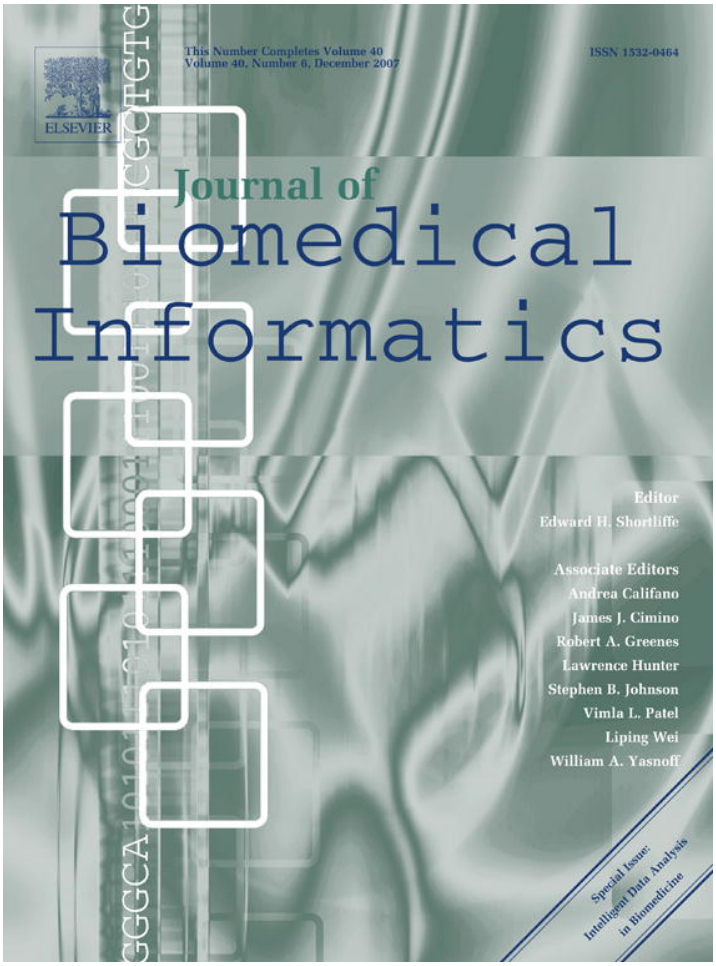

Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 40 (2007) 688–697
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

# Effects of SVM parameter optimization on discrimination and calibration for post-procedural PCI mortality

Michael E. Matheny [a,b,c,*], Frederic S. Resnic [a,b,d], Nipun Arora [d], Lucila Ohno-Machado [a,b]

[a] Decision Systems Group, Brigham & Women's Hospital, 75 Francis Street, Boston, MA 02115, USA
[b] Harvard-MIT Division of Health Sciences & Technology, Cambridge, MA, USA
[c] Division of General Medicine, Brigham & Women's Hospital, Boston, MA, USA
[d] Division of Cardiovascular Medicine, Brigham & Women's Hospital, Boston, MA, USA

Received 22 September 2006
Available online 18 May 2007

**Abstract**

Support vector machines (SVM) have become popular among machine learning researchers, but their applications in biomedicine have been somewhat limited. A number of methods, such as grid search and evolutionary algorithms, have been utilized to optimize model parameters of SVMs. The sensitivity of the results to changes in optimization methods has not been investigated in the context of medical applications.

In this study, radial-basis kernel SVM and polynomial kernel SVM mortality prediction models for percutaneous coronary interventions were optimized using (a) mean-squared error, (b) mean cross-entropy error, (c) the area under the receiver operating characteristic, and (d) the Hosmer–Lemeshow goodness-of-fit test (HL $\chi^2$). A threefold cross-validation inner and outer loop method was used to select the best models using the training data, and evaluations were based on previously unseen test data. The results were compared to those produced by logistic regression models optimized using the same indices.

The choice of optimization parameters had a significant impact on performance in both SVM kernel types.


*Keywords:* Machine learning; Support vector machines; Logistic regression; Evolutionary computing; Genetic algorithms; Percutaneous coronary intervention

## 1. Introduction

In the last few decades, significant emphasis has been placed on the development of statistical and machine learning models to help predict risk in various patient populations. These models have been widely used to improve the quality of care [1], provide institutional quality scorecards [2], provide risk stratification [3], assist patient selection [4] in research, evaluate futility of care [5], and provide individual patient prognostications [6].

One of the most recent developments in machine learning modeling has been support vector machines (SVM) [7,8]. These models are based on the theory of risk minimization and are able to find an optimal separation hyperplane in a multi-dimensional space to perform classification of a dichotomous outcome. They have been compared to other methods, such as logistic regression [9–12], *k*-nearest neighbors [13], and neural networks [9,12–14]. The results have been heterogeneous, but in many published reports SVMs have been found to be equivalent to more traditional types of models [9,10,15,16]. In some domains, SVMs were claimed to have outperformed other methodologies [11–13,17]. However, limited use of appropriate comparative indices and potential publication bias may account for some of the reported differences. If SVMs improve on the

* Corresponding author. Address: Decision Systems Group, Brigham & Women's Hospital, 75 Francis Street, Boston, MA 02115, USA. Fax: +1 617 739 3672.
*E-mail address:* mmatheny@dsg.harvard.edu (M.E. Matheny).

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.05.008

predictions issued by currently used models, this could be clinically significant, with positive impact on patient and provider satisfaction, improved outcomes, as well as cost savings.

The primary challenges in applying SVM modeling methods to a given domain lies in the selection of the kernel and its parameters as well as the magnitude of the soft margin. Kernels allow SVMs, which are linear machines, to transform the feature space and behave as non-linear models. The parameters of the kernel determine the shape of the separating margin used to classify a set of features (variables). The soft margin is an addition to the SVM method that allows some level of training data misclassification [18].

Several methods have been applied to SVMs in order to provide an automated optimization process, or "tuning", for the selection of parameters. The standard method is grid-search, in which the parameters are varied by fixed steps-sizes through a range of values, and the performance of each set of parameters is measured and compared. Gradient-based approaches can also be used, such as simulated annealing, but require that the score function for assessing the performance of the parameters be differentiable with respect to all the parameters [19–21]. Evolutionary algorithms, a class of iterative, randomized, global optimization techniques, have also been applied to the process of tuning [22]. At each iteration, performance of the offspring (each offspring is a set of parameter values) are evaluated, and the most "fit" are chosen for the next iteration. This continues until a pre-defined termination criterion based on a "fitness function" is met.

Regardless of which method is used to tune a SVM, a performance index (or set of indices) must be selected to evaluate the model. In general, models should be evaluated for discrimination, calibration, and overall error. Discrimination is a measure of the ability of a model to correctly rank a population in terms of individual probability for an outcome, and calibration is the ability to accurately assign a probability of an outcome to individuals or small sub-groups within the population. While a commonly used performance index is classification error, the medical community sees value in obtaining continuous estimates for binary outcomes, such that alternatives include those that measure discrimination (area under the receiver operating characteristic [AUC]), calibration (indices based upon calibration plots), or components of both (Spiegelhalter $Z$ score [23] or cross-entropy error [CEE]).

There has been little exploration regarding the choice of a performance measure on SVM tuning and how the perceived differences between SVMs and other types of models depend on tuning. We compared a variety of performance indices using a grid-search tuning for SVMs in a clinical domain in which several predictive models based on logistic regression have already been developed and evaluated.

## 2. Methods

### 2.1. Source data

We chose to model mortality in percutaneous coronary interventions (PCI) in this study. PCI is one of the most common procedures in cardiology, and is associated with significant morbidity and mortality. Development and validation of risk prediction models in this domain have become popular over the last few decades. This has been due to a combination of the need to stratify patients because of highly variable risk, and opportunity provided by detailed, high quality data that are coded using a national standardized data dictionary [24] as well as increases in mandatory electronic data collection and reporting by some state agencies.

Data were collected from Brigham and Women's Hospital (BWH) (Boston, MA) containing all cases (7914) of percutaneous coronary intervention (PCI) performed at the institution from January 1, 2002 to December 31, 2005. The outcome of interest was post-procedural in-hospital death, and there were 124 (1.57%) events during the collection period. The cases were used to generate 100 random data sets. All cases were used in each set, and 5540 were allocated for training and 2374 were allocated for testing. For SVM evaluation, each training set was randomly divided into 3957 *kernel* training and 1583 *sigmoid* training portions.

Data element definitions were based on the American College of Cardiology—National Cardiovascular Data Registry (ACC-NCDR) data dictionary [24]. The BWH Institutional Review Board approved this study.

### 2.2. Variable selection

After careful literature review, all previously identified risk factors for PCI were selected for inclusion in this study [25–31]. Univariate analysis was performed with SAS (Version 9.1, Cary, NC). Variables not significantly associated with the outcome of death were removed. A total of 21

Table 1
A summary of risk factors associated with mortality after percutaneous coronary intervention

| | |
|---|---|
| Acute heart attack | Hx COPD |
| Age | Hx PVD |
| Body mass index | Hx stroke |
| CHF class | Hyperlipidemia |
| CHF on presentation | Hypertension |
| Creatinine >2.0 mg/dL | IABP |
| Diabetes | Prior PCI |
| Elective case | Shock |
| Emergent case | Unstable angina |
| Family Hx heart disease | Urgent case |
| Heart rate | |

Hx, history; COPD, chronic obstructive pulmonary disease; PVD, peripheral vascular disease; CHF, congestive heart failure; PCI, percutaneous coronary intervention; IABP, intra-aortic balloon pump.

variables were retained for use in model creation and analysis. These variables are listed in Table 1.

### 2.3. Performance measures

A number of commonly used model evaluation methods were used for optimization functions and performance measurements in this study. The area under the receiver operating characteristic (AUC) [32] was used to measure discrimination. The Hosmer–Lemeshow goodness-of-fit statistic (HL $\chi^2$) was used to measure calibration [33], mean-squared error (MSE), and mean cross-entropy error (CEE) [34] were also included as residual-based global indices. HL $\chi^2$, MSE, and CEE performance measurements improve with lower values, and AUC measurements improve with higher values.

### 2.4. Support vector machines

SVMs are able to calculate the maximum margin (separating hyper-plane) between data with and without the outcome of interest if that data are linearly separable [35]. The margin is calculated by solving a constrained optimization problem using the Lagrangian formulation.

However, such applications in real data sets are limited, and a number of adaptations have been applied to SVMs in order to improve their utility. The feature space can be modified using kernels in order to allow fitting of data that are not linearly separable. Individual kernels, such as polynomial and radial basis functions, transform the feature space in distinct ways, and kernel performance is dependent on characteristics of the source data.

Radial (SVM-R) and polynomial (SVM-P) based kernels were selected for evaluation in this study because of their good performance in other domains [9,11,36]. Polynomial kernels transform the feature matrix using the following equation:

$$K(\mathbf{x},\mathbf{y}) = ((\mathbf{x}\circ\mathbf{y})+1)^d$$

where the primary kernel parameter is the power, represented as $d$ in the above equation, and $\mathbf{x}$ and $\mathbf{y}$ are vectors of features. The output of the kernel is the dot product of these two vectors to the $d$ power of the kernel.

Gaussian radial-based kernels [37] transform the feature matrix using the following equation:

$$K(\mathbf{x},\mathbf{y}) = e^{\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{2w^2}}$$

where the primary kernel parameter is $w$, defined as the width of the Gaussian radial basis function, and $\mathbf{x}$ and $\mathbf{y}$ are vectors of features. The output of the kernel depends on the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$.

Real data sets have variable levels of noise, and are unable to capture all of the features relevant to an outcome. In these situations, SVMs cannot correctly classify all training data, and an optimal separating hyperplane does not exist. In order to address this limitation, a soft margin classifier can be used [18]. This assigns a cost penalty for misclassification during training, and the optimization then minimizes the cost function during training.

The SVM models in this evaluation were developed using GIST 2.2.1 (Columbia University, New York, NY). A 2-norm soft margin was employed, which uses slack variables and assigns each misclassified case with a penalty term that quadratically increases depending on the case's distance to the hyperplane. The soft margin is parameterized by a constant ($C$) [38].

Classification SVMs can give outputs as a binary classifier ($-1, 1$) and also as a continuous discriminant (distance from the hyperplane). A method described by Platt [39] allows the generation of a pseudo probabilistic outcome by fitting a sigmoid function to the discriminant using independent holdout training data. In this study, we used the corrected Platt algorithm (Appendix A) provided by Lin et al. [40].

The parameter of each kernel type ($d$ and $w$ for the polynomial and Gaussian kernels, respectively) and the magnitude of the constant applied to the soft margin were optimized on the *kernel training set* separately for AUC, HL $\chi^2$, MSE, and CEE indices by a grid search method, using threefold cross-validation [41]. The *sigmoid training set* was used to convert SVM results into probabilities.

The width function for the radial-based kernel ranged from $2^{-4}$ to $2^4$ ($2^{-4}$, $2^{-3}$, $2^{-2}$, etc.), and the power of the polynomial-based kernel ranged from 1 to 6 by integers. The soft margin constant values used were 0 (no margin), 0.1, 0.3, 0.5, and 0.7. There were no instances of perfect mortality classification in either type of SVM kernel in the training data sets. Using the training set cross-validation results for each of the performance measures, the best set of parameters for the radial and polynomial kernels were used to generate a model on the entire kernel training set, and a sigmoid for discriminant conversion was generated using the sigmoid training set. Each of the models was then evaluated using the respective test data set.

### 2.5. Logistic regression

Comparisons between SVMs and other modeling methods are most commonly done using $k$-nearest neighbors or naïve Bayes modeling methods [42]. However, logistic regression (LR) is almost exclusively used in patient stratification and risk estimation for PCI mortality [25–30]. As a result, LR was chosen to provide the benchmark for the SVM comparisons in this evaluation.

Model development for LR was performed using SAS (Version 9.1, Cary, NC). A standard backwards stepwise model selection method was used [43]. This method generally uses one parameter to perform model building: the threshold for retention of a feature in the model. In an identical fashion to that used for SVM model generation above, threefold cross-validation was performed on each

training data set to optimize the feature selection threshold for the AUC, HL $\chi^2$, MSE, and CEE performance measures. The thresholds evaluated were 0.05–0.50, in 0.05 increments. The optimized threshold parameters were then used to generate a model for each of the entire training data sets, and subsequently applied to the respective test data.

### 2.6. Model comparison

Pair-wise comparisons between model types and performance measures were performed using a one-way ANOVA test for summary values with known standard errors [44] within SAS (Version 9.1, Cary, NC).

## 3. Results

A summary of the test data evaluation for each model type and cross-validation optimization parameter is shown in Table 2. All four performance measures were evaluated (with 95% confidence intervals) for each possible combination of model and optimization parameter.

Pair-wise comparisons were performed between each of the optimization methods for the radial-based kernel SVMs (Table 3) and the polynomial kernel SVMs (Table 4). When the HL $\chi^2$ optimization was applied to the radial-basis kernel SVMs, higher AUC values were achieved relative to the other optimizations (AUC $p = 0.014$, MSE $p = 0.038$, CEE $p = 0.010$). The HL $\chi^2$ optimization for SVM-R also resulted in lower HL $\chi^2$ (AUC $p = <0.001$, MSE $p = <0.001$, CEE $p = 0.040$) and CEE (AUC $p = <0.001$, MSE $p = <0.001$, CEE $p = <0.001$) values when compared to the other optimizations.

The MSE optimization resulted in lower MSE values when compared to AUC ($p = 0.034$) and HL $\chi^2$ ($p = 0.002$). The CEE optimization resulted in lower HL $\chi^2$ values when compared to either the MSE ($p = <0.001$) or AUC ($p = <0.001$) optimization. In addition, the CEE optimization resulted in lower MSE values when compared to HL $\chi^2$ optimization ($p = 0.009$).

When the HL $\chi^2$ optimization was applied to the polynomial kernel SVMs, lower AUC (AUC $p = 0.014$, MSE $p = 0.038$, CEE $p = 0.010$) and higher CEE (AUC $p = 0.010$, MSE $p = 0.003$, CEE $p = 0.001$) values were obtained relative to the other optimizations. Higher MSE values were also obtained for HL $\chi^2$ when compared to MSE ($p = 0.030$) and CEE ($p = 0.017$) optimizations.

The LR models were insensitive to all of the optimization methods for all performance measurements.

Pair-wise analyses were also performed in order to provide comparisons across model types for a given optimization method. Among the models optimized for AUC, the AUC was lower in the radial-based kernel SVM when compared to either of the other two models ($p = <0.001$). For the HL $\chi^2$ performance measure, the LR model was higher than both of the SVM models (SVM-R $p = 0.002$, SVM-P $p < 0.001$). The LR model had higher MSE values than the radial-based SVM ($p = 0.003$). The polynomial kernel SVM had lower CEE values compared to the other two models (LR $p = <0.044$, SVM-R $p = <0.001$).

Among the models optimized for HL $\chi^2$, both SVM models were found to have lower HL $\chi^2$ values ($p = <0.001$), and the radial-based SVM was found to have lower MSE values ($p = 0.025$) than either of the others.

Among the models optimized for MSE, the radial-based kernel SVM had lower AUC values ($p = <0.001$) than either of the other models. The polynomial kernel SVM had lower HL $\chi^2$ values than the LR model ($p = 0.033$) as well as lower CEE values than either of the other models ($p = <0.001$). The radial-based kernel SVM had lower MSE values than either of the other models (LR $p = <0.001$, SVM-P $p = 0.002$).

Table 2
Analysis of the test data by model type and cross-validation optimization method

| Evaluation index | | | | |
|---|---|---|---|---|
| Model (opt method) | AUC (95% CI) | HL $\chi^2$ (95% CI) | MSE (95% CI) | CEE (95% CI) |
| LR (AUC) | 0.912 (0.906–0.917) | 94.7 (46.0–143.4) | 0.0129 (0.0125–0.0133) | 0.0239 (0.0232–0.0246) |
| LR (HL $\chi^2$) | 0.911 (0.905–0.916) | 99.1 (46.8–151.4) | 0.0130 (0.0126–0.0133) | 0.0240 (0.0232–0.0275) |
| LR (MSE) | 0.912 (0.906–0.917) | 89.8 (53.0–126.7) | 0.0129 (0.0125–0.0133) | 0.0239 (0.0231–0.0246) |
| LR (CEE) | 0.912 (0.907–0.917) | 99.8 (49.2–150.3) | 0.0129 (0.0125–0.0133) | 0.0239 (0.0231–0.0246) |
| SVM-R (AUC) | 0.892 (0.886–0.899) | 29.3 (26.7–31.9) | 0.0120 (0.0115–0.0125) | 0.0272 (0.0263–0.0280) |
| SVM-R (HL $\chi^2$) | 0.904 (0.897–0.910) | **14.7 (12.1–17.3)** | 0.0123 (0.0119–0.0127) | 0.0234 (0.0228–0.0241) |
| SVM-R (MSE) | 0.877 (0.872–0.882) | 30.5 (28.2–32.9) | **0.0113 (0.0109–0.0118)** | 0.0266 (0.0258–0.0275) |
| SVM-R (CEE) | 0.890 (0.884–0.897) | 18.4 (16.0–20.8) | 0.0115 (0.0111–0.0119) | 0.0241 (0.0233–0.0249) |
| SVM-P (AUC) | 0.914 (0.908–0.921) | 20.9 (14.9–26.9) | 0.0125 (0.0121–0.0128) | 0.0228 (0.0221–0.0234) |
| SVM-P (HL $\chi^2$) | 0.902 (0.894–0.909) | 15.7 (12.8–18.6) | 0.0129 (0.0125–0.0133) | 0.0240 (0.0233–0.0248) |
| SVM-P (MSE) | 0.912 (0.906–0.919) | 19.5 (13.1–25.9) | 0.0123 (0.0119–0.0126) | 0.0226 (0.0220–0.0232) |
| SVM-P (CEE) | **0.915 (0.908–0.922)** | 20.8 (14.4–27.3) | 0.0122 (0.0118–0.0126) | **0.0224 (0.0217–0.0230)** |

LR, backwards step-wise logistic regression model; SVM-R, radial-based kernel SVM; SVM-P, polynomial-based kernel SVM; opt method, cross-validation optimization method; AUC, area under the receiver operating characteristic curve; HL $\chi^2$, Hosmer–Lemeshow goodness-of-fit; MSE, average mean-squared error; CEE, average cross-entropy error. The best results are shown in bold.

Table 3
Pair-wise comparison of each optimization method on the performance measurements for the radial-basis kernel SVM

| Optimization method | Performance measure | Model A (95% CI) | Model B (95% CI) | $p$ |
|---|---|---|---|---|
| AUC vs **HL $\chi^2$** | AUC | 0.892 (0.886–0.899) | **0.904 (0.897–0.910)** | **0.014** |
| AUC vs MSE | AUC | 0.892 (0.886–0.899) | 0.877 (0.872–0.882) | 0.694 |
| AUC vs CEE | AUC | 0.892 (0.886–0.899) | 0.890 (0.884–0.897) | 0.905 |
| **HL $\chi^2$** vs MSE | AUC | **0.904 (0.897–0.910)** | 0.877 (0.872–0.882) | **0.038** |
| **HL $\chi^2$** vs CEE | AUC | **0.904 (0.897–0.910)** | 0.890 (0.884–0.897) | **0.010** |
| MSE vs CEE | AUC | 0.877 (0.872–0.882) | 0.890 (0.884–0.897) | 0.608 |
| AUC vs **HL $\chi^2$** | HL $\chi^2$ | 29.3 (26.7–31.9) | **14.7 (12.1–17.3)** | **<0.001** |
| AUC vs MSE | HL $\chi^2$ | 29.3 (26.7–31.9) | 30.5 (28.2–32.9) | 0.487 |
| AUC vs **CEE** | HL $\chi^2$ | 29.3 (26.7–31.9) | **18.4 (16.0–20.8)** | **<0.001** |
| **HL $\chi^2$** vs MSE | HL $\chi^2$ | **14.7 (12.1–17.3)** | 30.5 (28.2–32.9) | **<0.001** |
| **HL $\chi^2$** vs CEE | HL $\chi^2$ | **14.7 (12.1–17.3)** | 18.4 (16.0–20.8) | **0.040** |
| MSE vs **CEE** | HL $\chi^2$ | 30.5 (28.2–32.9) | **18.4 (16.0–20.8)** | **<0.001** |
| AUC vs HL $\chi^2$ | MSE | 0.0120 (0.0115– 0.0125) | 0.0123 (0.0119–0.0127) | 0.312 |
| AUC vs **MSE** | MSE | 0.0120 (0.0115–0.0125) | **0.0113 (0.0109–0.0118)** | **0.034** |
| AUC vs CEE | MSE | 0.0120 (0.0115–0.0125) | 0.0115 (0.0111–0.0119) | 0.107 |
| HL $\chi^2$ vs **MSE** | MSE | 0.0123 (0.0119–0.0127) | **0.0113 (0.0109–0.0118)** | **0.002** |
| HL $\chi^2$ vs **CEE** | MSE | 0.0123 (0.0119–0.0127) | **0.0115 (0.0111–0.0119)** | **0.009** |
| MSE vs CEE | MSE | 0.0113 (0.0109–0.0118) | 0.0115 (0.0111–0.0119) | 0.612 |
| AUC vs **HL $\chi^2$** | CEE | 0.0272 (0.0263–0.0280) | **0.0234 (0.0228–0.0241)** | **<0.001** |
| AUC vs MSE | CEE | 0.0272 (0.0263–0.0280) | 0.0266 (0.0258–0.0275) | 0.323 |
| AUC vs **CEE** | CEE | 0.0272 (0.0263–0.0280) | **0.0241 (0.0233–0.0249)** | **<0.001** |
| **HL $\chi^2$** vs MSE | CEE | **0.0234 (0.0228–0.0241)** | 0.0266 (0.0258–0.0275) | **<0.001** |
| **HL $\chi^2$** vs CEE | CEE | **0.0234 (0.0228–0.0241)** | 0.0241 (0.0233–0.0249) | **<0.001** |
| MSE vs **CEE** | CEE | 0.0266 (0.0258–0.0275) | **0.0241 (0.0233–0.0249)** | **<0.001** |

Statistical significance ($p < 0.05$) of the best performing method and performance measurement result are indicated in bold.

Table 4
Pair-wise comparison of each optimization method on the performance measurements for the polynomial kernel SVM

| Optimization method | Performance measure | Model A (95% CI) | Model B (95% CI) | $p$ |
|---|---|---|---|---|
| **AUC** vs HL $\chi^2$ | AUC | **0.914 (0.908–0.921)** | 0.902 (0.894–0.909) | **0.014** |
| AUC vs MSE | AUC | 0.914 (0.908–0.921) | 0.912 (0.906–0.919) | 0.694 |
| AUC vs CEE | AUC | 0.914 (0.908–0.921) | 0.915 (0.908–0.922) | 0.905 |
| HL $\chi^2$ vs **MSE** | AUC | 0.902 (0.894–0.909) | **0.912 (0.906–0.919)** | **0.038** |
| HL $\chi^2$ vs **CEE** | AUC | 0.902 (0.894–0.909) | **0.915 (0.908–0.922)** | **0.010** |
| MSE vs CEE | AUC | 0.912 (0.906–0.919) | 0.915 (0.908–0.922) | 0.608 |
| AUC vs HL $\chi^2$ | HL $\chi^2$ | 20.9 (14.9–26.9) | 15.7 (12.8–18.6) | 0.201 |
| AUC vs MSE | HL $\chi^2$ | 20.9 (14.9–26.9) | 19.5 (13.1–25.9) | 0.733 |
| AUC vs CEE | HL $\chi^2$ | 20.9 (14.9–26.9) | 20.8 (14.4–27.3) | 0.992 |
| HL $\chi^2$ vs MSE | HL $\chi^2$ | 15.7 (12.8–18.6) | 19.5 (13.1–25.9) | 0.349 |
| HL $\chi^2$ vs CEE | HL $\chi^2$ | 15.7 (12.8–18.6) | 20.8 (14.4–27.3) | 0.205 |
| MSE vs CEE | HL $\chi^2$ | 19.5 (13.1–25.9) | 20.8 (14.4–27.3) | 0.740 |
| AUC vs HL $\chi^2$ | MSE | 0.0125 (0.0121–0.0128) | 0.0129 (0.0125–0.0133) | 0.137 |
| AUC vs MSE | MSE | 0.0125 (0.0121–0.0128) | 0.0123 (0.0119–0.0126) | 0.490 |
| AUC vs CEE | MSE | 0.0125 (0.0121–0.0128) | 0.0122 (0.0118–0.0126) | 0.370 |
| HL $\chi^2$ vs **MSE** | MSE | 0.0129 (0.0125–0.0133) | **0.0123 (0.0119–0.0126)** | **0.030** |
| HL $\chi^2$ vs **CEE** | MSE | 0.0129 (0.0125–0.0133) | **0.0122 (0.0118–0.0126)** | **0.017** |
| MSE vs CEE | MSE | 0.0123 (0.0119–0.0126) | 0.0122 (0.0118–0.0126) | 0.836 |
| **AUC** vs HL $\chi^2$ | CEE | **0.0228 (0.0221–0.0234)** | 0.0240 (0.0233–0.0248) | **0.010** |
| AUC vs MSE | CEE | 0.0228 (0.0221–0.0234) | 0.0226 (0.0220–0.0232) | 0.723 |
| AUC vs CEE | CEE | 0.0228 (0.0221–0.0234) | 0.0224 (0.0217–0.0230) | 0.439 |
| HL $\chi^2$ vs **MSE** | CEE | 0.0240 (0.0233–0.0248) | **0.0226 (0.0220–0.0232)** | **0.003** |
| HL $\chi^2$ vs **CEE** | CEE | 0.0240 (0.0233–0.0248) | **0.0224 (0.0217–0.0230)** | **0.001** |
| MSE vs CEE | CEE | 0.0226 (0.0220–0.0232) | 0.0224 (0.0217–0.0230) | 0.675 |

Statistical significance ($p < 0.05$) of the best performing method and performance measurement result are indicated in bold.

Among the models optimized for CEE, the radial-based kernel SVM had lower AUC values ($p = <0.001$) than the other models. Both SVM models had lower HL $\chi^2$ values than the LR model ($p < 0.001$). The SVM-R model had lower MSE values than both the SVM-P ($p = 0.011$) and the LR ($p = <0.001$) models, and the SVM-P model had lower CEE values than the LR ($p = 0.004$) and the SVM-R ($p = 0.001$) models.

## 4. Discussion

Parameter tuning of SVM-R models based on the optimization of HL $\chi^2$ values provided benefits for all of the performance measurements except MSE relative to the other optimization methods, and was the only SVM-R optimization to retain similar AUC values to the SVM-P and LR models. SVM-R models also resulted in improved HL $\chi^2$ and MSE performance compared with LR models for all optimizations except MSE.

Among the SVM-P models, the HL $\chi^2$ optimization method resulted in lower AUC, and worse MSE and CEE relative to the other methods. In addition, HL $\chi^2$ performance indices were insensitive to the optimization method. However, the SVM-P models had improved HL $\chi^2$ and CEE performance values over the LR models for all optimization methods except HL $\chi^2$.

The MSE method is a common optimization method in regression, and attempts to minimize the overall bias and variance. Valentini and Dietterich conducted a bias-variance analysis of SVMs using a grid search method with kernel and soft margin parameters [45]. The expected trade-off between bias and variance was not found in some data sets using either radial-basis or polynomial kernels, and suggested that a bias-only optimization could result in increased classification performance without a corresponding increase in variance.

Comparisons between CEE and MSE as optimization parameters for other machine learning methods were done by other authors who did not report significant differences [46]. The AUC and HL $\chi^2$ values were experimental optimization parameters. To our knowledge, there have been no reports on their use in optimizing SVMs.

There are a number of limitations in this study. This is an exploratory work, and given the large number of evaluations, there is an increased potential for false positives. Due to the high computational times associated with the large data set and the search space, feature selection subsets were not evaluated in each of the modeling methods. While features useful for predicting post-procedural mortality in PCI have been extensively explored in the literature, this could limit the optimization results for this domain.

While the logistic regression results in this study were similar to those found in the past for this clinical domain [25–31,47], the optimization process was limited to backward variable selection using each of the four optimization methods. This limitation may have contributed to the insensitivity of the LR models to the optimization processes, and may have biased the findings that SVM models were superior to LR models.

The parameter selection process used a threefold CV method, and the model evaluation used a separate testing sample over 100 randomized data sets. This is related to the nested stratified 10-fold CV method as described by Statnikov et al. [48]. The small number of training folds (or inner loops) were utilized because of high computational times of GIST in the relatively large data sets. This may have increased the variance of the results in the parameter optimization methods, although the large number of data points in each fold likely minimized this concern.

In summary, each of the model types were optimized on a cross-validated training set based on the maximization (AUC) or minimization (HL $\chi^2$, MSE, CEE) of one of four performance measurements. The choice of a tuning optimization parameter had significant impact on the performances of both SVM kernel types. Evaluation of these methods in this clinical domain suggests that the use of HL $\chi^2$ values for optimization relative to the other methods may improve performance for radial-basis kernel SVMs, but degrade performance for polynomial kernel SVMs. Radial-basis kernel SVMs were associated with lower discrimination and improved calibration when compared to the LR models, but the latter were subject to less exploration and hence further investigation is necessary. Polynomial kernel SVMs performed in a similar manner to the LR models with respect to discrimination, and some of the calibration measures showed improvement. However, the HL $\chi^2$ has known limitations [49], and therefore this result needs further validation.

This comparative study shows that SVMs can be very sensitive to the selected optimization method, and that it would be inappropriate to state that in general they are superior to LR models. Omission of certain evaluation indices may also favor a certain type of model, and care must be taken in declaring superiority of any model without sufficient exploration of different optimization techniques. Finally, the findings from this exploratory evaluation require confirmation in other types of clinical data, and further work is underway to explore these aspects.

## Acknowledgments

The authors thank Anne Fladger and her staff for their assistance. This study was funded in part by Grants R01-LM-08142, R01-1LM-009520, and 1-T15-LM-07092 from the National Library of Medicine of the National Institutes of Health.

## Appendix A. Lin et al. pseudo-code of the corrected Platt algorithm (reproduced with permission) [40]

```
Input parameters:
       out = array of SVM outputs
       target = array of booleans: is ith example a positive example?
       prior1 = number of positive examples
       prior0 = number of negative examples
Outputs:
       A, B = parameters of sigmoid

//Parameter setting
maxiter=100 //Maximum number of iterations
minstep=1e-10 //Minimum step taken in line search
sigma=1e-3 //Set to any value > 0
//Construct initial values: target support in array t, initial function value in fval
hiTarget=(prior1+1.0)/(prior1+2.0), loTarget=1/(prior0+2.0)
len=prior1+prior0
for i = 1 to len {
       if (label[i] > 0)
              t[i]=hiTarget
       else
              t[i]=loTarget
}
A=0.0, B=log((prior0+1.0)/(prior1+1.0)), fval=0.0
for i = 1 to len {
       fApB=deci[i]*A+B
       if (fApB >= 0)
              fval += t[i]*fApB+log(1+exp(-fApB))
       else
              fval += (t[i]-1)*fApB+log(1+exp(fApB))
}
for it = 1 to maxiter {
       //Update Gradient and Hessian (use H' = H + sigma I)
       h11=h22=sigma, h21=g1=g2=0.0
       for i = 1 to len {
              fApB=deci[i]*A+B
              if (fApB >= 0)
       p=exp(-fApB)/(1.0+exp(-fApB)), q=1.0/(1.0+exp(-fApB))
else
       p=1.0/(1.0+exp(fApB)), q=exp(fApB)/(1.0+exp(fApB))
```

```
        d2=p*q
        h11 += deci[i]*deci[i]*d2
        h22 += d2
        h21 += deci[i]*d2
        d1=t[i]-p
        g1 += deci[i]*d1
        g2 += d1
}
if (abs(g1)<1e-5 && abs(g2)<1e-5) //Stopping criteria
        break
det=h11*h22-h21*h21
dA=-(h22*g1-h21*g2)/det, dB=-(-h21*g1+h11*g2)/det //Modified Newton direction
gd=g1*dA+g2*dB
stepsize=1
while (stepsize >= minstep){ //Line search
        newA=A+stepsize*dA, newB=B+stepsize*dB, newf=0.0
        for i = 1 to len {
                fApB=deci[i]*newA+newB
                if (fApB >= 0)
                        newf += t[i]*fApB+log(1+exp(-fApB))
                else
                        newf += (t[i]-1)*fApB+log(1+exp(fApB))
        }
        if (newf<fval+0.0001*stepsize*gd){ //Check sufficient decrease
                A=newA, B=newB, fval=newf
                        break
                }
                else
                        stepsize=stepsize/2.0
        }
        if (stepsize < minstep){ //Line search fails
                print 'Line search fails'
                break
        }
}
if (it >= maxiter)
        print 'Reaching maximum iterations'
return [A,B]
```

## References


[1] Randolph AG, Guyatt GH, Carlet J. Understanding articles comparing outcomes among intensive care units to rate quality of care. Evidence based medicine in critical care group. Crit Care Med 1998;26:773–81.

[2] Topol EJ, Block PC, Holmes DR, Klinke WP, Brinker JA. Readiness for the scorecard era in cardiovascular medicine. Am J Cardiol 1995;75:1170–3.

[3] Hunt JP, Meyer AA. Predicting survival in the intensive care unit. Curr Prob Surg 1997;34:527–99.

[4] Knaus WA, Wagner DP, Draper EA. The value of measuring severity of disease in clinical research on acutely ill patients. J Chronic Dis 1984;37:455–63.

[5] Mendez-Tellez PA, Dorman T. Predicting patient outcomes, futility, and resource utilization in the intensive care unit: the role of severity scoring systems and general outcome prediction models. Mayo Clin Proc 2005;80:161–3.

[6] Hariharan S, Zbar A. Risk scoring in perioperative and surgical intensive care patients: a review. Curr Surg 2006;63:226–36.

[7] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. Pittsburgh, PA: ACM Press; 1992.

[8] Vapnik VN. The nature of statistical learning theory. 2nd ed. New York, NY: Springer-Verlag; 1999.

[9] Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, Binder M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. J Biomed Inform 2001;34:28–36.

[10] Gwiggner C, Lanckriet G. Characteristics in flight data—estimation with logistic regression and support vector machines. In: International conference on research in air transportation, Zilina, Slovakia, 2004.

[11] Mocellin S, Ambrosi A, Montesco MC, Foletto M, Zavagno G, Nitti D, et al. Support vector machine learning model for the prediction of sentinel node status in patients with cutaneous melanoma. Ann Surg Oncol 2006;13:1113–22.

[12] Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, et al. Computational prediction of methylation status in human genomic sequences. Proc Natl Acad Sci USA 2006;103:10713–6.

[13] Mavroforakis ME, Georgiou HV, Dimitropoulos N, Cavouras D, Theodoridis S. Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers. Artif Intell Med 2006;37:145–62. Epub 2006 May, 2023.

[14] Lin TH, Chiu SH, Tsai KC. Supervised feature ranking using a genetic algorithm optimized artificial neural network. J Chem Inf Model 2006;46:1604–14.

[15] Bhattacharya B, Solomatine DP. Machine learning in soil classification. Neural Netw 2006;19:186–95.

[16] Gromiha MM, Suwa M. Discrimination of outer membrane proteins using machine learning algorithms. Proteins 2006;63:1031–7.

[17] Wang Y, Xue Z, Xu J. Better prediction of the location of alpha-turns in proteins with support vector machine. Proteins 2006;65:49–54.

[18] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.

[19] Chapelle O, Vapnik V, Bousquet O, Mukherjee S. Choosing multiple parameters for support vector machines. Mach Learn 2002;36:131–59.

[20] Chung K-M, Kao W-C, Sun C-L, Wang L-L, Lin C-J. Radius margin bounds for support vector machines with the RBF kernel. Neural Comput 2003;15:2643–81.

[21] Glasmachers T, Igel C. Gradient-based adaptation of general Gaussian kernels. Neural Comput 2005;17:2099–105.

[22] Friedrichs F, Igel C. Evolutionary tuning of multiple SVM parameters. Neurocomputing 2004;64:107–17.

[23] Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5:421–33.

[24] Cannon CP, Battler A, Brindis RG, Cox JL, Ellis SG, Every NR, et al. American College of Cardiology key data elements and definitions for measuring the clinical management and outcomes of patients with acute coronary syndromes. J Am Coll Cardiol 2001;38:2114–30.

[25] O'Connor GT, Malenka DJ, Quinton H, Robb JF, Kellett Jr MA, Shubrooks S, et al. Multivariate prediction of in-hospital mortality after percutaneous coronary interventions in 1994–1996. J Am Coll Cardiol 1999;34:681–91.

[26] Hannan EL, Arani DT, Johnson LW, Kemp Jr HG, Lukacik G. Percutaneous transluminal coronary angioplasty in New York State risk factors and outcomes. JAMA 1992;268:3092–7.

[27] Hannan EL, Racz M, Ryan TJ, McCallister BD, Johnson LW, Arani DT, et al. Coronary angioplasty volume-outcome relationships for hospitals and cardiologists. JAMA 1997;277:892–8.

[28] Moscucci M, Kline-Rogers E, Share D, O'Donnell M, Maxwell-Eward A, Meengs WL, et al. Simple bedside additive tool for prediction of in-hospital mortality after percutaneous coronary interventions. Circulation 2001;104:263–8.

[29] Shaw RE, Anderson HV, Brindis RG, Krone RJ, Klein LW, McKay CR, et al. Development of a risk adjustment mortality model using the American College of Cardiology-National Cardiovascular Data Registry (ACC-NCDR) experience: 1998–2000. J Am Coll Cardiol 2002;39:1104–12.

[30] Ellis SG, Weintraub W, Holmes D, Shaw R, Block PC, King 3rd SB. Relation of operator volume and experience to procedural outcome of percutaneous coronary revascularization at hospitals with high interventional volumes. Circulation 1997;95:2479–84.

[31] Resnic FS, Ohno-Machado L, Selwyn A, Simon DI, Popma JJ. Simplified risk score models accurately predict the risk of major in-hospital complications following percutaneous coronary intervention. Am J Cardiol 2001;88:5–9.

[32] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.

[33] Lemeshow S, Hosmer Jr DW. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982;115:92–106.

[34] Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat 1951;22:79–86.

[35] Pavlidis P, Wapinski I, Noble WS. Support vector machine classification on the web. Bioinformatics 2004;20:586–7.

[36] Song X, Mitnitski A, Cox J, Rockwood K. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. Medinfo 2004;11:736–40.

[37] Scholkopf B, Sung K, Burges C, Girosi F, Niyogi P, Poggio T, et al. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Trans Sig Proc 1997;45:2758–65.

[38] Cristianini N, Shawe-Taylor J. An introduction to support vector machines. Cambridge, MA: Cambridge University Press; 2000.

[39] Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola AJ, Bartlett P, Schoelkopf B, Schuurmans D, editors. Advances in large margin classiers. Cambridge, MA: MIT Press; 1999.

[40] Lin H-T, Lin C-J, Weng RC. A note on Platt's probabilistic outputs for support vector machines. <http://www.csie.ntu.edu.tw/~cjin/papers/plattprob.ps/> [accessed: 03.08.06].

[41] Platt J. Fast training of support vector machines using sequential minimal optimization. In: Scholkopf B, Burges C, Smola A, editors. Advances in kernel methods—Support vector learning, 1998.

[42] Russell S, Norvig P. Artificial intelligence: a modern approach. 2nd ed. Upper Saddle River, NJ: Pearson Education; 2003.

[43] Hosmer D, Lemeshow S. Applied logistic regression. New York, NY: Wiley; 1989.

[44] Larson DA. Analysis of variance with just summary statistics as input. Am Stat 1992;46:151–2.

[45] Valentini G, Dietterich TG. Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. J Mach Learn Res 2004;5:725–75.
[46] Lucas SM. Discriminative training of the scanning N-tuple classifier, vol. 2686. Heidelberg, Germany: Springer Berlin; 2003. p. 222–9.
[47] Matheny ME, Ohno-Machado L, Resnic FS. Discrimination and calibration of mortality risk prediction models in interventional cardiology. J Biomed Inform 2005;38:367–75. Epub 2005 March, 2026.
[48] Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 2005;21:631–43.
[49] Bertolini G, D'Amico R, Nardi D, Tinazzi A, Apolone G. One model, several results: the Paradox of the Hosmer–Lemeshow goodness-of-fit test for the logistic regression model. J Epidemiol Biostat 2000;5:251–3.