nested CV has an inner loop CV nested in an outer CV (Supplementary Figure 1); the inner loop is responsible for model selection and hyperparameter tuning, while the outer loop is used for error estimation (18).

Data composition of randomly split training and test sets

We compared the composition of training and test datasets to understand how a random split can produce widely varying results. Three trials (i.e., training-test set pairs) from the meningioma task were

selected as representative cases out of 1,000 random dataset splits: Two trials where the model showed significantly mismatched results and one trial where the model showed similar results between CV and testing. On a two-dimensional feature space using the top two ($k = 2$) radiomics features, each datapoint was plotted in different colors according to the class, and a decision boundary was drawn, along with the mean CV AUC in the training set and the AUC in the test set.

All analyses were performed using Python 3.0 with scikit-learn 0.23.2 and R 4.0.2. The 95% confidence interval (CI) of AUC was estimated using the DeLong method. The difference in value between the two groups was considered statistically significant when the two-sided probability determined using the t -test was less than 0.05. The Benjamini–Hochberg correction was used for multiple comparisons.

Results

Model stability and performance

Number of input features

When the results determined using the 1,000 different trials (i.e., training-test set pairs) were averaged, as the number of input features increased, the model performance increased in the beginning and later decreased (Fig. 2). However, without averaging, some trials did not follow this pattern; for example, the CV AUC did not decrease despite the increase in feature number but plateaued instead (Fig. 2, upper panels). Consequently, model stability decreased (i.e., the average AUC difference between CV and testing increased), as the number of input features increased beyond a certain point (Fig. 2, lower panels). The number of features for the highest performance and that for the highest stability matched almost exactly, which can be considered an optimal number of input features (Fig. 2). In this study, the optimal number of input features was 6 for the GBM vs. metastasis task and 13 for the meningioma grading task. Hereafter, only the results with these optimal feature numbers are presented.

Sample size and task difficulty

The overall model performance and stability were better with the simple task and without undersampling (Table 1 and Fig. 3). The mean test AUC (performance) and mean AUC difference between CV and testing (stability) were 0.947 and 0.029 for the simple task, and 0.795 and 0.053 for the difficult task, respectively. With the undersampled data, however, the mean test AUC and AUC difference between CV and testing decreased to 0.923 and 0.054 for the simple task, and 0.764 and 0.108 for the difficult task, respectively.

Comparison of CV methods

The four CV methods are listed in decreasing order of their estimated AUC values and in increasing order of RSD under the same conditions: CV without a repetition, CV with 100 repetitions, nested CV without a repetition, and nested CV with 100 repetitions. Therefore, the nested CV with 100 repetitions was the most

conservative method with the lowest AUC and highest RSD, though statistically significant differences were only found between the CV without repetition and the rest three methods (Table 2).

However, when there was a large discrepancy between the mean CV AUC in the training set and the AUC in the test set, none of the CV methods helped reduce the discrepancy; example cases are summarized in Supplementary Table 2.

Data composition of randomly split training and test sets

The AUC difference between CV and testing varied widely based on samples comprising the training and test sets. Consequently, there were significant discrepancies between expected and actual model performance in some of the trials. Three representative trials for each task are summarized in Table 3.

The mechanism behind the performance discrepancy between CV and testing is demonstrated by three representative cases chosen among 1,000 trials of classifying low- and high-grade meningioma using two radiomic features (Fig. 4). The mean CV AUC in the first case (Trial no. 602) was low (0.612 [SD: 0.044]); however, the test AUC was high (0.893 [95% CI: 0.814–0.972]). The fitted linear decision boundary did not separate the two classes well in the training set. However, when applied to the test set, it performed better because samples from each class were often located on the correct sides in the feature space (Fig. 4, right panel). In contrast, in the second case (Trial no. 518), the mean CV AUC was high (0.797 [SD: 0.046]) while the test AUC was low (0.549 [95% CI: 0.413–0.685]). In the third case (Trial No. 608), the performance discrepancy was negligible (mean CV AUC: 0.743 [SD: 0.061]; test AUC: 0.740 [95% CI: 0.612–0.859]) owing to similar distributions of datapoints in the feature space.

Discussion

The results of this study demonstrate that estimated model performance may vary significantly depending on the compositions of the training and test datasets, especially when the sample size is too small, or the task is too difficult with the existing data.

Therefore, the real-world ML performance should be validated using an external and independent validation set. However, external validation is often infeasible in radiomics ML studies despite its importance. A recent report showed that external validation was missing in 81.8% of radiomics studies published in high-impact journals (19). Thus, in a limited single-institutional dataset with only internal validation, an appropriate strategy is necessary for using ML to produce reliable and stable results; several methods have been proposed to obtain reliable estimates of model performance in this case (9, 18, 20, 21).

However, as the results of this study show, a significant data mismatch between training and test datasets can be challenging to overcome by any technique because computer algorithms are much worse at making a reliable inference than human beings when a new datapoint is very different or far

away in the feature space than those used for training (22). Thus, a single random training-test set split possess risks for creating a model that is not generalizable and for reporting unreliable results. Hence, in a limited dataset such as a single institutional dataset with small sample size, we suggest estimating model performance and model stability using the entire dataset by nested CV or bootstrapping with repetitions, instead of using a training dataset from a single random training-test set split.

Our results related to the number of input features, sample size, and the level of task difficulty can be better understood by imagining an n -dimensional feature space (in this study, two-dimensional spaces were used for the sake of visualization; Fig. 4). The Euclidean distance between two random points in the two-dimensional unit square is approximately 0.52 on average. In the three-dimensional unit cube, the average distance is 0.66, and in the ten-dimensional hypercube, it increases to approximately 3.16 (23). In contrast, the average distance becomes smaller with a larger number of datapoints. Therefore, as the dimension (i.e., the number of input features) increases, or the sample size decreases, the datapoints in the feature space are located farther from one another, leading to a greater chance that two randomly split subsets have significantly different data distributions. Additionally, as a classification task becomes more challenging, there are more areas in a feature space where datapoints belonging to different classes are mingled together; the datapoints in these areas are more critical in estimating a decision boundary than datapoints in other areas. Thus, in a difficult task with many such datapoints, a decision boundary may vary greatly according to how a dataset is split into training and test sets.

In conclusion, a single random split of a dataset into training and test sets may lead to an unreliable report of the performance of radiomics ML models, especially with small sample size or difficult tasks. When external validation is not feasible, and performance estimation is suspected to be unreliable, it is suggested to conduct nested CV or other techniques in the entire dataset, instead of developing a model after a single random training-test set split.

Declarations

This study was approved by Yonsei University Health System Institutional Review Board.

Competing interests: The authors declare no competing interests.

References

1. Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. *Eur J Nucl Med Mol I* 2019;46:2656–2672
2. National Library of Medicine, *PubMed*. <https://pubmed.ncbi.nlm.nih.gov/>. Accessed September 29, 2020
3. Halevy A, Norvig P, Pereira F. The Unreasonable Effectiveness of Data. *Ieee Intell Syst* 2009;24:8–12
4. Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA* 2020;323:305–306

5. Hutson M. Artificial intelligence faces reproducibility crisis. *Science* 2018;359:725–726
6. Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. *Korean J Radiol* 2019;20:1124
7. Liu Y, Chen P-HC, Krause J, Peng L. How to Read Articles That Use Machine Learning. *JAMA* 2019;322:1806–1816
8. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology* 2018;286:800–809
9. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *Plos One* 2019;14:e0224365
10. Harrington P de B. Multiple Versus Single Set Validation of Multivariate Models to Avoid Mistakes. *Crit Rev Anal Chem* 2017;48:00–00
11. Xu Y, Goodacre R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J Analysis Test* 2018;2:249–262
12. Bae S, An C, Ahn SS, Kim H, Han K, Kim SW, et al. Robust performance of deep learning for distinguishing glioblastoma from single brain metastasis using radiomic features: model development and validation. *Sci Rep* 2020;10:12110
13. Chen C, Ou X, Wang J, Guo W, Ma X. Radiomics-Based Machine Learning in Differentiation Between Glioblastoma and Metastatic Brain Tumors. *Frontiers Oncol* 2019;9:806
14. Banzato T, Causin F, Puppa AD, Cester G, Mazzai L, Zotti A. Accuracy of Deep Learning to Differentiate the Histopathological Grading of Meningiomas on MR Images: A Preliminary Study. *J Magn Reson Imaging* 2019;50:1152–1159
15. Chen C, Guo X, Wang J, Guo W, Ma X, Xu J. The Diagnostic Value of Radiomics-Based Machine Learning in Predicting the Grade of Meningiomas Using Conventional Magnetic Resonance Imaging: A Preliminary Study. *Frontiers Oncol* 2019;9:1338
16. Park YW, Oh J, You SC, Han K, Ahn SS, Choi YS, et al. Radiomics and machine learning may accurately predict the grade and histological subtype in meningiomas using conventional and diffusion tensor imaging. *Eur Radiol* 2019;29:4068–4076
17. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJWL. Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. *Frontiers Oncol* 2015;5:272
18. Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research* 2010:2079–2107
19. Park JE, Kim D, Kim HS, Park SY, Kim JY, Cho SJ, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. *Eur Radiol* 2020;30:523–536
20. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *Bmc Bioinformatics* 2006;7:91

21. Teschendorff AE. Avoiding common pitfalls in machine learning omic data science. *Nat Mater* 2019;18:422–427
22. Addressing data mismatch, lectured by Andrew Ng. Coursera.com Web site. <https://www.coursera.org/lecture/machine-learning-projects/addressing-data-mismatch-biLiY>. Published August 1, 2017. Accessed December 4, 2020
23. Weisstein EW. "Hypercube Line Picking." From *MathWorld*–A Wolfram Web Resources. <https://mathworld.wolfram.com/HypercubeLinePicking.html>. Accessed November 11, 2020

Tables

Table 1. Model performance and stability of LASSO according to the sample size and the level of task difficulty

	Simple task (GBM vs. Metastasis)		Difficult task (Low- vs. High-grade meningioma)	
	Without under-sampling	With under-sampling	Without under-sampling	With under-sampling
Performance: Mean ($\pm$ SD) test AUC	0.947 ($\pm$ 0.009)	0.923 ($\pm$ 0.020)	0.795 ($\pm$ 0.015)	0.764 ($\pm$ 0.034)
Stability: Mean ($\pm$ SD) CV-testing AUC difference	0.029 ($\pm$ 0.022)	0.054 ($\pm$ 0.044)	0.053 ($\pm$ 0.040)	0.108 ($\pm$ 0.079)

LASSO, least absolute shrinkage and selection operator; GBM, glioblastoma; AUC, area under the curve; CV, cross validation; SD, standard deviation

Table 2. Comparison of four CV methods: mean AUC and RSD averaged across 1,000 different training sets

	(1) 5-fold CV without a repetition	(2) CV with 100 repetitions	(3) Nested CV with a repetition	(4) CV with 100 repetitions	<i>P</i> -value				
					(1)	(1)	(1)	(2)	(3)
					vs. (2)	vs. (3)	vs. (4)	vs. (3)	vs. (4)
Simple task: GBM vs. brain metastasis									
Average mean AUC	0.948 (±0.013)	0.944 (±0.013)	0.943 (±0.015)	0.942 (±0.014)	0.161	0.037	0.037	0.420	0.912
Average RSD	3.947 (±1.509)	4.263 (±1.743)	4.001 (±1.749)	4.285 (±1.019)	0.326	0.828	0.326	0.395	0.326
Difficult task: Low- vs. high-grade in brain meningioma									
Average mean AUC	0.812 (±0.222)	0.806 (±0.238)	0.802 (±0.236)	0.800 (±0.218)	0.099	0.017	0.003	0.482	0.524
Average RSD	7.09 (±3.032)	8.10 (±3.367)	8.340 (±2.969)	8.68 (±0.841)	0.439	0.287	0.015	0.652	0.439

Number in parentheses are standard deviations. AUC, area under the curve; RSD, relative standard deviation

Table 3. Model performance of LASSO: comparison between mean AUC from five-fold CV and AUC from testing in some representative training-test set pairs

Training-test sets	Trial No.	Optimal <i>C</i> -value	Mean AUC in CV ($\pm$ SD)	AUC in testing (95% CI)	AUC difference
Simple task: GBM vs. brain metastasis					
Mismatch, high CV AUC	126	0.2	0.980 ($\pm$ 0.032)	0.874 (0.776–0.971)	-0.106
Mismatch, high test AUC	50	11	0.918 ($\pm$ 0.083)	0.993 (0.979–1.000)	+0.075
No mismatch	602	0.2	0.951 ($\pm$ 0.016)	0.955 (0.902–1.000)	+0.004
Difficult task: Low- vs. high-grade in brain meningioma					
Mismatch, high CV AUC	518	0.2	0.840 ($\pm$ 0.023)	0.650 (0.527–0.774)	-0.190
Mismatch, high test AUC	608	0.9	0.702 ($\pm$ 0.089)	0.836 (0.744–0.928)	+0.134
No mismatch	81	1.5	0.820 ($\pm$ 0.048)	0.804 (0.700–0.909)	-0.016

LASSO, least absolute shrinkage and selection operator; AUC, area under the curve; CV, cross validation; SD, standard deviation; CI, confidence interval.

Figures

Figure 1

Study flow

Figure 2

The model performance and stability of least absolute shrinkage and selection operator (LASSO) in binary classification according to the number of input features

Figure 3

The relationship between the task difficulty, sample size, and average of and difference between areas under the curve (AUCs) from cross validation (CV) and test. (GBM, glioblastoma)

Figure 4

Discrepant model performances from cross validations (CVs) in the training set and tests in the test set, explained by the distribution of datapoints in the feature space

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterials.docx](#)