Empirical Comparison between Cross-Validation and Mutation-Validation in Model Selection

Jinyang Yu[1[0009−0008−3041−8690]], Sami Hamdan[1,2[0000−0001−5072−542X]], Leonard Sasse[1,2,5[0000−0002−2400−3404]], Abigail Morrison[3,4[0000−0001−6933−797X]], and Kaustubh R. Patil[1,2[0000−0002−0289−5480]]

[1] Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Center Jülich, Jülich, Germany
[2] Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[3] Institute for Neurosciennce and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6), Research Center Jülich, Jülich, Germany
[4] Department of Computer Science 3 - Software Engineering, RWTH Aachen University, Aachen, Germany
[5] Max Planck School of Cognition, Stephanstrasse 1a, Leipzig, Germany

Abstract. Mutation validation (MV) is a recently proposed approach for model selection, garnering significant interest due to its unique characteristics and potential benefits compared to the widely used cross-validation (CV) method. In this study, we empirically compared MV and k-fold CV using benchmark and real-world datasets. By employing Bayesian tests, we compared generalization estimates yielding three posterior probabilities: practical equivalence, CV superiority, and MV superiority. We also evaluated the differences in the capacity of the selected models and computational efficiency. We found that both MV and CV select models with practically equivalent generalization performance across various machine learning algorithms and the majority of benchmark datasets. MV exhibited advantages in terms of selecting simpler models and lower computational costs. However, in some cases MV selected overly simplistic models leading to underfitting and showed instability in hyperparameter selection. These limitations of MV became more evident in the evaluation of a real-world neuroscientific task of predicting sex at birth using brain functional connectivity.

Keywords: model selection · mutation validation · cross-validation.

1 Introduction and Related Work

The model selection process aims to find a model from a pool of candidate models, taking into account a variety of performance criteria encompassing predictive accuracy and computational efficiency [12,9]. The estimated model generalization error, which represents the expected error on unseen data, is a commonly used criterion for model selection. Generalization error can be empirically estimated using resampling techniques like cross-validation (CV) [9]. Notably, while

holdout validation is particularly effective when a wealth of data is available, CV emerges as the preferred method when dealing with limited data [10]. Nonetheless, it is imperative to acknowledge that CV, especially the commonly used k-fold CV, can be computationally intensive due to the necessity of fitting multiple models. Additionally, research has demonstrated that CV, which relies on excessively reusing the validation set, might lead to overfitting [7].

To address these challenges, Zhang et al. (2023) introduced a novel model selection approach that does not partition the dataset. Unlike CV, this method uses the entire dataset for training and validation while injecting noise into the fitting process. This noise is generated by mutating the sample labels, hence the name mutation validation (MV). While an over-complex classifier with a large capacity can generate a flexible decision boundary resulting in high accuracies on both original and mutated data, an over-simple classifier shows poor performance on both. A good classifier, on the other hand, can detect the ground-truth pattern despite the noise in the training labels and thus should show good performance on the original data and perform worse on the mutated data. This behavior is the intuition behind MV [18].

Zhang et al. [18] conducted extensive experiments on a broad range of datasets. They demonstrated that MV consistently and effectively captured underlying data patterns, thereby offering successful recommendations for the most suitable machine learning algorithm. In contrast, CV occasionally struggled to deliver comparable results. Further experiments showed MV's consistent preference for less complex models, when the algorithms were configured with specific capacity-related hyperparameters. While Zhang et al. provided valuable insights into the MV method, they did not assess the generalization performance of models selected by MV using nested-CV, which is commonly used in many application domains. Our study aimed to conduct a comprehensive comparison of the generalization estimates of the selected models, seeking to fully evaluate and understand MV's performance in contrast to the more commonly used CV method on further benchmark and real-world datasets. It is worth noting that none of our benchmark datasets were previously utilized by Zhang et al. Additionally, we employed Bayesian inference to support our analysis, a step not taken by Zhang et al. Furthermore, our study delved into differences in runtime (energy consumption), expanding the evaluation beyond merely generalization performance.

In summary, our study provides insights into strengths and limitations of MV with a basis on generalization performance, which can aid machine learning researchers and practitioners in making informed decisions regarding the trade-offs associated with those model selection approaches.

2 Methods and Experimental Setup

CV and MV. We implemented CV in a standard way such that the samples were randomly split in equally sized folds and each fold was used as the test set once.

As MV is a relatively new method, we provide an overview here for completeness. MV injects noise in the labels to generate mutated data. The generation of noise is a crucial aspect of MV. To illustrate the mechanism of label mutation, let's consider a binary classification problem. First, a class list is created, encompassing all unique class labels, i.e., 0, 1. Then, η proportion of the original labels are randomly selected and swapped, resulting in mutated labels. That is, labels of class 0 are replaced with 1, class 1 is replaced with 0. We used the recommended value of $\eta = 0.2$ [18].

Two models f and f_η are then trained, on the original training data S and on the mutated training data S_η, respectively. The performance difference between the models provides information regarding model complexity which is then used for model selection. We use $\hat{T}_S(f)$ to refer to the accuracy of model f on the original training data S, $\hat{T}_{S_\eta}(f_\eta)$ for the accuracy of model f_η on the mutated training data S_η, and $\hat{T}_S(f\eta)$ for the accuracy of model f_η on the original training data S. An empirical scoring metric m is used to assess model performance based on the theoretical metamorphic relation in MV [18]:

$$m = (1 - 2\eta)\hat{T}_S(f_\eta) + \hat{T}_S(f) - \hat{T}_{S_\eta}(f_\eta) + \eta$$

The score m aims to capture the changes in training accuracies before and after mutation, forming the foundation of MV model selection. For a good classifier, the performance difference $(\hat{T}_S(f) - \hat{T}_{S_\eta}(f_\eta))$ is expected to be substantial, and the accuracy $\hat{T}_S(f_\eta)$ to be high, resulting in a high score m. Conversely, an over-complex classifier, characterized by a small performance difference and a low accuracy $\hat{T}_S(f_\eta)$, the score m is expected to be low. An over-simple classifier is also anticipated to yield a low m score as both models will perform poorly. Finally, the model with the highest m is selected.

Datasets. In this investigation, our emphasis was on binary classification problems. We utilized 12 benchmark datasets sourced from the OpenML platform and the UCI repository [16,5] (Table 1).

The brain functional magnetic resonance imaging (fMRI) datasets were taken from the Amsterdam Open MRI Collection (AOMIC) [15]. The collection comprises three datasets: ID1000, PIOP1, and PIOP2 (Table 2). We used the functional MRI data from all three datasets: PIOP1 and PIOP2 obtained from resting-state task, while ID1000 based on a movie-watching task.

For each of the fMRI datasets, the functional connectivity (FC) representing synchrony between brain regions across time was extracted. We employed standard preprocessing steps, including motion correction and registration to Montreal Neurological Institute (MNI) space with the fMRIPrep pipeline [6], denoising and feature extraction with xcpEngine [1]. The parcellation of the processed fMRI images was carried out using the Schaefer 100 parcellation scheme which partitions the whole brain into one hundred non-overlapping parcels [14], resulting in 100 time series (1 per brain region). Finally, the FC was calculated as Pearson's correlation coefficients between the time series of all pairs of brain regions. The lower triangle of this symmetrical matrix was vectorized resulting in

Table 1: Overview of the benchmark datasets. The datasets obtained from UCI are marked with an asterisk (*).

Index	Dataset name	Number of instances	Number of instances with label 0	Number of instances with label 1	Number of features
1	mfeat-fourier	2000	200	1800	76
2*	autism-screening	609	180	429	92
3	mfeat-karhunen	2000	200	1800	64
4	mammography	11183	260	10923	6
5	letter	20000	813	19187	16
6	satellite	5100	75	5025	36
7	fri-c2-1000-10	1000	420	580	10
8	segment	2310	330	1980	19
9*	sonar	208	97	111	60
10	qsar-biodeg	1055	356	699	41
11*	early-stage-diabetes-risk	520	200	320	16
12	ozone-level-8hr	2534	160	2374	72

4950 features. This process was done for all participants (i.e. samples) resulting in 2-dimensional tabular data.

All benchmark datasets and FC datasets in our study are presented in a tabular format, with features and a target associated with each sample. Specifically, the target variable of FC datasets was binary labels with female (F) and male (M) according to the participant's sex assigned at birth, an important task for basic and applied neuroscience [17].

Table 2: Overview of the FC datasets.

Dataset	ID1000	PIOP1	PIOP2
Subjects	764	158	186
Target	382 (F) / 382(M)	79 (F) / 79 (M)	93 (F) / 93 (M)
Features	4950	4950	4950
Age min	19	18.25	18.25
Age max	26	26.25	25.75
Age mean	22.862	22.081	21.958
Age std	1.713	1.809	1.787

Machine Learning Algorithms. Machine learning algorithms often have a set of hyperparameters which need to be adjusted during the learning process. In many cases this results in a set of candidate models with different capacity. The capacity of a model can be seen as a measure of its ability to capture the complex relationships present in the underlying data pattern [2]. Following the principle of Occam's razor, when dealing with multiple models exhibiting similar performance, a preference is given to simpler models, namely those with smaller

capacities [13]. Hence, comparing model capacity can provide crucial information regarding the behavior of the model selection methods.

Specific algorithms, such as decision trees (DT), are associated with capacity-related hyperparameters, meaning that the resulting model's capacity heavily relies on the specific hyperparameter configuration. For instance, trees with higher depth can be considered more complex. Similarly, support vector machines (SVM) and Kernel Ridge Classifiers (KRC), when configured with the polynomial kernel, namely the polynomial SVM and polynomial KRC, also exemplify this characteristic. Additionally, multi-layer perceptrons (MLP) with various dropout rates are well suited to investigate model capacity. Considering these factors and the findings of Zhang et al., we designed experiments involving the above algorithms to examine different capacity-related hyperparameters. Specifically, DT involves the hyperparameter of maximum depth, with a range from 1 to 30. For MLP, the dropout rate serves as the hyperparameter, varying between 0.2 and 0.8. In the case of Polynomial SVM and Polynomial KRC, the hyperparameter is the polynomial degree, spanning from 1 to 15.

Bayesian Analysis. Considering the importance of properly comparing generalization performance in model selection, the comparison of model capacity and computational efficiency were based on Bayesian probabilistic analysis. Bayesian analysis provides a valuable alternative to traditional Null Hypothesis Significance Testing (NHST), offering potentially richer and more informative insights [11]. The Bayesian framework involves modeling the posterior probability distribution across the parameter space based on the observed data.

To illustrate, in situations involving two groups of data with equal length, a difference vector is computed, and the posterior probability distribution of the parameter, represented as the mean difference and denoted as μ, is subsequently formulated. Crucially, in this approach it is feasible to accept the null value of the evaluated parameter by setting the Region of Practical Equivalence (ROPE). This region encompasses parameter values that are considered practically indistinguishable from the null value. The size of the ROPE is determined according to the characteristics of the applications. For our empirical comparison of the two validation methods, the ROPE was set to [-0.025, 0.025], corresponding to a 5% difference.

In our study, the comparison framework was applied to each dataset, resulting in a difference vector derived from the two sets of scores (CV and MV). Two variations of the Bayesian paradigm were employed to evaluate the performance of CV and MV. The Bayesian correlated t-test for a single dataset [3] offered a probabilistic comparison on each dataset using the difference vector. Additionally, the Bayesian hierarchical test for multiple datasets [4] provided a final estimation based on the concatenated difference vectors from all datasets. The Bayesian analysis yields three posterior probabilities:

1. P_{CV}: the probability that the model selected by CV outperforms the model selected by MV;
2. $P_{\text{P.E.}}$: the probability that both models selected by CV and MV perform practically equivalent;

3. P_{MV}: the probability that the model selected by MV outperforms the model selected by CV.

Comparison Framework. It is important to define an appropriate scheme to allow for one-to-one comparison between CV and MV. We devised a framework based on nested cross-validation, which is particularly suitable for real-world datasets with a limited sample size. The inner loop is utilized for model selection, e.g. by selecting particular hyperparameters, while the outer loop is employed to evaluate the generalization performance of the selected model. We used a ten-times repeated 10-fold CV following previous recommendation to obtain reliable and robust estimates [3]. This generated three types of results:

1. 100 validation scores for the chosen models from the outer iterations;
2. 100 hyperparameter settings of the corresponding selected models;
3. the final model with the best hyperparameter setting, linked to the highest validation score, trained on the entire dataset.

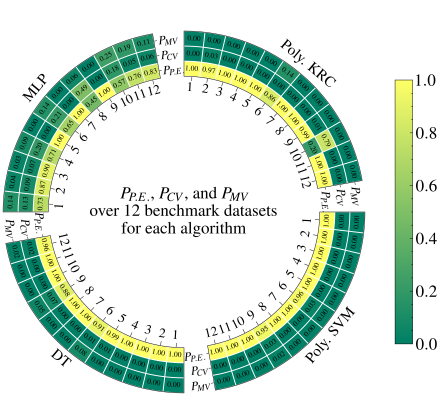
In the assessment of nested cross-validation results, we considered two primary aspects. First, we compared the generalization estimates through the application of Bayesian tests. Second, guided by the probabilities obtained from these tests, we compared two quantities; (1) model performance and capacity, as signified by the selected hyperparameter configurations, and (2) evaluation of computational efficiency, encompassing runtime and CO_2 emissions.

3 Results

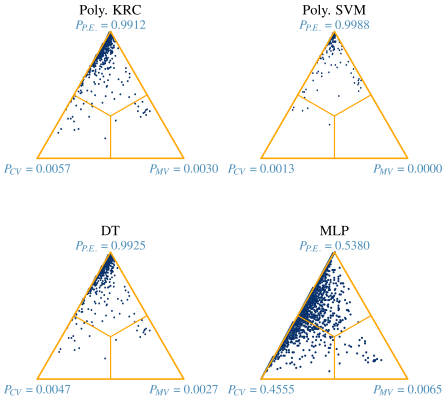
Comparison of Model Performance and Capacity. The comparative analysis of generalization performance, conducted through Bayesian correlated t-tests, showed that the models selected by both CV and MV exhibit practically equivalent performance (Fig. 1a, highlighted by inner cells in bright yellow), which is also confirmed by the results from the Bayesian hierarchical test (Fig. 1b).

This observed pattern holds across all three machine learning algorithms: polynomial KRC, polynomial SVM, and DT. On the majority of benchmark datasets examined, probabilities exceeding 90% affirm the practical equivalence between the two methods. In the case of MLP, it's worth noting that while certain specific datasets exhibit lower posterior probabilities regarding practical equivalence compared to the other three algorithms, the overarching trend still leans toward practical equivalence, remaining notably above chance level.

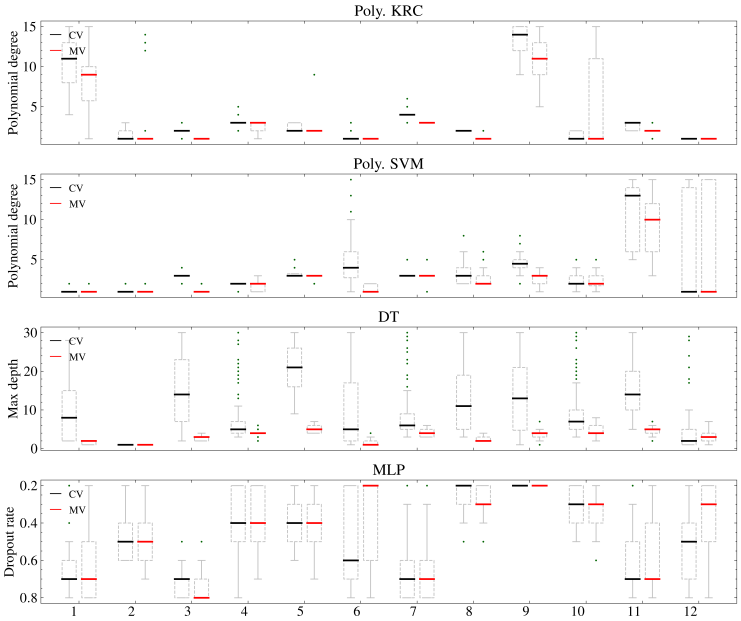
We compared the model capacities selected by both methods, as shown in Fig. 1c. A lower median line of the boxplots was interpreted as smaller model capacity. For KRC and SVM, MV-selected models exhibited similar or slightly lower median polynomial degrees. The interquartile range was small and comparable for both methods. In the case of DT, MV consistently chose models with notably lower maximum depths compared to CV, resulting in a more uniform and less varied selection. For MLP, the dropout rates chosen by MV were largely



(a) Bayesian correlated t-tests for the 12 benchmark datasets



(b) Bayesian hierarchical test for multiple datasets

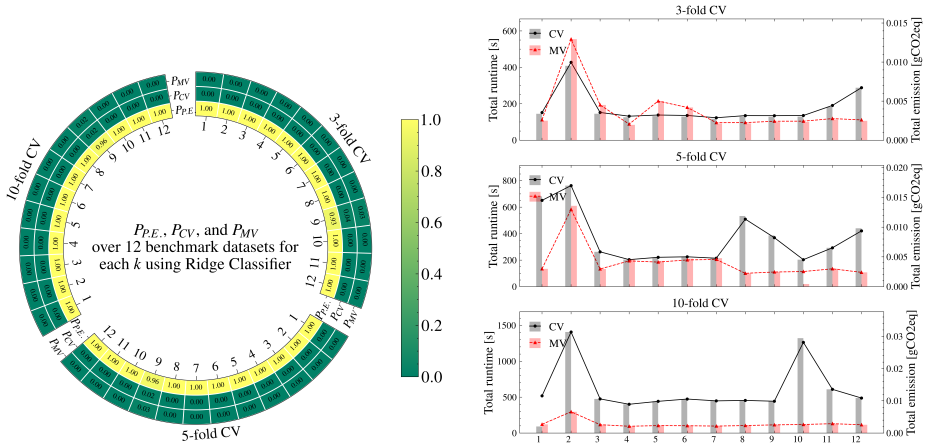


(c) Hyperparameter selection

Fig. 1: An overview of four algorithms evaluated on 12 benchmark datasets. Each subfigure consists of four sectors, one for each algorithm, with dataset indices cross-referenced in Table 1. (a) Each sector displays three tracks representing the posterior probabilities $P_{\text{P.E.}}$, P_{CV}, and P_{MV} for each case. These probabilities are presented as a heat-map. (b) The points represent samples drawn from the posterior probability distribution of 4000 samplings (default setting [4]). The final posterior probabilities $P_{\text{P.E.}}$, P_{CV}, and P_{MV} are located in the corners of each sector. (c) For each algorithm, boxplots indicate results obtained from the top 100 hyperparameter values generated by the comparison framework. Note that the dropout rate in the last sector is displayed on an inverted vertical axis, inline with the interpretation of capacity across all four sectors.

equivalent to those by CV. In summary, MV consistently favored models with lower capacity. Given the practical equivalence in performance between models selected by both methods, this suggests a preference for models determined by MV.

Comparison of Computational Efficiency. To achieve a balance between bias and variance, Kohavi [10] suggests using $k = 10$ folds for CV. However, selecting an appropriate value for k is not trivial. For instance, starting from version 0.22, the default value used by the `scikit-learn`[1] library was changed to $k = 5$ from previous $k = 3$. Hence, to analyse the effect of different values of k on CV and MV, we compared them using $k = 3, 5, 10$.



(a) Bayesian correlated t-tests for the 12 benchmark datasets

(b) Runtime and emission analysis

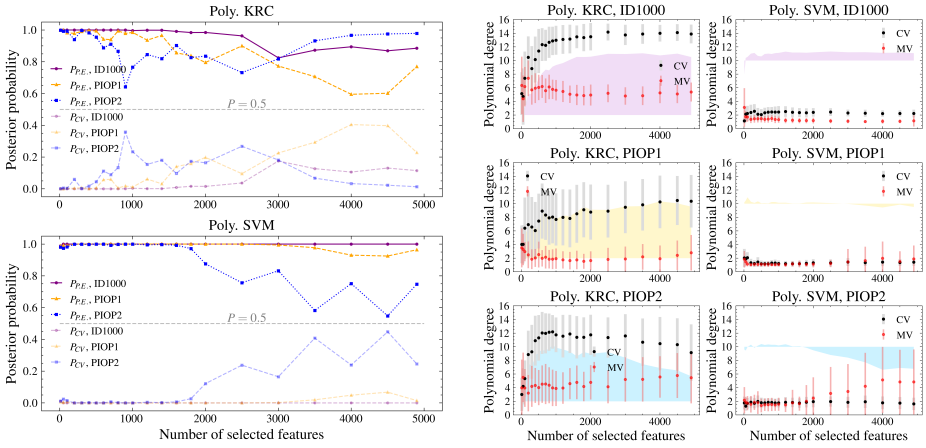
Fig. 2: The three sectors of each subfigure correspond to k = 3-, k = 5-, and k = 10-CV. (a) This subfigure contains results obtained from Bayesian correlated t-test across the 12 benchmark datasets. The indices of the datasets are listed in Table 1. In each sector, there are three tracks, representing the three posterior probabilities $P_{\text{P.E.}}$, P_{CV}, and P_{MV}. (b) The results from CV are shown in black and those from MV are shown in red. In each sector, the horizontal axis lists the indices of the benchmark datasets. The left vertical axis shows the total runtime of the procedure, and the right vertical axis shows the equivalent CO_2 emission.

Here we applied the algorithms to the benchmark datasets, KRC is presented as an example (Fig. 2a and Fig. 2b). The probability of $P_{\text{P.E.}}$ approaches 1, indicating practical equivalence in generalization performance between the two selection methods (Fig. 2a). The findings lead to further comparisons in computational efficiency. Overall, the computational efficiency of CV with varying k

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html`

compared to MV yielded a consistent pattern (Fig. 2b). With $k = 3$, the performance of both methods is comparable. However, as k increases to 5, MV shows higher efficiency than CV. Finally, at $k = 10$, MV demonstrated a noticeable advantage over CV in terms of efficiency and carbon emission.

Comparison on Brain FC Datasets. The brain FC datasets, like many other real-world data, contain numerous features (in our case 4950 Pearson's correlation coefficients per sample). In this context, feature selection can be a desirable preprocessing step [8]. Besides, the three FC datasets used differ largely in sample size (Table 2). In particular, ID1000 has over four times more samples than PIOP1 and PIOP2.



(a) Bayesian correlated t-tests for FC datasets (ID1000, PIOP1, and PIOP2)

(b) Hyperparameter selection

Fig. 3: (a) The Bayesian correlated t-test was used to calculate $P_{\text{P.E.}}$ and P_{CV} across subsets of the FC domain with varying numbers of selected best features. The above sector shows the results obtained from polynomial KRC, while the below sector displays those obtained from polynomial SVM. The probability curves in purple, yellow, and blue correspond to the datasets ID1000, PIOP1, and PIOP2, respectively. (b) The mean of the 100 best polynomial degrees across subsets of the FC domain ID1000, PIOP1, and PIOP2 for the polynomial KRC and SVM algorithms. Each point in the plot represents the mean of the polynomial degrees and the error bars demonstrate the standard deviation. The shaded areas in each sector shows the difference between the mean polynomial degrees from CV and MV.

A commonly used feature selection method is calculation of F-scores using an ANOVA, as provided by the `SelectKBest` in `scikit-learn`[2] which selects

[2] `https://scikit-learn.org/stable/modules/feature_selection.html`

the top K informative features corresponding to K highest F-scores. We generated several subsets of the FC datasets, each by selecting a different number of important features. This allowed us to compare CV and MV in a controlled manner on a large number of datasets each with different number of informative features. Each FC dataset was analyzed separately.

Two popular kernelized algorithms in neuroscience were investigated in this experiment. The probability values of polynomial KRC and polynomial SVM across the range of K (from 0 to 4950) are illustrated in Fig. 3a, respectively. Notably, the trend reflected a consistent linear decline in $P_{\text{P.E.}}$ in four cases, presenting a diminishing level of confidence in the practical equivalence between the two methods. This decline is mirrored by the probability P_{CV}, indicating a shift towards better performance of CV associated with generalization estimates as the number of features increased.

Exploring the relative relationship between the hyperparameters selected by CV and MV can aid in a deeper interpretation of these results. On the ID1000 dataset, the mean polynomial degree selected by CV was higher than that of MV for both cases (Fig. 3b, Poly. KRC, ID1000 and Poly. SVM, ID1000). The variance of the polynomial degrees selected by MV was generally lower or similar to that of CV, indicating more stable model selection by MV. Overall, MV selected models with lower complexity, consistent with the observations and results obtained on the benchmark datasets.

The results on the PIOP1 dataset were similar to those on ID1000 (Fig. 3b, Poly. KRC, PIOP1). Generally, the models selected by MV were less complex than those selected by CV. The relative performance of MV worsened as indicated by declining $P_{\text{P.E.}}$ (Fig. 3a, Poly. KRC) which might suggest that MV might tend to penalize complex models excessively and may encounter some level of underfitting. The mean hyperparameter values selected by both methods were similar (Fig. 3b, Poly. SVM, PIOP1). However, here MV displayed a more substantial variance, implying potential instability in the tuning process.

The PIOP1 dataset is challenging due to its limited sample size, a problem that PIOP2 also shares. However, the results on the PIOP2 dataset exposed further inconsistencies. As the number of features increased, MV selected models with higher capacity and $P_{\text{P.E.}}$ decreased (Fig. 3b, Poly. SVM, PIOP2). Furthermore, the variance of MV also showed an increase. In this specific instance, MV appears to have forfeited all of its advantages and performed worse than CV.

4 Discussion and Conclusion

Our systematic evaluation of generalization estimates involved the use of Bayesian correlated t-tests and Bayesian hierarchical tests, revealing that CV and MV exhibited practical equivalence in performance across the benchmark datasets. Building upon this observation, further experiments unveiled distinctions in terms of both model capacity and computational efficiency. First, comparison of hyperparameters indicating model capacity revealed that MV generally tended to select lower complexity models compared to CV, which is desirable in the light

of Occam's razor. Second, MV consistently demonstrated advantages in runtime and a reduction in carbon emissions, particularly when the number of CV folds k exceeded 3. Comparing the two methods on the neuroscientific FC datasets, with number of selected most informative features ranging from low to high, the results on the largest dataset (ID1000), revealed that MV remained a practical alternative to CV.

Collectively, the results suggest that MV could function as a valuable complement to CV. In particular, MV can be leveraged as a preliminary tool to augment the efficiency of hyperparameter tuning, particularly in resource-constrained environments. This would facilitate a more thorough exploration of the hyperparameter space while maintaining an acceptable runtime and a lower carbon footprint. Nevertheless, it is important to acknowledge limitations of MV. As our experiments on the FC data showed, MV may be susceptible to underfitting and instability leading to suboptimal model selection.

There remain opportunities for further enhancements that warrant exploration in future research endeavors. While this study primarily focused on binary classification problems, future investigations could extend this comparative analysis to encompass multiclass classification and regression tasks. Furthermore, considering the widespread prevalence of neural networks in diverse domains, it would be intriguing to examine how MV fares in comparison to CV within the realm of deep learning architectures. Finally, examining MV's behavior with respect to data characteristics could provide further insights.

5 Acknowledgements

This work was partly supported by the Helmholtz Portfolio Theme "Supercomputing and Modelling for the Human Brain" and by the Max Planck School of Cognition supported by the Federal Ministry of Education and Research (BMBF) and the Max Planck Society (MPG).

References

1. xcpengine-container 1.0.1, `https://pypi.org/project/xcpengine-container/`
2. Barbiero, P., Squillero, G., Tonda, A.P.: Modeling generalization in machine learning: A methodological and computational study. CoRR **abs/2006.15680** (2020)
3. Corani, G., Benavoli, A.: A bayesian approach for comparing cross-validated algorithms on multiple data sets. Machine Learning **100** (09 2015)
4. Corani, G., Benavoli, A., Demšar, J., Mangili, F., Zaffalon, M.: Statistical comparison of classifiers through bayesian hierarchical modelling. Machine Learning **106**(11), 1817–1837 (11 2017)
5. Dua, D., Graff, C.: UCI machine learning repository (2017), `http://archive.ics.uci.edu/ml`
6. Esteban, O., Markiewicz, C.J., Goncalves, M., Provins, C., Kent, J.D., DuPre, E., Salo, T., Ciric, R., Pinsard, B., Blair, R.W., Poldrack, R.A., Gorgolewski, K.J.: fMRIPrep: a robust preprocessing pipeline for functional MRI (07 2022)

7. Feldman, V., Frostig, R., Hardt, M.: The advantages of multiple classes for reducing overfitting from test set reuse. CoRR **abs/1905.10360** (2019)
8. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research **3**, 1157–1182 (03 2003)
9. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer series in statistics, Springer (2009)
10. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. p. 1137–1143. IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1995)
11. Kruschke, J.: Bayesian estimation supersedes the t test. Journal of experimental psychology: General **142**, 573–603 (07 2012)
12. Mitchell, T.M.: Machine learning. McGraw-hill New York (1997)
13. Raschka, S.: Model evaluation, model selection, and algorithm selection in machine learning. CoRR **abs/1811.12808** (2018)
14. Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.T.: Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cerebral Cortex (2018)
15. Snoek, L., van der Miesen, M.M., Beemsterboer, T., van der Leij, A., Eigenhuis, A., Steven Scholte, H.: The amsterdam open mri collection, a set of multimodal mri datasets for individual difference analyses. Scientific Data **8**(1), 85 (03 2021)
16. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: Openml: networked science in machine learning. SIGKDD Explorations **15**(2), 49–60 (2013)
17. Weis, S., Patil, K.R., Hoffstaedter, F., Nostro, A., Yeo, B.T.T., Eickhoff, S.B.: Sex Classification by Resting State Brain Connectivity. Cerebral Cortex **30**(2), 824–835 (06 2019)
18. Zhang, J.M., Harman, M., Guedj, B., Barr, E.T., Shawe-Taylor, J.: Model validation using mutated training labels: An exploratory study. Neurocomput. **539**(C) (06 2023). https://doi.org/10.1016/j.neucom.2023.02.042