Radiomics machine learning study with a 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: machine learning, magnetic resonance imaging, radiomics, brain tumor

Posted Date: November 19th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-105766/v1

Abstract

Objective: To determine how the estimated performance of a machine learning model varies according to how a dataset is split into training and test sets using brain tumor radiomics data, under different conditions.

Materials and Methods: Two binary tasks with different levels of difficulty ('simple' task, glioblastoma [GBM, n=109] vs. brain metastasis [n=58]; 'difficult' task, low- [n=163] vs. high grade [n=95] meningiomas) were performed using radiomics features from magnetic resonance imaging (MRI). For each trial of the 1,000 different training-test set splits with a ratio of 7:3, a least absolute shrinkage and selection operator (LASSO) model was trained by 5-fold cross-validation (CV) in the training set and tested in the test set. The model stability and performance was evaluated according to the number of input features (from 1 to 50), the sample size (full vs. undersampled), and the level of difficulty. In addition to 5-fold CV without a repetition, three other CV methods were compared: 5-fold CV with 100 repetitions, nested CV, and nested CV with 100 repetitions.

Results: The highest mean cross-validated area under the receiver operating characteristics curve (AUC) and the higher stability (lower AUC differences between training and testing) was achieved with 6 and 13 features from the GBM and meningioma task, respectively. For the simple task, simple task with undersampling, difficult task, and difficult task with undersampling, average mean AUCs were 0.947, 0.923, 0.795, and 0.764, and average AUC differences between training and testing were 0.029, 0.054, 0.053, and 0.108, respectively. Among four CV models, the most conservative method (i.e., lowest AUC and highest relative standard deviation [RSD]) was nested CV with 100 repetitions.

Conclusions: A single random split of a dataset into training and test sets may lead to an unreliable report of model performance in radiomics machine learning studies, and reporting the mean and standard deviation of model performance metrics by performing nested and/or repeated CV on the entire dataset is suggested.

Introduction

Since the advent of precision or personalized medicine, machine learning has attracted great interest as a framework that could revolutionize how we identify the best diagnosis and treatment for an individual patient. Machine learning research has expanded rapidly in many fields including radiomics, a method that extracts and uses a large number of features from medical images to uncover disease characteristics that fail to be appreciated by the naked eye.[1] According to the PubMed database with the search term of "(machine learning OR deep learning) AND radiomics", the number of published papers per year increased from 2 to 308 between 2015 and 2019.[2]

Machine learning is highly 'data hungry'. For a machine learning model to be robust, it may require millions of observations to reach acceptable performance levels.[3] Radiomics-based machine learning studies, however, are often conducted with a limited number of samples, especially when dealing with rare diseases. Nonetheless, many published papers reported fair or good performances. However, the reproducibility—which has been an intensely debated topic in science for the past few decades—of machine learning is a big concern; machine learning algorithms have a large number of parameters to train or manually set, and its training typically involves a lot of randomness, all of which pose unique challenges to the reproducibility by machine

learning.[4,5] Thus, caution should be taken when interpreting a reported model performance, as it may be over-optimistic, especially in the lack of external validation.[6,7]

As preliminary experiments, we tried performing two binary classification tasks, one of which was to differentiate meningioma grades using radiomics features. Single-institutional data were randomly split into training and test sets. In the training set, 5-fold cross-validation (CV) was performed for hyperparameter optimization. The optimized model showed a cross-validated area under the receiver operating characteristics curve (AUC) of 0.820 in the training set and 0.804 in the test set. However, in another trial with the same data and methods, we failed to reproduce the results; the cross-validated AUC in the training set was 0.840, but the AUC in the test set was far worse, 0.650.

In fact, not every method was the same in the above two trials; the compositions of the training and test sets were different because the dataset was split 'randomly'. From our experiences, the results often varied greatly depending on how datasets were split into training and test sets, especially with a small sample size or a difficult task with a suboptimal model performance. Previous studies have also suggested that the model performance measured with a single split of training and test sets may be over-optimistic.[8–10] However, to the best of our knowledge, the impact of random dataset splitting on the reliability of model performance in radiomics machine learning studies has not been investigated.

Therefore, the purpose of this study was to determine how the estimated performance of a machine learning model varies according to how a dataset is split into training and test sets using real-world brain tumor radiomics data, under different conditions: the number of input features (from 1 to 50), the sample size (full vs. undersampled), and the level of difficulty of machine learning task (simple vs. difficult).

Materials And Methods

Subjects

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

Two binary classification tasks with different levels of difficulty were performed using radiomics features extracted from brain magnetic resonance imaging (MRI) as input data. The first task was a relatively 'simple' task of differentiating between glioblastoma (GBM) and single metastasis, with reported accuracies of up to 89%.[11,12] The first dataset consisted of 167 adult patients with a single GBM (n=109) or brain metastasis (n=58) that were pathologically confirmed following brain MRI from January 2014 to December 2017. The second task was a 'difficult' task of differentiating between low- vs. high-grade meningioma, with reported accuracies of less than 76% by conventional MRI.[13,14] The second dataset consisted of 258 adult patients with a low grade (n=163) or high grade (n=95) meningioma diagnosed from February 2008 to September 2018. Both the datasets were from the same tertiary academic hospital, and some subsets of these patients were used in our previous reports.[11,15] MRI acquisition, image preprocessing, and radiomics feature extraction are described in Supplementary appendix.

Random split of datasets into training and test sets

The dataset was randomly split into training and test sets with a ratio of 7:3, while maintaining the proportions of the two classes. To examine how a result changes according to the data composition by a random training-test set split, the split was repeated 1,000 times by changing random state numbers from 0 to 999. By setting a random state number, we can reproduce the same results although it is called ‘random’ (See Discussion for more detailed explanation).

Feature selection

In the 1,000 different training sets, radiomics features were repeatedly selected based on the coefficient or feature importance by four machine learning models: least absolute shrinkage and selection operator (LASSO), linear support vector machine (SVM), adaptive boosting, and random forest. The frequencies of each feature being selected out of 1,000 trials were calculated for each model and were averaged. The top *k* features were selected in descending order of the average frequency, where *k* was a hyperparameter.

Model stability and performance

We investigated how the model stability, that is, the degree of change in results by the random training-test set split, as well as the model performance are affected by the number of input features, sample size, and task difficulty. In this study, we used LASSO for machine learning, which is one of the least flexible algorithms, in order to minimize the effect of model selection on the results. For each trial of the 1,000 different training-test set splits, a LASSO model was trained following optimization by 5-fold CV without a repetition in the training set and tested in the test set, and the mean cross-validated AUC and the test AUC were calculated. In this part of our experiments, a model was considered more stable when the difference between the mean CV AUC and the test AUC was smaller.

Number of input features

The process of repeating training and testing 1,000 times was repeated by increasing *k* (i.e., the number of input features) from 1 to 50 by 1. Based on the results of this experiment, the optimal number of features that achieved the best performance and stability was determined for each of the two tasks and was used for the following analyses.

Sample size and task difficulty

For each trial of training-test set splits, the average of and the difference between the mean CV AUC and the test AUC were calculated for the simple (GBM) task and difficult (meningioma) task. In addition, to determine the effects of sample size, 50% of the GBM and meningioma datasets were randomly sampled (with a random state number of 2020), on which all the processes were repeated.

Visualization of the effect of randomly splitting training-test sets

We attempted to visualize how the composition of training and test datasets determined by a random splitting can affect the fitting and evaluation of machine learning models. Of the 1,000 random dataset splits, three trials from the meningioma task were selected as representative cases: two trials where the training and test sets showed significant mismatch and one trial where the datasets showed similar compositions. On a two-dimensional feature space by two most robust radiomics features (k = 2), each case was plotted in different colors according to the class, and a decision boundary was drawn, along with the cross-validated mean AUC in training set and the AUC in the test.

Comparison of CV methods

In addition to 5-fold CV without a repetition, three other CV methods were conducted and compared: 5-fold CV with 100 repetitions, nested CV, and nested CV with 100 repetitions (Fig. 1). In CV with n repetitions, CV is repeated n times with shuffling data for each combination of hyperparameters. The nested CV has an inner loop CV nested in an outer CV; the inner loop is responsible for model selection and hyperparameter tuning, while the outer loop is for error estimation.[16] In addition to assessing the model performance by AUC, model stability was compared using a metric of relative standard deviation (RSD). RSD is calculated by dividing the standard deviation (SD) of a group of values by the average of the values.

All the analyses were performed using Python 3 with scikit-learn 0.23.2 or R 4.0.2. The 95% confidence interval (CI) of AUC was estimated by the DeLong method. The difference in values between two groups was considered statistically significant when two-sided probability by t-test was less than 0.05. Benjamini-Hochberg procedure was used to correct for multiple comparisons.

Results

Model stability and performance

Number of input features

Averaging the results from the 1,000 different training sets revealed the pattern that, as the number of input features increased, the mean cross-validated AUC increased at first but began to decrease at some point, more distinctly with the simple GBM task than the difficult meningioma task (Fig. 2). Without averaging, however, in some trials the mean AUC did not decrease despite the increase in the feature number but rather plateaued, whereas in other trials the mean AUC decreased more steeply and plateaued at a lower level (Fig. 2, the upper panels). Consequently, the deviation of mean AUCs across the different training-test set splits showed the reverse pattern; as the number of features increased, the average difference between mean AUC from CV in the training set and AUC in the test set decreased until it reached the lowest point around at the optimal feature number, and then increased again (Fig. 2, the lower panels). In our datasets, the graphs reached the peak for AUC and the base for stability when the number of features was 6 for the GBM task and 13 for the meningioma task, which were used as the optimal feature numbers for the following experiments.

Sample size and task difficulty

For the simple task, the overall model performance was better and more stable than the difficult task, indicated by higher AUCs and lower AUC differences between CV and testing. In the same task, undersampling a dataset resulted in diminished overall model performance and stability (Fig. 3). For the simple task, simple task with undersampling, difficult task, and difficult task with undersampling, average mean AUCs (±SDs) were 0.947 (±0.009), 0.923 (±0.020), 0.795 (±0.015), and 0.764 (±0.034), and average AUC differences between training and testing were 0.029 (±0.022), 0.054 (±0.044), 0.053 (±0.040), and 0.108 (±0.079), respectively.

Visualization of the effect of randomly splitting training-test sets

Depending on which samples comprised the training or test set, AUCs in CV and in testing varied widely. Consequently, in some of the trials there were significant discrepancies between the expected and actual model performance; three representative trials for each task are summarized in Table 1.

The mechanism behind the discrepancy is demonstrated in Figure 4, which depict three representative results chosen among 1,000 trials of classifying low- vs. high-grade in meningioma using two radiomics features. The first one (Trial no. 602) is the case where mean AUC in CV was low (0.612 [SD, 0.044]), but AUC in testing 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, but when applied to the test set, it could perform better because samples from each class were located more often on the correct sides in the feature space (Fig. 4, the right panel). The second one (Trial no. 518) is the opposite situation where mean AUC in CV was high (0.797 [SD, 0.046]), but AUC in testing was low (0.549 [95% CI, 0.413–0.685]). The third one (Trial no. 608) is the case where AUCs in CV and testing were

similar (mean CV AUC, 0.743 [SD, 0.061]; test AUC, 0.740 [95% CI, 0.612–0.859]) because of the similar distributions of data points in the feature space.

Comparison of CV methods

The estimated AUC was the highest with CV without a repetition, followed by CV with 100 repetitions, nested CV without a repetition, and nested CV with 100 repetitions in decreasing order. In contrast, the RSD was in reverse order. Therefore, the most conservative method (i.e., lowest AUC and highest RSD) was nested CV with 100 repetitions, although the differences were not statistically significant ($p > 0.326$) among the three methods–CV with repetitions and nested CV with or without repetitions (Table 2). However, in cases of severe mismatch in performance between training and testing, none of the CV methods was helpful to reduce the gap between the estimated AUC by CV and the test AUC; the results of example trials are summarized in Table 3.

Discussion

The results of this study demonstrate that the model performance and optimal hyperparameters estimated by cross validation in a training set may vary considerably according to the composition of the dataset, especially when the sample size is too small, or the task is too difficult to tackle with available data. In other words, if we are allowed to choose the distribution of samples in the feature spaces of training and test sets, we can obtain what we think would be ideal results, which is also called ‘cherry picking’.

A single random training-test set split allows for this cherry picking. Researchers may first find a random state number that produces the best possible training-test set split, and then fix the number to reproduce the same results. Although we call it ‘random’, the split we perform on computers is not truly random. In fact, nothing computers do is random when they are functioning correctly, because computers are deterministic devices and always behave predictably by design. Instead, to produce results that are ‘random enough’, computers use a ‘pseudorandom’ number generator which uses a mathematical algorithm to simulate randomness.[17] Thus, computers can reproduce the exact same results, given a pseudorandom number, which is essential to many computer or academic applications, including reproducible research.

Therefore, the real-world performance of machine learning should be validated using an independent external validation set. However, despite the importance of external validation to generalize the study result, external validation is often infeasible in radiomics machine learning studies; a recent report showed that external validation was missing in 81.8% of radiomics studies published in high-impact journals.[18] Thus, in a limited single-institutional dataset enabling only internal validation, an appropriate strategy is necessary in machine learning to produce reliable and stable results.

Several methods have been proposed to obtain more generalizable and reliable estimates of model performance.[8,16,19,20] However, a significant mismatch between training and testing datasets is difficult to overcome by any methods, because computer algorithms, unlike humans, cannot make a correct inference if a

new datapoint is very different, or far away in the feature space, from the datapoints used for training. Thus, to reduce the risk of reporting over-optimistic results in machine learning studies without an external validation set, we suggest that a metric for model stability such as RSD should also be reported following nested and/or repeated CVs in the entire dataset, instead of a single random training-test set split.

Our results related to the optimal number of input features, task difficulty, and sample size can be better understood by imagining a *n*-dimensional feature space, although here we presented only two-dimensional spaces for the sake of visualization (Fig. 4). If we pick two random points in the two-dimensional unit square, the Euclidian distance between two points is roughly 0.52 on average. In the three-dimensional unit cube, the average distance is 0.66, and in the 10-dimensional hypercube, it increases up to around 3.16.[21] Therefore, in a high-dimensional space, datapoints are located much farther than we can possibly imagine based on the real-world space. A larger number of input features mean a higher dimensionality of feature space. Similarly, a classification task is harder when the datapoints belonging to the same class are not close together in the feature space, but rather scattered and mixed with other class datapoints. Lastly, it is more likely that the distance between datapoints is farther with a smaller sample size. Thus, the impact of which datapoints are sampled (i.e., the results of random training-test set splitting) on the overall results can be larger with a larger number of input features, a harder task, and a smaller sample size, as under these circumstances.

In conclusion, a single random split of a dataset into training and test sets may lead to an unreliable report of model performance in radiomics machine learning studies, especially with a small sample size or a difficult task for available data. Thus, in a radiomics machine learning study with a limited number of sample or without an independent external validation set, we suggest reporting the mean and deviation of model performance metrics by performing nested and/or repeated CV on the entire dataset rather than simply reporting the results following a single random training-test set split.

Abbreviations

AUC = area under the receiver operating characteristics curve; CI = confidence interval; LASSO = least absolute shrinkage and selection operator; MRI = magnetic resonance imaging; RSD = relative standard deviation; SVM = support vector machine

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, et al. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. *Eur J Nucl Med Mol I*. 2019;46(13):2656–2672.
2. Medicine NL of. PubMed. n.d. Available at: 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(2):8–12.
4. Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. *Jama*. 2020;323(4):305–306.
5. Hutson M. Artificial intelligence faces reproducibility crisis. *Science*. 2018;359(6377):725–726.
6. Liu Y, Chen P-HC, Krause J, et al. How to Read Articles That Use Machine Learning. *Jama*. 2019;322(18):1806–1816.
7. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology*. 2018;286(3):800–809.
8. Vabalas A, Gowen E, Poliakoff E, et al. Machine learning algorithm validation with a limited sample size. *Plos One*. 2019;14(11):e0224365.
9. Harrington P de B. Multiple Versus Single Set Validation of Multivariate Models to Avoid Mistakes. *Crit Rev Anal Chem*. 2017;48(1):00–00.
10. 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(3):249–262.
11. Bae S, An C, Ahn SS, et al. Robust performance of deep learning for distinguishing glioblastoma from single brain metastasis using radiomic features: model development and validation. *Sci Rep-uk*. 2020;10(1):12110.
12. Chen C, Ou X, Wang J, et al. Radiomics-Based Machine Learning in Differentiation Between Glioblastoma and Metastatic Brain Tumors. *Frontiers Oncol*. 2019;9:806.
13. Banzato T, Causin F, Puppa AD, et al. Accuracy of Deep Learning to Differentiate the Histopathological Grading of Meningiomas on MR Images: A Preliminary Study. *J Magn Reson Imaging*. 2019;50(4):1152–1159.
14. Chen C, Guo X, Wang J, et al. 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.
15. Park YW, Oh J, You SC, 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(8):4068–4076.
16. 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.
17. Móri TF, Székely GJ. Pseudorandom processes. *Stoch Proc Appl*. 2019.
18. Park JE, Kim D, Kim HS, 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(1):523–536.
19. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *Bmc Bioinformatics*. 2006;7(1):91.
20. Teschendorff AE. Avoiding common pitfalls in machine learning omic data science. *Nat Mater*. 2019;18(5):422–427.

21. Weisstein EW. Hypercube Line Picking. 2020. Available at: https://mathworld.wolfram.com/HypercubeLinePicking.html. Accessed October 11, 2020.

Tables

Table 1. Performance of LASSO: comparison between mean AUC from 5-fold cross validation and AUC from testing in some representative training/test sets

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

Table 2. Comparison of four cross validation 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 without a repetition	(4) CV with 100 repetitions	*P*-value				
					(1) vs. (2)	(1) vs. (3)	(1) vs. (4)	(2) vs. (3)	(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 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 receiver operating characteristics curve; RSD, relative standard deviation; CV, cross validation; GBM, glioblastoma.

Table 3. Comparison of four cross validation methods in a discrepant training/test sets for each task

Task	CV method	Mean CV AUC (±SD)	RSD in CV	Test AUC (95% CI)
GBM task (Trial No. 126)	(1) 5-fold CV without a repetition	0.980 (±0.032)	7.086	0.874 (0.776–0.971)
	(2) CV with 100 repetitions	0.972 (±0.045)	8.191	0.870 (0.773–0.968)
	(3) Nested CV without a repetition	0.970 (±0.041)	8.518	0.874 (0.776–0.971)
	(4) Nested CV with 100 repetitions	0.971 (±0.024)	9.707	0.875 (0.778–0.971)
Meningioma task (Trial No. 518)	(1) 5-fold CV without a repetition	0.840 (±0.023)	2.784	0.650 (0.527–0.774)
	(2) CV with 100 repetitions	0.835 (±0.069)	5.297	0.648 (0.526–0.775)
	(3) Nested CV without a repetition	0.827 (±0.046)	5.611	0.646 (0.525–0.773)
	(4) Nested CV with 100 repetitions	0.826 (±0.038)	7.779	0.646 (0.524–0.771)

CV, cross validation; AUC, area under the receiver operating characteristics curve; SD, standard deviation; RSD, relative standard deviation; CI, confidence interval; GBM, glioblastoma.

Figures

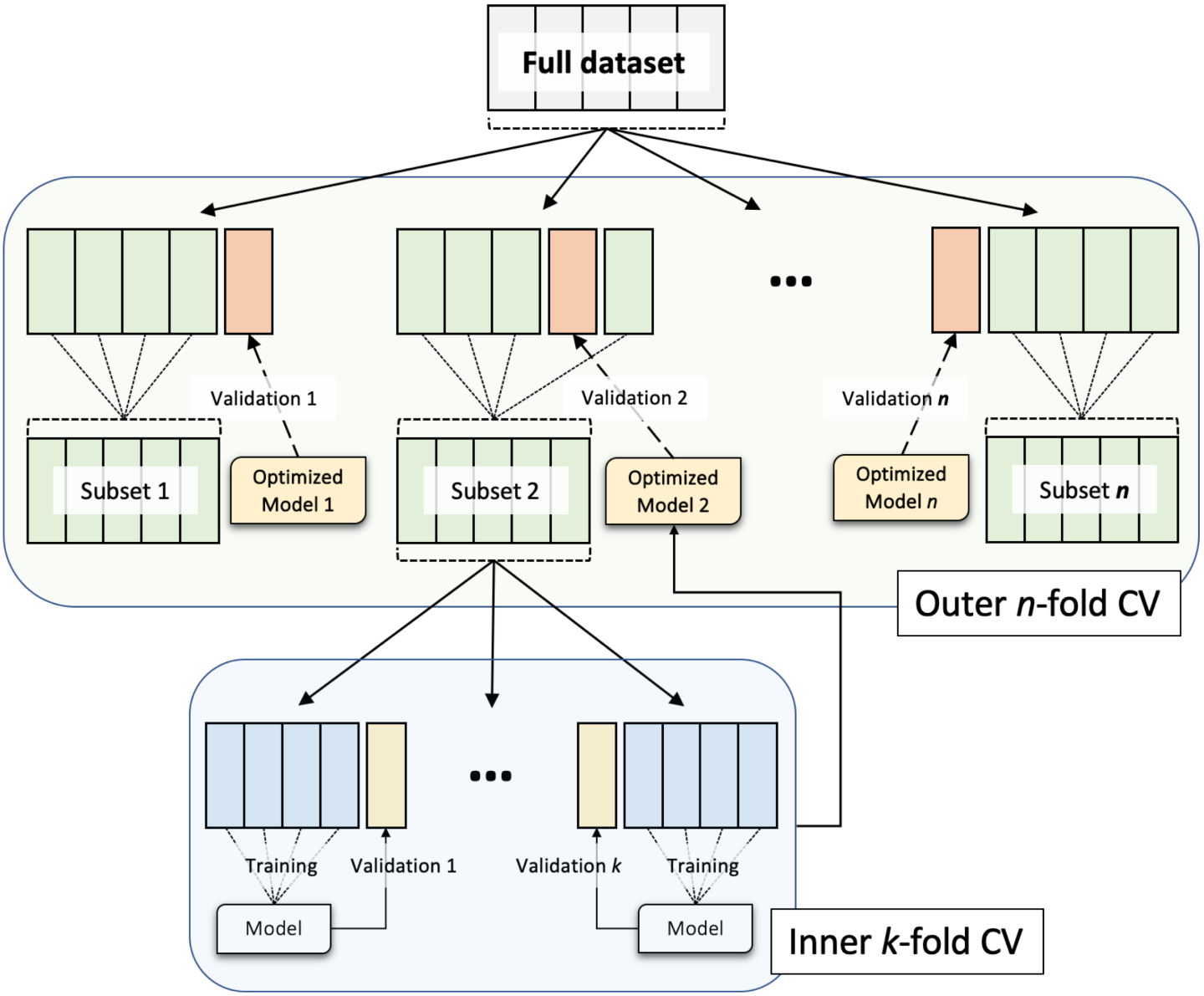

Figure 1

Nested cross-validation. CV = cross-validation

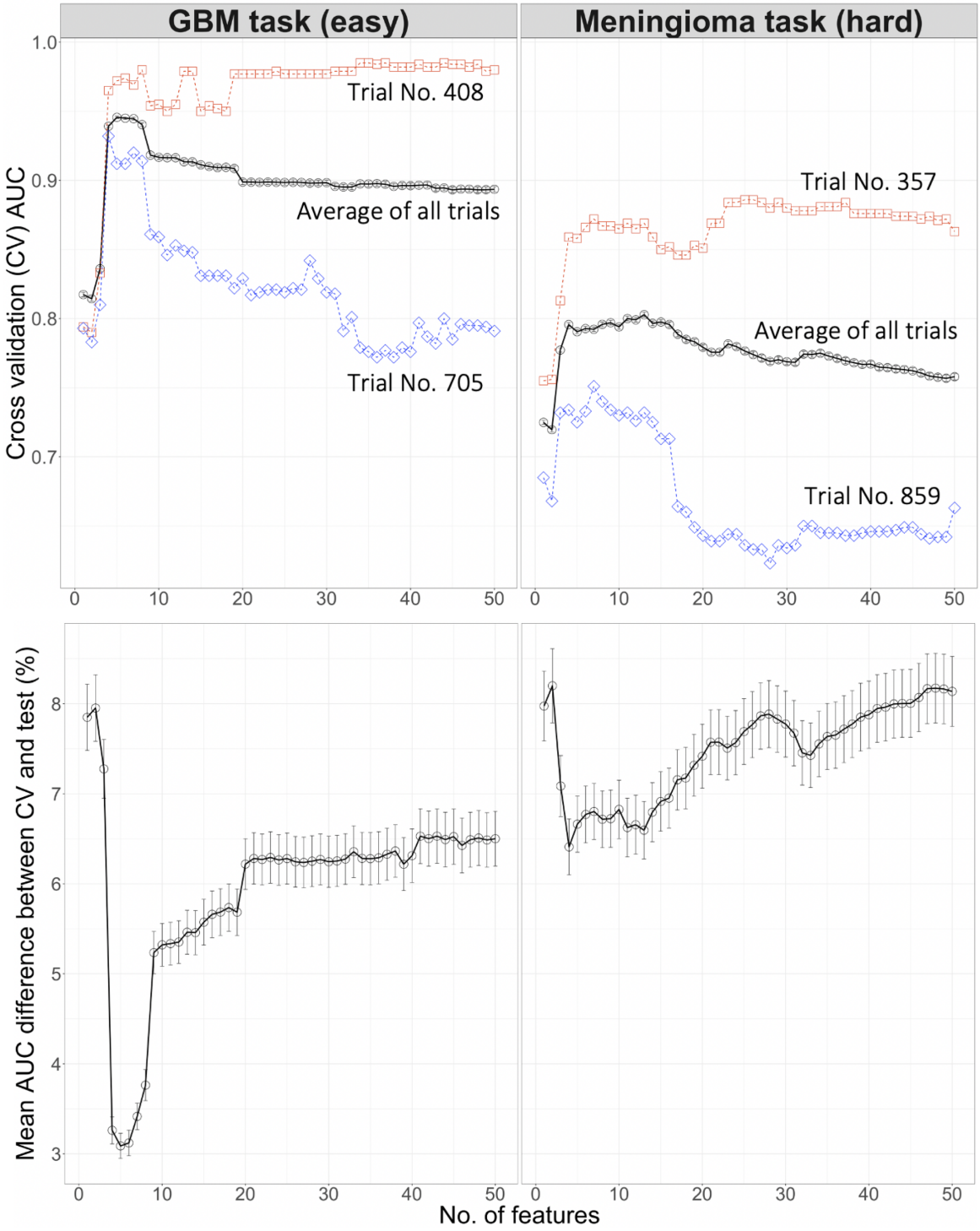

Figure 2

The performance and stability of least absolute shrinkage and selection operator (LASSO) in binary classification according to the number of input features. AUC = area under the receiver operating characteristics curve.

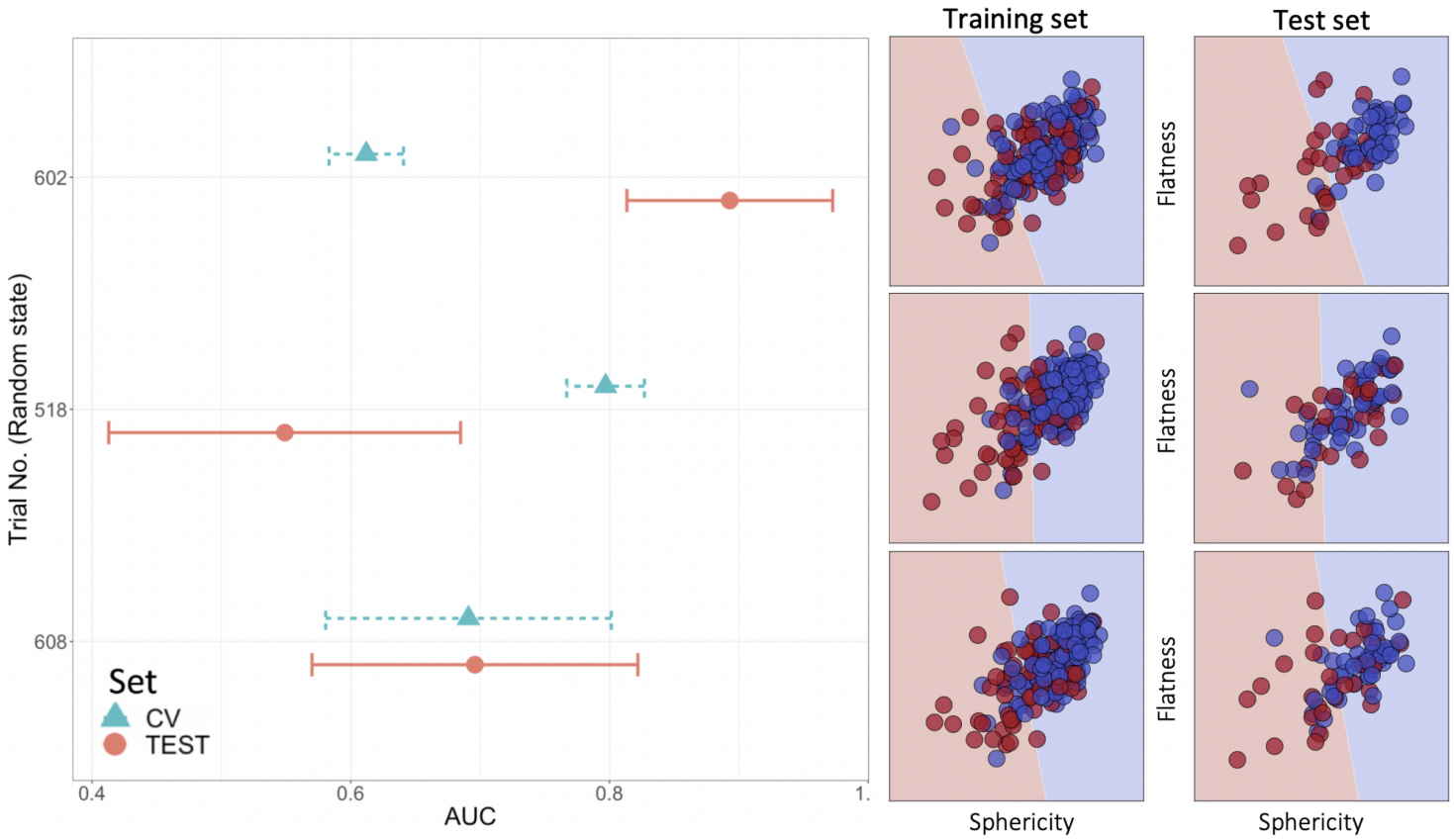

Figure 3

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

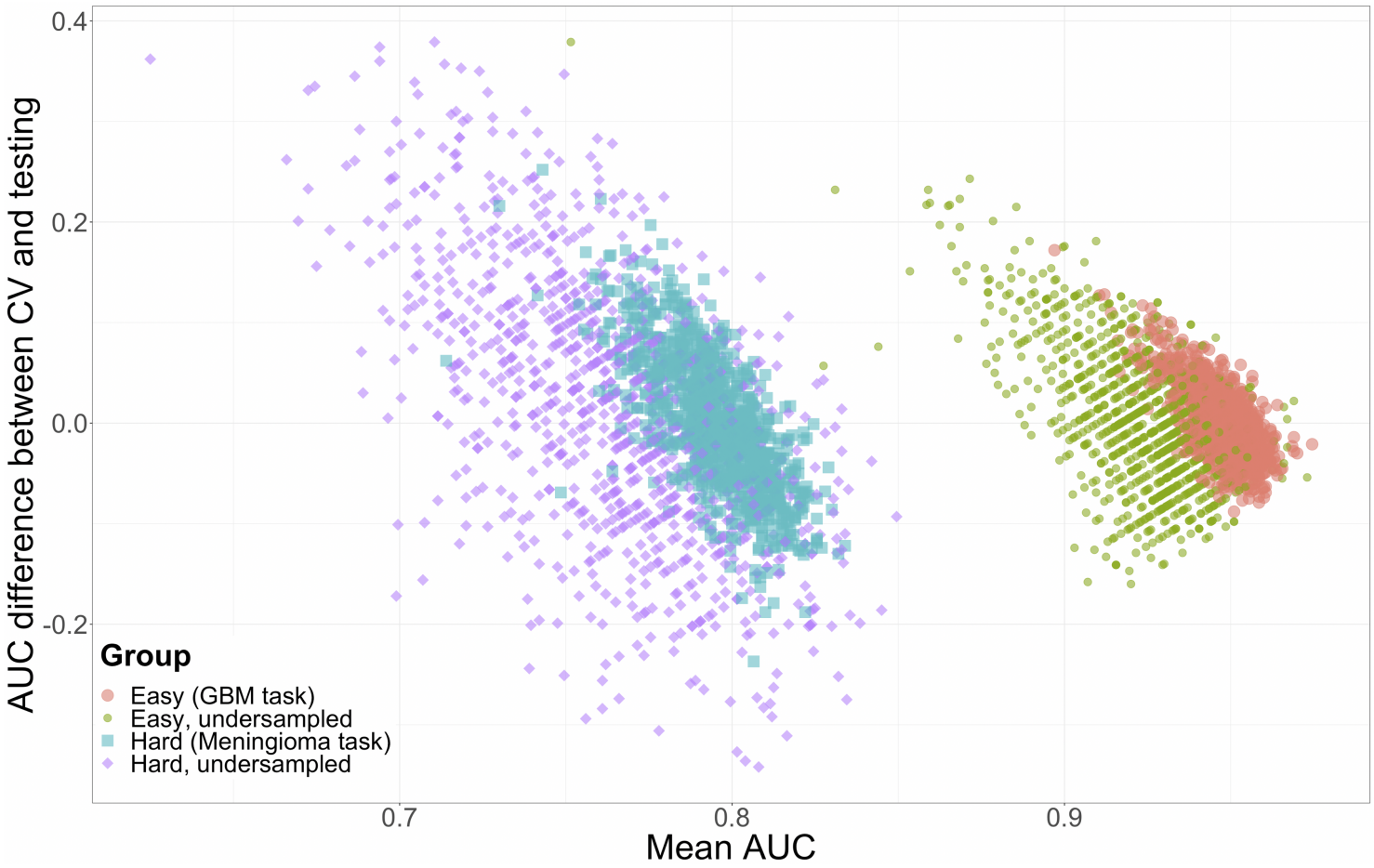

Figure 4

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

Supplementary Files

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

- InvestRadiolSupplementaryFileCA.docx