Statistical Variability in Comparing Accuracy of Neuroimaging Based Classification Models via Cross Validation

Bahram Jafrasteh[1,*], Ehsan Adeli[2,3], Kilian M. Pohl[2], Amy Kuceyeski[1], Mert R. Sabuncu[1,4], and Qingyu Zhao[1,*]

[1]Department of Radiology, Weill Cornell Medicine, New York, NY
[2]Department of Psychiatry & Behavioral Sciences, Stanford University, Stanford, CA
[3]Department of Computer Science, Stanford University, Stanford, CA
[4]School of Electrical and Computer Engineering, Cornell University and Cornell Tech, New York, NY
[*]baj4003@med.cornell.edu; qiz4006@med.cornell.edu

ABSTRACT

Machine learning (ML) has significantly transformed biomedical research, leading to a growing interest in model development to advance classification accuracy in various clinical applications. However, this progress raises essential questions regarding how to rigorously compare the accuracy of different ML models. In this study, we highlight the practical challenges in quantifying the statistical significance of accuracy differences between two neuroimaging-based classification models when cross-validation (CV) is performed. Specifically, we propose an unbiased framework to assess the impact of CV setups (e.g., the number of folds) on the statistical significance. We apply this framework to three publicly available neuroimaging datasets to re-emphasize known flaws in current computation of p-values for comparing model accuracies. We further demonstrate that the likelihood of detecting significant differences among models varies substantially with the intrinsic properties of the data, testing procedures, and CV configurations of choice. Given that many of the above factors do not typically fall into the evaluation criteria of ML-based biomedical studies, we argue that such variability can potentially lead to p-hacking and inconsistent conclusions on model improvement. The obtained results from this study underscore that more rigorous practices in model comparison are urgently needed in order to mitigate the reproducibility crisis in biomedical ML research.

Introduction

Machine learning (ML) has frequently been adopted in biomedical research in the past decade[1,2], with the number of ML-based investigations since 2010 being approximately 10 times greater than before 2010, according to data from Google Scholar (Search words "machine learning for biomedical"). Unlike classical statistical methods that are primarily bounded by univariate, population-level inference, data-driven ML bypasses the previous need for extensive prior assumptions on the data, explores complex multivariate relationships, and generates individual-level predictions[3–5]. As such, the fast-evolving AI field has witnessed thousands of new ML models developed each year, many of which demonstrate enhanced predictive capabilities in specific clinical applications[6–9]. While these endeavors have greatly advanced the frontiers of biomedical sciences, there is no common strategy to determine whether one model is indeed more accurate than another.

In light of the reproducibility crisis in biomedical research, researchers increasingly prioritize assessing the accuracy of models on external datasets across multiple independent studies[10–12]. However, several challenges arise when using external datasets, such as obtaining access to data specific to the clinical cohort of interest and aligning and standardizing data obtained from different studies[13–15]. Therefore, cross-validation (CV) based on a single data set remains a prevalent procedure for assessing ML models. In a CV setting, the data at hand are split into K folds, where $K-1$ folds are used to train the model and the remaining fold becomes the test data. The training and testing procedure repeats until all the folds have been used for testing. Compared to assessments based on a single test set, the CV procedure is particularly favorable in analyzing small-to-medium-sized datasets, such as those used in neuroimaging studies with $N < 1000$, to mitigate the potential high variance in accuracy associated with limited testing samples[16,17].

When developing a new ML model for a specific biomedical application, researchers often compare the accuracy of the proposed model with existing state-of-the-art methods. In reporting the results, researchers have a strong incentive to "bold" their proposed model in the table of comparison to highlight the improved accuracy. To statistically justify this improvement, researchers often use hypothesis testing to derive p-values quantifying the statistical significance in the accuracy difference. Several theoretical and practical challenges arise when performing hypothesis testing on accuracy scores from the

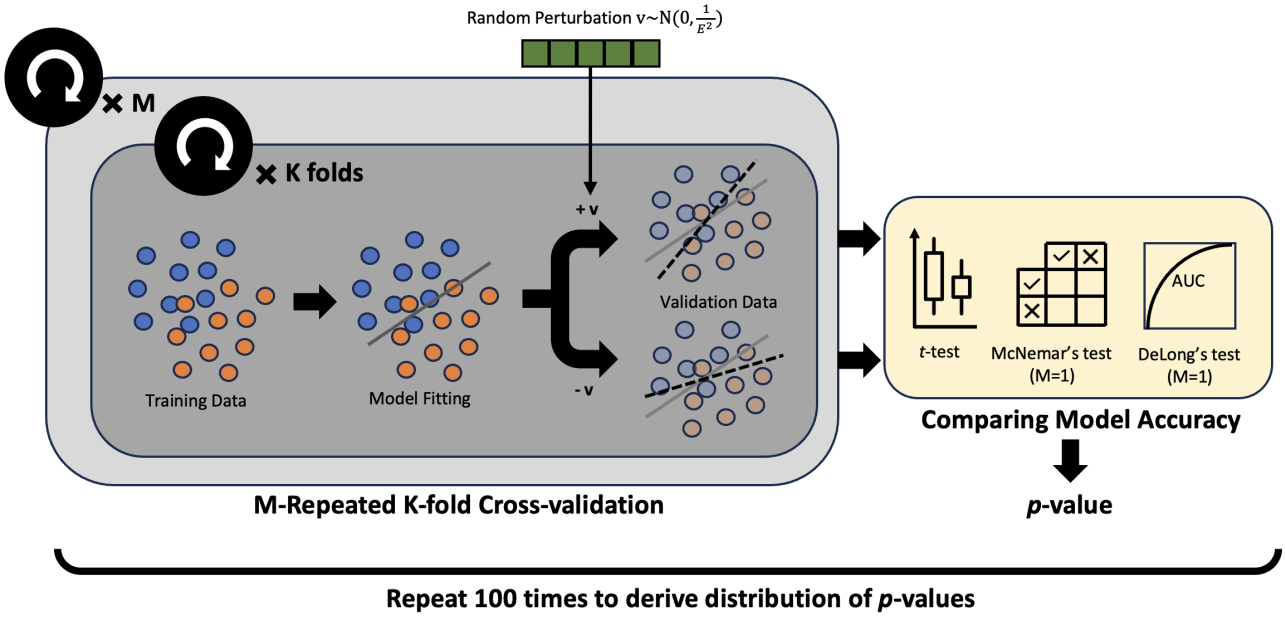

Figure 1. A framework for comparing two classifiers with the same intrinsic predictive power using a K-fold cross-validation repeated for M times. In each of the $K \times M$ training runs, the trained classifier undergoes two perturbations (perturbation level controlled by E), one along direction $+v$ and the other along $-v$, resulting in two perturbed models. The positively perturbed model and the negatively perturbed model are subsequently evaluated on the test dataset. After the $K \times M$ runs, a statistical hypothesis test compares the two sets of accuracy scores to yield a p-value quantifying the statistical difference between the two models. This entire procedure is repeated multiple times to derive the distribution of p-values.

CV setting[18,19]. For example, the overlap of training folds between different runs induces implicit dependency in accuracy scores, thereby violating the basic assumption of sample independence in most hypothesis testing procedures and may impact the normality of data distribution and the assumption of equal variance across groups. Although previously discussed in the literature, these issues still receive little attention among biomedical researchers or even ML practitioners. Moreover, largely unclear is to what extent the specific setup of CV (e.g. the choice of K) could potentially impact the test outcome.

Assessing the impact of CV setups on model comparison outcomes is inherently difficult, as whether a model is "superior" to another depends on other factors including model input dimension, training sample size, noise level, etc. We address this ambiguity by proposing an unbiased framework to disentangle the influences of these factors on model comparison and only focus on assessing the impact of CV setups on the statistical significance of accuracy differences between models. We apply this framework to compare models in classification tasks in three neuroimaging studies. Based on hundreds of thousands of training and testing runs of different ML classifiers on these datasets, we re-emphasize that some popular practices in deriving p-values for model comparison in the CV setting are fundamentally flawed. Moreover, we further reveal that the sensitivity of the statistical tests for model comparison varies with numerous factors that are often not considered as important in many ML studies. Overlooking these issues could lead to the potential for p-hacking. Given CV remains the primary model assessment procedure for a large number of biomedical studies[20], our results highlight the need to seek unified and unbiased testing procedures to avoid exacerbating the reproducibility crisis in the ML era[21].

Results

A Framework for Comparing Model Accuracy based on Cross-Validation

We first describe our framework for assessing the statistical significance of accuracy differences between two classification models evaluated by "repeated" CVs[22,23], a practice that has been shown to be problematic but is still frequently adopted by researchers. In repeated CV, the two models are both trained and evaluated using a K-fold (stratified) CV that is repeated for M times. The resulting $K \times M$ accuracy scores associated with either model are then compared by a statistical test. We now investigate whether this testing procedure can consistently quantify the statistical significance of the difference in classification accuracy with different choices of K and M.

In designing this framework, we first note that the accuracy of ML models generally depends on the dataset and sample size (e.g., training non-linear models generally requires more data than training linear models, so non-linear models are only

more predictive when training data is sufficient). It poses a challenge to disentangle the impact of CV setups on the accuracy difference between models. Therefore, we refrain from comparing models with different underlying algorithms but instead propose a framework to construct two classifiers with the same "intrinsic" predictive power (Fig. 1); that is, for any dataset, there is no theoretical algorithmic advantage of one model over another, and the observed accuracy difference between two models is only created by chance. Specifically, we create two classifiers by executing the following steps:

Step 1: Randomly choose N samples from each class;

Step 2: Create a random zero-centered Gaussian vector with standard deviation of $\frac{1}{E}$, where E is a predefined parameter called the perturbation level. The dimension of the vector equals to the number of features;

Step 3: In each of the $K \times M$ validation runs, train a linear Logistic Regression (LR) on the training data;

Step 4: Create a perturbed model by adding the random vector to the linear coefficients of its decision boundary;

Step 5: Create a second perturbed model by subtracting the random vector from the decision boundary;

Step 6: Evaluate the accuracy of two perturbed models on the testing data;

Step 7: Use a certain hypothesis testing procedure (e.g., paired t-test) to produce a p-value quantifying the significant difference in prediction accuracy across the $K \times M$ testing folds.

In this framework, the perturbations along two strictly opposite directions ensure that the magnitude of the discrepancy between the two models is strictly linked to the perturbation level E. In doing so, the observed accuracy differences between the two perturbed models is due simply to chance rather than to their intrinsic differences (e.g., one model has a superior algorithm design or is better suited to a specific sample size than the other model). Ideally, one would want to consistently quantify statistical significance of that difference regardless of the choices of K and M. In the following sections, we will demonstrate that in practice, one model can appear statistically significantly better than another based solely on variations in the choices of K and M.

Model Comparison Using Paired t-test

We applied the above framework to compare model accuracy in three neuroimaging-based classification tasks: 1) classifying 222 healthy control subjects vs. 222 patients with Alzheimer's disease based on T1-weighted MRI released by the Alzheimer's Disease Neuroimaging Initiative (ADNI)[24] study; 2) distinguishing 391 individuals with autism spectrum disorders (ASD) from 458 typically developing controls based on resting-state functional MRI released by the Autism Brain Imaging Data Exchange (ABIDE I) Dataset[25]; and identifying sex of 6125 boys and 5600 girls based on (head size corrected) T1-weighted MRI released by the Adolescent Brain Cognitive Development (ABCD) study[26]. Neuroimaging data of all three datasets were preprocessed into tabular measurements as the input features to the classification (See Section Methods).

A commonly misused procedure for comparing model accuracy is to use a paired t-test to compare the two sets of $K \times M$ accuracy scores from two models. To further illustrate this flaw, we applied the proposed framework (Fig. 1) to each of the three neuroimaging datasets to investigate the outcomes of the t-test based on various CV setups with different K, M combinations. In each K, M setup, we repeated the framework 100 times and recorded the average p-value of the corresponding statistical test.

In this experiment, we focused on balanced classification by setting the number of random samples $N = 500$ for ABCD, $N = 300$ for ABIDE, and $N = 222$ for ADNI. We chose E (in Step 2) for each dataset such that the resulting p-values were roughly on the same level. Supplement Fig. S1.1 a-f confirms that in all three classification tasks, the (unperturbed) Logistic Regression classifier achieved a classification accuracy significantly higher than chance in all K, M setups. Notably, changing K from 2-fold CV to 50-fold CV resulted in higher average classification and larger variance in accuracy over folds. Next, we compared the accuracy of the two perturbed classifiers in all three datasets. Based on the proposed comparison framework, Fig. 2a-c shows the range of p-values (quantifying significant accuracy differences) based on 2-fold or 50-fold CV, without repetition (M=1) or repeated for up to 10 times (M=10). We observe an undesired artifact that test sensitivity increased (lower p-values) with the number of CV repetitions M and the number of folds K. Furthermore, Fig. 3a-c shows the average p-value for more K, M combinations. If we used $p < 0.05$ as the significance threshold, Fig. 3d-f shows the "Positive Rate", i.e, how likely the two models have significantly different accuracy based on K-fold CV repeated for M times. We can observe that, despite applying two classifiers of the same intrinsic predictive power on the same dataset, the outcome of the model comparison largely depended on CV setups, with a higher likelihood of detecting a significant accuracy difference in a high K, M combination setting. For example, in the ABCD dataset, the positive rate increased on average by 0.49 from $M = 1$ to

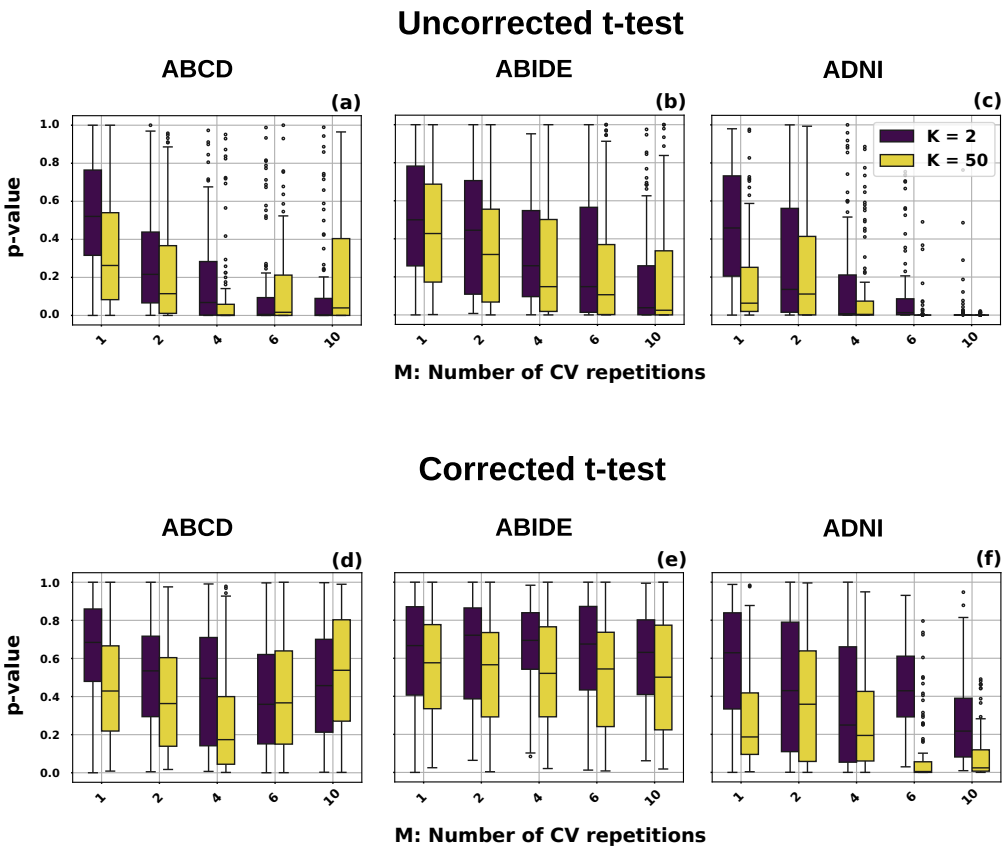

Figure 2. Statistical significance of comparing the accuracy of two Logistic Regression classifiers with the same intrinsic predictive power via cross-validation: (a-f) In each K, M setup, the framework of Fig. 1 was executed for 100 times. In each run, a paired t-test compared the $K \times M$ accuracy scores of the two perturbed Logistic Regression models. We record box-plots of the resulting p-values for (a-c) uncorrected t-test and (d-f) corrected t-test.

$M = 10$ across different K settings, and it increased on average by 0.07 from $K = 2$ to $K = 50$ across different M settings, which highlighted the dependence of the p-value on the choice of K and M settings.

As already pointed out in many studies, one major issue of repeated CV is that the $K \times M$ accuracy scores are highly dependent due to the overlap between the test (or training) folds of different validation runs[22,27]. This violates the assumption of sample independence in the t-test. To resolve this issue, a "corrected" version of the paired t-test[22,27] has been proposed to control for the dependency across accuracy scores. We then examined whether the corrected t-test could avoid the dependency of test sensitivity on K and M. Results in Fig. 2d-f show that the correction indeed resulted in more conservative p-values than the regular t-tests but still largely influenced test sensitivity. For example, 50-fold CV still resulted in lower range of p-values than 2-fold CV in all three datasets, and a large number of CV repetitions resulted in the lowest p-values in ADNI. Fig. 3j-l suggests that the highest positive rate occurred under a combination of large K and M.

Reproducibility of Results

First, we examined whether the observed pattern in Fig. 2 was due to the relatively low classification accuracy in the neuroimaging applications ($N < 1000$). We applied the proposed comparison framework of Fig. 1 to two synthetic classification datasets with $N = 10,000$ and $N = 100,000$, with a known Bayes error of 5% (See Methods for dataset creation). Supplement Figure S1.2 shows that the classifiers achieved high accuracy (92%) in both synthetic datasets. When $N = 10,000$ (Figure S1.2 a-d), the observed patterns aligned with the neuroimaging-based results, where higher K, M combinations resulted in lower p-values (greater chance of detecting significant accuracy differences between two perturbed classifiers). When increasing the sample size to $N = 100,000$, Supplement Figure S1.2 e-h suggests that compared to $N = 10,000$, the dependency of the p-value on M was less pronounced, but higher K still resulted in lower p-values in 80% of the time.

Next, to investigate whether the dependency of p-values on CV setups was specific to linear models, we repeated the neuroimaging-based experiments to compare accuracy scores between two perturbed Multi-Layer Perceptrons (MLP)

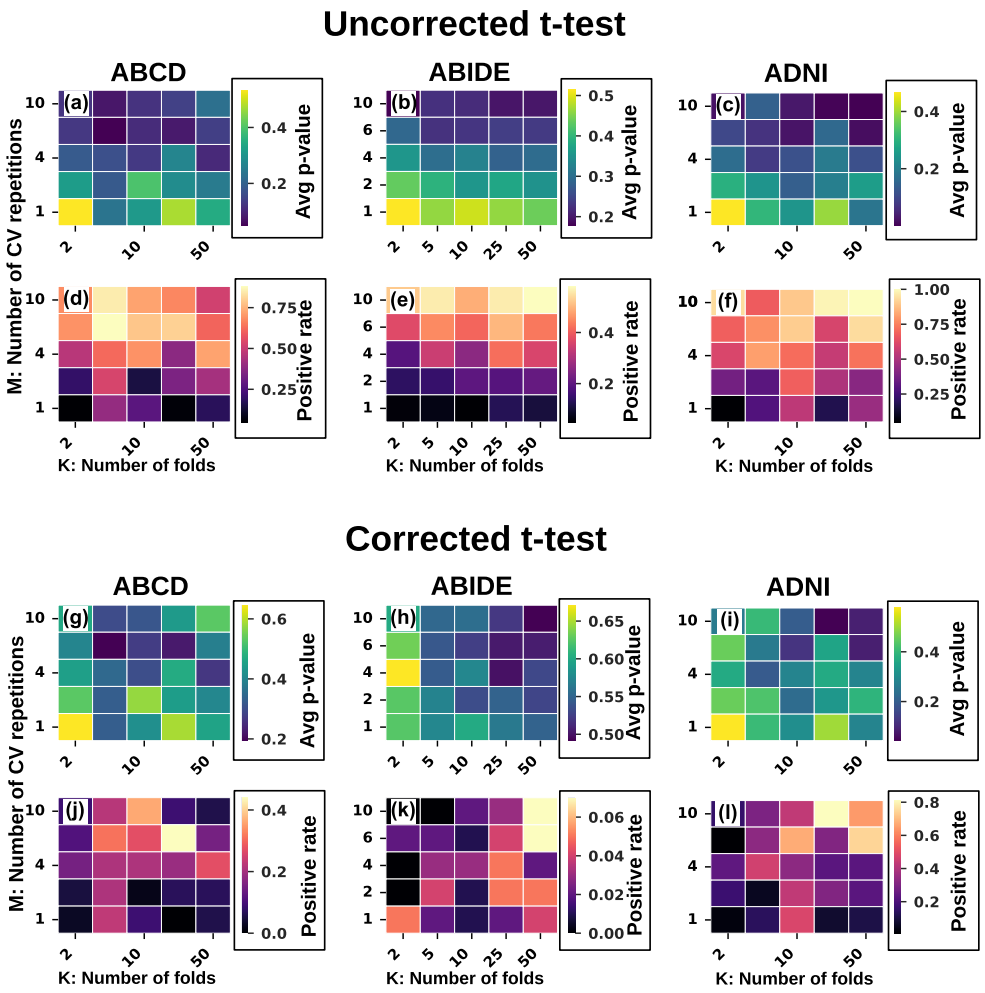

Figure 3. The average p-value and positive rate (how frequent the two perturbed Logistic Regression models had significant accuracy difference based on the threshold of $p < 0.05$ in the 100 runs) were recorded for each K, M combination for uncorrected t-test (a-f) and corrected t-test (g-l).

(Supplement Fig. S1.1 g-l and Fig. S1.3, where the perturbation (of Step 4) was applied by adding the random vector to the weights of the last fully connected layer. Supplement Fig. S1.3 suggests similar patterns as in Fig. 2, where the range of p-values depended on both K and M. Lastly, instead of applying positive or negative perturbations in the framework of Fig. 1, the model was perturbed by two totally different random Gaussian vectors with standard deviation $\frac{1}{E}$. We also replaced t-tests with permutation tests to handle the potential non-Gaussian distribution of accuracy scores. Supplement Figs. S1.4 and S1.5 largely replicate the findings that test sensitivity increased with K and M.

Rank of Sensitivity across t-tests, McNemar's test, and DeLong's test

In addition to the t-tests, two other commonly used testing procedures for comparing model accuracy are McNemar's test[28] and DeLong's test[29]. Unlike the t-tests, McNemar's and DeLong's tests typically only require one round of CV ($M = 1$). Specifically, the classification results are pooled together from the K runs, and the number of correctly classified samples and the area-under-the-ROC-curve (AUC) are compared between two models by Chi-squared statistics. Based on the framework of Fig. 1, we then examined whether the sensitivity of McNemar's test and DeLong's test depended on the number of folds K. To do so, we repeated the model comparison framework 100 times for each K setting (similar to the previous experiment) and recorded the distribution of p-values and positive rates of McNemar's test and DeLong's test. Supplementary Fig. S1.6 shows that these sensitivity levels were more invariant to K compared to the two variants of the t-test.

Next, we examined the relative sensitivity among t-test, McNemar's test, and DeLong's test (i.e., which procedure resulted in the most conservative p-values). We recorded the distribution of p-values and positive rates over all K, M settings in Fig. 4.

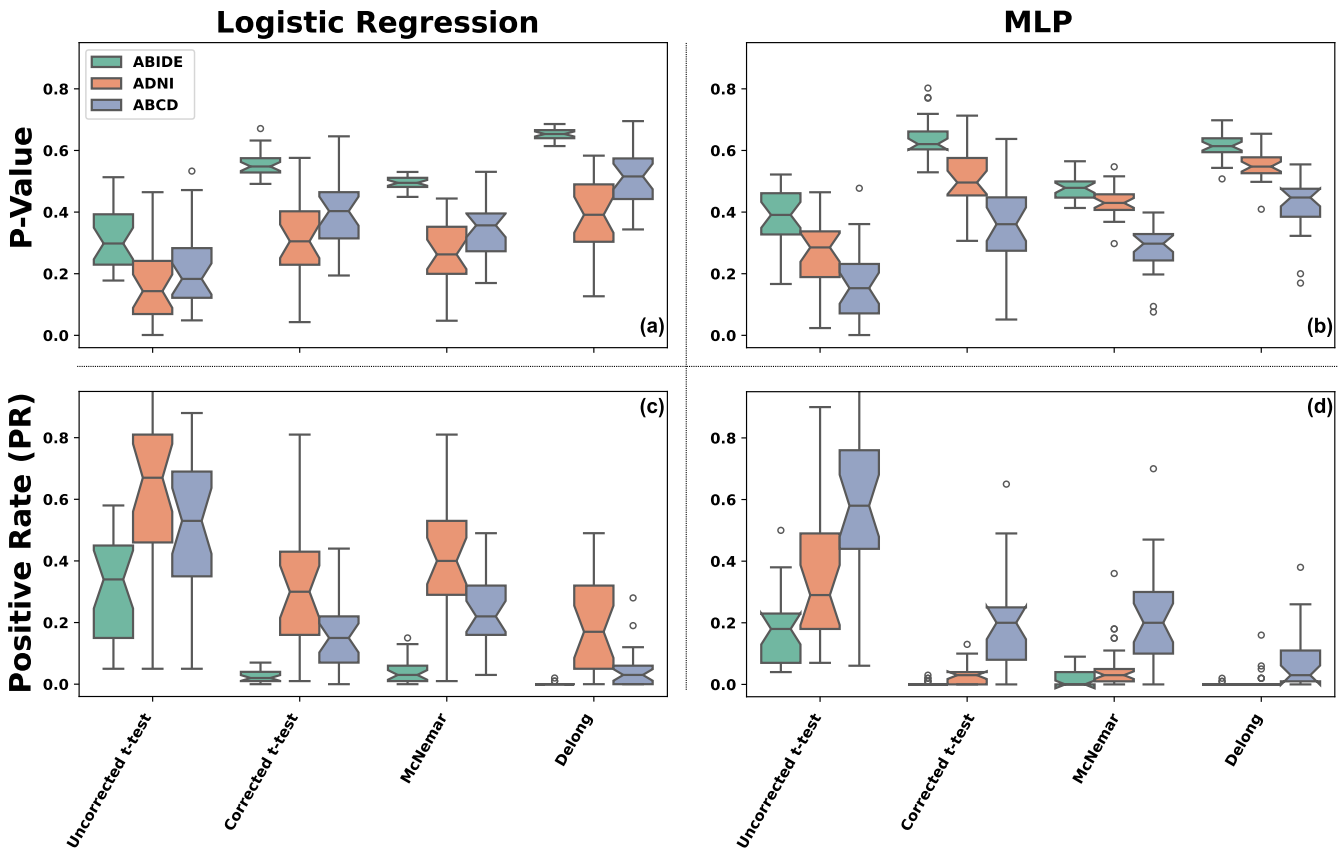

Figure 4. The distribution of p-values and positive rates when applying 4 testing procedures to compare two perturbed models in 3 neuroimaging-based classification tasks.

This figure shows that the average positive rate for the three datasets were not in the same range and also varied among different hypothesis testing procedures. For example, in the ABCD classification task, the difference in the average positive rate between the most sensitive test and the least sensitive test was 0.46 for the Logistic Regression model (Fig. 4c) and 0.51 for the MLP model (Fig. 4d). Another observation from Fig. 4 is that the overall rank of sensitivity among the 4 procedures was the same across the 3 datasets: uncorrected t-test was always the most sensitive procedure (highest positive rate), followed McNemar's test and corrected t-test, and the least sensitive procedure was DeLong's test.

We investigated whether the rank of sensitivity among the 4 testing procedures remained constant or, alternatively, varied with CV setups. To investigate this, we repeated the comparison of the two perturbed Logistic Regression models in ADNI under two perturbation levels. Fig. 5 plots the average p-value over 100 runs for each E,K,M combination, and the shape of each radar plot encodes the rank sensitivity of the 4 procedures. For example, Fig. 5a suggests that when choosing one-time 2-fold CV to compare two models perturbed at level $E = 6$, the rank of sensitivity was the same as in Fig. 4, with uncorrected t-test being the most sensitive procedure (smallest p-values closest to the center) and DeLong's test being the least sensitive (largest p-values farthest away from the center). According to the shape changes of radar plots in Fig. 5, the rank of sensitivity among the 4 test procedures was not constant. The two variants of the t-test were the most sensitive tests (lowest p-value) when $M = 1, K = 50$ (Fig. 5c) but were less sensitive than DeLong's test and McNemar's test under perturbation level $E = 3$ in Fig. 5a,b.

Variability of Test Outcomes in Comparing Different ML Models

Variations in the sensitivity of test procedures based on CV setups may contribute to p-hacking, where one could search through CV setups and testing procedures to pursue statistical significance of accuracy differences between two models with different methodological design. To show this, we estimated the accuracy of 5 classifiers, i.e. Multilayer Perceptron (MLP), Logistic regression (LR), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN), on the three datasets. For each classifier, we evaluated the accuracy using 25 different CV setups (based on different choices of K and M, see Methods for the details). The average classification accuracy for each classifier are shown in Table 1. Apart from KNN consistently achieving the lowest accuracy across all three datasets, the accuracy difference among the remaining 4 classifiers was small,

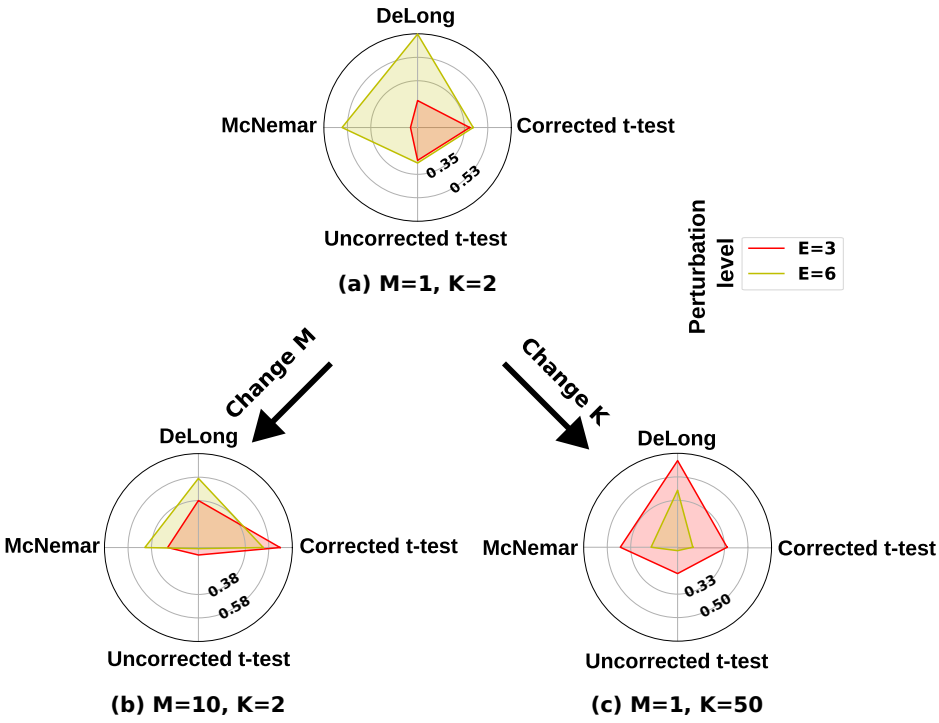

Figure 5. 4 testing procedures were applied to compare two perturbed MLP models for classifying 222 controls and 222 patients from the ADNI dataset. The test was repeated 100 times at two different perturbation levels (E), number of folds (K), and number of CV repetitions (M). Each radar plot records the average p-value over the 100 runs.

with the gap between the most and least accurate classifier being 6%, 3%, and 2%, for ABCD, ABIDE and ADNI datasets, respectively.

Next, we aimed to detect statistical differences in the accuracy of the 5 classifiers. There were 10 pairs of models to be compared. For each pair of models, we conducted 100 comparisons by combining the 4 testing procedures with the 25 CV setups. For each testing procedure, we recorded the positive rate, i.e., percentage of reaching a significance level of $p < 0.05$ out of the 25 CV setups (Fig. 6). Across all methods and datasets, McNemar's test, with an average positive rate of 0.44, was the most sensitive method, followed by the uncorrected t-test at 0.37, the corrected t-test at 0.33, and finally, DeLong's test, was the least sensitive, with an average of 0.30. Critically, the test outcomes significantly varied with CV setups and testing procedures, making it difficult to draw consistent conclusions about whether one classifier was significantly more accurate than another. For example, in the ABCD classification, only the comparison between RF and LR using DeLong's test was consistently insignificant (positive rate = 0). For all other model pairs, there was at least one CV setup and testing procedure combination that resulted in a statistically significant accuracy difference between the two models. However, none of the comparisons were significant for all 100 tests, although the comparison of 3 model pairs (KNN vs. RF, KNN vs. SVM, KNN vs. LR) had a positive rate>0.8 across all 4 testing procedures. These results on ABCD were largely replicated on ABIDE, where the majority of model comparisons showed inconsistent outcomes; i.e., only a subset of CV setups resulted in significant outcomes (0<positive rate<1). Lastly, for the ADNI classification, although the classification accuracy was similar across the 5 classifiers with only a 2% difference between the least accurate (e.g., KNN) and most accurate classifiers (e.g. RF), there were still CV setups and testing procedures that reached statistical significance for model comparisons.

	RF	SVM	KNN	MLP	LR
ABCD	0.68 (0.02)	0.72 (0.01)	0.59 (0.01)	0.66 (0.02)	0.69 (0.01)
ABIDE	0.76 (0.05)	0.79 (0.05)	0.62 (0.05)	0.76 (0.05)	0.79 (0.05)
ADNI	0.80 (0.01)	0.80 (0.01)	0.78 (0.02)	0.79 (0.01)	0.78 (0.01)

Table 1. Mean and standard deviation of classification accuracy of 5 classifiers applied to the 3 datasets based on 25 different CV setups.

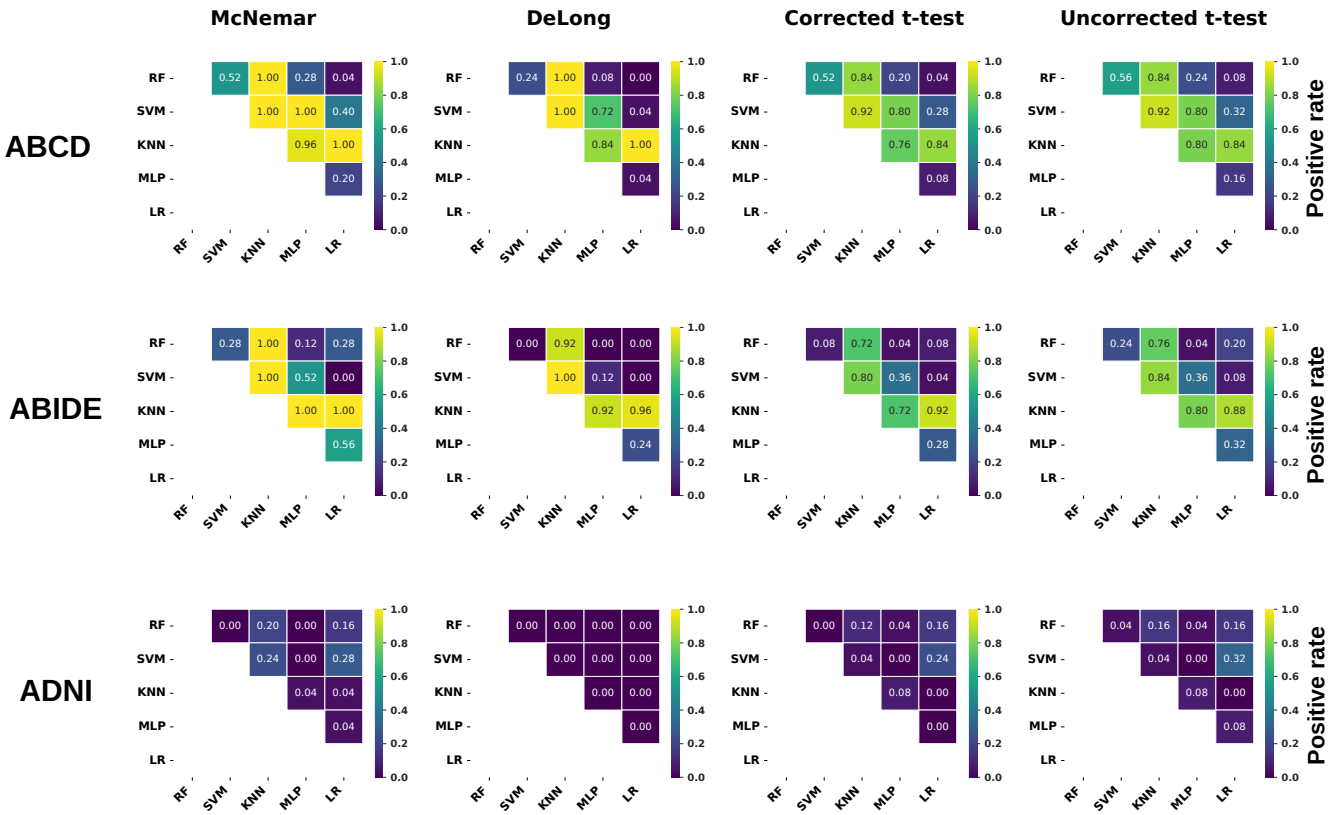

Figure 6. We compared classification accuracy across 5 classifiers on the 3 datasets. For a model pair, we conducted the comparison using 4 testing procedures in 25 different CV setups and recorded the positive rate at the $p < 0.05$ level for each procedure.

Discussion

Based on a novel framework that evaluates the statistical significance of model comparison in three neuroimaging-based classification tasks, we first re-iterated the pitfall that using t-test to compare accuracy scores in CV (repeated or not) could lead to inflated test sensitivity. More importantly, the relative sensitivity among test procedures, including t-tests, McNemar's test, and DeLong's test, was dependent on factors related to CV setup and characteristics of the data. This eventually led to contradictory conclusions when applying different testing procedures and setups to compare the accuracy of classification models in neuroimaging studies.

ML has been increasingly used in neuroimaging research to identify biomarkers of neurological and psychiatric disorders, predict individual differences in behavior and clinical outcomes, and uncover patterns in high-dimensional brain data that are not easily accessible through traditional statistical methods[30–32]. To date, CV remains the predominant evaluation strategy, largely due to the limited sample size in most clinical cohorts and the lack of established benchmark datasets in most neuroimaging studies. However, despite considerable efforts being invested in ML model design, how to rigorously formulate hypothesis tests to compare model performance has been largely neglected in the context of CV. Although commonly regarded as a standard evaluation approach, studies have shown that improper use of CV can lead to issues such as data leakage and inflated accuracy estimates in small-sized samples[33–37]. Moreover, there is no standard setup for CV: a search of keyword "cross-validation" on PubMed returned studies that used every single choice of K (number of folds) from 2 to 10, with many other choices beyond 20. In this study, we particularly illustrate that such variability in CV setups can influence the statistical significance of model comparisons.

Although prior studies have investigated the influence of CV setup on the performance of a single model[35], there has been limited investigation on the influence of CV on the "comparison" between two different models. The difficulty of validating model comparison outcome lies in the fact there is usually no ground truth in knowing which model works better on a particular dataset. Every model has its unique assumption about the data and becomes suboptimal when the assumption is violated. As a result, statistical significance of accuracy difference depends on testing procedures, model choices, and data characteristics. To alleviate this challenge, our proposed framework aimed to disentangle the impact of model/data choices (by comparing two

models with the same algorithm complexity and architectural design), so that the observed variance of statistical significance was only linked to CV setups.

Using this framework, we first re-emphasized a known pitfall that the accuracy scores in CV are not independent samples[22,27,38]. Instead, these scores arise from overlapping testing folds (when $M > 1$) or overlapping training folds (when $K > 2, M \geq 1$) and can have an arbitrary sample size by increasing the number of folds or the number of CV repetition (e.g., Monte-Carlo CV). This will eventually lead to significant accuracy differences by regular t-tests between any two models, even if they are practically identical. Moreover, increasing the number of folds not only inflates the number of accuracy scores, but also increases training set size (potentially improving model accuracy) and decreases test set size (potentially increasing variance of accuracy). These factors jointly influence the characteristics of test statistics in complex ways, making the sensitivity of model comparisons highly dependent on the CV setup. While overlooked by many researchers, there exists active discussions on how to resolve the dependency of test sensitivity on CV setups. One common recommendation is to use a fixed setting 5×2 CV, which was suggested in previous simulation studies to have balanced power and Type I error[22,27,39]. Our analysis, however, indicated the 5×2 was one particular setting from the whole $K \times M$ spectrum and its theoretical advantage over other settings seemed unjustified. Another approach is to use corrected paired t-tests to control for the dependency, but our analysis indicated it did not fundamentally address the problem and still led to substantial variation across test setups. Recently, studies have increasingly used non-parametric approaches[40–43] to quantify the uncertainty of accuracy of a model. However, these approaches are still primarily used for comparing accuracy with a random null model[44,45], and it is unclear how to optimally transfer these approaches to between-model comparison (see Supplement Fig. S1.7 for additional results using bootstrapping).

Based on the proposed framework, our study also indicated that the outcome of McNemar's and DeLong's tests exhibited greater invariant with respect to the number of folds compared to the t-tests. However, the remaining problem was that the rank of sensitivity across the four procedures was not stable, highlighting a persistent challenge. The variability of sensitivity makes it difficult for researchers to choose a testing procedure aligned with their analytic goals – whether prioritizing conservativeness or sensitivity. Critically, our analysis shows that conclusions about model comparisons can vary substantially depending on the chosen testing procedure and CV setup. For any two models, even with a small accuracy gap, there is likely to be a testing procedure and CV setup that yields statistical significance for the accuracy difference. Given that the choice of testing procedure/CV setup does not typically fall into the evaluation criteria of ML-based biomedical studies, this highlights a critical pitfall: the potential for cherry-picking test settings to favor a desired outcome.

Having revealed the problem, the remaining question is what should be the best practices for reducing the variability of test results related to testing procedures and CV setups? We first note that compared to the substantial literature on controlling Type I error in traditional statistical tests, we need more discussion on quantifying model difference in the ML setting. While we do not have a comprehensive solution, here we outline a few directions. When confined to CV on a single dataset, one should at least consistently use McNemar's or DeLong's tests (that are invariant to K) and avoid using different test procedures across studies to allow for comparable outcomes. Given the ongoing debate on the disadvantage of prioritizing reporting p-values in scientific research[46–48], an alternative view is that researchers should shift attention from p-values to actually effect sizes[49,50], which, in the context of model comparison, means a prerequisite threshold of accuracy increase (e.g. 5%) for determining meaningful performance difference. In doing so, one can compare models not only by CV but also based on a single accuracy score from a one-time validation run, e.g., using pre-registration[51] in the Open Science framework or testing on an independent study or a benchmark dataset. This strategy, however, requires sufficient samples in both the train and test data (e.g., in certain epidemiological studies) to ensure the generalizability of the trained model and robustness of the single accuracy score. Another strategy would be considering uncertainty in the model prediction alongside the accuracy. A model can present higher accuracy compared to other models but provide more uncertain predictions. Thus, using criteria such as the negative log-likelihood of the test dataset can help select the most appropriate model[52–56].

Lastly, there are a few limitations in the current analysis. While our findings are likely to generalize to a wide variety of biomedical studies, we focused our analysis on measurements from brain MRI data as the sample size is typically below 1,000, making cross-validation a more appropriate approach to get reliable accuracy estimates of ML models within a single dataset. A systematic validation across all data types and samples sizes is still needed. Moreover, the investigation needs to be further expanded to regression analysis where the evaluation metrics and model comparison approaches are substantially different from classification models. Lastly, our study focused on analyzing statistical tests for model comparison but did not optimize for data preprocessing, feature selection, and architecture/hyperparameter search. As a result, the trained models might not fully capture the discriminative signal in the datasets. The reported accuracy might not be on par with the most accurate models in the literature and should be interpreted with caution.

Methods

Data Description

The Adolescent Brain Cognitive Development (ABCD) Study: The ABCD dataset[26] was used to identify sex based on structural measurements derived from the T1-weighted MRI of 11,725 participants at their baseline visits (9-10 years old, 6125 boys/5600 girls). From the preprocessed data in the ABCD 5.0 release, we used cortical thickness and brain surface area of 34 bilateral regions defined by the Desikan–Killiany (DK)[57] atlas and volumes of 24 subcortical regions (aseg)[58], yielding 92 input features. Additionally, to account for the possible confounding effect of sex differences in head size, we removed the effect of head size from the features using linear regression[59]. This critical adjustment ensures that the observed variations in brain measurements reflect true morphological differences, rather than simply differences in head size.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) Dataset[24]: Baseline data from ADNI 1/2/3/GO was used to distinguish between normal controls and patients with Alzheimer's Disease (AD) based on preprocessed structural measurements from the UCSF Cross-Sectional FreeSurfer Data Release. After removing subjects with missing data, we only kept those that passed the overall quality control by the UCSF Freesurfer pipeline[60], resulting in 1088 controls (573 males with average age of 73.66 and 515 females with average age of 71.93) and 222 AD patients (120 males with average age of 74.60 and 102 females with average age of 73.37). From the 373 brain features capturing cortical thicknesses and volumes of various brain regions, we selected 59 features based on the criteria defined in[61]. Subsequently, features with missing values were removed, resulting in a final set of 51 features (Supplement Table S1.1). Similar to the ABCD experiments, a regression analysis was applied to the neuroimaging features to correct for head size variability.

The Autism Brain Imaging Data Exchange (ABIDE I) Dataset[25]: This dataset was employed to compare 391 individuals with autism spectrum disorders (ASD) to 458 typically developing controls based on resting-state functional MRI released in the ABIDE preprocessed repository (average age: 16.90, 133 males and 716 females). We processed the functional MRI data by creating a functional connectivity (correlation) matrix from the time series extracted from the CC200 atlas using the *cpac_nofilt_noglobal* pipeline[25]. Only the upper triangle of these symmetric matrices was utilized to avoid redundancy. We integrated neuroimaging data with demographic information (age at scan and gender) and applied principal component analysis (PCA) on the resulting 19902 features to reduce the data dimensions to 256.

Synthetic Datasets

We generated two simulated datasets with 10,000 and 100,000 samples to investigate the impact of sample size on the results. Each dataset was generated using the `make_classification` function from scikit-learn[62] library. In each dataset, samples consisted of 100 features, including 20 informative features that were directly predictive of the class label (*n_informative*=20), and 10 redundant features that were linear combinations of the informative ones (*n_redundant*=10). We introduced 5% label noise by randomly flipping a subset of the labels (*flip_y*=0.05), corresponding to a Bayes error of approximately 5%. These datasets allowed us to assess whether the observed results remain consistent with larger sample sizes, thus addressing potential limitations due to small datasets.

Classification Models

K-nearest neighbors (KNN): The number of neighbors was set to 5 in the KNN classifier, which employs a uniform weighting strategy, ensuring that each of the neighbors contributes equally to the voting process.

Logistic regression (LR): The logistic regression was optimized by the efficient L-BFGS solver, using an L2 regularization with its weight set to 1. The optimization runs for a maximum of 1000 iterations to ensure convergence.

Multilayer Perceptron (MLP): We used a two-layer MLP that utilizes a ReLU activation function and dropout regularization to enhance model performance and prevent overfitting. The hidden layer has a dimension of 128, and the dropout rate was 0.5. Another linear layer processes the data for the output stage where a sigmoid activation function was used in the last layer to enable binary classification. The loss function used is binary cross-entropy with logits, and class weights were applied to handle the imbalanced dataset. Additionally, an early stopping mechanism is implemented to optimize training efficiency and stop the training process when no further improvement in validation performance is observed. In all experiments, the learning rate is set to 0.001, the maximum number of epochs and batch sizes are 400, and 100, respectively.

Random Forest (RF): The RF model was configured with 100 trees, utilizing the Gini impurity criterion, which maximize information gain when splitting nodes. The model employs bootstrapping to enhance the diversity and robustness of the individual trees, thereby improving the overall performance and reducing overfitting.

Support Vector Classifier (SVC): This model employs the Support Vector Machine (SVM) framework, utilizing the Radial Basis Function (RBF) kernel to classify data. The SVC is configured with a regularization parameter of 1, and the kernel coefficient is determined to be the inverse of the number of features.

The Multilayer Perceptron model was implemented using the PyTorch[63] package, while the rest of the models were implemented using the scikit-learn[64] package in Python. Unless otherwise specified, default parameter values were used for both frameworks. The versions used were PyTorch 1.10.0 and scikit-learn 0.24.2.

Cross-Validation Setups for Model Comparison

To conduct model comparison among the 5 classifiers in Fig. 6, we used the same sample size as in Fig. 2, i.e., $N = 1000$ for the ABCD dataset, $N = 444$ for ADNI, and $N = 600$ for ABIDE. Comparison of any two models was repeated with 25 different CV setups, with the number of folds in CV set to 2, 5, 10, 25 and 50 folds, and the number of repetitions of CV set to 1, 2, 4, 6 and 10.

Data Availability

The ADNI data are public and shared through the LONI Image and Data Archive (IDA) https://ida.loni.usc.edu under the contingency on adherence to the ADNI Data Use Agreement. To access the data, one would need to apply through https://adni.loni.usc.edu/data-samples/adni-data/. We used the tabular data, UCSF - Cross-Sectional FreeSurfer (5.1) [ADNI1,GO,2] and UCSF - Cross-Sectional FreeSurfer (6.0) [ADNI3]. Subject and measurement selection is included in the Github code https://github.com/BahramJafrasteh/classifier_test.

The ABCD data are public and shared under authenticated access by The National Institute of Mental Health Data Archive (NDA). To access the data, one needs to visit https://nda.nih.gov/study.html?id=2147 and select the "ABCD 5.0 Tabulated Release Data" file in the Results section to download the data. We used the tabular data, mri_y_smr_thk_dsk.csv, mri_y_smr_area_dsk.csv, mri_y_smr_vol_aseg.csv, and abcd_p_demo.csv. Subject and measurement selection is included in the Github code. The preprocessed ABIDE data are freely available at http://preprocessed-connectomes-project.org/abide/. Our Github code automatically downloads the data and runs model training and testing.

References

1. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. medicine* **25**, 44–56 (2019).
2. He, J. *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nat. medicine* **25**, 30–36 (2019).
3. Ghassemi, M. *et al.* A review of challenges and opportunities in machine learning for health. *AMIA Summits on Transl. Sci. Proc.* **2020**, 191 (2020).
4. Davatzikos, C. Machine learning in neuroimaging: Progress and challenges. *NeuroImage* **197**, DOI: 10.1016/j.neuroimage.2018.10.003 (2018).
5. Makridakis, S., Spiliotis, E. & Assimakopoulos, V. Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one* **13**, e0194889 (2018).
6. Reddy, S., Fox, J. & Purohit, M. P. Artificial intelligence-enabled healthcare delivery. *J. Royal Soc. Medicine* **112**, 22–28 (2019).
7. Poalelungi, D. G. *et al.* Advancing patient care: how artificial intelligence is transforming healthcare. *J. personalized medicine* **13**, 1214 (2023).
8. Liu, X. *et al.* Advances in deep learning-based medical image analysis. *Heal. Data Sci.* **2021** (2021).
9. Jiang, F. *et al.* Artificial intelligence in healthcare: past, present and future. *Stroke vascular neurology* **2** (2017).
10. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533** (2016).
11. Gibney, E. Is ai fuelling a reproducibility crisis in science. *Nature* **608**, 250–1 (2022).
12. Youssef, A. *et al.* External validation of ai models in health should be replaced with recurring local validation. *Nat. Medicine* **29**, 2686–2687 (2023).
13. Yang, J., Soltan, A. A. & Clifton, D. A. Machine learning generalizability across healthcare settings: insights from multi-site covid-19 screening. *NPJ digital medicine* **5**, 69 (2022).
14. Barish, M., Bolourani, S., Lau, L. F., Shah, S. & Zanos, T. P. External validation demonstrates limited clinical utility of the interpretable mortality prediction model for patients with covid-19. *Nat. Mach. Intell.* **3**, 25–27 (2021).
15. Yu, A. C., Mohajer, B. & Eng, J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiol. Artif. Intell.* **4**, e210064 (2022).

16. 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 testing* **2**, 249–262 (2018).

17. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. *PloS one* **14**, e0224365 (2019).

18. Feng, J. *et al.* Sequential algorithmic modification with test data reuse. In *Uncertainty in Artificial Intelligence*, 674–684 (PMLR, 2022).

19. Wong, T.-T. & Yeh, P.-Y. Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowl. Data Eng.* **32**, 1586–1594, DOI: 10.1109/TKDE.2019.2912815 (2020).

20. Bradshaw, T. J., Huemann, Z., Hu, J. & Rahmim, A. A guide to cross-validation for artificial intelligence in medical imaging. *Radiol. Artif. Intell.* **5**, e220232 (2023).

21. Bausell, R. B. *The problem with science: the reproducibility crisis and what to do about it* (Oxford University Press, 2021).

22. Bouckaert, R. R. & Frank, E. Evaluating the replicability of significance tests for comparing learning algorithms. In Dai, H., Srikant, R. & Zhang, C. (eds.) *Advances in Knowledge Discovery and Data Mining*, 3–12 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2004).

23. Abrol, A. *et al.* Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nat. Commun.* **12**, DOI: 10.1038/s41467-020-20655-6 (2021).

24. Petersen, R. C. *et al.* Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. vol. 74, 201–209 (AAN Enterprises, 2010).

25. Craddock, C. *et al.* The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Front. Neuroinformatics* **7**, 5 (2013).

26. Casey, B. J. *et al.* The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Dev. cognitive neuroscience* **32**, 43–54 (2018).

27. Nadeau, C. & Bengio, Y. Inference for the generalization error. *Adv. neural information processing systems* **12** (1999).

28. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157, DOI: 10.1007/bf02295996 (1947).

29. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 837–845 (1988).

30. Abrol, A. *et al.* Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nat. communications* **12**, 353 (2021).

31. Yassin, W. *et al.* Machine-learning classification using neuroimaging data in schizophrenia, autism, ultra-high risk and first-episode psychosis. *Transl. psychiatry* **10**, 278 (2020).

32. Tejavibulya, L. *et al.* Predicting the future of neuroimaging predictive models in mental health. *Mol. psychiatry* **27**, 3129–3137 (2022).

33. Bates, S., Hastie, T. & Tibshirani, R. Cross-validation: what does it estimate and how well does it do it? *J. Am. Stat. Assoc.* **119**, 1434–1445 (2024).

34. Krstajic, D., Buturovic, L. J., Leahy, D. E. & Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. cheminformatics* **6**, 1–15 (2014).

35. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage* **180**, 68–77 (2018).

36. Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S. & Scheinost, D. Data leakage inflates prediction performance in connectome-based machine learning models. *Nat. Commun.* **15**, 1829 (2024).

37. Hosseini, M. *et al.* I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neurosci. & Biobehav. Rev.* **119**, 456–467 (2020).

38. Wong, T.-T. & Yang, N.-Y. Dependency analysis of accuracy estimates in k-fold cross validation. *IEEE Transactions on Knowl. Data Eng.* **29**, 2417–2427 (2017).

39. Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* **10**, 1895–1923 (1998).

40. Efron, B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. statistical association* **78**, 316–331 (1983).

41. Michelucci, U. & Venturini, F. Estimating neural network's performance with bootstrap: A tutorial. *Mach. Learn. Knowl. Extr.* **3**, 357–373 (2021).

42. Tsamardinos, I., Greasidou, E. & Borboudakis, G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Mach. learning* **107**, 1895–1922 (2018).

43. Janssen, A. & Pauls, T. How do bootstrap and permutation tests work? *The Annals statistics* **31**, 768–806 (2003).

44. Parkes, L. *et al.* Transdiagnostic dimensions of psychopathology explain individuals' unique deviations from normative neurodevelopment in brain structure. *Transl. psychiatry* **11**, 232 (2021).

45. Dhamala, E., Jamison, K. W., Jaywant, A., Dennis, S. & Kuceyeski, A. Distinct functional and structural connections predict crystallised and fluid cognition in healthy adults. *Hum. brain mapping* **42**, 3102–3118 (2021).

46. Amrhein, V., Greenland, S. & McShane, B. Scientists rise up against statistical significance. *Nature* **567**, 305–307 (2019).

47. Di Leo, G. & Sardanelli, F. Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. *Eur. radiology experimental* **4**, 1–8 (2020).

48. Wasserstein, R. L. & Lazar, N. A. The asa statement on p-values: context, process, and purpose (2016).

49. Rajput, D., Wang, W.-J. & Chen, C.-C. Evaluation of a decided sample size in machine learning applications. *BMC bioinformatics* **24**, 48 (2023).

50. Duffy-Deno, K. The curse of big data.

51. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl. Acad. Sci.* **115**, 2600–2606, DOI: 10.1073/pnas.1708274114 (2018).

52. Jafrasteh, B., Villacampa-Calvo, C. & Hernandez-Lobato, D. Input dependent sparse Gaussian processes. In *Proceedings of the 39th International Conference on Machine Learning*, 9739–9759 (PMLR, 2022).

53. Ding, J., Tarokh, V. & Yang, Y. Model selection techniques: An overview. *IEEE Signal Process. Mag.* **35**, 16–34 (2018).

54. Ding, Y., Liu, J., Xiong, J. & Shi, Y. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 4–5 (2020).

55. Rust, R. T., Simester, D., Brodie, R. J. & Nilikant, V. Model selection criteria: An investigation of relative accuracy, posterior probabilities, and combinations of criteria. *Manag. science* **41**, 322–333 (1995).

56. Zou, K. *et al.* A review of uncertainty estimation and its application in medical imaging. *Meta-Radiology* 100003 (2023).

57. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage* **31**, 968–980 (2006).

58. Fischl, B. *et al.* Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**, 341–355 (2002).

59. Adeli, E. *et al.* Chained regularization for identifying brain patterns specific to hiv infection. *Neuroimage* **183**, 425–437 (2018).

60. Hartig, M. *et al.* Ucsf freesurfer methods. *ADNI Alzheimers Dis. Neuroimaging Initiative: San Francisco, CA, USA* **5** (2014).

61. Seo, K., Pan, R., Lee, D., Thiyyagura, P. & Chen, K. Visualizing alzheimer's disease progression in low dimensional manifolds. heliyon 5 (8): e02216 (2019).

62. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. machine Learn. research* **12**, 2825–2830 (2011).

63. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).

64. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

The work was partly supported by the National Institute of Health (AA028840 (QZ), R61AG084471 (EA), DA057567 (KMP), AA021697 (KMP)), BBRF Young Investigator Grant, the Stanford University Jaswa Innovator Award, the 2024 Stanford HAI Hoffman-Yee Grant, the Stanford HAI Google Cloud Credit, and the DGIST Joint Research Project.

Author contributions statement

BJ and QZ carried out data preprocessing and planned the experiments. QZ, BJ, EA, KMP, AK and MRS contributed in writing the manuscript, advanced conception and design of the study, all authors contributed to the final version of the manuscript.

Additional information

The authors declare that they have no competing interests. All authors report no biomedical financial interests or potential conflicts of interest.