Pacific Symposium on Biocomputing 26:208-219 (2021)

Optimization of Genomic Classifiers for Clinical Deployment: Evaluation of Bayesian Optimization to Select Predictive Models of Acute Infection and In-Hospital Mortality*

Michael B. Mayhew†, Elizabeth Tran, Kirindi Choi, Uros Midic, Roland Luethy, Nandita Damaraju and Ljubomir Buturovic

Inflammatix, Inc.
Burlingame, California 94010, USA
†*E-mail: mmayhew@inflammatix.com*
www.inflammatix.com

Acute infection, if not rapidly and accurately detected, can lead to sepsis, organ failure and even death. Current detection of acute infection as well as assessment of a patient's severity of illness are imperfect. Characterization of a patient's immune response by quantifying expression levels of specific genes from blood represents a potentially more timely and precise means of accomplishing both tasks. Machine learning methods provide a platform to leverage this *host response* for development of deployment-ready classification models. Prioritization of promising classifiers is dependent, in part, on hyperparameter optimization for which a number of approaches including grid search, random sampling and Bayesian optimization have been shown to be effective. We compare HO approaches for the development of diagnostic classifiers of acute infection and in-hospital mortality from gene expression of 29 diagnostic markers. We take a deployment-centered approach to our comprehensive analysis, accounting for heterogeneity in our multi-study patient cohort with our choices of dataset partitioning and hyperparameter optimization objective as well as assessing selected classifiers in external (as well as internal) validation. We find that classifiers selected by Bayesian optimization for in-hospital mortality can outperform those selected by grid search or random sampling. However, in contrast to previous research: 1) Bayesian optimization is not more efficient in selecting classifiers in all instances compared to grid search or random sampling-based methods and 2) we note marginal gains in classifier performance in only specific circumstances when using a common variant of Bayesian optimization (i.e. automatic relevance determination). Our analysis highlights the need for further practical, deployment-centered benchmarking of HO approaches in the healthcare context.

Keywords: hyperparameter optimization; Bayesian optimization; acute infection; sepsis; disease severity; mortality; classification; molecular diagnostics; genomics.

1. Introduction

Patient lives depend on the swiftness and accuracy of 1) assessment of the severity of their illness and 2) detection of acute infection (when present). The COVID-19 pandemic has put this fact into stark relief. Currently, clinicians determine severity of illness by computing scores

*Supplementary material can be found at `https://arxiv.org/abs/2003.12310`

(e.g. SOFA[1]) based on patient physiological features associated with the risk of adverse events (e.g. in-hospital mortality, organ failure). Similarly, detection of acute infection generally involves evaluation of symptoms (e.g. cough, runny nose, fever) as well as laboratory tests for the presence of specific pathogens. However, these methods provide superficial and imprecise measures of patient illness. Recent work has highlighted the potential of using gene expression measurements from patient blood to detect the presence and type of infection to which the patient is responding[2–5] as well as the patient's severity of illness.[6]

Coupled with these host response signatures, advances in machine learning (ML) provide a platform for the development of robust, diagnostic classifiers of acute infection status (e.g. bacterial or viral) and in-hospital mortality from gene expression. An important step in this development is optimization of the classifier's hyperparameters (e.g. penalty coefficient in a LASSO logistic regression, learning rates for gradient descent). Hyperparameter optimization begins with specification of a search space and proceeds by generating a user-specified number of hyperparameter configurations, training the classifier models given by each configuration, and evaluating the performance of the trained classifier in *internal validation*. Internal validation performance is typically assessed either on a separate validation/tuning dataset or by cross-validation. Configurations are then ranked by this performance, with the top configuration selected and retained for *external validation* (application to a held-out dataset).

Multiple HO approaches have been proposed. For classifiers with relatively small hyperparameter spaces (e.g. support vector machines), optimizing over a pre-defined grid of hyperparameter values (grid search; GS) has proven effective. More recent work has shown that optimization by randomly sampling (RS) hyperparameter configurations can lead to better coverage of high-dimensional hyperparameter spaces and potentially better classifier performance.[7] Bayesian optimization (BO) is a global optimization procedure that has also proven effective for hyperparameter optimization in classical[8–12] and biomedical[13–16] ML applications. In BO, one uses a model (commonly a Gaussian process (GP)[17]) to approximate the objective function one wants to optimize; for hyperparameter optimization, the objective function maps from hyperparameter configurations to the internal validation performance of their corresponding classifiers. In contrast to GS/RS, BO proceeds by sequentially evaluating configurations with each newly visited configuration used to update the model of the objective function.

In this work, we compare GS/RS and BO methods for hyperparameter optimization of gene expression-based diagnostic classifiers for two clinical tasks: 1) detection of acute infection and 2) prediction of mortality within 30 days of hospitalization. We optimize and train three different types of classifiers using gene expression features from 29 diagnostic markers in a multi-study cohort of 3413 patient samples for acute infection detection (3288 for 30-day mortality prediction). Patient samples were assayed on a variety of technical platforms and collected from a range of geographical regions, healthcare settings, and disease contexts. Our extensive analysis evaluates the BO approach, in particular, under a range of computational budgets and optimization settings. Crucially, beyond assessing and comparing the performance of top classifiers in internal validation, we further evaluate top models selected by all HO approaches in a multi-cohort external validation set comprising nearly 300 patients profiled by a targeted diagnostic instrument (NanoString). Our analysis provides important

insights for diagnostic classifier development using genomic data, and, more generally, about the implementation and practical usage of HO methods in healthcare.

2. Related Work

Previous studies comparing HO approaches in the ML community have demonstrated that BO can select promising classifiers more efficiently (with fewer evaluations of hyperparameter configurations) than GS/RS methods.[8–12,15,16,18] However, these studies have focused on internal validation performance and on benchmark datasets whose composition and handling (i.e. partitioning into training-validation-test splits) doesn't necessarily reflect characteristics of healthcare settings (i.e. smaller, structured, and more heterogeneous datasets; high propensity for models to be applied to out-of-distribution samples at test time[19]).

Bayesian optimization has also found recent success in genomics and biomedical applications.[20–22] Ghassemi et al.[13] compare multiple HO approaches, including BO, for tuning parameters of the multi-scale entropy of heart rate time series to aid mortality prediction among sepsis patients. Colopy et al.[14] analyzed RS and BO methods for optimization of patient-specific GP regression models used in vital-sign forecasting. A study by Nishio et al.[15] evaluated both RBF SVM and XGBoost classifiers tuned by either RS or BO for detection of lung cancer from nodule CT scans. Borgli et al.[16] evaluated BO for tuning and transfer learning of pre-trained convolutional neural networks to detect gastrointestinal conditions from images. Again, however, these studies only reported either internal validation performance or performance on a test set partitioned from a full, relatively small and homogeneous (e.g. collected from a single hospital) dataset, making conclusions difficult to draw about the generalizability of selected models in other segments of the deployment population. Moreover, these studies focused on: 1) no more than two classifier types, 2) a narrow range of settings for BO, and 3) physiological or image data. To our knowledge, no studies have evaluated the external validation performance of selected models, an important pre-requisite for eventual model deployment. In addition, no comparison of HO approaches has yet been attempted for development of diagnostic classifiers using genomic data.

3. Methods

3.1. *Cohort & Feature Description*

To build our datasets, we combined gene expression data from public sources and in-house clinical studies designed for research in diagnosing acute infections and sepsis. We collected the publicly available studies from the NCBI GEO and EMBL-EBI ArrayExpress databases using a systematic search.[2] The public studies were profiled using a variety of different technical platforms (e.g. mostly microarrays). Samples from the in-house clinical studies were profiled on the NanoString nCounter platform using a custom codeset for 29 diagnostic genes of interest. All included studies consisted of samples from our target population: both adult and pediatric patients from diverse geographical regions and clinical settings. Each included study had measurements taken from patient blood for all 29 markers. To account for heterogeneity across studies, we performed co-normalization (see[5] and the Supplement).

The features we used in our analyses were based on the expression values of 29 genes pre-

viously found to accurately discriminate three different aspects of acute infection: 1) viral vs. bacterial infection (7 genes),[3] 2) infection vs. non-infectious inflammation (11 genes),[2] and 3) high vs. low risk of 30-day mortality (11 genes).[6] Building on our previous work,[5] we computed both the geometric means and arithmetic means of these six groups of genes, producing 12 features. We optimized and trained our classifiers on the combination of these 12 features and the expression values of all 29 genes (41 features in total). Labels for one of three classes of the acute infection detection or BVN task (**B**acterial infection, **V**iral infection, or **N**on-infectious inflammation) were determined differently for each of the training and validation studies depending on available data. For training set studies, we used the labels provided by each study, deferring to each study's criteria for adjudication which may have involved multi-clinician adjudication with or without positive pathogen identification or positive pathogen identification alone. When BVN adjudications were not directly provided by the study, we assigned class labels based on available pathogen test results from the study metadata/manuscripts. For validation data, one study was adjudicated by a panel of clinicians using all available clinical data (including pathogen test results) while all other validation studies were labeled by us using only pathogen test results. Non-infected determinations did not include healthy controls. Binary indicator labels of whether a patient died within 30 days of hospitalization were derived from study metadata (when available) and the associated study's manuscripts.

For both tasks, we separated studies into a training set and an external validation set. For the BVN task, the training set consisted of 43 studies (profiled outside Inflammatix) and 3413 patients (1087 with bacterial infection, 1244 with viral infection, and 1082 non-infected). The BVN external validation set consisted of six studies (profiled by Inflammatix) and 293 patients (153 with bacterial infection, 106 with viral infection, and 34 non-infected). For the mortality task, the training set consisted of 33 studies (profiled outside Inflammatix) and 3288 patients (175 30-day mortality events) while the mortality external validation set comprised four studies (profiled by Inflammatix) and 348 patients (80 30-day mortality events). A description of the publicly available studies in our training set appears in Supplementary Table 1.

3.2. *Grouped cross-validation*

Previous analyses by our group[5] suggested that alternative cross-validation strategies were preferable over conventional k-fold cross-validation (CV) for identifying classifiers able to generalize across heterogeneous patient populations. We use 5-fold grouped CV (full studies are allocated to one and only one of five folds) to rank and select hyperparameter configurations from GS/RS methods and as an objective function in BO.

3.3. *Classifier types and performance assessment*

We evaluated three types of classification models: 1) support vector machines with a radial basis function (RBF) kernel, 2) XGBoost (XGB[23]) and 3) multi-layer perceptrons (MLP). MLP models were trained with the Adam optimizer[24] with mini-batch size fixed at 128.

For the BVN task, we ranked and selected models based on multi-class AUC (mAUC).[25] For the mortality task, we selected models by binary AUC but report both AUC and average precision to account for class imbalance. To determine performance of models in grouped 5-

fold CV, we pooled the model's predicted probabilities for each fold and computed the relevant metric from the pooled probabilities. The top-performing hyperparameter configuration was then trained on the full training set and applied to the external validation set. We computed external validation performance for these top models using their predicted probabilities for the validation samples. We computed 95% bootstrap confidence intervals for differences in classification performance by sampling predicted probabilities with replacement 5000 times (using the same set of bootstrap sample IDs for both sets of predicted probabilities in the comparison), computing the relevant performance metric on each bootstrap sample, computing the difference between performance metrics for each bootstrap sample in a given comparison, and reporting the 2.5th and 97.5th quantiles of the 5000 differences.

3.4. *Hyperparameter optimization details*

For RBF SVM, we conduct a grid search over configurations of the cost, C, and bandwidth hyperparameters, γ. C values ranged from 1e-03 to 2.15 and γ values ranged from 1.12e-04 to 10. We generated RS samples for XGBoost and MLP uniformly and independently of one another from pre-specified ranges or from grids (Suppl. Tables 2 and 3).

For BO, the objective function maps from hyperparameter configurations to 5-fold grouped CV performance of the corresponding classifiers. The two main components of BO are: 1) a model that approximates the objective function, and 2) an acquisition function to propose the next configuration to visit. We use a GP regression model with Gaussian noise to approximate the objective function. To initialize construction of the objective function, we uniformly and independently sample configurations (either 5 or 25) from the hyperparameter space.

We investigate both the expected improvement and upper confidence bound acquisition functions. We use both standard and automatic relevance determination (ARD) forms of the Matern5/2 covariance function in BO's GP model of the objective (further details in Supplement). We also perform BO in the hyperparameters' native scales (*original* space) or in which continuous and discrete hyperparameter dimensions are searched in the continuous range 0 to 1 and transformed back to their native scales prior to their evaluation (*transformed*).

4. Results

We compared BO and GS/RS approaches for hyperparameter optimization of three types of classifiers for two clinical tasks. For the BVN task, we sought classifiers that could achieve high performance in predicting whether a patient had a bacterial or viral infection or was showing a non-infectious inflammatory response. For the mortality task, we sought high-performing classifiers of mortality events within 30 days of hospital admission. Though we considered BO at two initialization budgets (5 and 25 configurations), we did not see substantial differences in performance between classifiers with 5 and 25 initial configurations (Suppl. Table 4, Suppl. Figs. 3-6). We focus on BO results with 25 initial configurations and the expected improvement acquisition function for the remainder of this work (results for all runs in Supplement).

General comparison of classifier performance across tasks and HO approaches Across both tasks and HO approaches, we note distinct performance characteristics of the selected classifiers of each type. While RBF SVM classifiers performed similarly to the other

two classifier types on the BVN task, they were the worst performers on the mortality task. XGB classifiers selected by either RS or BO demonstrated competitive performance in both tasks and were remarkably consistent in their performance regardless of the number of hyperparameter configurations evaluated for HO. MLPs achieved the highest internal and external validation performance for both acute infection detection and mortality prediction (Table 1), suggesting potential benefits of learning latent features (hidden layers) for these tasks. We also find that, despite the considerable class imbalance in the mortality task, all classifier types selected by AUC still demonstrated average precision considerably higher than the respective baselines for internal ($\frac{175}{3288} \approx 0.053$) and external ($\frac{80}{348} \approx 0.230$) validation.

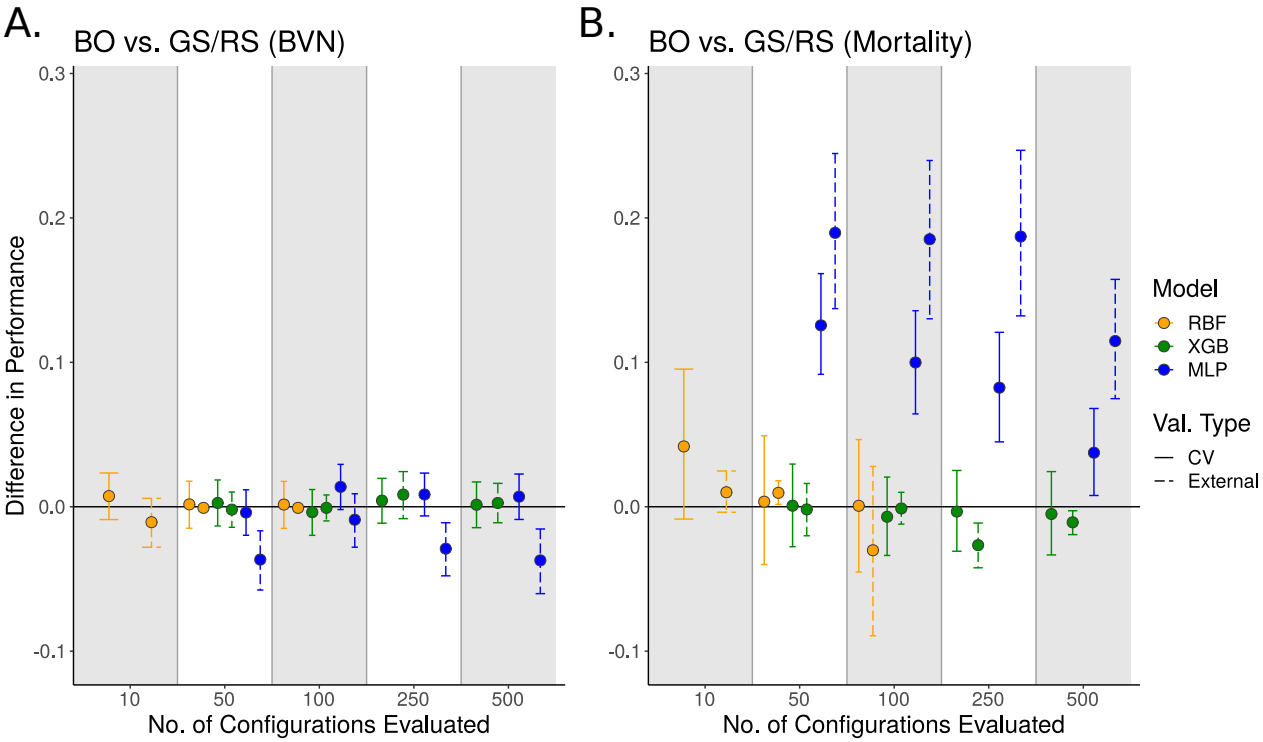

Fig. 1: **Differences in classification performance of models selected by either BO or GS/RS using BO evaluation budgets.** Performance differences greater than 0 on the BVN (A; mAUC) and mortality (B; AUC) tasks indicate better performance for the BO-selected classifier. Classifiers were selected with the indicated number of hyperparameter configurations evaluated. Automatic relevance determination was not enabled for BO. Points represent observed differences while error bars represent 95% bootstrap confidence intervals.

Evaluation of BO- and GS/RS-selected classifiers at evaluation budgets typical of BO. Previous studies have shown that BO can select promising classifiers more efficiently than GS/RS methods. Surprisingly, we find that at smaller numbers of configurations evaluated (more typical of BO), classifiers selected by GS/RS showed similar or better performance in both internal and external validation (Table 1 and Figs. 1) when compared with corresponding BO-selected classifiers. We observed similar trends when using the upper confidence

Table 1: Grouped 5-fold CV and external validation (Val.) performance of selected classifiers for the BVN and mortality tasks. BO results used the EI acquisition function and 25 initialization points. The ARD column indicates whether automatic relevance determination was enabled (Y/N) in BO's GP model of the objective function. **Bold** numbers indicate the best performance for a column. BVN column shows performance in mAUC; mortality column shows AUC performance with average precision in parentheses. *Grid specified only 4757 configurations.

Model	HO Type	No. of Evals.	ARD	BVN CV	BVN Val.	Mortality CV	Mortality Val.
RBF	GS	10	-	0.808	0.862	0.758 (0.182)	0.736 (0.375)
	GS	50	-	0.814	0.853	0.797 (0.169)	0.739 (0.372)
	GS	100	-	0.814	0.853	0.800 (0.192)	0.782 (0.533)
	GS	250	-	0.814	0.853	0.801 (0.191)	0.749 (0.386)
	GS	500	-	0.815	0.853	0.801 (0.191)	0.749 (0.386)
	GS	1000	-	0.815	0.853	0.839 (0.225)	0.708 (0.444)
	GS	5000*	-	0.815	0.853	0.839 (0.225)	0.708 (0.444)
	BO	10	Y	0.811	0.788	0.800 (0.190)	0.747 (0.383)
	BO	10	N	0.815	0.851	0.800 (0.187)	0.746 (0.381)
	BO	50	Y	0.816	0.852	0.801 (0.196)	0.752 (0.389)
	BO	50	N	0.816	0.852	0.801 (0.194)	0.749 (0.385)
	BO	100	Y	0.816	0.852	0.800 (0.197)	0.753 (0.392)
	BO	100	N	0.816	0.852	0.801 (0.196)	0.752 (0.389)
XGB	RS	50	-	0.809	0.830	0.880 (0.315)	0.819 (0.542)
	RS	100	-	0.813	0.827	0.885 (0.288)	0.819 (0.526)
	RS	250	-	0.812	0.826	0.885 (0.308)	0.829 (0.556)
	RS	500	-	0.810	0.829	0.885 (0.320)	0.826 (0.559)
	RS	1000	-	0.810	0.822	0.885 (0.311)	0.822 (0.552)
	RS	5000	-	0.813	0.830	0.888 (0.310)	0.823 (0.552)
	RS	25000	-	0.815	0.860	0.889 (0.303)	0.816 (0.532)
	BO	50	Y	0.818	0.865	0.887 (0.301)	0.814 (0.540)
	BO	50	N	0.812	0.828	0.881 (0.275)	0.817 (0.516)
	BO	100	Y	0.811	0.825	0.885 (0.314)	0.825 (0.559)
	BO	100	N	0.809	0.826	0.878 (0.288)	0.817 (0.521)
	BO	250	Y	0.818	0.865	0.886 (0.290)	0.826 (0.539)
	BO	250	N	0.816	0.834	0.882 (0.272)	0.802 (0.483)
	BO	500	Y	0.818	0.865	0.889 (0.346)	0.827 (0.591)
	BO	500	N	0.812	0.831	0.880 (0.313)	0.815 (0.538)
MLP	RS	50	-	0.818	0.860	0.763 (0.121)	0.631 (0.288)
	RS	100	-	0.814	0.863	0.785 (0.156)	0.640 (0.301)
	RS	250	-	0.824	0.861	0.807 (0.211)	0.625 (0.366)
	RS	500	-	0.819	0.859	0.853 (0.240)	0.691 (0.401)
	RS	1000	-	0.835	**0.872**	0.809 (0.158)	0.637 (0.333)
	RS	5000	-	0.837	0.835	0.826 (0.249)	0.796 (0.546)
	RS	25000	-	**0.840**	0.856	0.859 (0.267)	0.743 (0.428)
	BO	50	Y	0.816	0.820	0.888 (0.340)	0.823 (0.554)
	BO	50	N	0.814	0.824	0.888 (0.290)	0.820 (0.564)
	BO	100	Y	0.822	0.845	0.886 (0.296)	**0.847** (0.631)
	BO	100	N	0.828	0.854	0.884 (0.292)	0.825 (0.577)
	BO	250	Y	0.817	0.848	0.890 (0.312)	0.842 (0.614)
	BO	250	N	0.832	0.832	0.889 (0.335)	0.812 (0.566)
	BO	500	Y	0.837	0.855	**0.894** (0.304)	0.835 (0.593)
	BO	500	N	0.826	0.822	0.890 (0.330)	0.806 (0.561)

bound acquisition function (Suppl. Figs. 7 and 8, Suppl. Table 5) or the transformed hyperparameter space (Suppl. Figs. 11 and 12, Suppl. Table 6). However, we do note two instances in which BO-selected classifiers exceeded performance of GS/RS-selected classifiers: 1) XGBoost classifiers in external validation for the BVN task and 2) MLP classifiers for the mortality task. While these instances support prior findings of BO's efficiency, our results also suggest that simply committing to a single HO approach could miss models that generalize well and that performance of selected classifiers will depend on the task and classifier type.

Evaluation of BO- and GS/RS-selected classifiers at evaluation budgets typical of GS/RS. In the previous analysis, we compared BO- and GS/RS-selected classifiers at evaluation budgets typical of BO (i.e. fewer configurations evaluated). In Figure 2, we compare BO-selected classifiers from their highest evaluation budgets (100 evaluations for RBF and 500 evaluations for XGB and MLP) to classifiers selected by GS/RS at larger evaluation budgets. Interestingly, we find that the BO-selected MLP classifiers for the mortality task continue to outperform their corresponding RS-selected counterparts, even with 25000 configurations evaluated for RS. Similarly, we find that BO-selected XGBoost classifiers exceed external validation performance of RS-selected classifiers on the BVN task up to an evaluation budget of 25000 configurations (though the differences do not persist at 25000 configurations). We observe these differences when conducting BO with the upper confidence bound acquisition function or with a transformed hyperparameter space (Suppl. Figs. 9, 10, 13 and 14). These results indicate the relative efficiency of BO in candidate classifier selection in these two instances but also illustrate the competitiveness of GS/RS-selected classifiers in our setting.

Assessment of effects on classifier performance of automatic relevance determination in BO. For high-dimensional hyperparameter spaces, some hyperparameters may have a greater impact on the model's generalization performance than others. Automatic relevance determination (ARD;[26]) in the GP model of BO's objective provides the means to estimate effects of variations in hyperparameter dimensions on the objective's value and has been used in multiple implementations of BO (e.g. Snoek et al., 2012[8] and BoTorch, `https://botorch.org/docs/models`). We directly compare the internal and external validation performance of classifiers selected by BO with and without ARD. In Figure 3, we find that enabling ARD seems to lead to comparable if not slightly better internal validation performance at higher evaluation budgets. Moreover, enabling ARD seems to improve external validation performance for both XGB (BVN task) and MLP classifiers (both tasks). In fact, the highest external validation performance by XGB classifiers on the BVN task is only achieved with ARD enabled (Table 1). However, these differences in performance are not as evident when using the upper confidence bound acquisition function (Suppl. Fig. 15) or conducting BO in the transformed hyperparameter space (Suppl. Fig. 16). Thus, ARD may not be necessary to select top-performing diagnostic classifiers for these two clinical tasks.

5. Discussion & Conclusions

In this analysis, we compared HO approaches for diagnostic classifier development to determine what approach (if any) led to improvements in: 1) external validation performance or

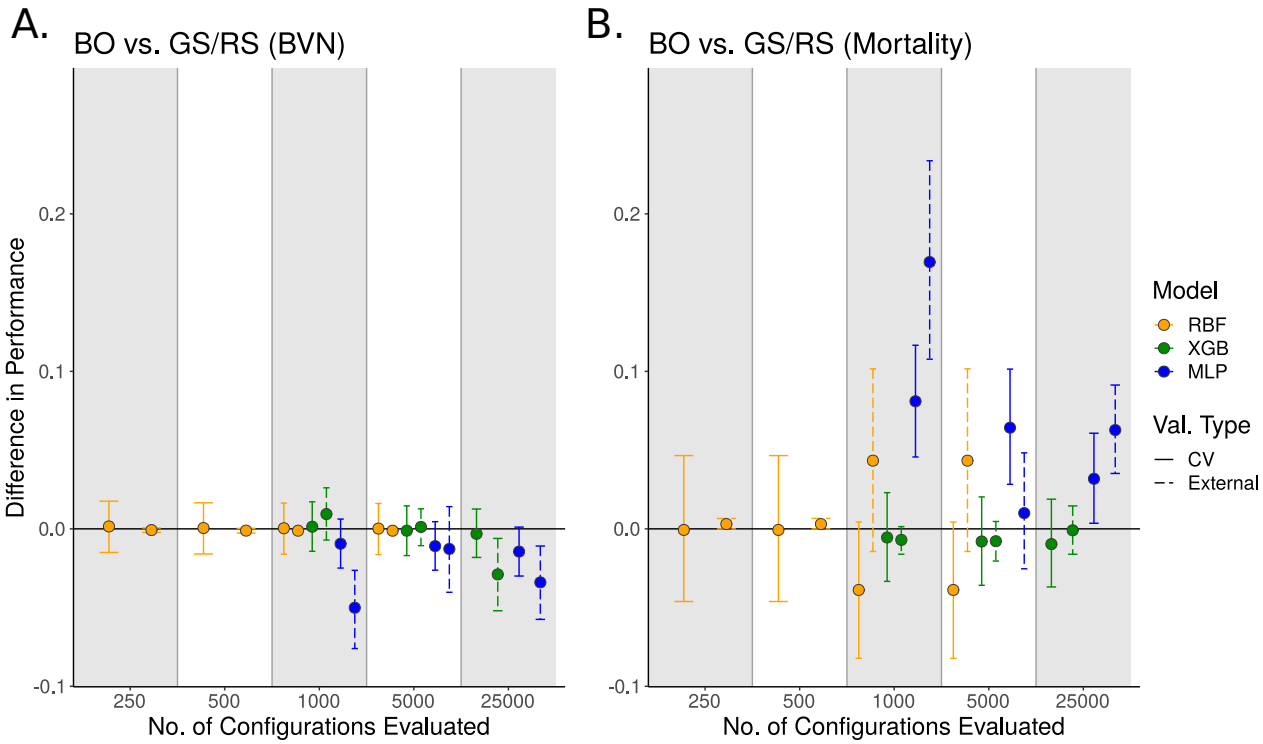

Fig. 2: **Differences in classification performance of models selected by either BO or GS/RS using GS/RS evaluation budgets.** Run settings and figure layout are the same as in Figure 1 except that here, indicated evaluation budgets apply to GS/RS-selected classifiers; BO-selected classifiers are taken from 100-evaluation (RBF) or 500-evaluation (XGB and MLP) runs.

2) computational efficiency. Consistent with previous findings, we found that BO was able to prioritize candidate classifiers for two tasks relevant to emergency and critical care with a fraction of the configurations evaluated using GS/RS. As embarrassingly parallel approaches like GS/RS can necessitate the use of commodity computing clusters, BO's efficiency makes the approach a potentially cost-effective solution. We also found that external validation performance of BO-selected MLPs for in-hospital mortality was consistently better across a range of HO evaluation budgets than that of GS/RS-selected classifiers, highlighting BO's potential to uncover diagnostic classifiers that generalize better to unseen patients.

However, and in contrast to previous comparisons of HO approaches, our analyses indicated that GS/RS methods could select classifiers for both tasks with evaluation budgets comparable to those used for BO. We also found mixed evidence in support of enabling ARD in the kernel of BO's GP model of the objective function. Thus, while we hoped we would uncover distinct and general differences between HO approaches in order to develop better guidelines about when (or even if) to use one approach over another, we did not identify such clear differences across tasks, classifier types, and optimization settings. Rather, our analysis suggests that both GS/RS and BO approaches should be investigated for classifier development.

We acknowledge limitations of our approach. For our RS runs, we sampled configurations

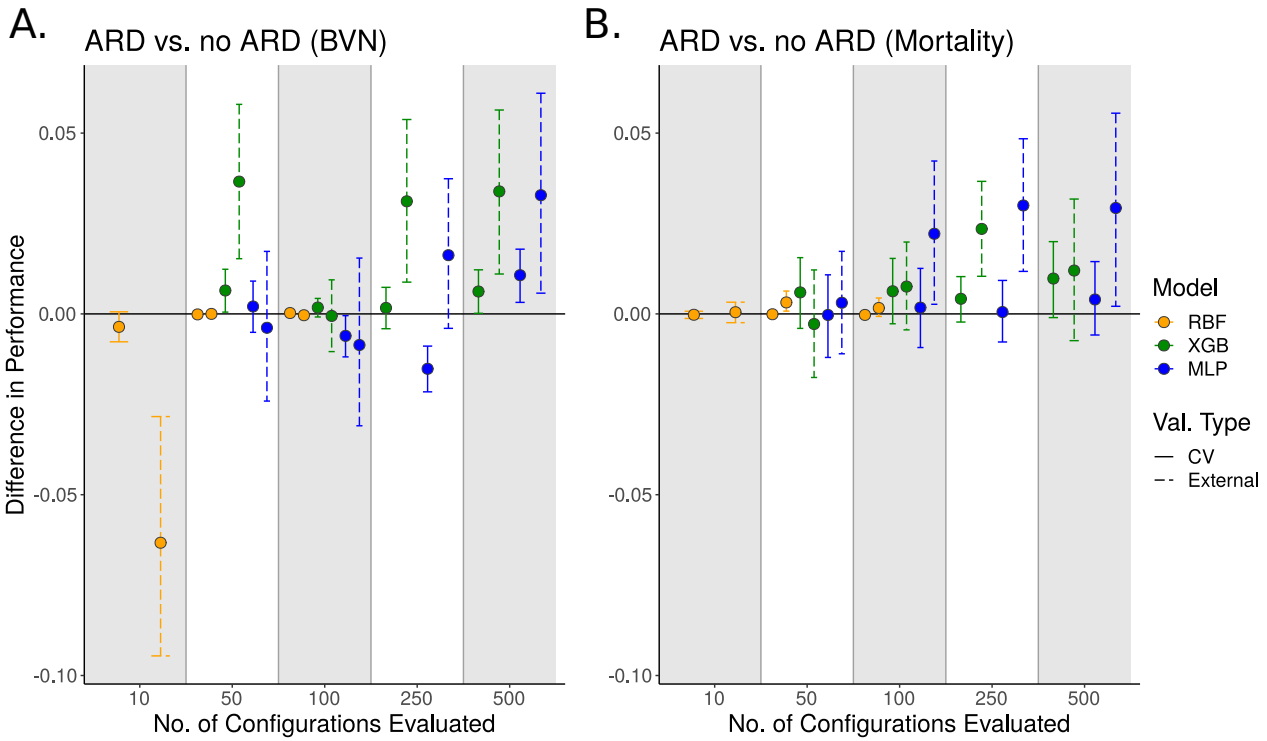

Fig. 3: **Differences in classification performance for BO-selected classifiers with or without automatic relevance determination (ARD) enabled.** Performance differences greater than 0 on the BVN (A; mAUC) and mortality (B; AUC) tasks indicate better performance for the classifier selected by ARD-enabled BO. Points represent observed differences while error bars represent 95% bootstrap confidence intervals.

uniformly and independently from pre-defined ranges or grids of values. Other random sampling approaches could've been used in which configurations are generated dependent on the values of previously generated configurations (e.g. Latin hypercube or low-discrepancy Sobol sequences) in order to encourage diversity of the resulting sample.[7] We felt that the similar performance we observed between BO and GS/RS-selected models using basic variants of GS/RS didn't necessarily justify further analysis with more sophisticated GS/RS variants. A second limitation is that we used a single set of features derived from a previously identified set of 29 gene expression markers. We chose these features based on previous analyses[5] and consistent with our goal of developing diagnostic classifiers from these specific markers for clinical deployment. We acknowledge our conclusions may not hold with other feature sets.

Throughout this work, we wanted our hyperparameter optimization to reflect our clinical deployment scenario: that classifiers would likely be evaluated on structured populations (e.g. from a given geographic region) not seen in training. A recent study by Google highlighted this challenge for deployment in healthcare: their AI system for breast cancer screening showed drops in predictive performance when trained on mammograms from the UK and applied to mammograms from the US.[27] However, our survey of ML studies comparing hyperparameter optimization approaches highlighted important differences from our setting in terms of dataset

partitioning and, consequently, in the choice of internal validation-based objective function. For example, we found that ML studies primarily focused on larger (N >∼100k) datasets composed mainly of natural images. These benchmarks were often constructed (e.g. MNIST; `http://yann.lecun.com/exdb/mnist/`) to satisfy the assumption that the distribution of training and external validation samples are similar if not the same. Internal validation was then performed on subsets of these 'mixed' datasets, with samples from the same structured group in the full dataset appearing in both the training and validation set. However, as patient data is known to be heterogeneous due to biological differences as well as differences in geography, healthcare delivery, and assay technologies used, that assumption of distributional similarity between training and external validation samples is likely to be violated. Indeed, our recent work found that standard k-fold cross-validation gives optimistically biased estimates of generalization error in our setting,[5] breaking the group structure in left-out folds by randomly distributing patients from the same study into different cross-validation folds (akin to test set contamination). Consequently, in difference to the ML studies we reviewed, we opted for grouped 5-fold cross-validation as our objective function as well as evaluation of performance in external validation to aid model selection.

In conclusion, we find that both GS/RS and BO remain promising avenues for hyper-parameter optimization and represent key components in the development of more effective diagnostics for emergency and critical care.

References

1. A. E. Jones, S. Trzeciak and J. A. Kline, The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation, *Critical care medicine* **37**, 1649 (May 2009), 19325482[pmid].
2. T. Sweeney, A. Shidham, H. R. Wong and P. Khatri, A comprehensive time-course–based multicohort analysis of sepsis and sterile inflammation reveals a robust diagnostic gene set, *Science Translational Medicine* **7** (2015).
3. T. Sweeney, H. R. Wong and P. Khatri, Robust classification of bacterial and viral infections via integrated host gene expression diagnostics, *Science Translational Medicine* **8** (2016).
4. T. Sweeney and P. Khatri, Benchmarking sepsis gene expression diagnostics using public data, *Critical care medicine* **45**, p. 1 (2017).
5. M. B. Mayhew, L. Buturovic, R. Luethy, U. Midic, A. R. Moore, J. A. Roque, B. D. Shaller, T. Asuni, D. Rawling, M. Remmel, K. Choi, J. Wacker, P. Khatri, A. J. Rogers and T. E. Sweeney, A generalizable 29-mrna neural-network classifier for acute bacterial and viral infections, *Nature Communications* **11**, p. 1177 (2020).
6. T. Sweeney, T. Perumal and R. e. a. Henao, A community approach to mortality prediction in sepsis via gene expression analysis, *Nat Commun* (2018).
7. J. Bergstra and Y. Bengio, Random search for hyper-parameter optimization, *Journal of Machine Learning Research* **13**, 281 (2012).
8. J. Snoek, H. Larochelle and R. P. Adams, Practical Bayesian optimization of machine learning algorithms, *In Advances in neural information processing systems* , 2951 (2012).
9. J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat and R. Adams, Scalable Bayesian optimization using deep neural networks, in *International conference on machine learning*, 2015.
10. A. Klein, S. Falkner, S. Bartels, P. Hennig and F. Hutter, Fast Bayesian Optimization of Machine Learning Hyperparameters on Large Datasets, in *Proceedings of the 20th International Confer-*

ence on Artificial Intelligence and Statistics, eds. A. Singh and J. Zhu, Proceedings of Machine Learning Research, Vol. 54 (PMLR, Fort Lauderdale, FL, USA, 20–22 Apr 2017).
11. S. Falkner, A. Klein and F. Hutter, BOHB: Fast and Efficient Hyperparameter Optimization at Scale, in *ICML*, 2018.
12. A. Klein, Z. Dai, F. Hutter, N. Lawrence and J. Gonzalez, Meta-Surrogate Benchmarking for Hyperparameter Optimization, in *Advances in Neural Information Processing Systems 32*, eds. H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox and R. Garnett (Curran Associates, Inc., 2019) pp. 6270–6280.
13. M. Ghassemi, L. Lehman, J. Snoek and S. Nemati, Global optimization approaches for parameter tuning in biomedical signal processing: A focus on multi-scale entropy, in *Computing in Cardiology 2014*, Sep. 2014.
14. G. W. Colopy, S. J. Roberts and D. A. Clifton, Bayesian Optimization of Personalized Models for Patient Vital-Sign Monitoring, *IEEE Journal of Biomedical and Health Informatics* **22**, 301 (March 2018).
15. M. Nishio, M. Nishizawa, O. Sugiyama, R. Kojima, M. Yakami, T. Kuroda and K. Togashi, Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization, *PloS one* **13**, e0195875 (Apr 2018), 29672639[pmid].
16. R. J. Borgli, H. Kvale Stensland, M. A. Riegler and P. Halvorsen, Automatic Hyperparameter Optimization for Transfer Learning on Medical Image Datasets Using Bayesian Optimization, in *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*, May 2019.
17. C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, 2006).
18. J. Bergstra, D. Yamins and D. D. Cox, Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures (2013).
19. S. Ben-David, J. Blitzer, K. Crammer and F. Pereira, Analysis of representations for domain adaptation, in *Advances in Neural Information Processing Systems 19*, eds. B. Schölkopf, J. C. Platt and T. Hoffman (MIT Press, 2007) pp. 137–144.
20. M. Thomas and R. Schwartz, A method for efficient Bayesian optimization of self-assembly systems from scattering data, *BMC Systems Biology* **12**, p. 65 (2018).
21. R. Tanaka and H. Iwata, Bayesian optimization for genomic selection: a method for discovering the best genotype among a large number of candidates., *Theor Appl Genet* **131**, 93 (2018).
22. S. Mao, Y. Jiang, E. B. Mathew and S. Kannan, BOAssembler: a Bayesian Optimization Framework to Improve RNA-Seq Assembly Performance (2019).
23. T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 (ACM, New York, NY, USA, 2016).
24. D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization (2014).
25. D. J. Hand and R. J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, *Machine learning* **45**, 171 (2001).
26. R. M. Neal, *Bayesian Learning for Neural Networks* (Springer-Verlag, Berlin, Heidelberg, 1996).
27. S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C. J. Kelly, D. King, J. R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K. C. Young, J. De Fauw and S. Shetty, International evaluation of an AI system for breast cancer screening, *Nature* **577**, 89 (2020).