2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Machine Learning Model Validation for Early Stage Studies with Small Sample Sizes

Robyn Larracy[1], Angkoon Phinyomark[1], and Erik Scheme[1], *Senior Member, IEEE*

***Abstract*— In early stage biomedical studies, small datasets are common due to the high cost and difficulty of sample collection with human subjects. This complicates the validation of machine learning models, which are best suited for large datasets. In this work, we examined feature selection techniques, validation frameworks, and learning curve fitting for small simulated datasets with known underlying discriminability, with the aim of identifying a protocol for estimating and interpreting early stage model performance and for planning future studies. Of a variety of examined validation configurations, a nested cross-validation framework provided the most accurate reflection of the selected features' discriminability, but the relevant features were often not properly identified during the feature selection stage for datasets with small sample sizes. Ultimately, we recommend that: (1) filter-based feature selection methods should be used to minimize overfitting to noise-based features, (2) statistical exploration should be conducted on datasets as a whole to estimate the level of discriminability and the feasibility of the classification problems, and (3) learning curves should be employed using nested cross-validation performance estimates for forecasting accuracy at larger sample sizes and estimating the required number of samples to converge towards best performance. This work should serve as a guideline for researchers incorporating machine learning in small-scale pilot studies.**

I. Introduction

Machine learning has enabled groundbreaking advances in the analysis of biomedical signals, images, and omics data in recent years. Due to the complexity of the physiological processes that produce these samples, data-driven pattern recognition techniques are often able to outperform conventional statistical tools [1]. Among the countless applications for machine learning in biomedical research are detecting disease biomarkers, monitoring injury rehabilitation, developing prostheses, and predicting predisposition for injury or illness.

However, while very large datasets are preferred for machine learning applications, these are often unattainable in biomedical studies. Due to the expense associated with data collection involving human participants or labeling of data by domain experts, pilot studies with small sample sizes are commonly used for determining feasibility, securing further funding, and/or for subsequent sample size planning. The future of the given projects may therefore hinge on these preliminary results meaning that, despite the small number of samples, performance estimates must reflect the true error rate of the problem.

Unfortunately, there are very real challenges facing performance estimation with small datasets. The most valuable measure of a classifier's performance is its generalization error, which reflects its performance for samples not seen at any point during the algorithm's creation. Simply splitting the available samples into random groups for training and testing, however, is generally unsuitable for low sample sizes [2]. With too few samples used for training, the quality of the model and its decision rules are weakened. Similarly, with too few test samples, performance estimates are unreliable. Hence, more efficient approaches are required to appropriately utilize the limited samples.

Endeavouring to exploit the available data as much as possible, on the other hand, can result in overly optimistic performance estimates for a problem. This occurs when the same or highly related samples that were used for feature selection, hyperparameter selection, and/or classifier training are reused for estimating a model's performance [3], [4], [5]. This information leakage makes it difficult to detect and prevent overfitting, where the model learns the noise in the dataset in addition to, or rather than, truly valuable patterns [6]. Overfitting can be especially severe with small datasets since strong spurious patterns may occur by chance, and perfect or near-perfect accuracy can be achieved by continually tuning model parameters and hyperparameters to the available samples. While many researchers are careful to avoid information leakage in the classifier training step, information leakage in the other stages of model development, especially feature selection [7], can result in similarly inflated performance estimates.

In this study, simulated datasets were created with varying degrees of discriminability to investigate performance estimation in small-sample binary classification problems. First, the impact of different feature selection techniques was investigated, based on their ability to identify relevant features and provide performance estimates that reflect the true underlying predictive power of the features. Second, these same metrics were then used to compare six configurations for developing and validating models, built on commonly used holdout, cross-validation, and bootstrapping techniques. Lastly, these validation configurations were further assessed using fitted learning curves to evaluate each configuration's ability to forecast model performance at larger sample sizes. Ultimately, the goal of this work is to provide an effective, data efficient framework for validating classification models with small sample sizes.

*This work was supported in part by Mitacs Canada.

[1]All authors are with the Institute of Biomedical Engineering, University of New Brunswick, Fredericton, NB, E3B 5A3, Canada, rlarracy@unb.ca, aphinyom@unb.ca, escheme@unb.ca

II. Methods

A. Data

The data used in this work were simulated, and can be categorized into four types based on their level of class discriminability: (1) low, (2) moderate, (3) high, and (4) varying. Each dataset consisted of two equally sized subsets, constituting a balanced binary classification problem. Fifty features per class were randomly drawn from unit variance Gaussian distributions. Forty of these features were drawn from the same zero-mean distribution for both classes, representing irrelevant features (or noise), while 10 features were made to be differentiable between the two classes by increasing the distribution's mean for only the positive group, i.e., (1) low, with a 0.1 difference in mean for the positive and negative class distributions for each feature; (2) moderate, with a 0.5 difference; (3) high, with a 0.9 difference; and (4) varying, with differences spanning from 0.1 to 1.0 in steps of 0.1 for the 10 features.

The size of the datasets ranged across 20 sample sizes: beginning with 10, samples were added to each dataset in increments of 10 until 100 samples were reached, and then in increments of 50 until 600 samples were reached. For each of the four discriminability cases, an additional dataset of 100,000 samples was simulated to be used for estimating the true classification ability afforded by the distributions of the 10 discriminative features, with 50% randomly selected for training and 50% for testing a linear support vector machine (SVM) classifier.

B. Feature Selection

For the first experimental aim, three feature selection methods were examined for small sample sizes. First was feature selection using simple variable ranking, a filter method that orders features by the value of a scoring function (here, the 10 highest ranked features according to independent t-tests). This method is based solely on independent feature relevance to the class labels so it may not produce optimal predictors, but it is computationally efficient. A second filter method, minimum redundancy maximum relevance feature selection (mRMR), uses a scoring function (a difference of Pearson correlation coefficients in this work) to sequentially select features that exhibit high relevance to the binary class labels but also low redundancy compared to higher ranked features [8]. Third, sequential forward selection (SFS), a wrapper method, uses classification accuracy to establish a well-performing subset of features from the dataset. Successively, features are added to the subset that exhibit the best improvement in accuracy of all remaining features. In this work, linear SVM classifiers were used to obtain resubstitution performance estimates for the SFS procedure.

These three methods were assessed using a holdout validation configuration for each of the twenty sample sizes, repeated for 100 iterations with the varying discriminability data. Linear SVM classifiers were used to estimate both training accuracy, for the 80% subset of samples that were used to train the algorithm, and test accuracy, for the 20% test set unseen during development or training. The feature selection methods were further evaluated based on their ability to identify the truly relevant features, those that were simulated with real underlying differences rather than those that exhibited coincidental noise-based differences, based on the percentage of correctly selected features.

C. Model Validation

For the second experimental aim, six model validation techniques were considered for assigning samples to training and test groups. The first two 'non-nested' techniques (A and B) utilized all available samples for model development, comprised of only a feature selection step in this work, while four 'nested' techniques (C, D, E and F) performed model development using only training subsets of the data.

A) Leave-one-out cross-validation (LOOCV): model development is performed using all samples, then the model's performance is evaluated using LOOCV where one sample is used as the model's test set and the remaining samples as the training set. This is repeated for each of the n samples, shifting the test exemplar each time, and the final performance estimate is averaged from all samples. The same selected features and hyperparameters are retained for each surrogate model.

B) Bootstrapping: model development is performed using all samples, then the model's performance is estimated using 50 bootstrap subsamples of n samples. Specifically, bootstrapping performs random subsampling with replacement for establishing the model's training set and all remaining samples (those 'out-of-bag') are used as test samples. This procedure is repeated for several iterations and the out-of-bag performance estimates are averaged across all surrogate models to determine the final estimate. As with the LOOCV configuration, the same selected features and hyperparameters are retained for each surrogate model.

C) Holdout: 20% of samples are randomly selected for testing, the other 80% are used for both model development and training.

D) Nested k-fold cross-validation (Nested 10-CV): samples are randomly partitioned into k approximately equal sized groups (in this work, $k = 10$). In turn, each group is used as the test set while all remaining groups are used for model development and training, creating k surrogate models. The final performance estimate is the average test accuracy across all surrogate models. Note that model development is repeated in each iteration using only training samples.

E) Nested LOOCV: n partitions are used for the nested cross-validation procedure. Again, model development is performed independently for each surrogate model.

F) Nested Bootstrapping: n samples are subsampled with replacement for 50 iterations. Model development is repeated in each iteration using only training samples.

For each sample size and discriminability level, the six validation configurations were applied to one hundred iterations of simulated data. The t-test variable ranking method

was used for feature selection and linear SVM classifiers were used to estimate classification performance.

D. Learning Curve

For the third experimental aim, learning curves were used to evaluate the forecasting capability of each of the validation configurations. The commonly-used inverse power law model was adopted for this purpose, as in (1) where Y is the fitted curve, n is sample size, a is the best achievable error rate, b is the learning rate, and c is the decay rate [9].

$$Y = (1 - a) - b \cdot n^c \tag{1}$$

Briefly, the learning curve fitting procedure is as follows:

1) Randomly select a stratified subset of n_0 samples from the dataset.
2) Apply model development and validation to obtain a classification performance estimate for the subset.
3) Add m stratified samples to the existing subset and re-apply model development and validation to estimate the new performance. Repeat this step until all available samples have been added, yielding a sequence of classification performance estimates for sample sizes n_0 to the total number of samples, n_{max}.
4) Use least-squares regression to fit an inverse power law to the sequence of performance estimates.
5) Infer the model's performance at larger sample sizes from the fitted curve.

For each of the one hundred iterations performed using varying discriminability data in Section II-C, learning curves were fit to the first ten measured performance estimates, ranging in sample sizes from 10 (n_0) to 100 (n_{max}) in increments of 10 (m) samples. An n_{max} of 100 was selected to exemplify a typical small-scale biomedical dataset. The remaining estimates, ranging in sample sizes from 150 to 600, were used to determine the fit's root-mean-square error (RMSE) at larger sample sizes, and thus, the quality of the forecast. While the value of c was unconstrained during the fitting procedure, a was bounded between 0 and 1 and b was made to be positive to enforce the desired convergence to $1 - a$ (the stabilized accuracy).

III. Results and Discussion

A. Feature Selection

Fig. 1(a) illustrates the average performance of the three feature selection methods using the holdout validation method for 100 simulated datasets of varying discriminability. The training and test curves are revealing: first, models built using small sample sizes clearly suffer from training set overfitting, considering the extremely high training accuracies but poor test accuracies. In these cases, generalization ability can be improved, at least to a certain degree, with the addition of more training data. This exemplifies the well-established advantages of larger sample sizes in machine learning model development. Second, SFS, the wrapper method, suffered from overfitting more than both the variable ranking and mRMR filter methods, especially for small

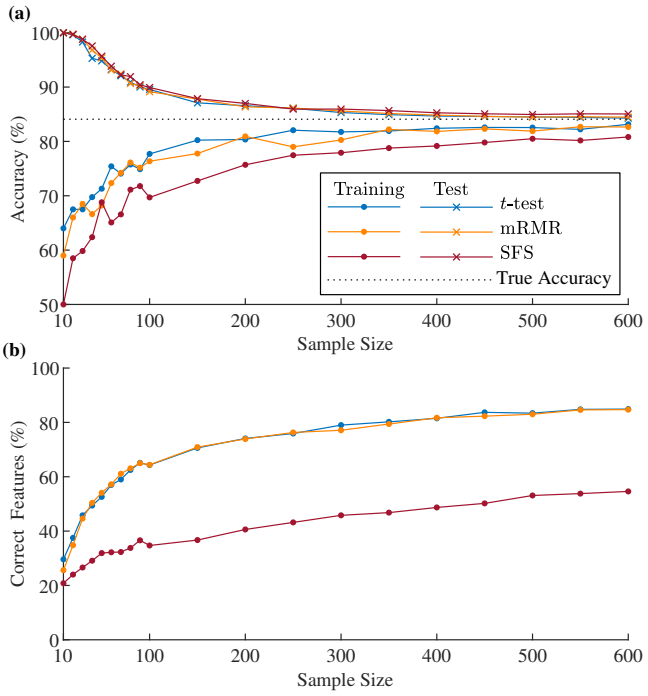

Fig. 1. (a) Average training and test classification accuracies and (b) percentage of correctly selected features using the holdout validation technique with the feature ranking based on t-test, mRMR, and SFS methods for sample sizes of 10-600.

sample sizes where a 10%-15% disparity in testing accuracy was evident. Correspondingly, it can also be observed from the percentage of correct features selected (Fig. 1(b)) that SFS was least effective in identifying the features with true underlying differences, selecting on average over 20% more random noise features than the filter methods. These results suggest that filter-based feature selection methods may be most appropriate in early stage studies with small sample sizes, and more advanced wrapper-based methods should be reserved for larger sample sizes. Lastly, the simple variable ranking technique based on t-tests had the best performance in this study. This may be due in part to the fact that the simulated datasets consisted of normally distributed and uncorrelated features which are ideal conditions for this method. For complex real-world data where correlated, non-normal features can be expected, more advanced filter-based feature selection methods like mRMR, measuring both relevance and redundancy, may be a better fit.

B. Model Validation

The average classification performance estimates for each of the six validation techniques over 100 trials are presented in Fig. 2, for all four levels of dataset discriminability. As in the comparison of feature selection methods, the proportions of correctly selected features are also presented.

For the low discriminability case (Fig. 2(a)), the underlying predictive power was barely better than random data (50% for binary classification), exhibiting a true accuracy of only 56%. With such a small effect, the truly relevant features can be indistinguishable from noise-based features.

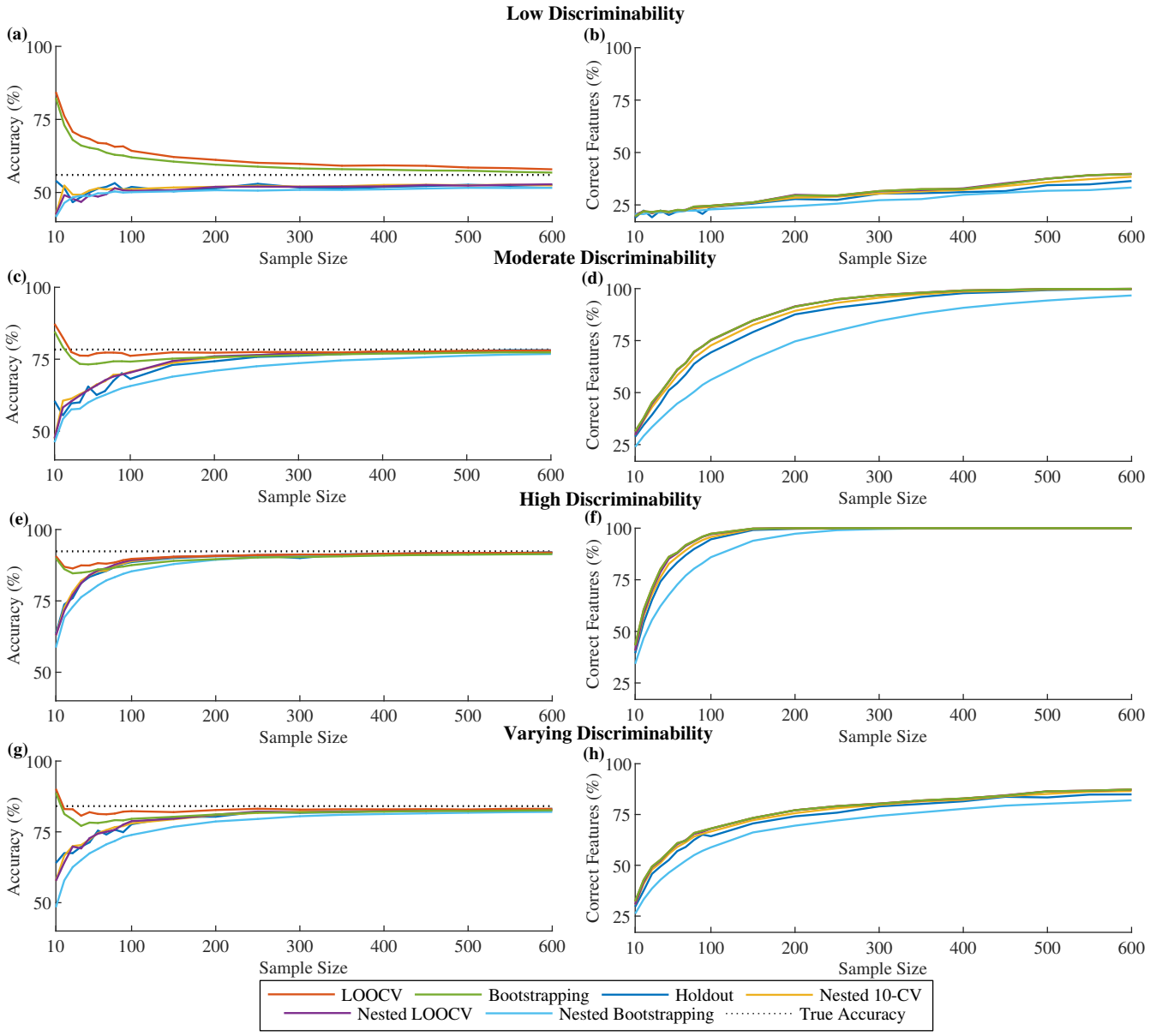

Fig. 2. (a, c, e, g) Average test accuracies and (b, d, f, h) percentage of correctly selected features for each validation configuration and level of data discriminability: (a, b) low, (c, d) moderate, (e, f) high, and (g, h) varying. The feature ranking based on t-testing was applied for feature selection and the linear SVM classifier was used for determining test accuracies.

This was evident during the simulations given the low proportions of correctly selected features with this data, which remained below 50% across all examined sample sizes and validation configurations. Due to information leakage during model development, the non-nested validation techniques, LOOCV and bootstrapping, yielded highly over-optimistic performance estimates for this mostly random data. The problem was most apparent with sample sizes below 100, but the optimism remained even up to 600 samples. The nested validation techniques provided much more accurate performance estimates, better reflecting the large amount of noise-based features and the low effect size of the truly differing distributions. For a real-world problem, designers should therefore perform some initial exploration of the data to assess the quality of the features for classification. Data that does not demonstrate a reasonable level of discriminability in an early stage study (considering both the anticipated learning curve trajectory and the desired end stage performance) should likely not be pursued further.

The moderate discriminability features provide a more viable classification problem (Fig. 2(c)). In this case, however, overfitting was once again evident with the non-nested configurations. Despite their performance estimates more closely exemplifying the true accuracy (78.4%) than the nested configurations for sample sizes less than 200, the proportions of correctly selected features indicate that these estimates were based partially on noise-based features. The four nested configurations, on the other hand, had lower estimates that converged after only a few hundred samples to the true accuracy. This is the gradual improvement in

accuracy expected as the sample size, and therefore the model's ability to learn the appropriate features and class associations, is increased. Since the true underlying accuracy of a real problem is unknown, and cannot be estimated without a sufficient number of samples, it may be beneficial to implement both nested and non-nested configurations to provide a rough idea of the accuracy range when sample size is small. A non-nested configuration will almost certainly experience overfitting and may closely match or exceed the true accuracy, while a nested configuration will provide a conservative estimate that is generally lower than or equal to the true accuracy.

Notably, with the effect sizes of the features increased to a high discriminability case, all six of the validation configurations required fewer samples to converge to the true accuracy (92.4%, Fig. 2(e)). The improved discriminability enabled the models to better identify the correct features and exploit useful patterns, with each configuration reaching an average accuracy estimate within 2% of the true accuracy by 300 samples. With fewer than 300 samples, incidentally, all of the validation methods had pessimistic estimates. Even the non-nested configurations, which still showed evidence of overfitting at very low sample sizes (< 50 samples), did not provide estimates that exceeded the true accuracy as they had in the low and moderate discriminability cases. A real world dataset, however, may likely consist of features with diverse effect sizes. This condition was represented by the varying discriminability case in Fig. 2(g), which exhibited similar results to the moderate and high discriminability cases.

Regardless of the level of discriminability, certain trends were evident in the proportions of correctly selected features (Fig. 2(b), 2(d), 2(f), and 2(h)). The non-nested configurations, LOOCV and bootstrapping, generally yielded the highest percentage of correct features among all configurations, reflecting the increased statistical power afforded during the feature ranking by utilizing all available samples. However, this also allowed the model to find the most convenient noise-based features for the test samples, promoting overfitting and inflating performance estimates. Of the nested configurations, the nested LOOCV configuration, which uses all samples but one for feature selection, was the best at identifying the appropriate features. This was followed closely by the nested 10-CV and then holdout configurations, which used 90% and 80% of samples for feature selection, respectively. The nested bootstrapping configuration had the worst performance in this regard, presumably since it uses the fewest unique samples for model development and training (on average, 63.2% [2]) and repetitions in the bootstrap subsamples can emphasize noise-based patterns. This aspect of the bootstrap procedure explains the more conservative model performance estimates with the bootstrapping and nested bootstrapping configurations compared to their cross-validation counterparts.

While the level of discriminability of the features was varied, this work considered only a fixed number of features that were all drawn from unit standard deviation Gaussian distributions. Assorted feature distributions, correlated features, and highly disproportionate feature-to-sample size ratios are expected for many real-world biomedical problems, so a tailored simulation experiment using dataset-specific characteristics may be a valuable step for future implementations. The number of irrelevant features, in particular, has been shown to greatly disrupt the feature selection procedure and widen the gap between estimates from the nested and non-nested frameworks [10]. Further, there are several additional aspects of a classification model and its development that can affect performance and performance estimates. Though this work examined a handful of feature selection methods, the suitability of the selected algorithms and techniques for pre-processing, feature extraction, and classification can have a large impact. Future studies should incorporate a variety of methods in each of the classification stages to gain a greater understanding of the effect of these methods on model validation.

C. Learning Curve

Fig. 3 depicts the fitted learning curves for the averaged varying discriminability performance estimates from the results shown in Fig. 2. The non-nested configurations clearly did not fit well with the inverse power law model and showed the worst forecasting ability by greatly underestimating performance at higher sample sizes. The nested configurations performed much better in this regard. Nested bootstrapping, in particular, provided the best forecasting ability overall with a mean RMSE of 5.18%. This is, however, the most computationally intensive method, and its RMSE was not significantly improved ($p > 0.05$) compared to the less costly nested 10-CV and nested LOOCV. Though the holdout configuration was the least computationally intensive method, it exhibited the highest RMSE for projected performance of all the nested methods, reaffirming the necessity for more efficient sample use with small datasets. Hence, the nested CV configurations provided the best trade-off between forecasting ability and computation time.

While previous works examining learning curves for sample size planning have assumed fixed feature sets [9], [11], the protocol employed in this work incorporated feature selection to accommodate changes in the optimal feature set for the model as the sample size was varied. To further improve the flexibility of the learning curve procedure, future works should allow the number of selected features to vary at each sample size as well, since in practice, the required number will be unknown. Additionally, though standard techniques were adopted in this work, there are several other models and fitting procedures that could be adopted for learning curve analysis [12]. Certain methods may outperform others given the specific dataset, validation framework and set of utilized algorithms, so these factors should be further explored. Likewise, it would be beneficial to assess the presented techniques with real datasets.

IV. Conclusions

Ultimately, this work has shown that when sample size is sufficiently large, the selected model validation techniques

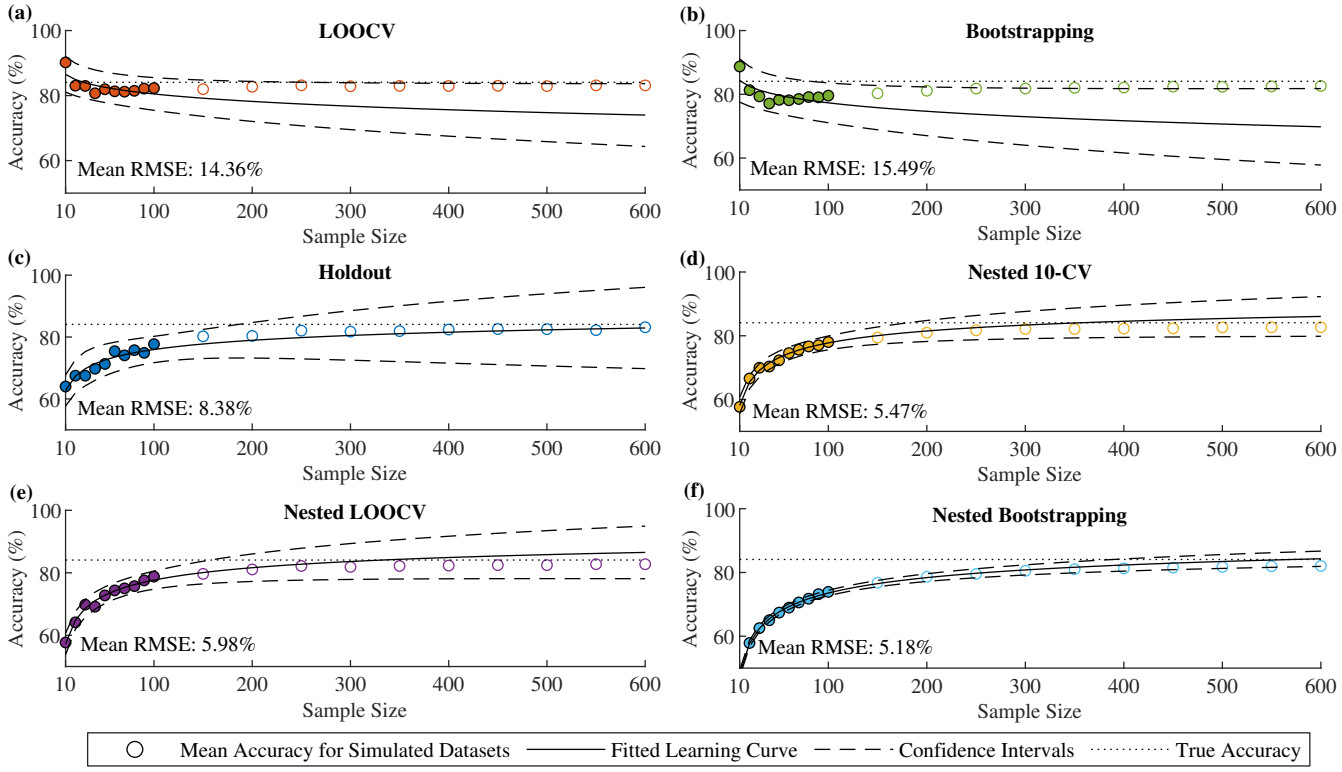

Fig. 3. Fitted learning curves for the averaged varying discriminability performance estimates using (a) LOOCV, (b) bootstrapping, (c) holdout, (d) nested 10-CV, (e) nested LOOCV, and (f) nested bootstrapping. For each of the 100 datasets, the first 10 points (filled) were used for curve fitting and the final 10 points (unfilled) were used for estimating RMSE. The presented RMSE values were averaged across all 100 datasets.

are largely inconsequential, with all techniques eventually converging to the same or similar results. When sample size is small, however, these methods have a much larger impact. Of all the compared validation configurations, the nested CV frameworks had the most success in reflecting the true accuracy of the selected features. The performance estimates were limited by the quality of the selected feature sets, however, which included a large proportion of noise-based features when sample sizes were low. Hence, simply applying nested CV to obtain a single performance estimate may not be adequate, and additional considerations are necessary. First, rather than wrapper-based techniques, filter-based feature selection techniques should be used for small sample sizes to avoid overfitting to irrelevant features. We also recommend that researchers perform an initial exploration of their early stage dataset to assess the level of discriminability of the features, which will aid in the interpretation of performance estimates and gauge the project's potential. Lastly, to estimate the maximum achievable accuracy of the problem and the sample size required to reach this accuracy, we suggest a learning curve fitting procedure using nested CV performance estimates. These recommendations should serve as a practical starting point for researchers performing small-scale feasibility studies and sample size planning.

References

[1] I. Inza, B. Calvo, R. Armañanzas, E. Bengoetxea, P. Larranaga, and J. Lozano, "Machine learning: An indispensable tool in bioinformatics," *Methods Mol Biol*, vol. 593, pp. 25–48, 2010.

[2] C. Beleites, R. Baumgartner, C. Bowman, R. Somorjai, G. Steiner, R. Salzer, and M. G. Sowa, "Variance reduction in estimating classification error using sparse datasets," *Chemom Intell Lab Syst*, vol. 79, pp. 91–100, 2005.

[3] R. G. Brereton, "Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data," *Trends Anal Chem*, vol. 25, no. 11, pp. 1103–1111, 2006.

[4] M. Hosseini, M. Powell, J. Collins, C. Callahan-flintoft, W. Jones, H. Bowman, and B. Wyble, "I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data," *Neurosci Biobehav Rev*, vol. 119, pp. 456–467, 2020.

[5] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J Mach Learn Res*, vol. 11, pp. 2079–2107, 2010.

[6] G. S. Handelman, H. K. Kok, R. V. Chandra, A. H. Razavi, S. Huang, M. Brooks, M. J. Lee, and H. Asadi, "Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods," *Am J Roentgenol*, vol. 212, pp. 38–43, 2019.

[7] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS ONE*, vol. 14, pp. 1–20, 2019.

[8] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy," *IEEE Trans Pattern Anal Mach Intell*, vol. 27, no. 8, pp. 1226–1238, 2005.

[9] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance," *BMC Med Inform Decis Mak*, vol. 12, 2012.

[10] G. Aldehim and W. Wang, "Determining appropriate approaches for using data in feature selection," *Int J Mach Learn Cybern*, vol. 8, no. 3, pp. 915–928, 2017.

[11] K. R. Hess and C. Wei, "Learning curves in classification with microarray data," *Semin Oncol*, vol. 37, no. 1, pp. 65–68, 2015.

[12] T. Viering and M. Loog, "The shape of learning curves: a review," 2021, arXiv preprint arXiv:2103.10948, March 2021.