Journal of Machine Learning Research 23 (2022) 1-53 Submitted 6/20; Revised 10/22; Published 11/22

Stable Classification

Dimitris Bertsimas dbertsim@mit.edu
Sloan School of Management and Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Jack Dunn jack@interpretable.ai
Interpretable AI
1 Broadway, 14th Floor
Cambridge, MA 02142, USA

Ivan Paskov ipaskov@mit.edu
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Editor: Philipp Hennig

Abstract

We address the problem of instability of classification models: small changes in the training data leading to large changes in the resulting model and predictions. This phenomenon is especially well established for single tree based methods such as CART, however it is present in all classification methods. We apply robust optimization to improve the stability of four of the most commonly used classification methods: Random Forests, Logistic Regression, Support Vector Machines, and Optimal Classification Trees. Through experiments on 30 data sets with sizes ranging between 10^2 and 10^4 observations and features, we show that our approach (a) leads to improvements in stability, and in some cases accuracy, compared to the original methods, with the gains in stability being particularly significant (even, surprisingly, for those methods that were previously thought to be stable, such as Random Forests) and (b) has computational times comparable with (and indeed in some cases even faster than) the original methods allowing the method to be very scalable.

Keywords: stability, optimal decision trees, robustness, interpretability, logistic regression, support vector machines, classification

1. Introduction

We address the problem of instability of classification models: small changes in the training data leading to large changes in the resulting model and predictions. Such instability arises due to two primary sources: (a) Training Instability: variability arising due to the choice of training/validation split, and (b) Temporal Instability: variability arising due to receiving new data over time. Decision tree based methods such as CART are well known to exhibit both such forms of instability and high variance. Indeed, it was this very issue that motivated Breiman (1996a) to develop Bagging and Breiman (2001) to further refine Bagging with Random Forests, which are explicitly designed to reduce such instability via averaging. While certainly more stable than CART, the cost of increasing stability was high: Random Forests are by and large uninterpretable, and Breiman (1996b) asks "whether there is a more stable single-tree version of CART."

In this paper, we answer this question in the affirmative. Moreover, despite Random Forests being more stable with respect to the choice of training/validation split, it still suffers from temporal instability, and in general it is still an open question whether its overall stability can be improved. The same applies to logistic regression, as it is well known to suffer from instability of its parameter estimates, especially when the classes are well separated, see Hastie et al. (2001) for more information. These questions too we answer in the affirmative in this paper. More precisely, we generalize the robust optimization based approach for constructing stable linear regression models developed by Bertsimas and Paskov (2020) to general classification methods. Specifically, we develop a methodology for building classification models that are robust with respect to how the data set is split into training and validation sets. We apply this approach to four popular classification methods: Random Forests (RF), Logistic Regression (LR), Support Vector Machines (SVM), and Optimal Classification Trees (OCT). Through experiments on 30 data sets with sizes ranging between 10^2 and 10^4 observations and features, we show that our approach (a) leads to improvements in stability, and in some cases accuracy, compared to the original methods, with the gains in stability being particularly significant (even, surprisingly, for those methods that were previously thought to be stable, such as Random Forests) and (b) has computational times comparable with the original methods allowing the method to be very scalable.

1.1 Literature

There are many different notions of stability and robustness in the literature. For example, Bousquet and Elisseeff (2002) consider the problem of quantifying the stability of learning algorithms with respect to perturbation or removal of any single point in the training set. A different but often-considered notion of model stability is the robustness of the model predictions in the face of adversarial attacks (for example, see Madry et al., 2017). In this work, we will focus on the stability of the model in the sense of typical machine learning workflow. Specifically, when it is required to split a data set into different subsets (i.e., for training, validation and testing), we are interested in developing approaches that increase the stability of the resulting model with respect to the particular split of the data set.

The idea of using optimization (over randomization) to build regression models that are robust with respect to the subsample of data they are trained upon was first developed by Bertsimas and Paskov (2020) building on the theme of using optimization versus randomization in machine learning models, see Chapters 15-18 in Bertsimas and Dunn (2019). Bertsimas and Paskov (2020) use robust optimization techniques to formulate the problem of finding a linear regression that is robust with respect to the choice of training split. They demonstrate that such an approach constructs linear regression models that have both improved performance and improved stability compared to their non-robust counterparts, while also remaining tractable. In this paper, we extend this methodology beyond linear regression to some of the most popular classification methods: RF, LR, SVM, and OCT introduced in Breiman (2001), Cox (1966), Vapnik (1963), and Bertsimas and Dunn (2017), respectively. We also extend this work by developing alternate solution methodologies for this class of problems that allow us to solve the robust optimization problem even when reformulation into a robust counterpart via duality is either not feasible or impossible.

CART (Breiman et al., 1984) has long held a reputation of instability. One reason for this is that small changes in the training data can easily lead to different split decisions being made early in the tree training process, which in turn changes how the algorithm proceeds recursively, and can result in large changes in the final tree. Another source of instability is the challenge of finding the "right-sized tree" through hyperparameter valididation, as Breiman (1996b) shows that the regularization process of CART is unstable, meaning that small changes to the training set can lead to large changes in the selected hyperparameter value. To address this, there have many approaches aimed at improving the stability of tree-based methods such as bagging developed by Breiman (1996a), boosting developed by Freund and Schapire (1995) and Random Forests, developed by Breiman (2001). All three of these methods aim to stabilize the output of the final trained predictor by combining the predictions of multiple sub-models, thus minimizing the impact the instability of any one particular sub-model can have on the stability of the overall process. Additionally, the trees in a Random Forest are usually trained as deeply as possible, which obviates the need for the unstable hyperparameter validation used by CART to find the right-sized tree. Indeed, Breiman (1996b) proposed averaging as a means of stabilizing any general method, albeit at the cost of interpretability, and while approaches like Random Forests have better stabilty and performance than CART, interpretability is sacrificed. Last et al. (2002) develop a different approach for stabilizing CART, where they attempt to use statistical significance testing and pruning to produce stable trees. While more stable than CART, their approach unfortunately suffers from poor accuracy.

In a different stream of work, Duchi and Namkoong (2021), Shafieezadeh-Abadeh et al. (2015), Mohajerin Esfahani and Kuhn (2018), and Duchi et al. (2021) approach a related problem from a distributionally robust framework. Compared to these approaches, the advantage of our method is that it is nonparametric and thus more flexible, as well as significantly more computationally efficient, due to these approaches being posed as distributionally robust optimization problems whereas ours reduces to a convex optimization problem.

Finally, to the best of our knowledge, no prior work exists attempting to stabilize RF or SVM, likely because these methods are already widely believed to be stable. Indeed, RF was explicitly designed to further stabilize the bagging procedure by averaging uncorrelated trees (see Breiman, 2001 for more detail) and empirically such models are generally significantly more stable than CART, so the lack of work attempting further stabilization may simply be because RF are usually "stable enough". SVM are also generally considered stable, which could be explained by the fact that changing any point in the training data that is not a support vector will not affect the solution, and so such models may not appear as susceptible to data perturbations.

1.2 Contributions and Structure

In this paper, we extend the approach of Bertsimas and Paskov (2020) to general classification problems. We develop a robust optimization framework for stabilizing any classification method, and apply it to RF, LR, SVM, and OCT. We present three approaches: Robust Counterpart, Cutting Planes and Monte Carlo. Through experiments on 30 data sets, we show that the stable methods improve stability, and in some cases accuracy, compared to the original methods, with the gains in stability being particularly significant. We also demonstrate empirically that surprisingly this approach benefits methods that are generally thought of as stable already, such as Random Forests.

In Section 2, we describe the general stable methodology, as well as how to quantify the stability of a method. In Section 3, we discuss how to efficiently compute stable solutions. In Section 4, we present computational results comparing four classification methods to their stable counterparts. In Section 5, we present benchmarks of the runtimes of the stable algorithms. In Section 6, we present a convergence analysis of behavior as the number of iterations of the algorithms increases. In Section 7, we summarize our results and report our conclusions.

2. The Stable Methodology

In this section, we describe a way to quantify the stability of a method, and then use this measure to derive the general stable methodology.

As a motivating example, consider a hypothetical scenario in healthcare where we are constructing a system for automatically emitting alerts when a patient is at risk of sepsis. In this setting, not only are we concerned about the accuracy of the predictions, but also their stability as the model is updated over time. It would be undesirable that retraining the model might cause a large number of patients to suddenly receive alerts because the predictions have changed significantly. Ideally, we would have a training process that results in models that generate similar predictions for any given patient regardless of the specific data set used for training.

Suppose that we are trying to select between two approaches for training logistic regression models for this problem (for instance different regularization schemes). A typical approach is to split the data set into multiple pairs of training and validation sets (e.g., with cross

Approach	Result	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Logistic Regression #1	Accuracy	0.840	0.840	0.850	0.860	0.860
	Coefficients	[0.0, 1.0, 3.1]	[0.0, 1.1, 3.0]	[0.0, 1.2, 3.1]	[0.0, 0.9, 3.1]	[0.0, 1.0, 3.0]
Logistic Regression #2	Accuracy	0.835	0.850	0.850	0.850	0.865
	Coefficients	[0.2, 0.8, 2.9]	[0.3, 1.0, 2.5]	[0.2, 1.0, 3.2]	[0.1, 1.2, 2.8]	[0.4, 0.6, 3.0]

Table 1: Synthetic example comparing the results of two logistic regression approaches across multiple folds of a data set.

validation), use each training set to construct a model, and use the corresponding validation set to evaluate the model. We can then average the performance across each of the training/validation pairs to compare the different approaches and select the winner. Table 1 shows a synthetic example of such a setup, reporting both the validation accuracy and fitted model coefficients for each approach on each of the cross-validation folds.

The conventional way of selecting between these approaches would be to consider the mean or even the standard deviation of accuracy across the folds, however in this example these are identical between the methods. Despite these metrics being the same, upon further inspection we can see that there are significant differences between the approaches. The accuracy of the first approach is tighter around 0.85, but has more frequent deviation from this mean value, whereas the second approach has less frequent but larger deviations from the mean. We also see that the model coefficients have higher volatility in the second approach, while the first approach is relatively much more stable and only ever selects the same two features. In the context of our sepsis alert system, it seems clear that each of the five models in the first approach are likely to generate alerts for the same patients, whereas the five models of the second approach might generate alerts for very different sets of patients due to the high variability in coefficients. For these reasons, it seems more likely that the first approach would lead to more stable predictions and performance as we generalize to new data, as it is reliably generating similar models.

The point of this example is to demonstrate that looking only at the mean and variance of the performance across the folds provides a limited view into the characteristics of each approach. Instead, if we are aiming to develop an approach that we are confident will generalize well to new data, we might seek to compare the approaches not only by performance, but also other metrics that assess the stability of the trained models. In this way, the problem is really a multi-objective problem where we should consider both stability and performance simultaneously.

2.1 Measuring Stability

If our motivation is to construct more stable models, we must first have some way of quantifying the stability of a model. In this section, we describe a number of such measures.

Each of the meaures presented require s different models trained on different variants of the training data each time. These models are constructed through the following process:

1. Split the data into training and testing sets.
2. Repeat s times:
   (a) Take a bootstrap sample from the training set.
   (b) Train a model on this bootstrap sample.
   (c) Use the trained model to make predictions on the testing set.

This process results in s models, each trained on a different bootstrap sample of the training data, and used to make predictions on the same testing set. We note that the proposed measures involve calculating variances across different aspects of these s models, and thus it is important that we select s sufficiently high to generate enough models and arrive at meaningful variance calculations. As an example, in the experiments of Section 4 we use $s = 100$.

Next, we discuss how to use these results to calculate various measures of stability across the s models.

2.1.1 Output Stability

An important measure of the stability of a method is the variability of its outputs. For each point in the test set, we have s different predicted probabilities, one from each model. Intuitively, the variation among these s different predictions is related to the stability of the process, as an approach with higher stability would ideally lead to more consistent probability predictions.

To measure this, we calculate the variance of these predictions for each of the points in the test set, and then average these across all testing points. We will assume for simplicity that we are dealing with a binary classification problem and so the predicted probability can just be taken to be the probability of belonging to the second class in the problem (the extension to a multi-class setting is trivial by introducing another loop over all classes). Concretely, if our test set has n_{test} observations, and $\widehat{p}_{ij}$ is the predicted probability of model $j \in [s] = \{1, \ldots, s\}$ for point $i \in [n_{\text{test}}]$, then we define

$$\text{OutputStabilityScore} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left[\frac{1}{s-1} \sum_{j=1}^{s} \left(\widehat{p}_{ij} - \frac{1}{s} \sum_{j=1}^{s} \widehat{p}_{ij} \right)^2 \right].$$

This score quantifies the stability of the probability predictions of any model-training approach, with a lower score indicating higher stability. Indeed, if we consider an approach with no instability and where each of the s models generated the same predictions, we would have zero variance in the predictions for each point, and so the overall output stability score would be zero. On the other hand, if the predictions for any given point are significantly different between each of the s models, the point-wise variances would be high, and thus the overall output stability score would also be high. Finally, we note that the quality of this metric depends on the size of the test set n_{test}, and so it is important that our test set is sufficiently large to ensure we are averaging these variances over a representative sample of datapoints.

2.1.2 Structural Stability

We can also quantify the stability by assessing the similarity in structure across the trained models, to give some indication of how much the underlying model is changing structurally in response to data changes.

In the case of parameterized models (such as LR and SVM), we can simply measure the standard deviation of the parameters in the model over the s models. For non-parametric models (such as RF and OCT), we propose calculating the structural stability by first calculating the variable importance scores for each model, and then measuring the standard deviation in the importance score for each feature across the s models.

2.1.3 Hyperparameter Stability

A third way to quantify stability is to assess the variability of the final values of any hyperparameters that are tuned during the training and validation process, as a procedure that consistently estimates the same tuned value should lead to more stable performance. We propose measuring this using the standard deviation in the tuned hyperparameter values across the s models.

2.2 The Stable Methodology

With a measure of stability defined, we now proceed to derive a methodology for building stable models. At a high-level, what we would like to do is construct a model that is robust with respect to the specific data set that is used for training the model. One way to think about this is to view the training data set as a sample from the true data distribution, and then require that the resulting model be robust with respect to the specific sample that was received. Viewing the partitioning of the data into training/validation sets as a sampling mechanism from this true data distribution (because for a given choice of split, we get one training set), we desire to build models that are robust with respect to the choice of training/validation split.

Method	Model m	Model Class $\mathcal{M}$	Loss $f(m,x,y)$	Algorithmic Comments
SVM	(β, β_0)	$\mathbb{R}^p \times \mathbb{R}$	$\max\{0, 1 - y_i(\beta^T x_i + \beta_0)\}$	Linear Optimization Problem
LR	(β, β_0)	$\mathbb{R}^p \times \mathbb{R}$	$\log(1 + e^{-y_i(\beta^T x_i + \beta_0)})$	Convex Optimization Problem
OCT	tree of fixed depth	set of all tree models	$1 - p_{\{i,y_i\}}$	Solved to Optimality
RF	set of trees of fixed depth	set of all tree models	$1 - p_{\{i,y_i\}}$	Bagging De-correlated Trees

Table 2: List of the Model, Model Class, Loss Function, and Algorithmic details for the four methods considered in this paper. Note that below $p_{\{i,k\}}$ is the predicted probability of getting label k for point i.

We begin by considering a general model formulation:

$$\min_{m \in \mathcal{M}} \sum_{i=1}^{n} f(m, x_i, y_i), \tag{1}$$

where m is a model optimized over a class of models $\mathcal{M}$, and $f(m,x,y)$ gives the cost of applying model m to a given datapoint (x,y). We list in Table 2 the corresponding model class and loss function for the four classification problems considered in this paper. At this time we additionally note that this formulation is compatible with any basis function expansion of the given covariates (i.e. polynomials, step functions, splines (using the truncated power basis representation, see Hastie et al., 2001 for more information), kernels, wavelets, Fourier series, etc).

Now, we would like to find a model that is robust with respect to the choice of training/validation split. A way to achieve this is to associate each observations (x_i, y_i) to a binary variable z_i, $i \in [n]$ that indicates whether or not (x_i, y_i) participates in the training set. We can then train a given classification algorithm over all possible allocations of these z_i's, resulting in a model that is explicitly built to do well not just over one training set, as is typical, but over all possible training sets. This can be formalized as the following problem:

$$\min_{m \in \mathcal{M}} \max_{z \in \mathcal{Z}} \sum_{i=1}^{n} z_i f(m, x_i, y_i), \tag{2}$$

where $\mathcal{Z}$ is the so-called uncertainty set in the language of robust optimization. In this way, we must now optimize a model that minimizes the worst-case training error across elements of $\mathcal{Z}$.

A natural choice of uncertainty set is all subsets of size k:

$$\mathcal{Z} = \left\{ z : \ \sum_{i=1}^{n} z_i = k, \quad z_i \in \{0,1\},\ i \in [n] \right\}.$$

At an optimal solution of (2), each z_i will be equal to either 0 or 1, with the interpretation that if $z_i = 1$, then point (x_i, y_i) is assigned to the training set, otherwise it it is assigned to the validation set. The number k indicates the desired proportion between the size of the training and validations sets. Namely, by setting $k = 0.7n$ we recover the typical 70/30 training/validation split and by setting $k = 0.5n$ we recover the 50/50 training/validation split, etc.

In this way, the above formulation is a faithful translation of our earlier intuition: find a model m that does the best against the hardest subset of size k in the data. Our choice to minimize over the worst-case training error rather than an average-case is primarily motivated by computational efficiency; as we will discuss in Section 3, this robust optimization formulation allows us to optimize over the worst-case without meaningfully changing the complexity of the problem. In contrast, optimizing over an average-case would require us to explicitly form a large-enough set of cases over which to optimize, resulting in a significant increase in the number of variables in the optimization problem and thus likely affecting the tractability.

3. Computing Stable Solutions

In this section, we describe how to compute stable solutions by solving Problem 2. As we described in the previous section, our formulation belongs to the class of robust optimization (RO) problems, see Bertsimas and den Hertog (2022). The two most frequently described methods in the literature for solving such problems are reformulation to a deterministic optimization problem (often called the robust counterpart) or an iterative cutting-plane method. Bertsimas et al. (2015) show that both approaches are tractable. In this section, we also develop a third approach based on Monte Carlo simulation that applies widely (in particular to all four problems we consider), while remaining competitive in terms of performance.

In what follows, we first derive the robust counterpart for Problem 2. We then describe how to apply the cutting plane algorithm for Problem 2. Finally, we introduce our third approach for solving RO problems and show how to apply it to Problem 2.

3.1 Tractable Robust Counterpart

Consider again the stable formulation:

$$\min_{m \in \mathcal{M}} \max_{z \in \mathcal{Z}} \sum_{i=1}^{n} z_i f(m, x_i, y_i) \quad \text{with} \quad \mathcal{Z} = \left\{ z : \ \sum_{i=1}^{n} z_i = k, \quad z_i \in \{0,1\}, \ i \in [n] \right\}. \tag{3}$$

As the inner maximization problem is linear in z, the problem is equivalent to optimizing over the convex hull of $\mathcal{Z}$

$$\text{conv}(\mathcal{Z}) = \left\{ z : \ \sum_{i=1}^{n} z_i = k, \quad 0 \le z_i \le 1, \ i \in [n] \right\}.$$

Thus, Problem 3 is equivalent to

$$\min_{m \in \mathcal{M}} \max_{z \in \text{conv}(\mathcal{Z})} \sum_{i=1}^{n} z_i f(m, x_i, y_i) \quad \text{with} \quad \text{conv}(\mathcal{Z}) = \left\{ z : \sum_{i=1}^{n} z_i = k, 0 \leq z_i \leq 1,\ i \in [n] \right\}. \tag{4}$$

Problem 4 belongs to the class of robust optimization problems, see Bertsimas and den Hertog (2022) and Bertsimas et al. (2011) for a review. We leverage techniques from RO to solve Problem 4 efficiently. Namely, to alleviate the multiplication of variables (i.e., the product of z_i with $f(m, x_i, y_i)$) we take the linear optimization dual of the inner maximization problem

$$\max_{z_i} \sum_{i=1}^{n} z_i f(m, x_i, y_i) \quad \text{subject to} \quad \sum_{i=1}^{n} z_i = k, \quad 0 \leq z_i \leq 1,\ i \in [n]$$

by introducing the dual variable θ for the first constraint and the dual variables $u_i,\ i \in [n]$ for the second set of constraints to arrive at:

$$\min_{\theta, u_i} \ k\theta + \sum_{i=1}^{n} u_i \quad \text{subject to} \quad \theta + u_i \geq f(m, x_i, y_i), \quad u_i \geq 0,\ i \in [n].$$

Substituting this minimization problem back into the outer minimization we arrive at the following problem:

$$\min_{\substack{m \in \mathcal{M}; \\ \theta, u_i \in \mathbb{R}}} k\theta + \sum_{i=1}^{n} u_i \quad \text{subject to} \quad \theta + u_i \geq f(m, x_i, y_i), \quad u_i \geq 0,\ i \in [n]. \tag{5}$$

This is a convex optimization problem for $f(\cdot)$ convex, and hence can be solved by commercial optimization software in very high dimensions. Using the formulas for $f(\cdot)$ from Table 2 we have that the stable robust counterparts for SVM and LR

$$\min_{\beta, \beta_0, \theta, u_i} \ k\theta + \sum_{i=1}^{n} u_i \quad \text{subject to} \quad \theta + u_i \geq max\{0, 1 - y_i(\beta^T x_i + \beta_0)\}), \quad u_i \geq 0,\ i \in [n], \tag{6}$$

$$\min_{\beta, \beta_0, \theta, u_i} \ k\theta + \sum_{i=1}^{n} u_i \quad \text{subject to} \quad \theta + u_i \geq \log(1 + e^{-y_i(\beta^T x_i + \beta_0)}), \quad u_i \geq 0,\ i \in [n], \tag{7}$$

respectively. Note that the robust counterpart of Stable SVM (Problem 6) is a linear optimization problem, easily solvable for very large dimensions, see Bertsimas and Tsitsiklis (1997) for more details, while the robust counterpart of Stable LR (Problem 7) is a convex optimization problem, easily solvable for large dimensions, see Boyd and Vandenberghe (2004) for more details. We remark that the robust counterpart method only applies for SVM and LR.

3.2 Cutting Plane Algorithm

We next describe how to apply the cutting plane algorithm to Problem 2. We start with the stable formulation:

$$\min_{m \in \mathcal{M}} \max_{z \in \mathcal{Z}} \sum_{i=1}^{n} z_i f(m, x_i, y_i) \quad \text{with} \quad \mathcal{Z} = \left\{ z : \sum_{i=1}^{n} z_i = k, \quad z_i \in \{0,1\},\ i \in [n] \right\}.$$

Re-expressing this in an equivalent epigraph formulation we obtain

$$\min_{\substack{m \in \mathcal{M};\\ t \in \mathbb{R}}} t \quad \text{s.t.} \quad t \geq \max_{z \in \mathcal{Z}} \sum_{i=1}^{n} z_i f(m, x_i, y_i),\ \mathcal{Z} = \left\{ z : \sum_{i=1}^{n} z_i = k, \quad z_i \in \{0,1\},\ i \in [n] \right\}, \tag{8}$$

which is equivalent to:

$$\min_{\substack{m \in \mathcal{M};\\ t \in \mathbb{R}}} t \quad \text{s.t.} \quad t \geq \sum_{i=1}^{n} z_i f(m, x_i, y_i),\ \forall z \in \mathcal{Z} = \left\{ z : \sum_{i=1}^{n} z_i = k, \quad z_i \in \{0,1\},\ i \in [n] \right\}. \tag{9}$$

We now begin with some random subset $\mathcal{Z}_1 \subset \mathcal{Z}$ and solve

$$\min_{\substack{m \in \mathcal{M};\\ t \in \mathbb{R}}} t \quad \text{s.t.} \quad t \geq \sum_{i=1}^{n} z_i f(m, x_i, y_i) \quad \forall z \in \mathcal{Z}_1. \tag{10}$$

We let m_1^*, t_1^* denote minimizers of Problem 10 and search for a violated constraint in the original problem by computing: $\max_{z \in \mathcal{Z}} \sum_{i=1}^{n} z_i f(m_1^*, x_i, y_i)$. Denote the optimum value of this c^* and the maximizing z by z^*. If $t_1^* \geq c^*$, then m_1^* is optimal for the original problem and we are done. If $t_1^* < c^*$, then the constraint $t \geq \sum_{i=1}^{n} z_i^* f(m_1^*, x_i, y_i)$ is violated in the original problem. In this case, we need to add this constraint to Problem 10 and repeat, i.e., let $\mathcal{Z}_2 = \mathcal{Z}_1 \cup \{z^*\}$ and then solve:

$$\min_{\substack{m \in \mathcal{M};\\ t \in \mathbb{R}}} t \quad \text{s.t.} \quad t \geq \sum_{i=1}^{n} z_i f(m, x_i, y_i) \quad \forall z \in \mathcal{Z}_2, \tag{11}$$

and then repeat this procedure until we find an optimum solution. The algorithm converges as discussed in Fletcher and Leyffer (1994). The method applies to all four classification problems we consider in this paper.

3.3 Monte Carlo

While the cutting plane algorithm described in the previous section is theoretically guaranteed to eventually discover the optimal solution, in practice it may be very slow, especially if Problem 10 is not easy to solve, as is the case with OCT. The reason for the difficulty is the need to solve nested versions of Problem 10 in a loop potentially many times. Instead, we introduce the idea to randomly sample a number ζ of points without replacement from $\mathcal{Z}$, denote this collection $\mathcal{Z}_\zeta$ and solve:

$$\min_{\substack{m\in\mathcal{M};\\ t\in\mathbb{R}}} t \quad \text{s.t.} \quad t \geq \sum_{i=1}^{n} z_i f(m, x_i, y_i) \quad \forall z \in \mathcal{Z}_\zeta, \tag{12}$$

and return the resulting $m \in \mathcal{M}$. The method was introduced in Calafiore and Campi (2006) and Campi et al. (2018), where probabilistic guarantees are derived for the solution to be feasible with high probability.

The advantages of this approach are:

(a) it is very fast as we only need to solve Problem 12 once;

(b) it applies to all four classification methods we consider in this paper;

(c) its performance is comparable with the robust counterpart and the cutting planes methods.

While the solution is random as it is dependent on the random sample chosen, we can eliminate the randomness in the solution by employing a scheme similar to that derived in Wyner (1967), wherein the user constructs deterministic sequences to model uniformly distributed points.

4. Computational Experiments:

In this section, we present computational results comparing the four classification methods to their stable counterparts. We compare these methods along the metrics of accuracy and stability. For accuracy, we report accuracy (we also computed Area Under the Curve (AUC) and saw that the results were similar). For stability, we report output stability, structural stability, and hyperparameter stability. We include average results averaged across the 30 data sets. Note that for hyperparameter stability, we take the geometric average as the hyperparamater is tuned over a range of different orders of magnitude. Finally, we also replicate the above experimental setup by tuning the hyperparameter using cross-validation rather than validation. The full results at the individual data set level can be found in the appendix.

4.1 Testing Methodology

To compare the classification methods to their stable counterparts, we collected 30 data sets from the UCI Machine Learning Repository (Dua and Taniskidou, 2017). The exact list of data sets can be found in the appendix. For each data set, we employ the methodology of Section 2.1 as follows:

1. We split the data randomly into 90% training and 10% testing set.
2. We repeat the following process $s = 100$ times:
   (a) Take a bootstrap sample of the training data.
   (b) For each method, train a model on this bootstrap sample.
3. Using the resulting 100 models for each method, calculate the average accuracy on the test set, along with the output, structural, and hyperparameter stability for all methods.

We consider the following methods:

- **SVM**: ℓ_2-regularized Support Vector Machines, tuning the regularization parameter.
- **LR**: ℓ_2-regularized Logistic Regression, tuning the regularization parameter.
- **RF**: Random Forests with 100 trees, tuning the minbucket parameter.
- **OCT**: Optimal Classification Trees, tuning the complexity parameter.

We compare the following variants for each method:

- **Original:** The nominal approach.
- **SMC:** The Stable Monte Carlo approach with $\zeta = 20$ in all cases, as we observed this was typically enough iterations for the metrics to stabilize (to illustrate this, Figures 1 and 2 show a representative example of these metrics when solving Problem 12 for each $\zeta \in \{1, \ldots, 20\}$).
- **SCP:** We run the Stable Cutting Plane approach until convergence.
- **SRC:** Where available, we solve the Stable Robust Counterpart directly (Problems 6 and 7, for SVM and LR, respectively).

We repeat the experiments using each of the following approaches to tune hyperparameter values:

- **Single split:** We split the bootstrap sample into 70% training and 30% validation, and select the hyperparameter value that leads to the best validation performance.
- **Cross-validation:** We perform 5-fold cross-validation on the bootstrap sample and select the hyperparameter value with the best average out-of-fold performance.

	Accuracy	Output Stability	Structural Stability	Hyperparameter Stability
Original	0.773 (1.67)	8.493×10^{-3} (2.43)	0.821 (2.57)	1.860×10^{1} (1.23)
SMC	0.761 (2.43)	9.567×10^{-3} (2.57)	0.761 (1.97)	1.878×10^{1} (2.60)
SCP	0.795 (2.03)	5.654×10^{-3} (2.63)	0.806 (2.83)	1.871×10^{1} (1.73)
SRC	0.807 (1.67)	5.672×10^{-3} (2.37)	0.823 (2.63)	1.869×10^{1} (1.57)

Table 3: Comparison of accuracy, output stability, structural stability, and hyperparameter stability for original, SMC, SCP, and SRC versions of SVM.

	Accuracy	Output Stability	Structural Stability	Hyperparameter Stability
Original	0.779 (1.6)	7.508×10^{-3} (2.63)	0.753 (2.8)	1.842×10^{1} (1.5)
SMC	0.775 (2.6)	7.132×10^{-3} (2.60)	0.681 (2.1)	1.865×10^{1} (2.2)
SCP	0.804 (1.7)	4.059×10^{-3} (2.60)	0.752 (2.8)	1.842×10^{1} (1.7)
SRC	0.813 (1.5)	3.862×10^{-3} (2.17)	0.734 (2.3)	1.842×10^{1} (1.5)

Table 4: Comparison of cross-validated accuracy, output stability, structural stability and hyperparameter stability for original, SMC, SCP, and SRC versions of SVM.

4.2 Support Vector Machines

In Table 3, we report the accuracy, output stability, structural stability, and hyperparameter stability for SVM, Stable-Monte Carlo (SMC), Stable - Cutting Plane (SCP), and Stable - Robust Counterpart (SRC). Each entry in Table 3 represents the average metric value for the corresponding method/metric pair over the 30 data sets from the UCI Machine Learning Repository. For accuracy higher numbers are desirable as they indicate greater predictive accuracy. For output stability, structural stability, and hyperparameter stability, lower numbers are desirable as they indicate greater stability. We also include (in parenthesis) the average rank achieved by that method/metric pair across the 30 data sets, where lower numbers are desirable.

The same is repeated in Table 4 for the cross-validation experiments, with an additional column recording the average hyperparameter stability over the 30 data sets. As with the other stability measures, lower numbers are desirable.

Tables 3 and 4 both indicate that the stable methodology improves both the accuracy of the original method as well as its stability; indeed we see improvements across accuracy, output stability, and structural stability. On hyperparameter stability, the methods are similar, with perhaps a slight edge given to the nominal. Interestingly, we observe strong performance from SMC on structural stability, but lags behind the other stable variants on the other metrics. Indeed, generally SRC achieves the strongest performance, with SCP performing fairly similarly, as is to be expected.

	Accuracy	Output Stability	Structural Stability	Hyperparameter Stability
Original	0.672 (1.43)	1.004×10^{-2} (2.87)	3.258 (2.90)	1.848×10^{1} (2.83)
SMC	0.666 (2.53)	1.042×10^{-2} (3.23)	2.975 (2.33)	1.857×10^{1} (2.60)
SCP	0.657 (3.40)	8.083×10^{-3} (1.77)	2.941 (2.07)	1.849×10^{1} (1.77)
SRC	0.668 (2.63)	9.283×10^{-3} (2.13)	3.292 (2.70)	1.822×10^{1} (1.73)

Table 5: Comparison of accuracy, output stability, structural stability, and hyperparameter stability for original, SMC, SCP, and SRC versions of LR.

	ACC	Output Stability	Structural Stability	Hyperparamter Stability
Original	0.671 (1.43)	8.586×10^{-3} (2.70)	3.128 (2.77)	1.808×10^{1} (2.57)
SMC	0.663 (2.60)	8.694×10^{-3} (3.00)	2.793 (2.30)	1.842×10^{1} (2.30)
SCP	0.657 (3.33)	7.469×10^{-3} (2.00)	2.831 (2.37)	1.813×10^{1} (1.77)
SRC	0.663 (2.63)	7.387×10^{-3} (2.30)	2.909 (2.57)	1.799×10^{1} (1.43)

Table 6: Comparison of cross-validated accuracy, output stability, structural stability, and hyperparameter stability for original, SMC, SCP, and SRC versions of LR.

4.3 Logistic Regression

In Table 5, we report the accuracy, output stability, structural stability, and hyperparameter stability for LR, Stable-Monte Carlo (SMC), Stable - Cutting Plane (SCP), and Stable - Robust Counterpart (SRC). The structure of Table 5 is identical to that of Table 3. Table 5 indicates that the stable methodology improves the stability but not the accuracy of the original method. In particular, we see that the original, SMC and SRC are about the same in terms of accuracy, with SCP slightly lower than the others. In contrast to the effect on accuracy, the stable methodology provides a strong improvement on all three stability metrics (output, structural, and hyperparameter), particularly for SCP. Table 6 indicates a similar story, with the stable methods improving upon the original in terms of stability, but this time at the cost of a small decrease in accuracy.

4.4 Random Forests

In Table 7, we report results on RF. We observe very modest improvements in accuracy, and larger improvements in output and structural stability. In particular, we see that SMC has a small edge in terms of accuracy over the original, that SCP has a small edge in terms of output stability, and that in terms of hyperparameter stability, both SMC and SCP show an improvement over the original, with the difference being the greatest for SMC. This latter point on stability is particularly significant as RF is generally regarded as stable, given this was a goal of its design. Table 8 indicates a similar story, with the additional detail that SMC and the original now appear tied in terms of accuracy.

	Accuracy	Output Stability	Structural Stability	Hyperparameter Stability
Original	0.847 (2.0)	0.016 (2.1)	0.004 (1.7)	5.924 (2.0)
SMC	0.849 (1.8)	0.016 (2.1)	0.003 (1.6)	5.605 (1.8)
SCP	0.838 (2.1)	0.014 (1.8)	0.012 (2.7)	6.044 (2.1)

Table 7: Comparison of accuracy, output stability, structural stability, and hyperparameter stability for original, SMC, and SCP versions of RF.

	Accuracy	Output Stability	Structural Stability	Hyperparameter Stability
Original	0.854 (1.7)	0.0138 (2.2)	0.0032 (1.5)	4.292 (1.7)
SMC	0.854 (1.8)	0.0136 (1.9)	0.0030 (1.6)	3.818 (1.5)
SCP	0.842 (2.4)	0.0128 (1.9)	0.0141 (2.9)	4.513 (2.2)

Table 8: Comparison of cross-validated accuracy, output stability, structural stability, and hyperparameter stability for original, SMC, and SCP versions of RF.

4.5 Optimal Classification Trees

In Table 9, we report results on OCT. We observe that in terms of accuracy, the original has an edge. In terms of output stability, SMC again has sizable edge over both the original and SCP. Finally, in terms of hyperparameter stability, the original performs the best. Table 10 indicates a similar story, with again SMC achieving the strongest output stability, however now in terms of accuracy, SMC and the original are closer than before.

5. Computational Times

In this section, we compare the computational times of the Original, SMC, SCP, and SRC versions of the four methods, averaged across the 30 data sets. We note that the hardware used for all the experiments was a computer equipped with an Intel Core i9-9900K processor, while for the Software we used Julia 1.3.1, Ipopt 3.13.2 for LR, and Gurobi 9.0.0 for SVM.

The results can be found in Table 11, which is organized as follows: each row corresponds to an implementation, each column to a classification method. Entry (i, j) then corresponds to the average computational time for implementation i of classification method j. Note that

	Accuracy	Output Stability	Structural Stability	Hyperparameter Stability
Original	0.828 (1.6)	0.029 (2.0)	0.028 (1.6)	2.751 (1.8)
SMC	0.825 (2.0)	0.026 (1.8)	0.029 (2.2)	2.893 (2.3)
SCP	0.821 (2.3)	0.029 (2.1)	0.031 (2.2)	2.811 (1.9)

Table 9: Comparison of accuracy, output stability, structural stability, and hyperparameter stability for original, SMC, and SCP versions of OCT.

	Accuracy	Output Stability	Structural Stability	Hyperparameter Stability
Original	0.829 (1.7)	0.021 (2.0)	0.024 (1.9)	1.116 (1.4)
SMC	0.826 (1.8)	0.019 (1.6)	0.026 (2.0)	1.359 (1.9)
SCP	0.824 (2.3)	0.022 (2.3)	0.028 (2.1)	1.485 (2.1)

Table 10: Comparison of cross-validated accuracy, output stability, structural stability, and hyperparameter stability for original, SMC, and SCP versions of OCT.

	SVM	LR	RF	OCT
Original	1.000	1.000	1.000	1.000
SMC	0.398	0.625	**3.149**	**2.176**
SCP	1.621	4.093	8.920	2.880
Stable - Reformulation	**0.312**	**0.488**	NA	NA

Table 11: Comparison of the computational times of the SMC, SCP, and SRC versions of the four methods relative to the runtime of the original method, averaged across the 30 data sets. The best stable run time for each method has been bolded.

the times are first scaled so that the original method has time 1, so that the other method times indicate the overhead factor for that method, i.e., 2 means takes twice as long as the original method, 0.5 means takes half as long.

Overall, we note that the stable versions of each classification method have computational times comparable with the original methods, suggesting that the stable methodology is scalable. Indeed, we even see in a few cases the approach offers a speed improvement over the original (i.e., whenever a reformulation is possible, as well as for the SMC versions of all the methods except for RF and OCT). This may seem surprising, as one might expect the runtime to increase with additional constraints, however we believe that a plausible explanation is that the robust constraints make the optimal solution "more obvious" in some sense and thus able to be found faster. Finally, as expected, the SCP approach has the longest runtimes, in the worst case 8.9 times slower than the original, and in the best case 1.6 times slower than the original.

6. Convergence Analysis

Finally, to provide deeper insight into the fast runtimes offered by the SMC versions of each method, we present two representative plots of the evolution of accuracy and stability as a function of the number of iterations. Specifically, in Figure 1 we plot accuracy as a function of the number of iterations for the three stable variants of logistic regression, and then do the same in Figure 2 for output stability. We observed similar convergence behavior for the other methods and stability metrics.

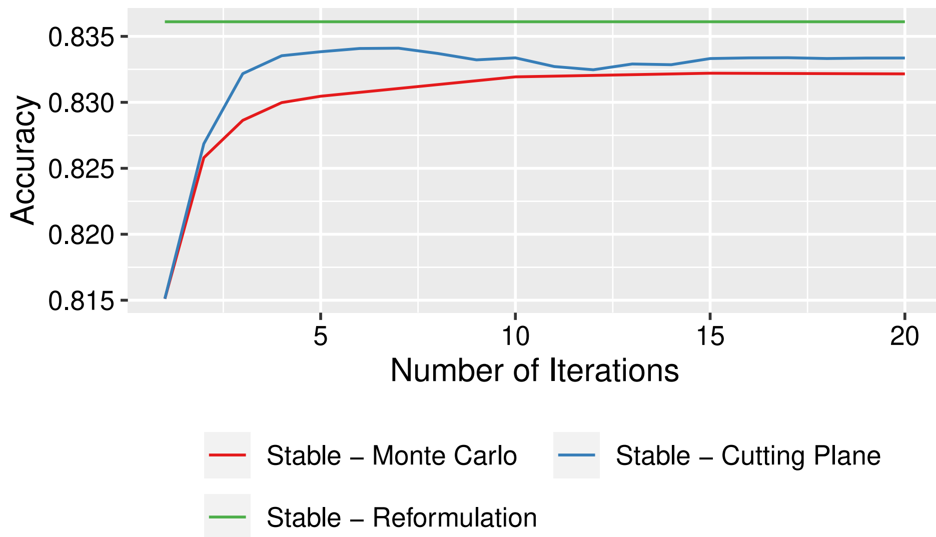

Figure 1: Comparison of LR accuracy between original, SCP, SMC and SRC as a function of the number of iterations.

For SMC, we repeatedly resolve the problem while progressively increasing the number of sets sampled from $\mathcal{Z}$ to show how the outcome varies with the number of sets sampled. For SCP, we show the outcome as a function of the number of iterations of the cutting plane algorithm (where each cut adds a new set from $\mathcal{Z}$ to the problem). The robust counterpart is solved in a single step, so is shown as a horizontal line.

We see that both SMC and SCP seem to converge in performance within five iterations. This indicates that SMC is able to approximate the set $\mathcal{Z}$ with relatively few samples, and that SCP only needs to consider a small number of "hardest" training sets to train a model that works well across all such training sets.

7. Conclusion

In this paper, we propose a robust optimization based framework for stabilizing any classification method and derive efficient algorithms that scale the approach to very large problem sizes. The approach is generally applicable to general classification problems. Through experiments on 30 data sets with sizes ranging between 10^2 and 10^4 observations and features, we show that our approach (a) leads to improvements in stability, and in some cases accuracy, compared to the original methods, with the gains in stability being particularly significant and (b) has computational times comparable with (and indeed in some cases even faster than) the original methods, allowing the approach to be very scalable.

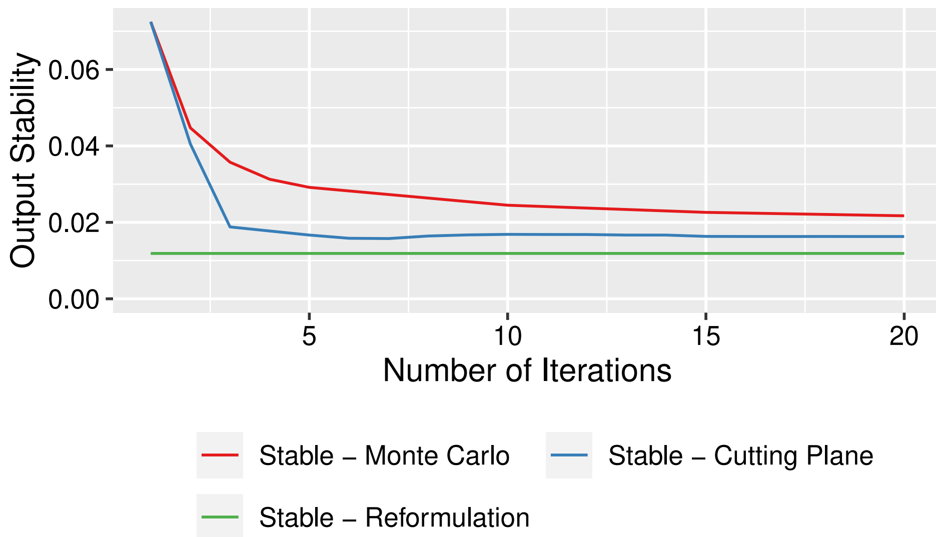

Figure 2: Comparison of LR output stability between original, SCP, SMC and SRC as a function of the number of iterations.

In the case of SVM and LR, we have the ability to derive tractable exact robust counterparts, and the results suggest that this approach is preferable as it leads to better performance over the SMC and SCP approaches, and surprisingly, faster run times than even the original method. In the case of RF and OCT, both the SCP and SMC approaches often showed improvements in stability. For these methods, the SMC approach was significantly faster than the SCP approach, while the performance and stability characteristics were similar, making the SMC approach more attractive.

What is perhaps most exciting, is that all of these benefits accrue to even the simplest implementation of stability: the Monte Carlo approach. In this approach, practitioners have a conceptually simple prescription for how to train models that barely increases the computational complexity over their un-stabilized counterparts. The fact that it leads to improvements in both stability and accuracy suggest that perhaps the current approaches to training algorithms have been operating at an interior point with respect to the performance/stability Pareto curve. The results, especially in the case of SVM, suggest that we can in fact make improvements in both accuracy and stability, without paying much of a computational cost, leaving the practitioner little reason not to employ the methodology.

Acknowledgments

We thank the action editor and the reviewers of the paper for many helpful comments that improved the paper significantly.

References

Dimitris Bertsimas and Dick den Hertog. *Robust and Adaptive Optimization*. Dynamic Ideas, 2022.

Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106: 1039–1082, 2017.

Dimitris Bertsimas and Jack Dunn. *Machine Learning under a Modern Optimization Lens*. Dynamic Ideas, 2019.

Dimitris Bertsimas and Ivan Paskov. Stable regression: On the power of optimization over randomization in training regression problems. *Journal of Machine Learning Research*, 21(230):1–25, 2020.

Dimitris Bertsimas and John Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

Dimitris Bertsimas, Iain Dunning, and Miles Lubin. Reformulation versus cutting-planes for robust optimization a computational study. *Computational Management Science*, 13 (2):195–217, 2015.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996a.

Leo Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383, 1996b.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

Giuseppe Calafiore and Marco Campi. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51(5):742–753, 2006.

Marco Campi, Simone Garatti, and Federico Ramponi. A general scenario theory for non-convex optimization and decision making. *IEEE Transactions on Automatic Control*, 63 (12):4067–4078, 2018.

David R. Cox. Some procedures associated with the logistic qualitative response curve. *Research papers in statistics: Festschrift for J. Neyman*, pages 55–71, 1966.

Dheeru Dua and Efi Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

John C. Duchi, Peter W. Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969, 2021.

Roger Fletcher and Sven Leyffer. Solving mixed integer nonlinear programs by outer approximation. *Mathematical Programming*, 66(1):327–349, 1994.

Yoav Freund and Robert Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. *European Conference on Computational Learning Theory*, pages 23–37, 1995.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2001.

Mark Last, Oded Maimon, and Einat Minkov. Improving stability of decision trees. *International journal of pattern recognition and artificial intelligence*, 16(2):145–159, 2002.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166, 2018.

Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1576–1584, 2015.

Vladimir Vapnik. Pattern recognition using generalized portrait method. *Automation and remote control*, 24:774–780, 1963.

Aaron Wyner. Random packings and coverings of the unit n-sphere. *The Bell System Technical Journal*, 46(9):2111–2118, 1967.

Appendix A. Individual data set results

In this section, we present computational results at an individual data set level comparing SVM, LR, RF, and OCT to their stable counterparts. We apply SRC, SCP and SMC for SVM and LR. We apply SCP and SMC for RF, and OCT. We compare these methods in performance and stability. For performance, we report accuracy (we also computed Area Under the Curve (AUC) and saw that the results were similar). For stability, we report output stability and hyperparameter stability. For SVM's and LR, we additionally report structural stability.

	Original	SMC	SCP	SRC
acute-inflammations-1	0.907	0.744	0.866	0.904
acute-inflammations-2	0.583	0.583	0.825	0.813
banknote-authentication	0.556	0.556	0.869	0.957
blood-transfusion-service-center	0.763	0.763	0.763	0.763
breast-cancer	0.932	0.934	0.932	0.932
breast-cancer-wisconsin-diagnostic	0.966	0.921	0.959	0.963
breast-cancer-wisconsin-original	0.757	0.758	0.757	0.757
breast-cancer-wisconsin-prognostic	0.769	0.770	0.769	0.769
climate-model-simulation-crashes	0.950	0.940	0.950	0.950
congressional-voting-records	0.929	0.909	0.931	0.929
connectionist-bench-sonar	0.629	0.622	0.645	0.719
credit-approval	0.828	0.830	0.822	0.828
fertility	0.863	0.865	0.863	0.863
haberman-survival	0.736	0.736	0.736	0.736
hepatitis	0.833	0.833	0.833	0.833
indian-liver-patient	0.713	0.713	0.713	0.713
ionosphere	0.837	0.826	0.805	0.814
mammographic-mass	0.828	0.830	0.821	0.826
monks-problems-1	0.781	0.766	0.777	0.777
monks-problems-2	0.591	0.613	0.558	0.571
monks-problems-3	0.841	0.731	0.814	0.835
parkinsons	0.849	0.842	0.849	0.849
planning-relax	0.709	0.709	0.709	0.709
qsar-biodegradation	0.872	0.866	0.872	0.872
seismic-bumps	0.934	0.934	0.934	0.934
spect-heart	0.500	0.500	0.578	0.686
spectf-heart	0.500	0.500	0.656	0.654
statlog-project-german-credit	0.739	0.736	0.739	0.739
thoracic-surgery	0.850	0.851	0.850	0.850
tic-tac-toe-endgame	0.653	0.653	0.650	0.656

Table 12: Comparison of Accuracy for Original, SMC, SCP and SRC versions of SVM. The results indicate that the original, SMC, SCP, and SRC versions of SVM achieve an average accuracy rate of 0.773, 0.761, 0.795, 0.807, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	3.930×10^{-3}	1.369×10^{-2}	3.507×10^{-3}	2.724×10^{-3}
acute-inflammations-2	2.084×10^{-18}	1.409×10^{-18}	2.250×10^{-3}	3.159×10^{-3}
banknote-authentication	3.405×10^{-16}	1.441×10^{-15}	1.794×10^{-16}	8.314×10^{-18}
blood-transfusion-service-center	1.170×10^{-16}	2.953×10^{-17}	1.628×10^{-16}	1.055×10^{-16}
breast-cancer	4.428×10^{-3}	4.440×10^{-3}	4.428×10^{-3}	4.428×10^{-3}
breast-cancer-wisconsin-diagnostic	2.930×10^{-3}	1.698×10^{-2}	9.649×10^{-3}	4.752×10^{-3}
breast-cancer-wisconsin-original	5.218×10^{-3}	1.089×10^{-3}	5.218×10^{-3}	5.218×10^{-3}
breast-cancer-wisconsin-prognostic	7.645×10^{-3}	3.557×10^{-3}	7.645×10^{-3}	7.645×10^{-3}
climate-model-simulation-crashes	5.710×10^{-3}	7.592×10^{-3}	5.710×10^{-3}	5.710×10^{-3}
congressional-voting-records	3.003×10^{-3}	8.205×10^{-3}	7.440×10^{-3}	9.681×10^{-3}
connectionist-bench-sonar	2.817×10^{-2}	2.639×10^{-2}	1.612×10^{-2}	1.616×10^{-2}
credit-approval	6.959×10^{-3}	6.685×10^{-3}	8.818×10^{-3}	6.959×10^{-3}
fertility	2.129×10^{-3}	1.245×10^{-3}	2.129×10^{-3}	2.129×10^{-3}
haberman-survival	6.617×10^{-4}	7.968×10^{-5}	6.617×10^{-4}	6.617×10^{-4}
hepatitis	1.422×10^{-16}	4.147×10^{-17}	1.771×10^{-16}	1.353×10^{-16}
indian-liver-patient	1.532×10^{-16}	6.124×10^{-16}	8.165×10^{-17}	1.295×10^{-16}
ionosphere	1.660×10^{-2}	1.810×10^{-2}	2.837×10^{-2}	2.868×10^{-2}
mammographic-mass	3.451×10^{-3}	2.109×10^{-3}	7.031×10^{-3}	4.981×10^{-3}
monks-problems-1	7.257×10^{-3}	7.579×10^{-3}	7.591×10^{-3}	7.005×10^{-3}
monks-problems-2	1.483×10^{-2}	5.201×10^{-3}	9.270×10^{-3}	1.141×10^{-2}
monks-problems-3	1.148×10^{-2}	3.293×10^{-2}	9.191×10^{-3}	1.364×10^{-2}
parkinsons	1.082×10^{-2}	1.034×10^{-2}	1.082×10^{-2}	1.082×10^{-2}
planning-relax	3.587×10^{-21}	1.470×10^{-12}	9.562×10^{-19}	1.143×10^{-19}
qsar-biodegradation	4.643×10^{-3}	5.327×10^{-3}	4.643×10^{-3}	4.643×10^{-3}
seismic-bumps	4.377×10^{-19}	6.467×10^{-19}	2.798×10^{-19}	4.168×10^{-19}
spect-heart	5.384×10^{-2}	5.384×10^{-2}	7.570×10^{-5}	3.666×10^{-4}
spectf-heart	5.384×10^{-2}	5.384×10^{-2}	7.798×10^{-5}	4.035×10^{-4}
statlog-project-german-credit	6.473×10^{-3}	7.188×10^{-3}	6.473×10^{-3}	6.473×10^{-3}
thoracic-surgery	7.788×10^{-4}	5.932×10^{-4}	7.788×10^{-4}	7.788×10^{-4}
tic-tac-toe-endgame	8.466×10^{-20}	9.237×10^{-19}	1.173×10^{-2}	1.173×10^{-2}

Table 13: Comparison of Output Stability for Original, SMC, SCP and SRC versions of SVM. The results indicate that the original, SMC, SCP, and SRC versions of SVM achieve an output stability of 8.493×10^{-3}, 9.567×10^{-3}, 5.654×10^{-3}, 5.672×10^{-3}, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	2.056×10^{-1}	2.134×10^{-1}	3.829×10^{-1}	2.874×10^{-1}
acute-inflammations-2	3.432×10^{-10}	2.892×10^{-10}	3.022×10^{-10}	2.116×10^{-10}
banknote-authentication	5.687×10^{-9}	5.957×10^{-9}	5.013×10^{-9}	1.519×10^{-9}
blood-transfusion-service-center	1.439×10^{-9}	8.233×10^{-10}	1.631×10^{-9}	1.349×10^{-9}
breast-cancer	1.356×10^{0}	1.375×10^{0}	1.356×10^{0}	1.356×10^{0}
breast-cancer-wisconsin-diagnostic	9.501×10^{-2}	1.202×10^{-1}	1.026×10^{-1}	9.655×10^{-2}
breast-cancer-wisconsin-original	7.450×10^{-1}	6.140×10^{-1}	7.450×10^{-1}	7.450×10^{-1}
breast-cancer-wisconsin-prognostic	9.881×10^{-1}	6.048×10^{-1}	9.881×10^{-1}	9.881×10^{-1}
climate-model-simulation-crashes	2.989×10^{0}	3.258×10^{0}	2.989×10^{0}	2.989×10^{0}
congressional-voting-records	5.228×10^{-1}	5.627×10^{-1}	5.616×10^{-1}	5.958×10^{-1}
connectionist-bench-sonar	2.898×10^{0}	2.503×10^{0}	2.891×10^{0}	2.895×10^{0}
credit-approval	1.060×10^{0}	1.079×10^{0}	1.080×10^{0}	1.060×10^{0}
fertility	3.989×10^{-1}	2.900×10^{-1}	3.989×10^{-1}	3.989×10^{-1}
haberman-survival	8.906×10^{-3}	1.400×10^{-3}	8.906×10^{-3}	8.906×10^{-3}
hepatitis	1.178×10^{-9}	7.537×10^{-10}	1.331×10^{-9}	1.130×10^{-9}
indian-liver-patient	1.344×10^{-9}	1.653×10^{-9}	7.655×10^{-10}	1.610×10^{-9}
ionosphere	4.210×10^{0}	4.170×10^{0}	3.936×10^{0}	4.216×10^{0}
mammographic-mass	4.835×10^{-1}	3.820×10^{-1}	5.235×10^{-1}	4.919×10^{-1}
monks-problems-1	6.811×10^{-1}	6.467×10^{-1}	6.797×10^{-1}	6.646×10^{-1}
monks-problems-2	9.223×10^{-1}	3.919×10^{-1}	3.999×10^{-1}	4.178×10^{-1}
monks-problems-3	5.956×10^{-1}	8.874×10^{-1}	6.711×10^{-1}	1.017×10^{0}
parkinsons	1.977×10^{0}	1.754×10^{0}	1.977×10^{0}	1.977×10^{0}
planning-relax	2.110×10^{-10}	4.665×10^{-5}	4.515×10^{-10}	2.111×10^{-10}
qsar-biodegradation	3.042×10^{0}	2.669×10^{0}	3.042×10^{0}	3.042×10^{0}
seismic-bumps	5.315×10^{-10}	3.820×10^{-10}	4.937×10^{-10}	5.090×10^{-10}
spect-heart	2.865×10^{-10}	2.103×10^{-10}	2.962×10^{-10}	2.136×10^{-10}
spectf-heart	9.522×10^{-9}	5.956×10^{-9}	5.324×10^{-9}	5.157×10^{-9}
statlog-project-german-credit	1.220×10^{0}	1.119×10^{0}	1.220×10^{0}	1.220×10^{0}
thoracic-surgery	2.370×10^{-1}	1.801×10^{-1}	2.370×10^{-1}	2.370×10^{-1}
tic-tac-toe-endgame	1.035×10^{-9}	7.451×10^{-10}	1.200×10^{-9}	1.127×10^{-9}

Table 14: Comparison of Structural Stability for Original, SMC, SCP and SRC versions of SVM. The results indicate that the original, SMC, SCP, and SRC versions of SVM achieve an average structural stability of 0.821, 0.761, 0.806, 0.823, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	1.374×10^{1}	4.134×10^{18}	1.704×10^{19}	8.163×10^{18}
acute-inflammations-2	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
banknote-authentication	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
blood-transfusion-service-center	2.939×10^{18}	1.894×10^{19}	2.939×10^{18}	2.939×10^{18}
breast-cancer	1.018×10^{10}	1.117×10^{8}	1.018×10^{10}	1.018×10^{10}
breast-cancer-wisconsin-diagnostic	1.240×10^{6}	6.408×10^{18}	4.734×10^{18}	1.980×10^{18}
breast-cancer-wisconsin-original	2.327×10^{19}	2.424×10^{19}	2.327×10^{19}	2.327×10^{19}
breast-cancer-wisconsin-prognostic	2.509×10^{19}	2.509×10^{19}	2.509×10^{19}	2.509×10^{19}
climate-model-simulation-crashes	1.000×10^{18}	1.980×10^{18}	1.000×10^{18}	1.000×10^{18}
congressional-voting-records	9.614×10^{2}	1.443×10^{18}	2.017×10^{16}	5.640×10^{18}
connectionist-bench-sonar	2.403×10^{19}	2.403×10^{19}	2.403×10^{19}	2.403×10^{19}
credit-approval	1.624×10^{3}	1.774×10^{3}	1.000×10^{18}	1.624×10^{3}
fertility	5.697×10^{18}	8.273×10^{18}	5.697×10^{18}	5.697×10^{18}
haberman-survival	1.358×10^{19}	2.380×10^{19}	1.358×10^{19}	1.358×10^{19}
hepatitis	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
indian-liver-patient	1.000×10^{18}	5.697×10^{18}	1.000×10^{18}	1.000×10^{18}
ionosphere	3.064×10^{2}	2.524×10^{18}	1.609×10^{18}	1.995×10^{18}
mammographic-mass	3.547×10^{2}	2.374×10^{2}	1.625×10^{18}	4.283×10^{13}
monks-problems-1	2.515×10^{-1}	1.867×10^{-1}	3.830×10^{15}	2.515×10^{-1}
monks-problems-2	6.576×10^{18}	1.067×10^{19}	5.859×10^{18}	5.826×10^{18}
monks-problems-3	2.373×10^{18}	6.302×10^{18}	9.748×10^{18}	1.211×10^{19}
parkinsons	2.420×10^{8}	2.619×10^{8}	2.420×10^{8}	2.420×10^{8}
planning-relax	0.000×10^{0}	1.980×10^{18}	0.000×10^{0}	0.000×10^{0}
qsar-biodegradation	4.771×10^{1}	8.218×10^{2}	4.771×10^{1}	4.771×10^{1}
seismic-bumps	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
spect-heart	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
spectf-heart	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
statlog-project-german-credit	1.980×10^{18}	1.980×10^{18}	1.980×10^{18}	1.980×10^{18}
thoracic-surgery	1.288×10^{19}	1.425×10^{19}	1.288×10^{19}	1.288×10^{19}
tic-tac-toe-endgame	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}

Table 15: Comparison of Hyperparameter Stability for Original, SMC, SCP and SRC versions of SVM. The results indicate that the original, SMC, SCP, and SRC versions of SVM achieve an average hyperparameter stability of 1.860×10^{1}, 1.878×10^{1}, 1.871×10^{1}, 1.869×10^{1}, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	0.741	0.740	0.741	0.741
acute-inflammations-2	0.549	0.549	0.542	0.542
banknote-authentication	0.715	0.636	0.529	0.714
blood-transfusion-service-center	0.462	0.458	0.433	0.434
breast-cancer	0.656	0.655	0.656	0.656
breast-cancer-wisconsin-diagnostic	0.803	0.802	0.803	0.803
breast-cancer-wisconsin-original	0.713	0.711	0.701	0.702
breast-cancer-wisconsin-prognostic	0.731	0.730	0.727	0.727
climate-model-simulation-crashes	0.513	0.510	0.513	0.513
congressional-voting-records	0.737	0.736	0.737	0.737
connectionist-bench-sonar	0.587	0.584	0.583	0.583
credit-approval	0.625	0.621	0.624	0.624
fertility	0.822	0.822	0.815	0.816
haberman-survival	0.690	0.684	0.662	0.662
hepatitis	0.819	0.819	0.811	0.817
indian-liver-patient	0.688	0.686	0.661	0.661
ionosphere	0.730	0.729	0.730	0.730
mammographic-mass	0.621	0.616	0.610	0.610
monks-problems-1	0.608	0.602	0.598	0.598
monks-problems-2	0.577	0.577	0.561	0.561
monks-problems-3	0.667	0.656	0.665	0.665
parkinsons	0.774	0.769	0.773	0.773
planning-relax	0.649	0.649	0.625	0.625
qsar-biodegradation	0.578	0.574	0.578	0.578
seismic-bumps	0.912	0.911	0.907	0.907
spect-heart	0.514	0.550	0.547	0.541
spectf-heart	0.500	0.500	0.500	0.500
statlog-project-german-credit	0.696	0.693	0.688	0.688
thoracic-surgery	0.804	0.804	0.798	0.798
tic-tac-toe-endgame	0.669	0.611	0.581	0.727

Table 16: Comparison of Accuracy for Original, SMC, SCP and SRC versions of LR. The results indicate that the original, SMC, SCP, and SRC versions of LR achieve an average accuracy rate of 0.773, 0.761, 0.795, 0.807, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	1.453×10^{-4}	1.387×10^{-4}	1.503×10^{-4}	1.490×10^{-4}
acute-inflammations-2	1.204×10^{-3}	1.447×10^{-3}	4.611×10^{-5}	7.163×10^{-5}
banknote-authentication	1.580×10^{-2}	3.896×10^{-2}	3.844×10^{-5}	1.260×10^{-2}
blood-transfusion-service-center	6.332×10^{-4}	9.205×10^{-4}	1.319×10^{-3}	1.192×10^{-3}
breast-cancer	4.406×10^{-3}	4.634×10^{-3}	4.405×10^{-3}	4.405×10^{-3}
breast-cancer-wisconsin-diagnostic	1.165×10^{-3}	1.380×10^{-3}	1.156×10^{-3}	1.156×10^{-3}
breast-cancer-wisconsin-original	1.102×10^{-2}	1.031×10^{-2}	1.125×10^{-2}	1.132×10^{-2}
breast-cancer-wisconsin-prognostic	1.477×10^{-2}	1.520×10^{-2}	1.381×10^{-2}	1.327×10^{-2}
climate-model-simulation-crashes	3.596×10^{-3}	4.262×10^{-3}	3.595×10^{-3}	3.595×10^{-3}
congressional-voting-records	1.095×10^{-2}	1.110×10^{-2}	1.094×10^{-2}	1.094×10^{-2}
connectionist-bench-sonar	2.653×10^{-2}	2.470×10^{-2}	2.550×10^{-2}	2.513×10^{-2}
credit-approval	7.901×10^{-3}	9.212×10^{-3}	7.677×10^{-3}	7.677×10^{-3}
fertility	1.702×10^{-2}	1.521×10^{-2}	2.015×10^{-2}	1.963×10^{-2}
haberman-survival	1.801×10^{-3}	1.937×10^{-3}	1.203×10^{-3}	1.211×10^{-3}
hepatitis	3.651×10^{-2}	3.543×10^{-2}	3.667×10^{-2}	3.485×10^{-2}
indian-liver-patient	3.685×10^{-3}	3.830×10^{-3}	3.688×10^{-3}	3.709×10^{-3}
ionosphere	2.398×10^{-2}	2.425×10^{-2}	2.404×10^{-2}	2.405×10^{-2}
mammographic-mass	1.636×10^{-3}	2.448×10^{-3}	1.356×10^{-3}	1.356×10^{-3}
monks-problems-1	1.556×10^{-2}	1.602×10^{-2}	1.487×10^{-2}	1.487×10^{-2}
monks-problems-2	1.366×10^{-2}	9.534×10^{-3}	7.382×10^{-4}	9.363×10^{-4}
monks-problems-3	1.295×10^{-2}	1.535×10^{-2}	1.214×10^{-2}	1.222×10^{-2}
parkinsons	7.474×10^{-3}	8.750×10^{-3}	7.351×10^{-3}	7.347×10^{-3}
planning-relax	2.948×10^{-3}	2.424×10^{-3}	2.386×10^{-3}	2.300×10^{-3}
qsar-biodegradation	5.778×10^{-3}	6.868×10^{-3}	5.725×10^{-3}	5.728×10^{-3}
seismic-bumps	3.925×10^{-4}	4.846×10^{-4}	3.065×10^{-4}	3.068×10^{-4}
spect-heart	1.919×10^{-2}	2.273×10^{-2}	1.800×10^{-2}	3.131×10^{-2}
spectf-heart	2.437×10^{-3}	2.385×10^{-3}	3.429×10^{-6}	1.276×10^{-5}
statlog-project-german-credit	7.644×10^{-3}	9.014×10^{-3}	6.239×10^{-3}	6.242×10^{-3}
thoracic-surgery	6.625×10^{-3}	6.305×10^{-3}	5.336×10^{-3}	5.341×10^{-3}
tic-tac-toe-endgame	2.375×10^{-2}	7.326×10^{-3}	2.402×10^{-3}	1.556×10^{-2}

Table 17: Comparison of Output Stability for Original, SMC, SCP and SRC versions of LR. The results indicate that the original, SMC, SCP, and SRC versions of LR achieve an output stability of 1.004×10^{-2}, 1.042×10^{-2}, 8.083×10^{-3}, 9.283×10^{-3}, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}
acute-inflammations-2	0.000×10^{0}	1.980×10^{18}	1.000×10^{18}	0.000×10^{0}
banknote-authentication	2.362×10^{8}	9.091×10^{18}	2.521×10^{19}	2.774×10^{4}
blood-transfusion-service-center	3.130×10^{4}	5.478×10^{3}	3.765×10^{15}	3.859×10^{9}
breast-cancer	6.901×10^{-1}	1.631×10^{0}	6.901×10^{-1}	6.901×10^{-1}
breast-cancer-wisconsin-diagnostic	3.094×10^{0}	1.470×10^{0}	1.788×10^{0}	1.818×10^{0}
breast-cancer-wisconsin-original	9.090×10^{18}	4.798×10^{18}	1.000×10^{18}	1.980×10^{18}
breast-cancer-wisconsin-prognostic	5.722×10^{17}	1.000×10^{18}	4.329×10^{5}	1.246×10^{-2}
climate-model-simulation-crashes	1.350×10^{-5}	3.112×10^{-6}	1.350×10^{-5}	1.350×10^{-5}
congressional-voting-records	5.588×10^{-7}	6.576×10^{-7}	5.588×10^{-7}	5.588×10^{-7}
connectionist-bench-sonar	1.000×10^{18}	1.000×10^{18}	4.641×10^{4}	1.000×10^{18}
credit-approval	1.396×10^{-2}	5.104×10^{-3}	2.521×10^{-3}	2.521×10^{-3}
fertility	2.454×10^{19}	2.036×10^{19}	9.091×10^{18}	1.216×10^{19}
haberman-survival	2.385×10^{3}	1.100×10^{2}	3.040×10^{-36}	3.040×10^{-36}
hepatitis	2.230×10^{19}	1.788×10^{19}	2.079×10^{19}	1.216×10^{19}
indian-liver-patient	3.766×10^{7}	3.883×10^{7}	4.613×10^{0}	4.613×10^{0}
ionosphere	2.897×10^{-3}	5.391×10^{-4}	8.942×10^{-4}	8.942×10^{-4}
mammographic-mass	1.466×10^{-3}	1.270×10^{-3}	2.635×10^{-4}	2.635×10^{-4}
monks-problems-1	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}
monks-problems-2	1.980×10^{18}	7.051×10^{-2}	1.000×10^{18}	1.000×10^{18}
monks-problems-3	2.106×10^{-5}	2.366×10^{-5}	1.760×10^{-5}	2.494×10^{-5}
parkinsons	4.639×10^{-5}	9.629×10^{-6}	4.639×10^{-5}	4.639×10^{-5}
planning-relax	2.355×10^{19}	2.509×10^{19}	1.288×10^{19}	2.120×10^{19}
qsar-biodegradation	4.150×10^{-3}	1.510×10^{-3}	1.814×10^{-3}	1.815×10^{-3}
seismic-bumps	1.072×10^{9}	2.009×10^{8}	4.543×10^{-2}	4.401×10^{-2}
spect-heart	1.872×10^{18}	1.368×10^{-2}	1.141×10^{0}	9.118×10^{-1}
spectf-heart	0.000×10^{0}	2.939×10^{18}	0.000×10^{0}	0.000×10^{0}
statlog-project-german-credit	4.192×10^{-2}	3.498×10^{-2}	1.364×10^{-3}	1.374×10^{-3}
thoracic-surgery	6.576×10^{18}	2.939×10^{18}	7.205×10^{-3}	7.210×10^{-3}
tic-tac-toe-endgame	1.233×10^{15}	2.424×10^{19}	2.080×10^{19}	5.719×10^{-1}

Table 18: Comparison of Structural Stability for Original, SMC, SCP and SRC versions of LR. The results indicate that the original, SMC, SCP, and SRC versions of LR achieve an average structural stability of 3.258, 2.975, 2.941, 3.292, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	1.374×10^{1}	4.134×10^{18}	1.704×10^{19}	8.163×10^{18}
acute-inflammations-2	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
banknote-authentication	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
blood-transfusion-service-center	2.939×10^{18}	1.894×10^{19}	2.939×10^{18}	2.939×10^{18}
breast-cancer	1.018×10^{10}	1.117×10^{8}	1.018×10^{10}	1.018×10^{10}
breast-cancer-wisconsin-diagnostic	1.240×10^{6}	6.408×10^{18}	4.734×10^{18}	1.980×10^{18}
breast-cancer-wisconsin-original	2.327×10^{19}	2.424×10^{19}	2.327×10^{19}	2.327×10^{19}
breast-cancer-wisconsin-prognostic	2.509×10^{19}	2.509×10^{19}	2.509×10^{19}	2.509×10^{19}
climate-model-simulation-crashes	1.000×10^{18}	1.980×10^{18}	1.000×10^{18}	1.000×10^{18}
congressional-voting-records	9.614×10^{2}	1.443×10^{18}	2.017×10^{16}	5.640×10^{18}
connectionist-bench-sonar	2.403×10^{19}	2.403×10^{19}	2.403×10^{19}	2.403×10^{19}
credit-approval	1.624×10^{3}	1.774×10^{3}	1.000×10^{18}	1.624×10^{3}
fertility	5.697×10^{18}	8.273×10^{18}	5.697×10^{18}	5.697×10^{18}
haberman-survival	1.358×10^{19}	2.380×10^{19}	1.358×10^{19}	1.358×10^{19}
hepatitis	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
indian-liver-patient	1.000×10^{18}	5.697×10^{18}	1.000×10^{18}	1.000×10^{18}
ionosphere	3.064×10^{2}	2.524×10^{18}	1.609×10^{18}	1.995×10^{18}
mammographic-mass	3.547×10^{2}	2.374×10^{2}	1.625×10^{18}	4.283×10^{13}
monks-problems-1	2.515×10^{1}	1.867×10^{1}	3.830×10^{15}	2.515×10^{1}
monks-problems-2	6.576×10^{18}	1.067×10^{19}	5.859×10^{18}	5.826×10^{18}
monks-problems-3	2.373×10^{18}	6.302×10^{18}	9.748×10^{18}	1.211×10^{19}
parkinsons	2.420×10^{8}	2.619×10^{8}	2.420×10^{8}	2.420×10^{8}
planning-relax	0.000×10^{0}	1.980×10^{18}	0.000×10^{0}	0.000×10^{0}
qsar-biodegradation	4.771×10^{1}	8.218×10^{2}	4.771×10^{1}	4.771×10^{1}
seismic-bumps	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
spect-heart	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
spectf-heart	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
statlog-project-german-credit	1.980×10^{18}	1.980×10^{18}	1.980×10^{18}	1.980×10^{18}
thoracic-surgery	1.288×10^{19}	1.425×10^{19}	1.288×10^{19}	1.288×10^{19}
tic-tac-toe-endgame	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}

Table 19: Comparison of Hyperparameter Stability for Original, SMC, SCP and SRC versions of LR. The results indicate that the original, SMC, SCP, and SRC versions of LR achieve an average hyperparameter stability of 1.848×10^{1}, 1.857×10^{1}, 1.849×10^{1}, 1.822×10^{1}, respectively.

	Original	SMC	SCP
acute-inflammations-1	1.000	1.000	0.999
acute-inflammations-2	1.000	1.000	1.000
banknote-authentication	0.990	0.990	0.989
blood-transfusion-service-center	0.770	0.769	0.761
breast-cancer	0.742	0.749	0.720
breast-cancer-wisconsin-diagnostic	0.954	0.954	0.955
breast-cancer-wisconsin-original	0.968	0.967	0.967
breast-cancer-wisconsin-prognostic	0.729	0.730	0.738
climate-model-simulation-crashes	0.936	0.936	0.923
congressional-voting-records	0.990	0.991	0.989
connectionist-bench-sonar	0.850	0.852	0.847
credit-approval	0.891	0.890	0.891
fertility	0.861	0.870	0.853
haberman-survival	0.713	0.714	0.718
hepatitis	0.902	0.900	0.895
indian-liver-patient	0.679	0.682	0.694
ionosphere	0.922	0.924	0.924
mammographic-mass	0.838	0.840	0.837
monks-problems-1	0.772	0.763	0.773
monks-problems-2	0.627	0.638	0.618
monks-problems-3	0.826	0.826	0.827
parkinsons	0.925	0.920	0.914
planning-relax	0.659	0.669	0.676
qsar-biodegradation	0.847	0.847	0.847
seismic-bumps	0.934	0.933	0.934
spect-heart	0.740	0.747	0.751
spectf-heart	0.791	0.795	0.788
statlog-project-german-credit	0.757	0.753	0.696
thoracic-surgery	0.847	0.846	0.848
tic-tac-toe-endgame	0.963	0.962	0.766

Table 20: Comparison of Accuracy for Original, SMC, and SCP versions of RF. The results indicate that the original, SMC, and SCP versions of RF achieve an average accuracy rate of 0.847, 0.849, 0.838, respectively.

	Original	SMC	SCP
acute-inflammations-1	1.876×10^{-4}	1.774×10^{-4}	1.409×10^{-3}
acute-inflammations-2	5.281×10^{-4}	3.382×10^{-4}	8.821×10^{-4}
banknote-authentication	1.790×10^{-3}	1.789×10^{-3}	1.939×10^{-3}
blood-transfusion-service-center	1.282×10^{-2}	1.309×10^{-2}	1.532×10^{-2}
breast-cancer	1.441×10^{-2}	1.784×10^{-2}	7.791×10^{-3}
breast-cancer-wisconsin-diagnostic	7.302×10^{-3}	7.932×10^{-3}	6.879×10^{-3}
breast-cancer-wisconsin-original	6.149×10^{-3}	6.624×10^{-3}	5.674×10^{-3}
breast-cancer-wisconsin-prognostic	2.833×10^{-2}	2.913×10^{-2}	2.785×10^{-2}
climate-model-simulation-crashes	1.113×10^{-2}	1.083×10^{-2}	1.174×10^{-2}
congressional-voting-records	9.588×10^{-3}	9.596×10^{-3}	9.614×10^{-3}
connectionist-bench-sonar	2.823×10^{-2}	3.068×10^{-2}	2.931×10^{-2}
credit-approval	1.212×10^{-2}	1.124×10^{-2}	1.159×10^{-2}
fertility	2.176×10^{-2}	2.068×10^{-2}	9.595×10^{-3}
haberman-survival	1.914×10^{-2}	1.734×10^{-2}	1.351×10^{-2}
hepatitis	2.032×10^{-2}	1.944×10^{-2}	1.855×10^{-2}
indian-liver-patient	1.696×10^{-2}	1.556×10^{-2}	2.147×10^{-2}
ionosphere	1.206×10^{-2}	1.207×10^{-2}	1.271×10^{-2}
mammographic-mass	3.884×10^{-3}	3.893×10^{-3}	4.891×10^{-3}
monks-problems-1	2.847×10^{-2}	2.771×10^{-2}	2.558×10^{-2}
monks-problems-2	3.474×10^{-2}	3.393×10^{-2}	2.000×10^{-2}
monks-problems-3	1.464×10^{-2}	1.436×10^{-2}	1.392×10^{-2}
parkinsons	2.114×10^{-2}	2.183×10^{-2}	1.972×10^{-2}
planning-relax	2.267×10^{-2}	2.433×10^{-2}	2.253×10^{-2}
qsar-biodegradation	1.345×10^{-2}	1.254×10^{-2}	1.391×10^{-2}
seismic-bumps	4.208×10^{-3}	4.238×10^{-3}	1.430×10^{-3}
spect-heart	3.118×10^{-2}	2.995×10^{-2}	2.917×10^{-2}
spectf-heart	3.447×10^{-2}	3.490×10^{-2}	3.263×10^{-2}
statlog-project-german-credit	1.310×10^{-2}	1.210×10^{-2}	8.470×10^{-3}
thoracic-surgery	1.145×10^{-2}	1.242×10^{-2}	4.800×10^{-3}
tic-tac-toe-endgame	9.515×10^{-3}	9.395×10^{-3}	1.030×10^{-2}

Table 21: Comparison of Output Stability for Original, SMC, and SCP versions of RF. The results indicate that the original, SMC, and SCP versions of RF achieve an output stability of 0.016, 0.016, 0.014, respectively.

	Original	SMC	SCP
acute-inflammations-1	1.443×10^{-3}	1.453×10^{-3}	1.854×10^{-3}
acute-inflammations-2	2.250×10^{-3}	1.961×10^{-3}	2.890×10^{-3}
banknote-authentication	1.675×10^{-4}	1.826×10^{-4}	2.399×10^{-4}
blood-transfusion-service-center	8.870×10^{-3}	6.981×10^{-3}	1.244×10^{-1}
breast-cancer	2.857×10^{-3}	2.364×10^{-3}	6.169×10^{-3}
breast-cancer-wisconsin-diagnostic	2.436×10^{-3}	2.880×10^{-3}	2.452×10^{-3}
breast-cancer-wisconsin-original	5.144×10^{-3}	6.586×10^{-3}	6.335×10^{-3}
breast-cancer-wisconsin-prognostic	6.254×10^{-3}	5.308×10^{-3}	8.981×10^{-3}
climate-model-simulation-crashes	1.059×10^{-3}	8.922×10^{-4}	5.616×10^{-3}
congressional-voting-records	1.151×10^{-3}	1.242×10^{-3}	1.276×10^{-3}
connectionist-bench-sonar	2.020×10^{-3}	2.558×10^{-3}	3.274×10^{-3}
credit-approval	4.672×10^{-4}	4.406×10^{-4}	4.435×10^{-4}
fertility	3.083×10^{-3}	2.585×10^{-3}	2.053×10^{-2}
haberman-survival	1.348×10^{-2}	1.073×10^{-2}	3.190×10^{-2}
hepatitis	4.905×10^{-3}	3.944×10^{-3}	6.901×10^{-3}
indian-liver-patient	7.780×10^{-3}	6.537×10^{-3}	3.344×10^{-2}
ionosphere	1.445×10^{-3}	1.534×10^{-3}	1.761×10^{-3}
mammographic-mass	1.535×10^{-3}	1.530×10^{-3}	1.923×10^{-3}
monks-problems-1	2.422×10^{-3}	3.306×10^{-3}	1.986×10^{-3}
monks-problems-2	1.413×10^{-2}	1.086×10^{-2}	9.493×10^{-3}
monks-problems-3	6.468×10^{-4}	7.871×10^{-4}	5.716×10^{-4}
parkinsons	3.752×10^{-3}	5.252×10^{-3}	4.540×10^{-3}
planning-relax	1.669×10^{-2}	1.319×10^{-2}	1.841×10^{-2}
qsar-biodegradation	5.609×10^{-4}	5.493×10^{-4}	9.844×10^{-4}
seismic-bumps	3.978×10^{-4}	3.855×10^{-4}	2.277×10^{-2}
spect-heart	2.735×10^{-3}	2.877×10^{-3}	3.511×10^{-3}
spectf-heart	3.610×10^{-3}	3.687×10^{-3}	4.574×10^{-3}
statlog-project-german-credit	8.091×10^{-4}	5.940×10^{-4}	1.027×10^{-2}
thoracic-surgery	8.763×10^{-4}	6.997×10^{-4}	2.289×10^{-2}
tic-tac-toe-endgame	3.884×10^{-4}	3.556×10^{-4}	6.399×10^{-3}

Table 22: Comparison of Structural Stability for Original, SMC, and SCP versions of RF. The results indicate that the original, SMC, and SCP versions of RF achieve an average hyperparameter stability of 0.004, 0.003, 0.012, respectively.

	Original	SMC	SCP
acute-inflammations-1	0.000	0.000	1.200
acute-inflammations-2	0.900	0.000	0.600
banknote-authentication	1.077	1.386	1.998
blood-transfusion-service-center	11.497	8.287	5.933
breast-cancer	8.123	8.223	10.671
breast-cancer-wisconsin-diagnostic	12.994	15.779	13.735
breast-cancer-wisconsin-original	5.915	10.442	6.340
breast-cancer-wisconsin-prognostic	6.645	5.344	3.647
climate-model-simulation-crashes	2.374	1.773	0.757
congressional-voting-records	4.520	4.428	4.585
connectionist-bench-sonar	5.620	7.159	8.029
credit-approval	9.903	11.126	5.499
fertility	0.618	0.574	0.239
haberman-survival	9.960	8.154	9.645
hepatitis	2.342	1.993	2.401
indian-liver-patient	18.917	15.466	13.100
ionosphere	4.423	3.883	6.134
mammographic-mass	20.358	21.435	24.111
monks-problems-1	3.112	3.270	2.801
monks-problems-2	8.997	7.344	5.273
monks-problems-3	1.874	2.011	1.820
parkinsons	4.009	5.716	5.075
planning-relax	6.905	5.597	4.897
qsar-biodegradation	5.015	2.577	11.993
seismic-bumps	2.169	1.807	0.400
spect-heart	2.772	2.947	3.043
spectf-heart	3.287	3.185	3.345
statlog-project-german-credit	11.568	6.833	15.241
thoracic-surgery	1.354	1.022	2.601
tic-tac-toe-endgame	0.463	0.386	6.195

Table 23: Comparison of Hyperparameter Stability for Original, SMC, and SCP versions of RF. The results indicate that the original, SMC, and SCP versions of RF achieve an average hyperparameter stability of 5.92, 5.60, 6.044, respectively.

	Original	SMC	SCP
acute-inflammations-1	1.000	1.000	1.000
acute-inflammations-2	0.983	0.984	0.983
banknote-authentication	0.978	0.977	0.977
blood-transfusion-service-center	0.769	0.769	0.765
breast-cancer	0.740	0.742	0.741
breast-cancer-wisconsin-diagnostic	0.926	0.920	0.924
breast-cancer-wisconsin-original	0.953	0.952	0.953
breast-cancer-wisconsin-prognostic	0.729	0.727	0.708
climate-model-simulation-crashes	0.911	0.911	0.907
congressional-voting-records	0.980	0.978	0.981
connectionist-bench-sonar	0.762	0.756	0.733
credit-approval	0.882	0.879	0.879
fertility	0.859	0.861	0.854
haberman-survival	0.706	0.703	0.697
hepatitis	0.842	0.848	0.845
indian-liver-patient	0.675	0.680	0.673
ionosphere	0.875	0.875	0.875
mammographic-mass	0.833	0.835	0.825
monks-problems-1	0.814	0.822	0.841
monks-problems-2	0.621	0.623	0.601
monks-problems-3	0.846	0.845	0.849
parkinsons	0.874	0.855	0.862
planning-relax	0.670	0.654	0.634
qsar-biodegradation	0.809	0.809	0.806
seismic-bumps	0.930	0.930	0.934
spect-heart	0.734	0.726	0.704
spectf-heart	0.695	0.684	0.685
statlog-project-german-credit	0.718	0.718	0.714
thoracic-surgery	0.839	0.833	0.846
tic-tac-toe-endgame	0.880	0.863	0.834

Table 24: Comparison of Accuracy for Original, SMC, and SCP versions of OCT. The results indicate that the original, SMC, and SCP versions of OCT achieve an average accuracy rate of 0.828, 0.825, 0.821, respectively.

	Original	SMC	SCP
acute-inflammations-1	2.357×10^{-4}	3.499×10^{-4}	3.499×10^{-4}
acute-inflammations-2	1.752×10^{-3}	1.752×10^{-3}	2.748×10^{-3}
banknote-authentication	5.066×10^{-3}	6.267×10^{-3}	5.565×10^{-3}
blood-transfusion-service-center	1.448×10^{-2}	1.339×10^{-2}	1.438×10^{-2}
breast-cancer	2.646×10^{-2}	2.227×10^{-2}	2.421×10^{-2}
breast-cancer-wisconsin-diagnostic	2.082×10^{-2}	2.099×10^{-2}	2.191×10^{-2}
breast-cancer-wisconsin-original	1.439×10^{-2}	1.557×10^{-2}	1.538×10^{-2}
breast-cancer-wisconsin-prognostic	3.654×10^{-2}	3.173×10^{-2}	4.169×10^{-2}
climate-model-simulation-crashes	2.455×10^{-2}	2.269×10^{-2}	2.086×10^{-2}
congressional-voting-records	1.056×10^{-2}	1.180×10^{-2}	1.156×10^{-2}
connectionist-bench-sonar	8.922×10^{-2}	7.583×10^{-2}	8.291×10^{-2}
credit-approval	1.867×10^{-2}	1.539×10^{-2}	1.975×10^{-2}
fertility	2.255×10^{-2}	2.178×10^{-2}	2.164×10^{-2}
haberman-survival	3.093×10^{-2}	2.415×10^{-2}	2.684×10^{-2}
hepatitis	4.241×10^{-2}	3.840×10^{-2}	3.922×10^{-2}
indian-liver-patient	3.531×10^{-2}	2.568×10^{-2}	3.111×10^{-2}
ionosphere	3.087×10^{-2}	3.234×10^{-2}	3.069×10^{-2}
mammographic-mass	1.243×10^{-2}	1.074×10^{-2}	2.096×10^{-2}
monks-problems-1	4.312×10^{-2}	4.237×10^{-2}	4.322×10^{-2}
monks-problems-2	4.713×10^{-2}	3.747×10^{-2}	4.864×10^{-2}
monks-problems-3	1.467×10^{-2}	1.518×10^{-2}	1.497×10^{-2}
parkinsons	4.749×10^{-2}	4.919×10^{-2}	4.816×10^{-2}
planning-relax	3.216×10^{-2}	3.322×10^{-2}	4.307×10^{-2}
qsar-biodegradation	3.638×10^{-2}	3.166×10^{-2}	3.381×10^{-2}
seismic-bumps	3.695×10^{-3}	4.573×10^{-3}	2.792×10^{-4}
spect-heart	3.093×10^{-2}	2.834×10^{-2}	4.220×10^{-2}
spectf-heart	8.941×10^{-2}	8.428×10^{-2}	8.920×10^{-2}
statlog-project-german-credit	2.512×10^{-2}	2.008×10^{-2}	1.599×10^{-2}
thoracic-surgery	9.076×10^{-3}	1.161×10^{-2}	4.062×10^{-3}
tic-tac-toe-endgame	4.251×10^{-2}	4.258×10^{-2}	4.517×10^{-2}

Table 25: Comparison of Output Stability for Original, SMC, and SCP versions of OCT. The results indicate that the original, SMC, and SCP versions of OCT achieve an output stability of 0.029, 0.026, 0.029 respectively.

	Original	SMC	SCP
acute-inflammations-1	4.890×10^{-2}	4.867×10^{-2}	4.860×10^{-2}
acute-inflammations-2	2.865×10^{-2}	2.862×10^{-2}	2.893×10^{-2}
banknote-authentication	1.107×10^{-2}	9.966×10^{-3}	1.278×10^{-2}
blood-transfusion-service-center	1.070×10^{-1}	1.018×10^{-1}	1.389×10^{-1}
breast-cancer	1.366×10^{-2}	1.574×10^{-2}	1.599×10^{-2}
breast-cancer-wisconsin-diagnostic	2.728×10^{-2}	2.730×10^{-2}	2.745×10^{-2}
breast-cancer-wisconsin-original	7.178×10^{-2}	7.274×10^{-2}	7.912×10^{-2}
breast-cancer-wisconsin-prognostic	1.427×10^{-2}	1.269×10^{-2}	1.983×10^{-2}
climate-model-simulation-crashes	3.112×10^{-2}	3.426×10^{-2}	2.697×10^{-2}
congressional-voting-records	2.223×10^{-2}	2.179×10^{-2}	2.147×10^{-2}
connectionist-bench-sonar	1.374×10^{-2}	1.352×10^{-2}	1.402×10^{-2}
credit-approval	1.118×10^{-2}	1.145×10^{-2}	1.105×10^{-2}
fertility	2.681×10^{-2}	1.896×10^{-2}	1.561×10^{-2}
haberman-survival	1.148×10^{-1}	1.254×10^{-1}	1.192×10^{-1}
hepatitis	2.365×10^{-2}	2.561×10^{-2}	2.712×10^{-2}
indian-liver-patient	4.243×10^{-2}	4.448×10^{-2}	5.432×10^{-2}
ionosphere	1.425×10^{-2}	1.571×10^{-2}	1.386×10^{-2}
mammographic-mass	1.724×10^{-2}	2.793×10^{-2}	2.937×10^{-2}
monks-problems-1	8.398×10^{-3}	1.075×10^{-2}	1.095×10^{-2}
monks-problems-2	2.912×10^{-2}	3.162×10^{-2}	3.844×10^{-2}
monks-problems-3	7.661×10^{-3}	8.837×10^{-3}	6.037×10^{-3}
parkinsons	3.009×10^{-2}	3.088×10^{-2}	3.228×10^{-2}
planning-relax	2.285×10^{-2}	3.218×10^{-2}	4.446×10^{-2}
qsar-biodegradation	1.613×10^{-2}	1.727×10^{-2}	1.628×10^{-2}
seismic-bumps	1.485×10^{-2}	2.077×10^{-2}	2.832×10^{-3}
spect-heart	1.505×10^{-2}	1.632×10^{-2}	1.590×10^{-2}
spectf-heart	1.688×10^{-2}	1.575×10^{-2}	1.656×10^{-2}
statlog-project-german-credit	8.081×10^{-3}	9.744×10^{-3}	9.563×10^{-3}
thoracic-surgery	7.289×10^{-3}	1.286×10^{-2}	8.757×10^{-3}
tic-tac-toe-endgame	1.773×10^{-2}	2.014×10^{-2}	1.987×10^{-2}

Table 26: Comparison of Structural Stability for Original, SMC, and SCP versions of OCT. The results indicate that the original, SMC, and SCP versions of OCT achieve an average structural stability of 0.028, 0.029, 0.031 respectively.

	Original	SMC	SCP
acute-inflammations-1	0.048	0.065	0.000
acute-inflammations-2	1.196	0.754	0.775
banknote-authentication	2.871	2.912	2.916
blood-transfusion-service-center	2.819	3.050	3.019
breast-cancer	3.331	3.383	3.335
breast-cancer-wisconsin-diagnostic	3.145	3.188	3.076
breast-cancer-wisconsin-original	3.004	3.109	3.042
breast-cancer-wisconsin-prognostic	2.427	2.708	3.170
climate-model-simulation-crashes	3.477	3.445	3.224
congressional-voting-records	3.098	3.034	3.044
connectionist-bench-sonar	3.320	3.304	3.067
credit-approval	2.496	2.724	2.841
fertility	3.470	2.660	2.917
haberman-survival	3.240	3.175	3.202
hepatitis	3.576	3.423	3.413
indian-liver-patient	2.412	2.886	3.113
ionosphere	2.969	3.275	3.027
mammographic-mass	2.320	3.043	2.720
monks-problems-1	3.402	3.237	3.208
monks-problems-2	3.334	3.420	3.274
monks-problems-3	2.609	2.773	2.822
parkinsons	3.425	3.354	3.221
planning-relax	2.518	2.950	3.172
qsar-biodegradation	2.357	2.665	2.510
seismic-bumps	1.837	2.693	1.221
spect-heart	3.352	3.085	3.375
spectf-heart	3.559	3.567	3.515
statlog-project-german-credit	1.550	2.652	2.505
thoracic-surgery	2.441	3.255	2.707
tic-tac-toe-endgame	2.923	3.000	2.897

Table 27: Comparison of Hyperparameter Stability for Original, SMC, and SCP versions of OCT. The results indicate that the original, SMC, and SCP versions of OCT achieve an average hyperparameter stability of 2.75, 2.89, 2.81, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	0.933	0.910	0.826	0.906
acute-inflammations-2	0.583	0.813	0.583	0.825
banknote-authentication	0.556	0.957	0.556	0.869
blood-transfusion-service-center	0.763	0.763	0.763	0.763
breast-cancer	0.935	0.935	0.936	0.935
breast-cancer-wisconsin-diagnostic	0.968	0.968	0.967	0.968
breast-cancer-wisconsin-original	0.757	0.757	0.759	0.757
breast-cancer-wisconsin-prognostic	0.766	0.766	0.771	0.766
climate-model-simulation-crashes	0.953	0.953	0.945	0.953
congressional-voting-records	0.948	0.948	0.948	0.948
connectionist-bench-sonar	0.639	0.726	0.647	0.650
credit-approval	0.840	0.840	0.840	0.840
fertility	0.867	0.867	0.867	0.867
haberman-survival	0.737	0.737	0.736	0.737
hepatitis	0.833	0.833	0.833	0.833
indian-liver-patient	0.713	0.713	0.713	0.713
ionosphere	0.840	0.840	0.836	0.839
mammographic-mass	0.833	0.833	0.831	0.833
monks-problems-1	0.779	0.777	0.768	0.779
monks-problems-2	0.586	0.563	0.609	0.566
monks-problems-3	0.927	0.927	0.915	0.906
parkinsons	0.860	0.860	0.857	0.860
planning-relax	0.709	0.709	0.709	0.709
qsar-biodegradation	0.875	0.875	0.870	0.875
seismic-bumps	0.934	0.934	0.934	0.934
spect-heart	0.500	0.686	0.500	0.578
spectf-heart	0.500	0.654	0.500	0.656
statlog-project-german-credit	0.744	0.744	0.742	0.744
thoracic-surgery	0.851	0.851	0.851	0.851
tic-tac-toe-endgame	0.653	0.656	0.653	0.650

Table 28: Comparison of Cross-validated Accuracy for Original, SMC, SCP and SRC versions of SVM. The results indicate that the original, SMC, SCP, and SRC versions of SVM achieve an average accuracy rate of 0.779, 0.775, 0.804, 0.813, respectively.

.1 Cross-Validation Results

	Original	SMC	SCP	SRC
acute-inflammations-1	2.484×10^{-3}	8.307×10^{-4}	7.166×10^{-3}	7.548×10^{-4}
acute-inflammations-2	2.084×10^{-18}	3.159×10^{-3}	1.409×10^{-18}	2.250×10^{-3}
banknote-authentication	3.405×10^{-16}	8.314×10^{-18}	1.441×10^{-15}	1.794×10^{-16}
blood-transfusion-service-center	1.168×10^{-16}	1.057×10^{-16}	4.841×10^{-17}	1.626×10^{-16}
breast-cancer	2.458×10^{-3}	2.458×10^{-3}	2.617×10^{-3}	2.458×10^{-3}
breast-cancer-wisconsin-diagnostic	1.160×10^{-3}	1.160×10^{-3}	1.337×10^{-3}	1.160×10^{-3}
breast-cancer-wisconsin-original	8.929×10^{-4}	8.929×10^{-4}	1.227×10^{-4}	8.929×10^{-4}
breast-cancer-wisconsin-prognostic	7.542×10^{-3}	7.542×10^{-3}	5.436×10^{-3}	7.542×10^{-3}
climate-model-simulation-crashes	4.540×10^{-3}	4.540×10^{-3}	5.398×10^{-3}	4.540×10^{-3}
congressional-voting-records	3.400×10^{-3}	3.311×10^{-3}	3.272×10^{-3}	3.373×10^{-3}
connectionist-bench-sonar	2.770×10^{-2}	1.623×10^{-2}	2.489×10^{-2}	1.617×10^{-2}
credit-approval	5.548×10^{-3}	5.548×10^{-3}	4.662×10^{-3}	5.548×10^{-3}
fertility	9.221×10^{-11}	8.132×10^{-11}	2.371×10^{-6}	1.303×10^{-10}
haberman-survival	9.107×10^{-4}	9.107×10^{-4}	1.129×10^{-4}	9.107×10^{-4}
hepatitis	1.422×10^{-16}	1.353×10^{-16}	4.147×10^{-17}	1.771×10^{-16}
indian-liver-patient	1.523×10^{-16}	1.274×10^{-16}	2.487×10^{-16}	8.194×10^{-17}
ionosphere	1.357×10^{-2}	1.415×10^{-2}	1.171×10^{-2}	1.474×10^{-2}
mammographic-mass	1.459×10^{-3}	1.457×10^{-3}	1.761×10^{-3}	1.468×10^{-3}
monks-problems-1	7.575×10^{-3}	7.370×10^{-3}	7.292×10^{-3}	7.345×10^{-3}
monks-problems-2	2.046×10^{-2}	1.550×10^{-2}	8.811×10^{-3}	1.290×10^{-2}
monks-problems-3	2.555×10^{-3}	2.986×10^{-3}	4.404×10^{-3}	1.254×10^{-2}
parkinsons	6.248×10^{-3}	6.248×10^{-3}	6.523×10^{-3}	6.248×10^{-3}
planning-relax	3.587×10^{-21}	1.143×10^{-19}	7.644×10^{-20}	9.562×10^{-19}
qsar-biodegradation	4.554×10^{-3}	4.554×10^{-3}	5.241×10^{-3}	4.554×10^{-3}
seismic-bumps	4.377×10^{-19}	4.168×10^{-19}	6.467×10^{-19}	2.798×10^{-19}
spect-heart	5.384×10^{-2}	3.666×10^{-4}	5.384×10^{-2}	7.570×10^{-5}
spectf-heart	5.384×10^{-2}	4.035×10^{-4}	5.384×10^{-2}	7.798×10^{-5}
statlog-project-german-credit	4.390×10^{-3}	4.390×10^{-3}	5.508×10^{-3}	4.390×10^{-3}
thoracic-surgery	9.830×10^{-5}	9.830×10^{-5}	6.425×10^{-13}	9.830×10^{-5}
tic-tac-toe-endgame	8.466×10^{-20}	1.173×10^{-2}	9.237×10^{-19}	1.173×10^{-2}

Table 29: Comparison of Cross-validated Output Stability for Original, SMC, SCP, and SRC versions of SVM. The results indicate that the Original, SMC, SCP, and SRC versions of SVM achieve an average output stability of 7.508×10^{-3}, 7.132×10^{-3}, 4.059×10^{-3}, 3.862×10^{-3}, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	1.622×10^{-1}	1.282×10^{-1}	1.430×10^{-1}	1.243×10^{-1}
acute-inflammations-2	3.432×10^{-10}	2.116×10^{-10}	2.892×10^{-10}	3.022×10^{-10}
banknote-authentication	5.687×10^{-9}	1.519×10^{-9}	5.957×10^{-9}	5.013×10^{-9}
blood-transfusion-service-center	1.437×10^{-9}	1.370×10^{-9}	8.098×10^{-10}	1.625×10^{-9}
breast-cancer	1.236×10^{0}	1.236×10^{0}	1.101×10^{0}	1.236×10^{0}
breast-cancer-wisconsin-diagnostic	1.102×10^{-1}	1.102×10^{-1}	1.109×10^{-1}	1.102×10^{-1}
breast-cancer-wisconsin-original	3.442×10^{-2}	3.442×10^{-2}	6.670×10^{-3}	3.442×10^{-2}
breast-cancer-wisconsin-prognostic	1.150×10^{0}	1.150×10^{0}	8.750×10^{-1}	1.150×10^{0}
climate-model-simulation-crashes	2.821×10^{0}	2.821×10^{0}	2.746×10^{0}	2.821×10^{0}
congressional-voting-records	6.757×10^{-1}	6.604×10^{-1}	6.688×10^{-1}	6.696×10^{-1}
connectionist-bench-sonar	2.982×10^{0}	2.978×10^{0}	2.577×10^{0}	2.973×10^{0}
credit-approval	8.281×10^{-1}	8.281×10^{-1}	7.429×10^{-1}	8.281×10^{-1}
fertility	1.285×10^{-5}	1.224×10^{-5}	6.337×10^{-3}	1.540×10^{-5}
haberman-survival	1.075×10^{-2}	1.075×10^{-2}	1.551×10^{-3}	1.075×10^{-2}
hepatitis	1.178×10^{-9}	1.130×10^{-9}	7.537×10^{-10}	1.331×10^{-9}
indian-liver-patient	5.885×10^{-10}	6.215×10^{-10}	1.135×10^{-8}	6.686×10^{-10}
ionosphere	3.511×10^{0}	3.515×10^{0}	2.887×10^{0}	3.521×10^{0}
mammographic-mass	3.031×10^{-1}	3.024×10^{-1}	3.837×10^{-1}	3.045×10^{-1}
monks-problems-1	6.977×10^{-1}	6.860×10^{-1}	6.378×10^{-1}	6.846×10^{-1}
monks-problems-2	1.142×10^{0}	6.302×10^{-1}	5.332×10^{-1}	6.025×10^{-1}
monks-problems-3	3.552×10^{-1}	3.731×10^{-1}	4.652×10^{-1}	9.242×10^{-1}
parkinsons	2.186×10^{0}	2.186×10^{0}	2.243×10^{0}	2.186×10^{0}
planning-relax	2.110×10^{-10}	2.111×10^{-10}	1.422×10^{-8}	4.515×10^{-10}
qsar-biodegradation	3.329×10^{0}	3.329×10^{0}	3.251×10^{0}	3.329×10^{0}
seismic-bumps	5.315×10^{-10}	5.090×10^{-10}	3.820×10^{-10}	4.937×10^{-10}
spect-heart	2.865×10^{-10}	2.136×10^{-10}	2.103×10^{-10}	2.962×10^{-10}
spectf-heart	9.522×10^{-9}	5.157×10^{-9}	5.956×10^{-9}	5.324×10^{-9}
statlog-project-german-credit	1.001×10^{0}	1.001×10^{0}	1.063×10^{0}	1.001×10^{0}
thoracic-surgery	5.161×10^{-2}	5.161×10^{-2}	4.034×10^{-5}	5.161×10^{-2}
tic-tac-toe-endgame	1.035×10^{-9}	1.127×10^{-9}	7.451×10^{-10}	1.200×10^{-9}

Table 30: Comparison of Cross-validated Structural Stability for Original, SMC, SCP, and SRC versions of SVM. The results indicate that the Original, SMC, SCP, and SRC versions of SVM achieve an average structural stability of 0.753, 0.681, 0.752, 0.734, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	2.886×10^{0}	2.886×10^{0}	2.272×10^{0}	2.886×10^{0}
acute-inflammations-2	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
banknote-authentication	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
blood-transfusion-service-center	2.939×10^{18}	2.939×10^{18}	2.403×10^{19}	2.939×10^{18}
breast-cancer	1.398×10^{4}	1.398×10^{4}	9.903×10^{3}	1.398×10^{4}
breast-cancer-wisconsin-diagnostic	1.719×10^{3}	1.719×10^{3}	3.537×10^{3}	1.719×10^{3}
breast-cancer-wisconsin-original	1.067×10^{19}	1.067×10^{19}	1.842×10^{19}	1.067×10^{19}
breast-cancer-wisconsin-prognostic	1.358×10^{19}	1.358×10^{19}	1.288×10^{19}	1.358×10^{19}
climate-model-simulation-crashes	3.900×10^{-2}	3.900×10^{-2}	3.293×10^{-2}	3.900×10^{-2}
congressional-voting-records	2.315×10^{0}	2.310×10^{0}	1.461×10^{0}	2.313×10^{0}
connectionist-bench-sonar	2.461×10^{19}	2.461×10^{19}	2.516×10^{19}	2.461×10^{19}
credit-approval	3.704×10^{1}	3.704×10^{1}	1.840×10^{1}	3.704×10^{1}
fertility	1.000×10^{18}	1.000×10^{18}	1.980×10^{18}	1.000×10^{18}
haberman-survival	2.161×10^{19}	2.161×10^{19}	2.161×10^{19}	2.161×10^{19}
hepatitis	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
indian-liver-patient	0.000×10^{0}	0.000×10^{0}	1.894×10^{19}	0.000×10^{0}
ionosphere	1.655×10^{1}	2.848×10^{11}	8.415×10^{0}	1.606×10^{12}
mammographic-mass	1.094×10^{2}	1.091×10^{2}	7.748×10^{1}	1.140×10^{2}
monks-problems-1	5.171×10^{-2}	5.171×10^{-2}	6.257×10^{-2}	5.171×10^{-2}
monks-problems-2	9.934×10^{-3}	3.054×10^{10}	2.939×10^{18}	1.120×10^{-2}
monks-problems-3	4.787×10^{-1}	1.520×10^{12}	1.233×10^{15}	2.184×10^{16}
parkinsons	2.356×10^{-3}	2.356×10^{-3}	2.930×10^{-3}	2.356×10^{-3}
planning-relax	0.000×10^{0}	0.000×10^{0}	1.980×10^{18}	0.000×10^{0}
qsar-biodegradation	4.018×10^{0}	4.018×10^{0}	3.898×10^{0}	4.018×10^{0}
seismic-bumps	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
spect-heart	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
spectf-heart	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}
statlog-project-german-credit	1.883×10^{1}	1.883×10^{1}	1.798×10^{1}	1.883×10^{1}
thoracic-surgery	4.798×10^{18}	4.798×10^{18}	5.697×10^{18}	4.798×10^{18}
tic-tac-toe-endgame	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}	0.000×10^{0}

Table 31: Comparison of Cross-validated Hyperparameter Stability for Original, SMC, SCP, and SRC versions of SVM. The results indicate that the Original, SMC, SCP, and SRC versions of SVM achieve an average hyperparameter stability of 1.842×10^{1}, 1.865×10^{1}, 1.842×10^{1}, 1.842×10^{1}, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	0.741	0.740	0.741	0.741
acute-inflammations-2	0.549	0.549	0.542	0.542
banknote-authentication	0.725	0.547	0.529	0.619
blood-transfusion-service-center	0.462	0.458	0.434	0.435
breast-cancer	0.658	0.657	0.658	0.658
breast-cancer-wisconsin-diagnostic	0.803	0.802	0.803	0.803
breast-cancer-wisconsin-original	0.710	0.709	0.705	0.706
breast-cancer-wisconsin-prognostic	0.733	0.731	0.729	0.729
climate-model-simulation-crashes	0.513	0.511	0.513	0.513
congressional-voting-records	0.736	0.736	0.736	0.736
connectionist-bench-sonar	0.596	0.592	0.593	0.593
credit-approval	0.625	0.621	0.624	0.624
fertility	0.820	0.820	0.816	0.816
haberman-survival	0.689	0.683	0.661	0.661
hepatitis	0.814	0.814	0.808	0.809
indian-liver-patient	0.689	0.686	0.661	0.661
ionosphere	0.732	0.731	0.733	0.733
mammographic-mass	0.621	0.616	0.610	0.610
monks-problems-1	0.607	0.601	0.599	0.599
monks-problems-2	0.577	0.577	0.561	0.561
monks-problems-3	0.667	0.657	0.666	0.666
parkinsons	0.774	0.769	0.773	0.773
planning-relax	0.649	0.649	0.624	0.625
qsar-biodegradation	0.579	0.575	0.578	0.578
seismic-bumps	0.912	0.911	0.907	0.907
spect-heart	0.521	0.549	0.547	0.525
spectf-heart	0.500	0.500	0.500	0.500
statlog-project-german-credit	0.697	0.695	0.689	0.689
thoracic-surgery	0.805	0.804	0.799	0.799
tic-tac-toe-endgame	0.621	0.603	0.578	0.674

Table 32: Comparison of Cross-validated Accuracy for Original, SMC, SCP and SRC versions of LR. The results indicate that the original, SMC, SCP, and SRC versions of LR achieve an average accuracy rate of 0.671, 0.663, 0.657, 0.663, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	1.453×10^{-4}	1.387×10^{-4}	1.503×10^{-4}	1.490×10^{-4}
acute-inflammations-2	1.204×10^{-3}	1.510×10^{-3}	4.056×10^{-5}	7.163×10^{-5}
banknote-authentication	1.163×10^{-2}	2.558×10^{-3}	3.568×10^{-5}	1.035×10^{-3}
blood-transfusion-service-center	6.003×10^{-4}	9.089×10^{-4}	1.062×10^{-3}	1.050×10^{-3}
breast-cancer	4.173×10^{-3}	4.426×10^{-3}	4.173×10^{-3}	4.173×10^{-3}
breast-cancer-wisconsin-diagnostic	1.305×10^{-3}	1.507×10^{-3}	1.343×10^{-3}	1.343×10^{-3}
breast-cancer-wisconsin-original	4.166×10^{-3}	9.538×10^{-3}	1.038×10^{-2}	1.033×10^{-2}
breast-cancer-wisconsin-prognostic	1.181×10^{-2}	1.323×10^{-2}	1.114×10^{-2}	1.116×10^{-2}
climate-model-simulation-crashes	4.131×10^{-3}	4.659×10^{-3}	4.130×10^{-3}	4.130×10^{-3}
congressional-voting-records	9.750×10^{-3}	9.727×10^{-3}	9.754×10^{-3}	9.754×10^{-3}
connectionist-bench-sonar	2.505×10^{-2}	2.365×10^{-2}	2.366×10^{-2}	2.380×10^{-2}
credit-approval	8.117×10^{-3}	9.676×10^{-3}	8.046×10^{-3}	8.043×10^{-3}
fertility	1.735×10^{-2}	1.788×10^{-2}	2.007×10^{-2}	2.047×10^{-2}
haberman-survival	1.892×10^{-3}	2.022×10^{-3}	1.376×10^{-3}	1.398×10^{-3}
hepatitis	2.479×10^{-2}	2.329×10^{-2}	2.418×10^{-2}	2.385×10^{-2}
indian-liver-patient	2.903×10^{-3}	3.315×10^{-3}	3.600×10^{-3}	3.619×10^{-3}
ionosphere	2.823×10^{-2}	2.547×10^{-2}	2.610×10^{-2}	2.610×10^{-2}
mammographic-mass	1.638×10^{-3}	2.459×10^{-3}	1.358×10^{-3}	1.358×10^{-3}
monks-problems-1	1.785×10^{-2}	1.845×10^{-2}	1.637×10^{-2}	1.637×10^{-2}
monks-problems-2	1.364×10^{-2}	9.661×10^{-3}	7.606×10^{-4}	9.672×10^{-4}
monks-problems-3	1.318×10^{-2}	1.578×10^{-2}	1.241×10^{-2}	1.248×10^{-2}
parkinsons	7.506×10^{-3}	8.814×10^{-3}	7.366×10^{-3}	7.365×10^{-3}
planning-relax	9.064×10^{-4}	1.130×10^{-3}	2.109×10^{-3}	1.921×10^{-3}
qsar-biodegradation	5.901×10^{-3}	7.059×10^{-3}	5.843×10^{-3}	5.840×10^{-3}
seismic-bumps	3.564×10^{-4}	4.581×10^{-4}	3.149×10^{-4}	3.152×10^{-4}
spect-heart	1.552×10^{-2}	2.449×10^{-2}	1.626×10^{-2}	1.785×10^{-3}
spectf-heart	2.064×10^{-3}	2.030×10^{-3}	2.320×10^{-12}	3.139×10^{-12}
statlog-project-german-credit	7.877×10^{-3}	9.204×10^{-3}	6.492×10^{-3}	6.492×10^{-3}
thoracic-surgery	6.236×10^{-3}	6.452×10^{-3}	5.404×10^{-3}	5.410×10^{-3}
tic-tac-toe-endgame	7.663×10^{-3}	1.332×10^{-3}	1.450×10^{-4}	1.082×10^{-2}

Table 33: Comparison of Cross-validated Output Stability for Original, SMC, SCP, and SRC versions of LR. The results indicate that the Original, SMC, SCP, and SRC versions of LR achieve an average output stability of 8.586×10^{-3}, 8.694×10^{-3}, 7.469×10^{-3}, 7.387×10^{-3}, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	4.994×10^{-1}	5.221×10^{-1}	5.068×10^{-1}	5.059×10^{-1}
acute-inflammations-2	1.448×10^{-10}	1.184×10^{-2}	1.488×10^{-10}	1.063×10^{-10}
banknote-authentication	2.151×10^{0}	5.371×10^{-2}	4.761×10^{-3}	3.517×10^{-2}
blood-transfusion-service-center	1.239×10^{-2}	1.552×10^{-2}	1.081×10^{-2}	1.078×10^{-2}
breast-cancer	5.900×10^{0}	5.578×10^{0}	5.899×10^{0}	5.900×10^{0}
breast-cancer-wisconsin-diagnostic	5.223×10^{-1}	5.253×10^{-1}	5.499×10^{-1}	5.500×10^{-1}
breast-cancer-wisconsin-original	7.527×10^{-1}	2.496×10^{0}	4.023×10^{0}	3.874×10^{0}
breast-cancer-wisconsin-prognostic	3.654×10^{0}	3.692×10^{0}	3.394×10^{0}	3.401×10^{0}
climate-model-simulation-crashes	4.389×10^{0}	4.329×10^{0}	4.391×10^{0}	4.391×10^{0}
congressional-voting-records	4.304×10^{0}	4.194×10^{0}	4.305×10^{0}	4.305×10^{0}
connectionist-bench-sonar	1.262×10^{1}	1.123×10^{1}	1.213×10^{1}	1.216×10^{1}
credit-approval	4.530×10^{0}	4.454×10^{0}	4.680×10^{0}	4.676×10^{0}
fertility	4.677×10^{0}	4.310×10^{0}	4.654×10^{0}	4.741×10^{0}
haberman-survival	3.732×10^{-2}	3.526×10^{-2}	1.344×10^{-2}	1.351×10^{-2}
hepatitis	3.884×10^{0}	3.689×10^{0}	3.905×10^{0}	3.884×10^{0}
indian-liver-patient	1.128×10^{0}	1.100×10^{0}	5.653×10^{-1}	5.698×10^{-1}
ionosphere	1.664×10^{1}	8.262×10^{0}	8.945×10^{0}	8.946×10^{0}
mammographic-mass	7.277×10^{-1}	8.637×10^{-1}	5.612×10^{-1}	5.615×10^{-1}
monks-problems-1	1.666×10^{0}	1.542×10^{0}	1.830×10^{0}	1.830×10^{0}
monks-problems-2	9.568×10^{-1}	7.653×10^{-1}	7.535×10^{-2}	9.803×10^{-2}
monks-problems-3	3.313×10^{0}	3.301×10^{0}	3.380×10^{0}	3.374×10^{0}
parkinsons	4.163×10^{0}	4.011×10^{0}	4.096×10^{0}	4.102×10^{0}
planning-relax	2.740×10^{-2}	1.872×10^{-1}	4.497×10^{-1}	3.821×10^{-1}
qsar-biodegradation	6.724×10^{0}	6.505×10^{0}	6.640×10^{0}	6.638×10^{0}
seismic-bumps	2.407×10^{0}	3.145×10^{0}	3.409×10^{0}	3.411×10^{0}
spect-heart	7.813×10^{-1}	3.161×10^{0}	1.950×10^{0}	3.353×10^{-1}
spectf-heart	4.229×10^{-9}	5.205×10^{-4}	4.229×10^{-9}	2.528×10^{-9}
statlog-project-german-credit	2.163×10^{0}	2.388×10^{0}	1.840×10^{0}	1.840×10^{0}
thoracic-surgery	3.525×10^{0}	3.238×10^{0}	2.727×10^{0}	2.730×10^{0}
tic-tac-toe-endgame	1.692×10^{0}	1.842×10^{-1}	1.468×10^{-4}	3.997×10^{0}

Table 34: Comparison of Cross-validated Structural Stability for Original, SMC, SCP, and SRC versions of LR. The results indicate that the Original, SMC, SCP, and SRC versions of LR achieve an average structural stability of 3.128, 2.793, 2.831, 2.909, respectively.

	Original	SMC	SCP	SRC
acute-inflammations-1	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}
acute-inflammations-2	0.000×10^{0}	3.879×10^{18}	0.000×10^{0}	0.000×10^{0}
banknote-authentication	2.029×10^{4}	3.876×10^{18}	2.352×10^{19}	7.734×10^{3}
blood-transfusion-service-center	3.040×10^{-36}	3.040×10^{-36}	8.967×10^{0}	3.040×10^{-36}
breast-cancer	8.273×10^{-4}	2.269×10^{-5}	8.273×10^{-4}	8.273×10^{-4}
breast-cancer-wisconsin-diagnostic	2.180×10^{-1}	1.266×10^{-1}	1.910×10^{-1}	1.900×10^{-1}
breast-cancer-wisconsin-original	1.240×10^{11}	4.741×10^{10}	9.499×10^{0}	3.165×10^{2}
breast-cancer-wisconsin-prognostic	1.588×10^{-3}	6.425×10^{-4}	9.732×10^{-4}	9.556×10^{-4}
climate-model-simulation-crashes	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}
congressional-voting-records	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}
connectionist-bench-sonar	1.498×10^{-4}	1.339×10^{-5}	8.416×10^{-5}	1.153×10^{-4}
credit-approval	9.387×10^{-5}	7.351×10^{-5}	1.382×10^{-6}	1.903×10^{-6}
fertility	2.524×10^{19}	1.322×10^{19}	1.980×10^{18}	3.879×10^{18}
haberman-survival	3.040×10^{-36}	3.040×10^{-36}	1.716×10^{-6}	3.040×10^{-36}
hepatitis	5.697×10^{18}	1.023×10^{6}	8.325×10^{5}	8.162×10^{5}
indian-liver-patient	4.021×10^{-1}	4.144×10^{-1}	3.040×10^{-36}	3.040×10^{-36}
ionosphere	1.145×10^{-5}	8.351×10^{-6}	9.414×10^{-7}	9.414×10^{-7}
mammographic-mass	2.458×10^{-4}	1.618×10^{-4}	3.040×10^{-36}	3.040×10^{-36}
monks-problems-1	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}
monks-problems-2	3.040×10^{-36}	1.881×10^{-5}	3.040×10^{-36}	3.040×10^{-36}
monks-problems-3	9.414×10^{-7}	3.047×10^{-7}	9.414×10^{-7}	9.414×10^{-7}
parkinsons	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}	3.040×10^{-36}
planning-relax	4.065×10^{18}	1.789×10^{19}	1.425×10^{19}	2.513×10^{19}
qsar-biodegradation	4.120×10^{-5}	7.161×10^{-6}	2.049×10^{-5}	2.049×10^{-5}
seismic-bumps	3.100×10^{-2}	5.825×10^{-2}	3.033×10^{-3}	3.033×10^{-3}
spect-heart	1.241×10^{18}	2.397×10^{-1}	1.301×10^{0}	2.008×10^{-2}
spectf-heart	0.000×10^{0}	2.160×10^{19}	0.000×10^{0}	0.000×10^{0}
statlog-project-german-credit	2.268×10^{-3}	6.932×10^{-4}	3.344×10^{-36}	3.344×10^{-36}
thoracic-surgery	5.738×10^{-3}	4.715×10^{-4}	1.716×10^{-6}	1.716×10^{-6}
tic-tac-toe-endgame	2.402×10^{5}	1.892×10^{19}	1.000×10^{18}	2.219×10^{0}

Table 35: Comparison of Cross-validated Hyperparameter Stability for Original, SMC, SCP, and SRC versions of LR. The results indicate that the Original, SMC, SCP, and SRC versions of LR achieve an average hyperparameter stability of 1.808×10^{1}, 1.842×10^{1}, 1.813×10^{1}, 1.799×10^{1} , respectively.

	Original	SMC	SCP
acute-inflammations-1	1.000	1.000	1.000
acute-inflammations-2	1.000	1.000	1.000
banknote-authentication	0.990	0.991	0.991
blood-transfusion-service-center	0.778	0.776	0.763
breast-cancer	0.744	0.749	0.718
breast-cancer-wisconsin-diagnostic	0.954	0.956	0.956
breast-cancer-wisconsin-original	0.968	0.969	0.968
breast-cancer-wisconsin-prognostic	0.746	0.748	0.746
climate-model-simulation-crashes	0.937	0.937	0.923
congressional-voting-records	0.988	0.988	0.987
connectionist-bench-sonar	0.862	0.860	0.847
credit-approval	0.890	0.891	0.891
fertility	0.871	0.869	0.857
haberman-survival	0.733	0.733	0.727
hepatitis	0.904	0.904	0.900
indian-liver-patient	0.687	0.687	0.691
ionosphere	0.924	0.923	0.922
mammographic-mass	0.840	0.841	0.839
monks-problems-1	0.821	0.811	0.814
monks-problems-2	0.632	0.637	0.618
monks-problems-3	0.833	0.834	0.831
parkinsons	0.924	0.923	0.922
planning-relax	0.663	0.671	0.682
qsar-biodegradation	0.852	0.850	0.850
seismic-bumps	0.934	0.934	0.934
spect-heart	0.768	0.764	0.769
spectf-heart	0.802	0.806	0.798
statlog-project-german-credit	0.758	0.756	0.699
thoracic-surgery	0.850	0.849	0.848
tic-tac-toe-endgame	0.965	0.963	0.778

Table 36: Comparison of Cross-validated Accuracy for Original, SMC, and SCP versions of RF. The results indicate that the original, SMC, and SCP versions of RF achieve an average accuracy of 0.854, 0.854, 0.842, respectively.

	Original	SMC	SCP
acute-inflammations-1	1.845×10^{-4}	1.839×10^{-4}	8.560×10^{-4}
acute-inflammations-2	3.444×10^{-4}	3.382×10^{-4}	8.792×10^{-4}
banknote-authentication	1.728×10^{-3}	1.703×10^{-3}	1.774×10^{-3}
blood-transfusion-service-center	4.882×10^{-3}	5.328×10^{-3}	7.205×10^{-3}
breast-cancer	1.724×10^{-2}	1.605×10^{-2}	7.038×10^{-3}
breast-cancer-wisconsin-diagnostic	5.511×10^{-3}	5.522×10^{-3}	5.130×10^{-3}
breast-cancer-wisconsin-original	5.621×10^{-3}	5.738×10^{-3}	5.565×10^{-3}
breast-cancer-wisconsin-prognostic	8.450×10^{-3}	9.215×10^{-3}	1.466×10^{-2}
climate-model-simulation-crashes	1.015×10^{-2}	9.487×10^{-3}	1.167×10^{-2}
congressional-voting-records	1.066×10^{-2}	1.087×10^{-2}	1.042×10^{-2}
connectionist-bench-sonar	3.062×10^{-2}	2.852×10^{-2}	2.390×10^{-2}
credit-approval	1.473×10^{-2}	1.432×10^{-2}	1.459×10^{-2}
fertility	8.744×10^{-3}	8.532×10^{-3}	9.368×10^{-3}
haberman-survival	6.356×10^{-3}	7.581×10^{-3}	1.108×10^{-2}
hepatitis	2.093×10^{-2}	1.942×10^{-2}	1.921×10^{-2}
indian-liver-patient	1.119×10^{-2}	1.107×10^{-2}	2.149×10^{-2}
ionosphere	1.125×10^{-2}	1.142×10^{-2}	1.284×10^{-2}
mammographic-mass	3.657×10^{-3}	3.651×10^{-3}	4.589×10^{-3}
monks-problems-1	2.435×10^{-2}	2.327×10^{-2}	2.275×10^{-2}
monks-problems-2	3.645×10^{-2}	3.552×10^{-2}	8.714×10^{-3}
monks-problems-3	1.193×10^{-2}	1.236×10^{-2}	1.096×10^{-2}
parkinsons	2.071×10^{-2}	2.002×10^{-2}	1.898×10^{-2}
planning-relax	1.870×10^{-2}	1.861×10^{-2}	2.013×10^{-2}
qsar-biodegradation	1.488×10^{-2}	1.397×10^{-2}	1.348×10^{-2}
seismic-bumps	1.658×10^{-3}	1.941×10^{-3}	1.433×10^{-3}
spect-heart	4.995×10^{-2}	4.652×10^{-2}	4.242×10^{-2}
spectf-heart	4.028×10^{-2}	4.043×10^{-2}	3.758×10^{-2}
statlog-project-german-credit	7.575×10^{-3}	9.349×10^{-3}	6.253×10^{-3}
thoracic-surgery	6.418×10^{-3}	8.734×10^{-3}	4.704×10^{-3}
tic-tac-toe-endgame	9.412×10^{-3}	9.341×10^{-3}	1.353×10^{-2}

Table 37: Comparison of Cross-validated Output Stability for Original, SMC, and SCP versions of RF. The results indicate that the original, SMC, and SCP versions of RF achieve an average output stability of 0.0138, 0.0136, 0.0128, respectively.

	Original	SMC	SCP
acute-inflammations-1	1.453×10^{-3}	1.439×10^{-3}	1.829×10^{-3}
acute-inflammations-2	2.158×10^{-3}	1.961×10^{-3}	2.887×10^{-3}
banknote-authentication	1.708×10^{-4}	1.725×10^{-4}	1.800×10^{-4}
blood-transfusion-service-center	5.256×10^{-3}	6.323×10^{-3}	1.553×10^{-1}
breast-cancer	1.250×10^{-3}	1.244×10^{-3}	6.632×10^{-3}
breast-cancer-wisconsin-diagnostic	2.023×10^{-3}	2.120×10^{-3}	2.211×10^{-3}
breast-cancer-wisconsin-original	5.018×10^{-3}	5.581×10^{-3}	6.325×10^{-3}
breast-cancer-wisconsin-prognostic	9.179×10^{-3}	7.422×10^{-3}	1.591×10^{-2}
climate-model-simulation-crashes	7.929×10^{-4}	9.241×10^{-4}	4.944×10^{-3}
congressional-voting-records	1.645×10^{-3}	1.582×10^{-3}	1.962×10^{-3}
connectionist-bench-sonar	1.500×10^{-3}	1.590×10^{-3}	2.019×10^{-3}
credit-approval	3.682×10^{-4}	3.396×10^{-4}	3.475×10^{-4}
fertility	3.641×10^{-3}	3.200×10^{-3}	1.969×10^{-2}
haberman-survival	4.077×10^{-3}	4.220×10^{-3}	5.983×10^{-2}
hepatitis	2.381×10^{-3}	2.373×10^{-3}	4.987×10^{-3}
indian-liver-patient	7.411×10^{-3}	6.620×10^{-3}	2.433×10^{-2}
ionosphere	1.361×10^{-3}	1.578×10^{-3}	1.914×10^{-3}
mammographic-mass	1.807×10^{-3}	1.671×10^{-3}	1.818×10^{-3}
monks-problems-1	7.645×10^{-4}	8.566×10^{-4}	8.564×10^{-4}
monks-problems-2	1.833×10^{-2}	1.455×10^{-2}	1.627×10^{-2}
monks-problems-3	5.195×10^{-4}	5.411×10^{-4}	5.673×10^{-4}
parkinsons	4.188×10^{-3}	4.432×10^{-3}	5.067×10^{-3}
planning-relax	1.501×10^{-2}	1.139×10^{-2}	1.936×10^{-2}
qsar-biodegradation	4.466×10^{-4}	4.745×10^{-4}	6.098×10^{-4}
seismic-bumps	5.490×10^{-4}	5.584×10^{-4}	2.277×10^{-2}
spect-heart	1.298×10^{-3}	1.373×10^{-3}	1.732×10^{-3}
spectf-heart	2.487×10^{-3}	2.696×10^{-3}	2.909×10^{-3}
statlog-project-german-credit	3.079×10^{-4}	3.119×10^{-4}	1.184×10^{-2}
thoracic-surgery	1.396×10^{-3}	1.152×10^{-3}	2.307×10^{-2}
tic-tac-toe-endgame	3.379×10^{-4}	3.255×10^{-4}	5.176×10^{-3}

Table 38: Comparison of Cross-validated Structural Stability for Original, SMC, and SCP versions of RF. The results indicate that the original, SMC, and SCP versions of RF achieve an average structural stability of 0.0032, 0.0030, 0.0141, respectively.

	Original	SMC	SCP
acute-inflammations-1	0.000	0.000	0.000
acute-inflammations-2	0.000	0.000	0.000
banknote-authentication	0.000	0.000	0.000
blood-transfusion-service-center	9.929	10.583	15.081
breast-cancer	1.389	1.493	1.739
breast-cancer-wisconsin-diagnostic	1.209	1.914	1.508
breast-cancer-wisconsin-original	4.224	2.311	5.524
breast-cancer-wisconsin-prognostic	5.224	4.222	6.136
climate-model-simulation-crashes	0.987	1.641	3.045
congressional-voting-records	8.366	8.264	9.315
connectionist-bench-sonar	1.879	1.661	1.783
credit-approval	2.850	2.707	6.861
fertility	1.660	1.316	1.337
haberman-survival	6.287	4.930	10.102
hepatitis	0.000	0.000	1.497
indian-liver-patient	16.728	14.038	4.142
ionosphere	2.859	4.007	7.635
mammographic-mass	31.557	27.335	25.466
monks-problems-1	0.000	0.000	0.171
monks-problems-2	13.416	10.541	6.805
monks-problems-3	1.606	1.770	1.894
parkinsons	4.415	3.590	5.127
planning-relax	5.173	4.070	3.938
qsar-biodegradation	1.196	1.350	1.435
seismic-bumps	3.552	3.043	0.000
spect-heart	0.000	0.000	0.000
spectf-heart	0.000	0.000	0.000
statlog-project-german-credit	2.260	1.651	11.663
thoracic-surgery	1.988	2.116	3.173
tic-tac-toe-endgame	0.000	0.000	0.000

Table 39: Comparison of Cross-validated Hyperparameter Stability for Original, SMC, and SCP versions of RF. The results indicate that the original, SMC, and SCP versions of RF achieve an average hyperparameter stability of 4.292, 3.818, 4.513, respectively.

	Original	SMC	SCP
acute-inflammations-1	1.000	1.000	1.000
acute-inflammations-2	0.984	0.984	0.983
banknote-authentication	0.980	0.980	0.979
blood-transfusion-service-center	0.775	0.774	0.767
breast-cancer	0.733	0.756	0.752
breast-cancer-wisconsin-diagnostic	0.925	0.922	0.923
breast-cancer-wisconsin-original	0.956	0.954	0.952
breast-cancer-wisconsin-prognostic	0.747	0.754	0.750
climate-model-simulation-crashes	0.914	0.915	0.915
congressional-voting-records	0.972	0.973	0.970
connectionist-bench-sonar	0.599	0.554	0.557
credit-approval	0.889	0.889	0.886
fertility	0.864	0.865	0.861
haberman-survival	0.723	0.715	0.711
hepatitis	0.848	0.849	0.845
indian-liver-patient	0.708	0.705	0.695
ionosphere	0.886	0.880	0.878
mammographic-mass	0.834	0.836	0.829
monks-problems-1	0.867	0.852	0.876
monks-problems-2	0.626	0.619	0.618
monks-problems-3	0.838	0.848	0.850
parkinsons	0.874	0.857	0.858
planning-relax	0.699	0.704	0.693
qsar-biodegradation	0.812	0.810	0.810
seismic-bumps	0.932	0.933	0.934
spect-heart	0.755	0.741	0.737
spectf-heart	0.686	0.682	0.687
statlog-project-german-credit	0.713	0.717	0.718
thoracic-surgery	0.843	0.845	0.847
tic-tac-toe-endgame	0.888	0.867	0.835

Table 40: Comparison of Cross-validated Accuracy for Original, SMC, and SCP versions of OCT. The results indicate that the original, SMC, and SCP versions of OCT achieve an average accuracy of 0.829, 0.826, 0.824, respectively.

	Original	SMC	SCP
acute-inflammations-1	2.357×10^{-4}	3.499×10^{-4}	3.499×10^{-4}
acute-inflammations-2	1.752×10^{-3}	1.752×10^{-3}	2.088×10^{-3}
banknote-authentication	4.954×10^{-3}	5.184×10^{-3}	5.084×10^{-3}
blood-transfusion-service-center	1.256×10^{-2}	1.121×10^{-2}	1.267×10^{-2}
breast-cancer	3.143×10^{-2}	2.022×10^{-2}	2.428×10^{-2}
breast-cancer-wisconsin-diagnostic	1.926×10^{-2}	1.973×10^{-2}	2.061×10^{-2}
breast-cancer-wisconsin-original	1.254×10^{-2}	1.350×10^{-2}	1.487×10^{-2}
breast-cancer-wisconsin-prognostic	7.811×10^{-3}	3.378×10^{-3}	7.711×10^{-3}
climate-model-simulation-crashes	1.940×10^{-2}	1.923×10^{-2}	1.581×10^{-2}
congressional-voting-records	1.343×10^{-2}	1.214×10^{-2}	1.452×10^{-2}
connectionist-bench-sonar	2.803×10^{-2}	2.012×10^{-2}	9.706×10^{-3}
credit-approval	7.497×10^{-3}	6.602×10^{-3}	1.347×10^{-2}
fertility	4.123×10^{-3}	5.231×10^{-3}	8.605×10^{-3}
haberman-survival	1.977×10^{-2}	1.645×10^{-2}	2.364×10^{-2}
hepatitis	3.846×10^{-2}	3.710×10^{-2}	3.621×10^{-2}
indian-liver-patient	5.486×10^{-3}	6.110×10^{-3}	1.636×10^{-2}
ionosphere	2.183×10^{-2}	2.073×10^{-2}	2.404×10^{-2}
mammographic-mass	1.069×10^{-2}	9.558×10^{-3}	1.860×10^{-2}
monks-problems-1	3.807×10^{-2}	4.030×10^{-2}	3.651×10^{-2}
monks-problems-2	4.494×10^{-2}	4.134×10^{-2}	4.739×10^{-2}
monks-problems-3	1.617×10^{-2}	1.336×10^{-2}	1.360×10^{-2}
parkinsons	4.377×10^{-2}	4.561×10^{-2}	4.625×10^{-2}
planning-relax	7.625×10^{-3}	3.223×10^{-3}	1.083×10^{-2}
qsar-biodegradation	3.661×10^{-2}	3.069×10^{-2}	3.347×10^{-2}
seismic-bumps	2.300×10^{-3}	1.819×10^{-3}	2.534×10^{-4}
spect-heart	3.338×10^{-2}	2.921×10^{-2}	4.335×10^{-2}
spectf-heart	9.073×10^{-2}	8.476×10^{-2}	9.656×10^{-2}
statlog-project-german-credit	1.915×10^{-2}	1.723×10^{-2}	1.540×10^{-2}
thoracic-surgery	5.441×10^{-3}	4.998×10^{-3}	3.228×10^{-3}
tic-tac-toe-endgame	4.114×10^{-2}	4.258×10^{-2}	4.615×10^{-2}

Table 41: Comparison of Cross-validated Output Stability for Original, SMC, and SCP versions of OCT. The results indicate that the original, SMC, and SCP versions of OCT achieve an average output stability of 0.021, 0.019, 0.022, respectively.

	Original	SMC	SCP
acute-inflammations-1	4.890×10^{-2}	4.867×10^{-2}	4.860×10^{-2}
acute-inflammations-2	2.804×10^{-2}	2.815×10^{-2}	2.779×10^{-2}
banknote-authentication	1.257×10^{-2}	1.341×10^{-2}	1.766×10^{-2}
blood-transfusion-service-center	1.101×10^{-1}	1.301×10^{-1}	1.414×10^{-1}
breast-cancer	1.582×10^{-2}	1.046×10^{-2}	1.271×10^{-2}
breast-cancer-wisconsin-diagnostic	2.723×10^{-2}	2.721×10^{-2}	2.662×10^{-2}
breast-cancer-wisconsin-original	6.549×10^{-2}	7.060×10^{-2}	7.301×10^{-2}
breast-cancer-wisconsin-prognostic	5.814×10^{-3}	1.862×10^{-3}	3.410×10^{-3}
climate-model-simulation-crashes	2.578×10^{-2}	3.170×10^{-2}	2.161×10^{-2}
congressional-voting-records	2.306×10^{-2}	2.189×10^{-2}	2.186×10^{-2}
connectionist-bench-sonar	4.623×10^{-3}	2.132×10^{-3}	2.118×10^{-3}
credit-approval	1.095×10^{-2}	1.064×10^{-2}	1.139×10^{-2}
fertility	4.942×10^{-3}	3.485×10^{-3}	4.199×10^{-3}
haberman-survival	1.037×10^{-1}	1.223×10^{-1}	1.273×10^{-1}
hepatitis	2.562×10^{-2}	2.965×10^{-2}	2.756×10^{-2}
indian-liver-patient	1.233×10^{-2}	1.732×10^{-2}	4.068×10^{-2}
ionosphere	9.213×10^{-3}	9.689×10^{-3}	9.082×10^{-3}
mammographic-mass	1.791×10^{-2}	2.064×10^{-2}	2.769×10^{-2}
monks-problems-1	1.034×10^{-2}	1.114×10^{-2}	9.770×10^{-3}
monks-problems-2	3.365×10^{-2}	3.730×10^{-2}	3.772×10^{-2}
monks-problems-3	9.005×10^{-3}	8.743×10^{-3}	8.427×10^{-3}
parkinsons	3.122×10^{-2}	3.208×10^{-2}	3.282×10^{-2}
planning-relax	7.399×10^{-3}	3.317×10^{-3}	9.855×10^{-3}
qsar-biodegradation	1.481×10^{-2}	1.693×10^{-2}	1.626×10^{-2}
seismic-bumps	1.381×10^{-2}	1.127×10^{-2}	2.441×10^{-3}
spect-heart	1.492×10^{-2}	1.635×10^{-2}	1.677×10^{-2}
spectf-heart	1.743×10^{-2}	1.592×10^{-2}	1.717×10^{-2}
statlog-project-german-credit	7.991×10^{-3}	9.279×10^{-3}	9.917×10^{-3}
thoracic-surgery	5.165×10^{-3}	6.194×10^{-3}	3.522×10^{-3}
tic-tac-toe-endgame	1.633×10^{-2}	1.929×10^{-2}	1.959×10^{-2}

Table 42: Comparison of Cross-validated Structural Stability for Original, SMC, and SCP versions of OCT. The results indicate that the original, SMC, and SCP versions of OCT achieve an average structural stability of 0.024, 0.026, 0.028, respectively.

	Original	SMC	SCP
acute-inflammations-1	0.000	0.000	0.000
acute-inflammations-2	0.000	0.000	0.000
banknote-authentication	0.000	0.000	0.000
blood-transfusion-service-center	0.704	1.415	1.733
breast-cancer	2.919	3.807	3.826
breast-cancer-wisconsin-diagnostic	2.453	2.907	2.407
breast-cancer-wisconsin-original	1.818	2.512	2.056
breast-cancer-wisconsin-prognostic	0.053	0.071	0.107
climate-model-simulation-crashes	2.446	2.389	3.004
congressional-voting-records	2.599	2.096	2.602
connectionist-bench-sonar	1.860	1.960	1.098
credit-approval	0.956	0.688	1.916
fertility	2.000	2.037	2.286
haberman-survival	1.786	2.102	2.642
hepatitis	0.000	0.327	0.234
indian-liver-patient	0.075	0.704	1.331
ionosphere	0.106	0.677	0.304
mammographic-mass	1.174	1.627	1.585
monks-problems-1	0.697	0.682	0.763
monks-problems-2	3.073	3.037	2.888
monks-problems-3	2.586	3.079	3.025
parkinsons	2.314	2.000	2.336
planning-relax	0.105	0.094	0.976
qsar-biodegradation	1.688	2.147	0.872
seismic-bumps	0.055	0.651	3.048
spect-heart	0.000	0.000	0.000
spectf-heart	0.000	0.000	0.000
statlog-project-german-credit	0.305	1.653	2.063
thoracic-surgery	1.702	2.109	1.444
tic-tac-toe-endgame	0.000	0.000	0.000

Table 43: Comparison of Cross-validated Hyperparameter Stability for Original, SMC, and SCP versions of OCT. The results indicate that the original, SMC, and SCP versions of OCT achieve an average hyperparameter stability of 1.116, 1.359, 1.485, respectively.