Optimal Parameter Selection of an SVM Model: Application to Hip Fracture Risk Prediction

Peng Jiang*, Samy Missoum†, Chengcheng Hu‡ and Zhao Chen§
University of Arizona, Tucson, Arizona, 85721, USA

This article presents a study of three (cross) validation metrics used for the selection of the optimal parameters of a support vector machine (SVM) classifier. The study focuses on problems for which the data is non-separable and unbalanced as is often the case for experimental and clinical data. The three metrics selected in this work are the area under the ROC curve, accuracy, and balanced accuracy. As a test example, the study investigates the optimal parameters for an SVM classification model for hip fracture. The hip fracture data is obtained from a finite element model that is fully parameterized. Because the data is computational, fully separable sets of data (fracture and safe) can be obtained. By projection onto a lower dimensional sub-space, the data becomes non-separable and is used to construct the SVM. The knowledge of the separable case provides a comparison metric (the weighted likelihood) that would be unknown if only clinical data is used. The performance of the various metrics are compared for several levels of separability, unbalance and size of training samples. A probabilistic SVM is used to compute the probability of fracture.

I. Introduction

In many areas of science and engineering, models are used to predict quantities such as the responses of a system or of a physical process. A model can be constructed in many ways. For instance, it can be purely computational such as finite element simulations or it can be solely constructed from experimental or clinical data. In any case, the most important characteristic of the model is its predictive capability. In other words, how is the model going to represent reality beyond the data that was used to construct it? For instance how can one validate a model for hip fracture risk that was constructed from clinical data ?

There exist several metrics to quantify the predictive ability of a model. These metrics can be used to construct the model through the use of cross-validation applied to a training set. In order to test these metrics more effectively, these metrics can be evaluated on a validation (or clinical) set that is different from the training data set. Examples of metrics are the accuracy, area under the Receiver Operating Characteristic (ROC) curve (AUC),[1,2] balanced accuracy,[3] F-score[4] and Matthews correlation coefficient.[5] Although some of these metrics have gained wide acceptance, care must be taken in some situations. For instance, considered a binary classification problem (e.g., failure or not). When the classification model is constructed on physical experiments or clinical data, two types of difficulties might appear. The first one is the non-separability of the data. The second problem stems from the fact that the data might be unbalanced. For instance, there is typically a much smaller number of failure than safe cases.

The focus of this article is to test the performance of three well-known validation metrics for a classification problem where both issues of non-separability and unbalance data occur. More specifically, the classification is base on an support vector machine (SVM) and the three metrics used are the AUC, basic accuracy and balanced accuracy. These metrics are used for both cross-validation and validation.

*Graduate Student, Aerospace and Mechanical Engineering Department,College of Engineering. AIAA Student Member.
†Associate Professor, Aerospace and Mechanical Engineering Department, College of Engineering. AIAA Senior Member.
‡Associate Professor, Mel and Enid Zuckerman College of Public Health.
§Professor, Mel and Enid Zuckerman College of Public Health.

In this work, computational data was used instead of experiments in order to have more control on the features of the data sets (non-separability and unbalance). In order to create non-separable cases, a set of separable data was created and then projected onto a sub-space thus leading to non-separable data. In addition, this approach provides an exact reference metric (based on a likelihood) to compare the validation metrics.

As part of an ongoing effort on the prediction of hip fracture, the test cases presented in this work are based on a fully parameterized finite element model of a femur. Given a failure criterion and a set of parameters (e.g., neck radius), an SVM separating failure and safe samples is constructed. Several scenarios of data sets for non-separability and unbalanced data are studied.

The paper is organized as follows: A review of SVM for balanced and unbalanced data is presented in the background Section II. The section also introduces three validation metrics: accuracy, balanced accuracy and AUC. Section III presents the details of the parameter selection strategy in case of non-separable and unbalanced data. It also provides the derivation for a likelihood-based reference metric. Finally, Section IV provides the results and conclusions based on various data sets with different levels of unbalance, separability, as well as sizes of training samples. In addition, an example of probability of hip fracture prediction using probabilistic SVM (PSVM)[6,7] is presented.

II. Background

II.A. Support vector machine (SVM) classification

In this paper, we are concerned with the predictive capability of a classification model. The classifier chosen in this work is referred to as a support vector machine (SVM).[8,9] SVM is now a widely accepted machine learning technique that has been used in many applications.[10,11,12,13,14,15]

An SVM is used to construct an explicit boundary that separates samples belonging to two classes labeled as $+1$ and -1. Given a set of N training samples $\mathbf{x}_i$ in a d-dimensional space, and the corresponding class labels, a linear SVM separation function is found through the solution of the following quadratic optimization problem shown in Equation 1.

$$\begin{aligned} \min_{\mathbf{w},\xi,b} \ & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ s.t. \ & y_i(\mathbf{w}\cdot\mathbf{x_i} - b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1,\ldots,N \end{aligned} \tag{1}$$

where b is a scalar referred to as the bias, C is the cost coefficient, ξ_i are slack variables which measure the degree of misclassification of each sample $\mathbf{x_i}$ in the case the data is non separable. SVM can be generalized by writing the dual problem and replacing the inner product by a kernel:

$$\begin{aligned} \min_{\lambda} \ & \frac{1}{2}\sum_{i,j}\lambda_i\lambda_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) - \sum_{i=1}^{N}\lambda_i \\ s.t. \ & \sum_{i=1}^{N}\lambda_i y_i = 0 \\ & 0 \leq \lambda_i \leq C, i = 1,\ldots,N \end{aligned} \tag{2}$$

where λ_i are Lagrange multipliers. The training samples for which the Lagrange multipliers are non-zero are referred to as the *support vectors*. The number of support vectors NSV is usually much smaller than N, and therefore, only a small fraction of the samples affect the SVM equation.

The corresponding SVM boundary is given as:

$$s(\mathbf{x}) = b + \sum_{i=1}^{N}\lambda_i y_i K(\mathbf{x_i},\mathbf{x}) = \mathbf{0} \tag{3}$$

The classification of any arbitrary point $\mathbf{x}$ is given by the sign of $s(\mathbf{x})$. The kernel function K in Equation 3 can have several forms, such as polynomial or Gaussian radial basis kernel, used in this article:

$$K(\mathbf{x}_i, \mathbf{x}) = \exp\left(-\gamma \left|\left|\mathbf{x}_i - \mathbf{x}\right|\right|^2\right), \gamma > 0 \tag{4}$$

where γ is the width parameter of the Gaussian kernel.

For some classification problems, especially when handling data collected for biomedical studies, the data is usually unbalanced. In other words, a class might be far more populated than the other one, as it is the case in hip fracture clinical studies. It order to balance the data, Osuna[16] and Vapnik[17] proposed using different cost coefficients (i.e., weights) for the different classes in the SVM formulation 5. The formulation for the linear case is given in Equation 5. it is generalized to the non-linear case using the kernel.

$$\begin{aligned} \min_{\mathbf{w},\xi,b} \ & \frac{1}{2}\|\mathbf{w}\|^2 + C^+ \sum_{i=1}^{N^+} \xi_i + C^- \sum_{i=1}^{N^-} \xi_i \\ s.t. \ & y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, \ldots, N \end{aligned} \tag{5}$$

where C^+ and C^- are cost coefficients for $+1$ and -1 class separately. N^+ and N^- are number of samples from $+1$ and -1 classes. The coefficients are typically chosen as:[18]

$$\begin{aligned} C^+ &= C \times w^+ \\ C^- &= C \times w^- \end{aligned} \tag{6}$$

where C is the common cost coefficient for both classes, w^+ and w^- are the weights for $+1$ and -1 class separately. In this article, the weights for different classes are selected as $w^+ = 1$ and $w^- = N^+/N^-$.

II.B. *k*-Fold cross validation

Cross validation is a commonly used technique to estimate how accurately a predictive model will perform in practice. In k-fold cross-validation, samples from both safe and failed classes in the training set are randomly divided into k subsets of equal size. Of all the k subsets, a single subset is used as validation samples for evaluating the model, and the remaining $k-1$ subsets are used as training samples. The cross-validation process is then repeated k times, with each of the k subsets used exactly once for the validation. The k results from the "folds" can be averaged to produce a single estimation of model performance. In this article, 10-fold cross-validation is used.[19,20]

II.C. Commonly used (cross) validation metrics

The SVM model depends on the cost coefficients and the coefficient of the kernel as presented in Equation 2 and 4. These parameters are typically found through cross-validation performed on a training set. In this article, three types of cross-validation metrics are used: accuracy, area under the ROC curve (AUC), and balanced accuracy. These metrics can also be used on validation (also called clinical trial) sets that are made of data not used in the training process. When these quantities are calculated on the clinical sets, they are referred to "validation" metrics as they quantify the predictive ability of the model.

II.C.1. Accuracy and Balanced Accuracy

For convenience, we use the following abbreviations for empirical quantities: P (# positive samples), N (# negative samples), TP (# true positives, correctly classified positive samples), TN (# true negatives, correctly classified negative samples), FP (# false positives, misclassified negative samples), FN (# false negatives, misclassified positive samples).

The criteria can be expressed as:

$$\text{Accuracy} = \frac{TP + TN}{P + N} \tag{7}$$

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right) \tag{8}$$

Accuracy is an intuitive and the most widely-used criterion for evaluating a classifier. And it works well if the number of samples in different classes are balanced. But when the two classes are highly unbalanced, the performance of this measure may lead to the phenomenon of "over-fitting" (see results Section IV). In the case the data is not balanced, the balanced accuracy should be used.

II.C.2. Area Under ROC Curve

The receiver operating characteristic (ROC) curve represents relation between true and false positive for all the possible thresholds of the model. In the SVM classification case, the thresholds are defined by the SVM value. More specifically, for each threshold, the number of predicted positive and negative samples vary and will lead to different pairs of True Positive Rate ($TPR = TP/(TP + FN)$), and False Positive Rate ($FPR = FP/(FP + TN)$). Graphed as coordinate pairs, these measures form the ROC curve. The ROC curve describes the performance of a model across the entire range of classification thresholds. An example ROC curve is depicted in Figure 1.

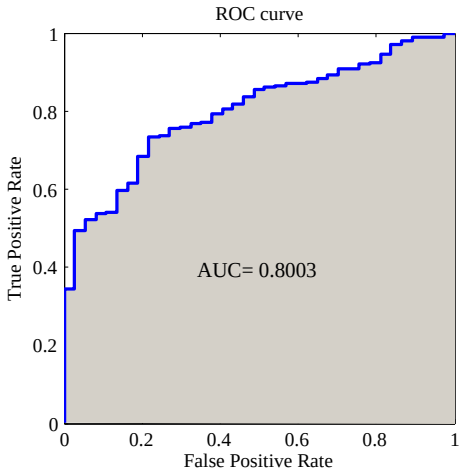


Figure 1. An example of ROC curve and its corresponding AUC.

The area under the ROC curve (AUC) is widely recognized as the measure of predicting ability of a model.[21] It is equal to the probability that a classifier will rank a randomly selected positive sample higher than a randomly chosen negative one.[2] The maximum value for the AUC is 1, indicating a "perfect" classifier that predicts all samples without misclassification. An AUC value of 0.5 indicates no discriminative ability between samples from different classes, as flipping a coin for decision-making.

II.D. Parameter selection strategy for SVM with Gaussian kernel based on grid search

The optimal parameters C and γ of the SVM are found by searching the values that maximize one of the cross-validation metrics described in the previous section. A typical approach consist of constructing a grid and choosing the maximizer out of the discrete set of points. Another approach is to use a global optimization method such as a Genetic Algorithm.[22] The ranges of the variables chosen in this work are: $C \in [2^{-10}, 2^{17}]$ and $\gamma \in [2^{-25}, 2^{10}]$. Within this range, the SVM can be a hard or soft classifier and the decision boundary can go from a hyperplane to a highly non-linear hypersurface.

II.E. Confidence intervals estimation

In order to obtain a confidence interval around the various validation metrics, bootstrapping can be used.[23,24]

For a dataset of size n, a bootstrap sample is created by selecting n instances uniformly from the pool of data with replacement along with their predicted scores from the SVM. The validation metric can be recalculated from this bootstrap sample. This process is repeated for a large number of times to form a distribution of validation metric values. From this distribution, 95% or 99% confidence intervals can be estimated using its empirical distribution.

III. Methodology

III.A. Manufactured cases with non separable data

In many engineering or biomedical problems, the data is not separable. This stems from the fact that the data is usually studied in a finite dimensional space which does not account for all the factors that might influence an outcome. For instance, when studying hip fracture data for a cohort of patients, the results are typically reported in a space made of parameters such as age, weight, bone mineral density et al. Even in the case when this space is highly dimensional, the data might still not be separable as the number of dimensions used might still be a fraction of the actual number of factors involved in the occurrence of hip fracture. In other words, non-separable data can be seen as the projection of otherwise separable data onto a space of lower dimensionality (Figure 2). The figure also depicts an SVM classifier in a three dimensional space where the data is separable and the classifier in the two dimensional space (projected space) where the data is not separable.

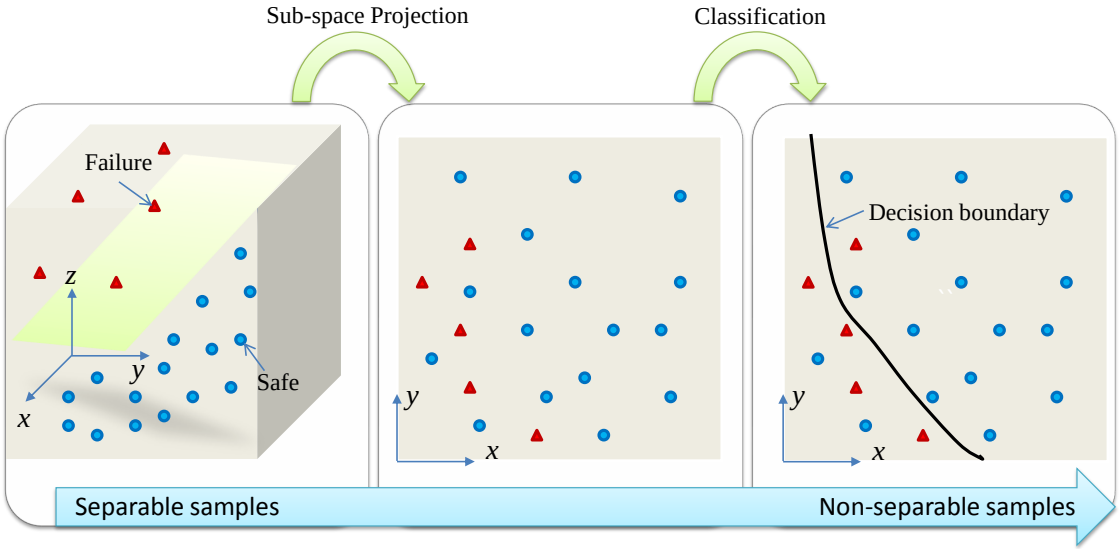


Figure 2. Manufactured non-separable samples based on separable ones.

Based on these observations, this article proposes to manufacture non-separable cases by projecting the data from separable space onto its sub-space. Using this approach, because the normally unknown separable case is available, a quantity can be derived to compare the various validation metrics to a "reference" value. The proposed likelihood-based metric in presented in the next section.

III.A.1. Calculation of weighted likelihood

For comparison of the various validation metrics, this section introduces a likelihood-based metric constructed from the separable case.

The metric is constructed by calculating the probability for a sample to belong to its actual class based on the (again normally unknown) separable case SVM. This is done for each training sample using Monte-Carlo simulations. Figure 3 provides an example of the methodology for a 3D case. Monte-Carlo samples are generated following normal distribution with mean and standard deviation based on variable "z". The probability of belonging to the safe class for any sample $\mathbf{x}_i$ can be calculated as:

$$P(+1|\mathbf{x}_i) = \frac{N_i^+}{N_{MC}}, i = 1, \ldots, N \tag{9}$$

where $\mathbf{x}_i$ is the ith sample in the sub-space, N_{MC} is the number of Monte Carlo samples and N_i^+ represents the number of Monte Carlo samples that are predicted as safe by the SVM constructed with separable data in the original space. And the corresponding probability of belonging to the failed class can be calculated

as:

$$P(-1|\mathbf{x}_i) = 1 - P(+1|\mathbf{x}_i), i = 1, \ldots, N \tag{10}$$

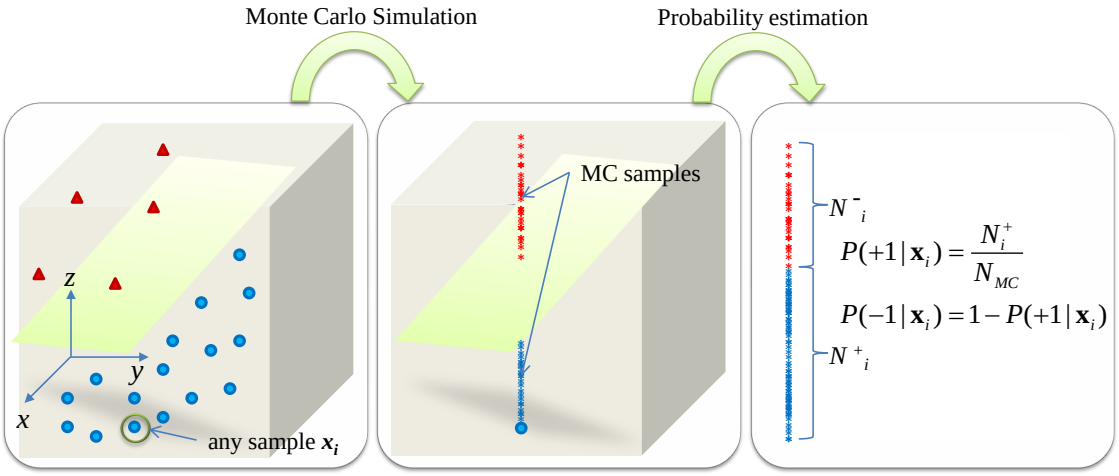


Figure 3. Calculation of the probability of belonging to a class for any sample x_i using Monte Carlo simulation. The sampling is performed based on the variables that were removed by projection.

Using these probabilities, a likelihood-based metric is used to quantify the similarity between the SVM in the subspace and in the separable space. The metric is implemented as a weighted likelihood which provides:

$$\overline{\mathcal{L}}_w = \frac{1}{N_{misc}} \sum_{i=1}^{N_{misc}} w_i \log(P_i) \tag{11}$$

$$P_i = \begin{cases} P(+1|\mathbf{x}_i) & \text{if } y_i = +1 \text{ and } s(\mathbf{x}_i)y_i < 0 \\ P(-1|\mathbf{x}_i) & \text{if } y_i = -1 \text{ and } s(\mathbf{x}_i)y_i < 0 \end{cases}, w_i = \begin{cases} 1 & \text{if } y_i = +1 \text{ and } s(\mathbf{x}_i)y_i < 0 \\ \frac{N^+}{N^-} & \text{if } y_i = -1 \text{ and } s(\mathbf{x}_i)y_i < 0 \end{cases}$$

where N_{misc} is the number of misclassified samples by the SVM constructed in the sub-space, P_i is the probability of being into the actual class for every misclassified sample, y_i is the actual class of sample $\mathbf{x}_i$ and w_i is weight of sample $\mathbf{x}_i$ involved in the calculation of $\overline{\mathcal{L}}_w$. The weights are used for the case where the data is unbalanced.

Basically, $\overline{\mathcal{L}}_w$ is the logarithm of weighted average of the probability of belonging to the actual class for every the misclassified sample using the SVM constructed in the sub-space. The unbalance between classes is taken into account by the weights w_i.

The larger the algebraic value of the weighted likelihood, the better is the SVM in comparison to the actual separable SVM.

III.A.2. Calculation of a reference weighted likelihood

In the case of manufactured data sets with separable data, it is possible to obtain a reference value for the weighted likelihood. This is done by increasing the number of samples to large values until the weighted likelihood converges to the reference weighted likelihood. In addition, the reference value is provided with a 95% confidence interval. An example is provided in Figure 9.

IV. Results

This section presents results of SVM parameter selection on various sample configurations using different validation metrics. As described in the methodology section, a non-separable data set is generated by projection of a separable case in higher dimension. Sample sets used in this section have different sizes, levels of separability as well as levels of unbalance. All scores of the three different (cross) validation metrics and

the weighted likelihood $\overline{\mathcal{L}}_w$ as well as their 95% confidence intervals are provided.

The following notations are used in this section:

- $\overline{\mathcal{L}}_w$: weighted likelihood.
- Ref. $\overline{\mathcal{L}}_w$: reference value of weighted likelihood for each sample set. The way of getting this reference value is shown in Section III.
- *AUC*: area under ROC curve.
- *Acc*: accuracy.
- *Bacc*: balanced accuracy.

As a "real world" test problem, we consider the case of "hip fracture" for which an SVM is used to classify the cases that are fractured and the cases that are not. The data is obtained from a fully parameterized finite element model as described in the following section. Once the SVM model is obtained, a probabilistic SVM (PSVM)[6,?] is constructed to obtain the probability of hip fracture.

IV.A. Fully parameterized finite element model of a femur

A fully parameterized finite element model of a femur is constructed in ANSYS using ANSYS Parametric Design Language (APDL).[25] The model parameters are listed in Table 1 and depicted in Figure 4 as well as the boundary conditions.[26,27]

Table 1. Parameters implemented in the finite element model of a femur.

Geometric Parameters		
Region	Name	Parameter
Neck	Outer diameter	*d_neck*
	Thickness of the cortical bone	*t_neck*
Intertrochanter	Outer width	*w_trochanter*
	Thickness of the cortical bone	*t_trochanter*
Shaft	Outer diameter	*d_shaft*
	Thickness of the cortical bone	*t_shaft*
	Neck-shaft angle	*alpha*
Other Parameters		
	Weight	*weight*
	Young's modulus	E_{Cortical}
	Poisson ratio	ν

IV.B. Sample generation and failure criteria

The data used in the experiments are obtained by sampling three variables: the Young's modulus of the cortical bone (E_{Cortical}), the thickness of cortical bone around the neck (T_{Neck}), and the weight of the individual. Each variable follows a normal distribution with means and standard deviations provided in Table 2. A total of 2000 samples were drawn and evaluated through the finite element model from which stress and strain information can be obtained. The samples generated in the three dimensional space where the data is separable are then projected onto a two dimensional space (E_{Cortical}, T_{Neck}) leading to non separable cases.

In this work, failure (i.e., fracture) is assessed using the maximum principal strain.[28] Other measures could be used,[29] however the choice of the measure does not remove any of the generality of the conclustions of the article.

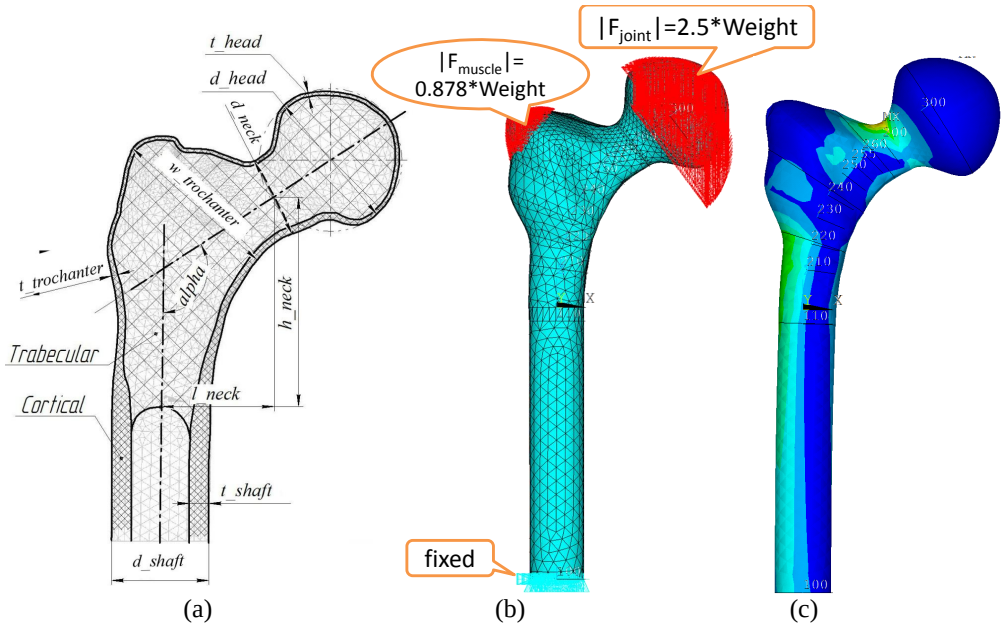


Figure 4. Fully parameterized finite element model of a femur: (a) Parameters. (b) Boundary condition. (c) Contour of principal strain.(Max principal strain is around the neck.)

Table 2. Distribution of 3 parameters for sample generation

Parameter	Distribution
E_{Cortical}	$N(17.80, 2.10)$ (Gpa)
T_{Neck}	$N(1.58, 0.26)$ (mm)
$weight$	$N(63.96, 15.90)$ (kg)

The thresholds chosen for the maximum principal strains are 1.04% in compression and 0.73% in tension.[30] Based on this failure criterion, samples can be classified into safe $(+1)$ or failed (-1) classes. As shown in Figure 5, the samples are clearly separable and an SVM that classifies these samples without misclassification is also provided. Projection of the three-dimensional samples onto the plane of E_{Cortical} and T_{Neck} leads to non-separable samples as shown in Figure 5.

IV.C. Parameter selection using *AUC*, *Acc*, and *Bacc*.

Based on the two-dimensional non-separable samples, the proposed strategy with different cross-validation metrics can be used for parameter selection. Both safe and failed samples in te projected sub-space are randomly split into training and validation sets of equal size as shown in Figure 6. Parameters of the SVM are selected using the training set and performance evaluation of the constructed SVMs is carried out on the validation set.

The three cross-validation metrics are used separately for SVM parameter selection. SVMs with highest scores are shown in Figure 7 and scores form grid search of (C, γ) are shown in Figure 8. Clearly, the SVM selected using accuracy leads to "over-fitting" and will have a poor predictive capability.

Table 3 also provides the results for the weighted likelihood. A comparison to the reference likelihood obtained as described in Section III is also provided. Figure 9 depicts the convergence of the reference likelihood as a function of the number of samples.

From this point on, this article will not consider the accuracy (Acc) since it does not (and this is a well-known problem) provide good results in the case of unbalanced data.

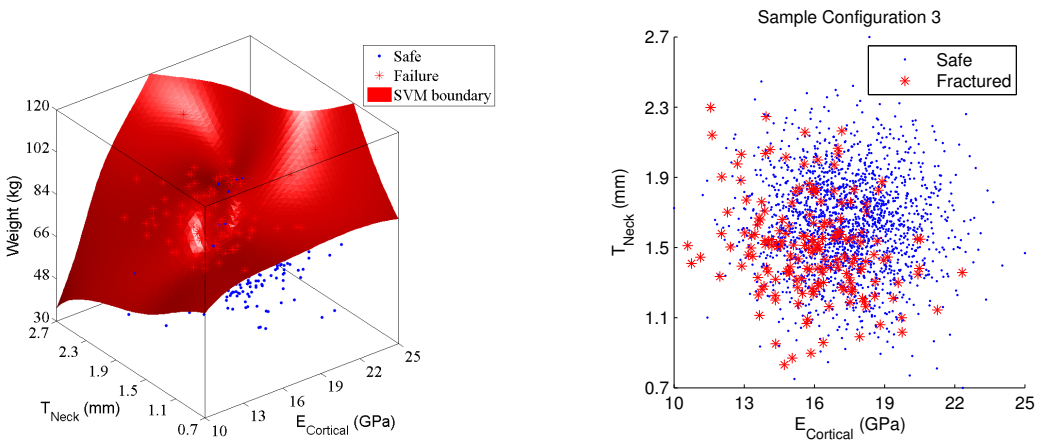
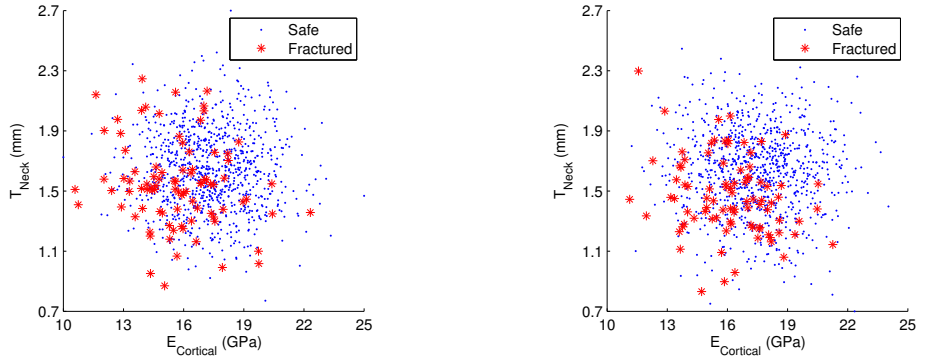
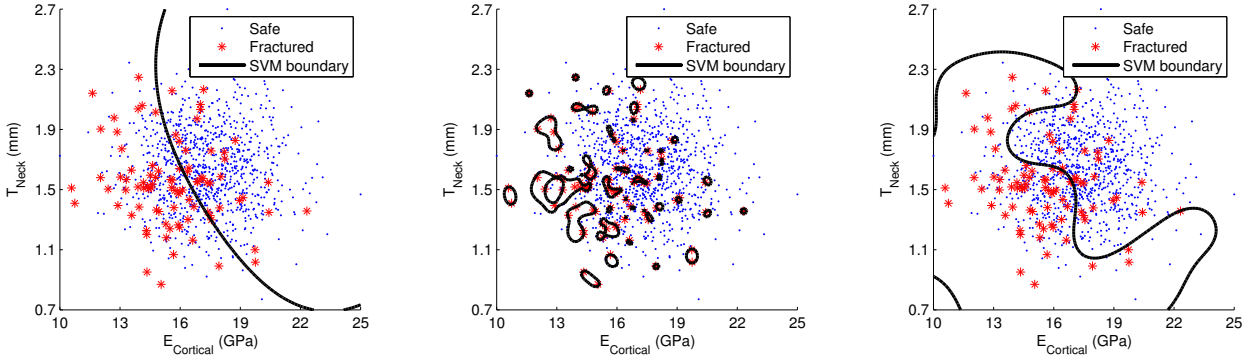


Figure 5. 3 dimensional SVM constructed using additional information of weight (left). Projection onto the ($E_{\mathbf{Cortical}}$ and $T_{\mathbf{Neck}}$) space (right).



(a) Training samples (b) Validation samples

Figure 6. Examples of training (a) and validation (b) sample sets.

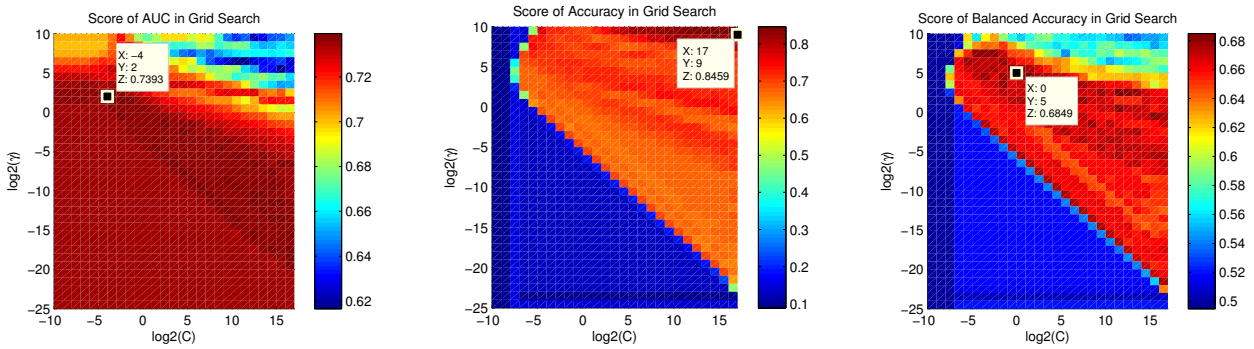


(a) SVM selected using AUC (b) SVM selected using accuracy (c) SVM selected using balanced accuracy

Figure 7. Training samples and SVMs selected based on different validation metrics.

IV.D. Influence of separability, unbalance, and training sample size

This section describes a systematic study by selecting hyper-parameters for SVM on sample sets with different levels of separability, unbalance, and size of training sample set.



(a) Score of AUC from grid search

(b) Score of accuracy from grid search

(c) Score of balanced accuracy from grid search

Figure 8. Scores from grid search using cross validation.

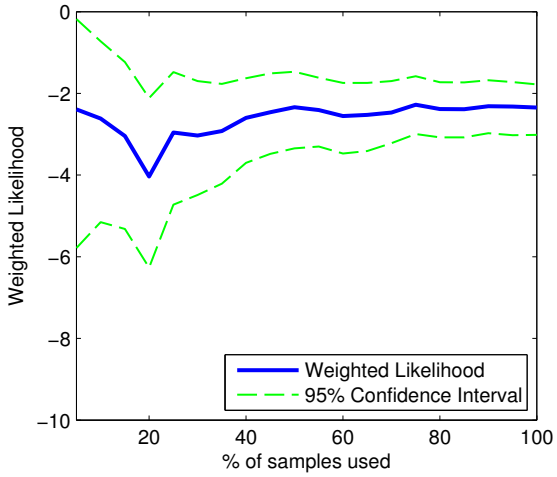


Figure 9. Evolution of $\overline{\mathcal{L}}_w$ and its 95% confidence interval for data in projected space by adding portions of samples for SVM training. Ref. $\overline{\mathcal{L}}_w$ is the value at convergence.

Table 3. Scores and 95% confidence intervals of SVMs selected using different criteria.

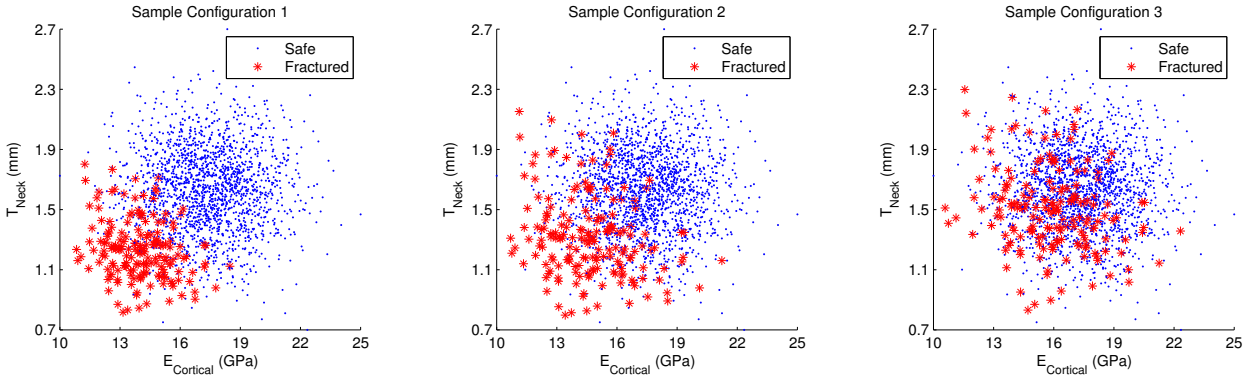
SVM selected using:	Score [95% CI]			(Ref.$\overline{\mathcal{L}}_w$=-2.35 [-3.02, -1.78])	Difference from Ref. $\overline{\mathcal{L}}_w$*
	AUC	*Acc*	*Bacc*	$\overline{\mathcal{L}}_w$ [95% CI]	
AUC	0.73 [0.68 0.78]	0.68 [0.65 0.71]	0.65 [0.60 0.71]	-3.02 [-4.03 -2.06]	28.78%
Acc	0.96 [0.94 0.97]	0.94 [0.92 0.95]	0.77 [0.72 0.82]	-8.98 [-11.87 -6.06]	353.40%
Bacc	0.67 [0.61 0.73]	0.70 [0.67 0.73]	0.63 [0.57 0.69]	-3.90 [-5.11 -2.77]	66.08%

IV.D.1. Level of separability

In order to create more sample configurations with different levels of separability, the authors used an isoprobabilist transformation between Normal and Weibull distribution (Equation 12) to control the spread of fractured samples in sub-space of E_{Cortical} and T_{Neck}. Including the original distribution (Figure 5), two other level of separability are introduced Figure 10 (a) and (b). These three levels of separability will be referred to as configuration 1,2, and 3. And Configuration 3 contains samples without transformation as shown in the previous section.

$$F(\mathbf{x_i}|a,b) = \Phi(\frac{\mathbf{x_i} - \mu_i}{\sigma_i}), i = 1,2 \tag{12}$$

where $F(\mathbf{x_i}|a,b)$ is the cumulative distribution function of Weibull distribution with parameter a and b. Φ is the cumulative distribution function of standard normal distribution. μ_i and σ_i are the empirical mean and standard deviation of variables E_{Cortical} and T_{Neck} separately.



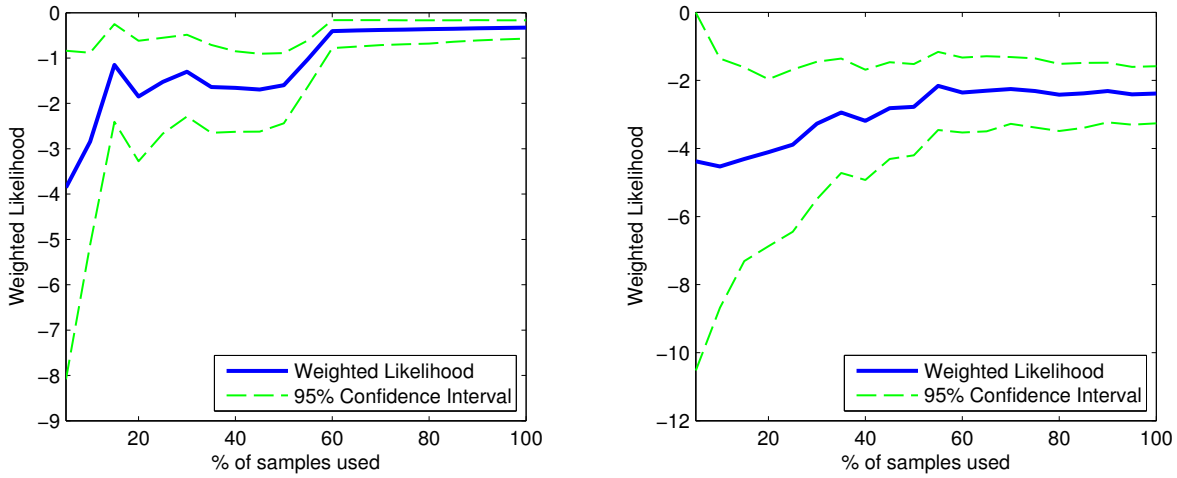
(a) Configuration 1 (b) Configuration 2 (c) Configuration 3

Figure 10. Different sample configurations used for parameter selection.

As shown in Figure 10, 3 sample configurations are generated for validation metrics comparison and the other 2 configurations used in this section. Results from accuracy are not shown in the following studies since it is not a good cross-validation metric, which was demonstrated in the previous section.

Figure 11 shows the evolution of weighted likelihood as well as its 95% confidence interval for both sample configuration 1 and 2. Reference values of weighted likelihood for sample configuration 1 & 2 as well as their 95% confidence intervals are given in Table 4 and 5.

Results on sample configuration 1 and 2 are listed in Table 4 and 5 separately.



(a) Configuration 1. (b) Configuration 2.

Figure 11. Evolution of $\overline{\mathcal{L}}_w$ and its 95% confidence interval for Configuration 1 & 2 by adding portions of samples for SVM training.

Table 4. Scores and 95% confidence intervals of SVMs selected using different criteria on Configuration 1.

SVM selected using:	Score [95% CI] *AUC*	*Acc*	*Bacc*	(Ref. $\overline{\mathcal{L}}_w$=-0.33 [-0.57, -0.17]) $\overline{\mathcal{L}}_w$ [95% CI]	Difference from Ref. $\overline{\mathcal{L}}_w{}^*$(%)
AUC	0.97 [0.96 0.98]	0.89 [0.87 0.91]	0.93 [0.91 0.95]	-0.35 [-0.69 -0.16]	6.05%
Bacc	0.97 [0.96 0.98]	0.83 [0.81 0.85]	0.90 [0.88 0.92]	-0.72 [-1.50 -0.20]	118.67%

Table 5. Scores and 95% confidence intervals of SVMs selected from different criteria for Configuration 2.

SVM selected using:	Score [95% CI] *AUC*	*Acc*	*Bacc*	(Ref.$\overline{\mathcal{L}}_w$=-2.39 [-3.26, -1.58]) $\overline{\mathcal{L}}_w$ [95% CI]	Difference from Ref. $\overline{\mathcal{L}}_w{}^*$(%)
AUC	0.92 [0.88 0.94]	0.82 [0.79 0.84]	0.83 [0.79 0.87]	-2.49 [-3.78 -1.31]	4.21%
Bacc	0.91 [0.88 0.94]	0.84 [0.82 0.86]	0.84 [0.80 0.88]	-2.92 [-4.51 -1.54]	22.33%

Figure 12 shows that as the samples becomes more separable, weighted likelihood from the SVM selected based on AUC is closer to the reference value and its 95% confidence interval is smaller than the SVM selected based on balanced accuracy.

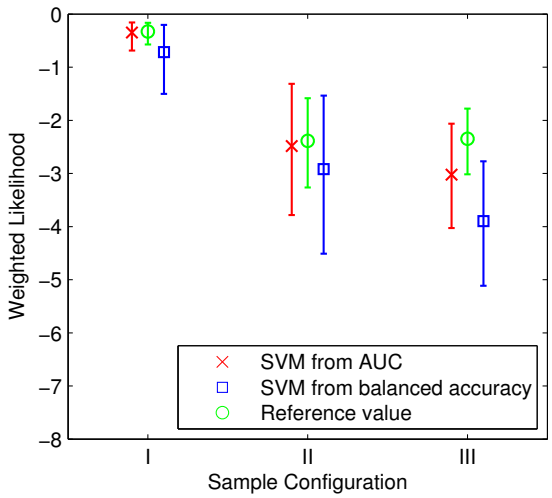
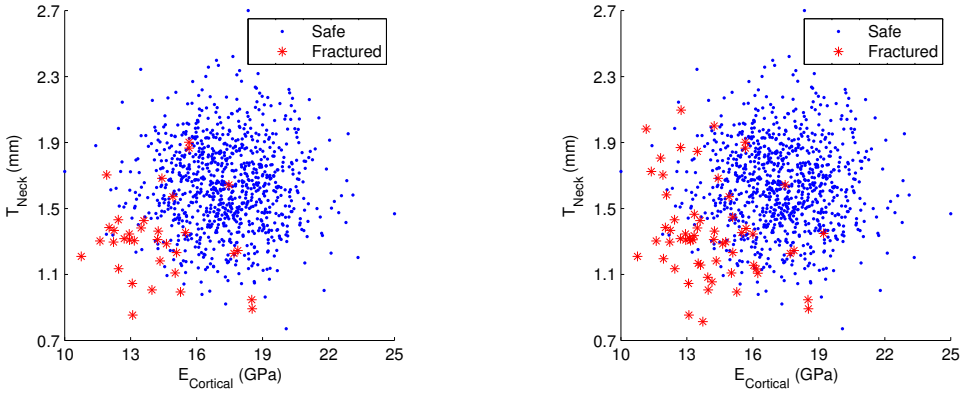


Figure 12. Comparison of weighted likelihood between different sample configurations.

IV.D.2. Level of unbalance

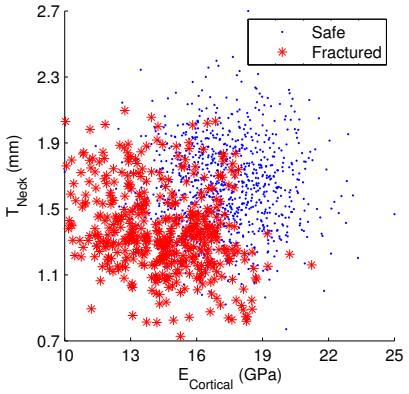
This section studies the change of weighted likelihood and its 95% confidence interval with different levels of unbalance by varying ratios between safe(+1) and failed(−1) classes.

Samples of different ratios between safe and failed classes are given in Figure 13.

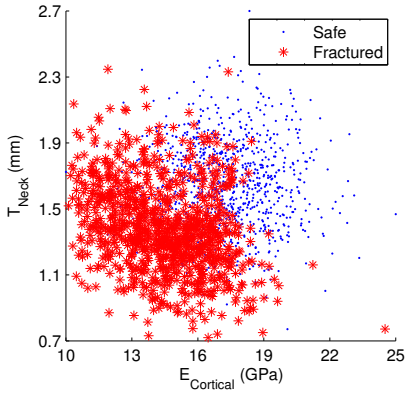


(a) Case 1 (34 failed samples)

(b) Case 2 (60 failed samples)



(c) Case 3 (455 failed samples)



(d) Case 4 (911 failed samples)

Figure 13. For sample configuration 2, study the influence of level of unbalance by changing the ratio between safe and failed samples.

Table 6. Performances of SVMs selected from different criteria on validation set for 4 cases with various level of unbalance.

	SVM selected using:	Score [95% CI]			Difference from Ref. $\overline{\mathcal{L}}_w$ (%) *
		AUC	*Bacc*	$\overline{\mathcal{L}}_w$	
Case 1	*AUC*	0.90 [0.87 0.93]	0.84 [0.77 0.90]	-3.40 [-6.17 -1.14]	16.80%
	Bacc	0.90 [0.87 0.94]	0.80 [0.72 0.87]	-5.30 [-8.92 -2.30]	82.20%
Case 2	*AUC*	0.90 [0.86 0.94]	0.81 [0.75 0.86]	-3.95 [-6.08 -2.16]	55.33%
	Bacc	0.90 [0.86 0.94]	0.83 [0.78 0.88]	-4.06 [-6.46 -1.92]	59.76%
Case 3	*AUC*	0.90 [0.89 0.92]	0.83 [0.81 0.85]	-1.75 [-2.12 -1.37]	20.76%
	Bacc	0.90 [0.89 0.92]	0.83 [0.80 0.85]	-1.78 [-2.15 -1.45]	23.28%
Case 4	*AUC*	0.91 [0.90 0.92]	0.8 [0.82 0.85]	-1.36 [-1.58 -1.15]	10.31%
	Bacc	0.89 [0.87 0.91]	0.84 [0.82 0.86]	-1.38 [-1.60 -1.15]	11.52%

[1] Reference $\overline{\mathcal{L}}_w$ for Case 1 is -2.91 [-4.63, -1.39].
[2] Reference $\overline{\mathcal{L}}_w$ for Case 2 is -2.54 [-3.79, -1.50].
[3] Reference $\overline{\mathcal{L}}_w$ for Case 3 is -1.45 [-1.67, -1.24].
[4] Reference $\overline{\mathcal{L}}_w$ for Case 4 is -1.23 [-1.36, -1.12].

Scores and relative differences of weighted likelihood to the reference value for SVMs selected based on different validation metrics are listed in Table 6.

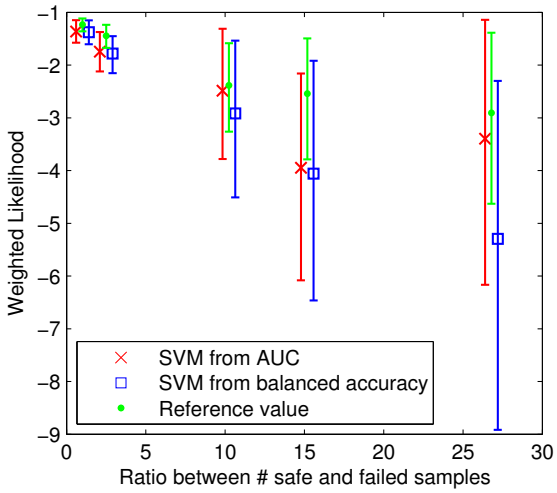


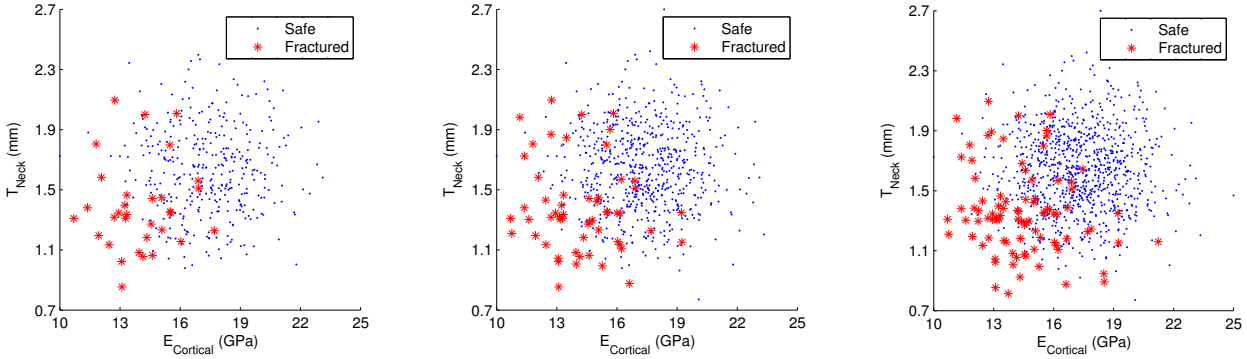
Figure 14. Weighted likelihood and its 95% confidence interval for different levels of unbalance.

Figure 14 shows that as the ratio between safe and failed classes grows larger, the 95% confidence interval of weighted likelihood becomes wider. And compared with SVMs selected based on balanced accuracy, the ones selected from AUC give closer weighted likelihood to the reference values and smaller confidence intervals if the level of unbalance is the same.

IV.D.3. Number of samples

This section studies the influence of number of samples used in training of SVM by using portions of samples in the training set while keeping unbalance between classes at the same level.

Three cases with different sizes of training samples are created as shown in Figure 15. Case 1 uses 40% of training samples, Case 2 uses 70% of training samples and Case 3 contains all samples available in the training set.



(a) Case 1 with 396 samples. (b) Case 2 with 693 samples. (c) Case 3 with 1000 samples.

Figure 15. Different sizes of training samples used for parameter selection.

Results of weighted likelihood and relative difference from the reference value are provided in Table 7. As the size of training samples grow larger, as shown in Figure 16 and Table 7, SVMs selected based on AUC have closer values of weighted likelihood to the reference value and smaller 95% confidence intervals compared with SVMs selected from balanced accuracy.

Figure 16 shows weighted likelihood increases if more samples are used in training the SVM. SVMs

Table 7. Performances of SVMs selected using different metrics on validation set for 3 cases with various sizes in training set.

	SVM selected using:	Score [95% CI] *AUC*	Score [95% CI] *Bacc*	(Ref.$\overline{\mathcal{L}}_w$=-2.39 [-3.26, -1.58]) $\overline{\mathcal{L}}_w$	Difference from Ref. $\overline{\mathcal{L}}_w{}^*$(%)
Case 1	*AUC*	0.91 [0.88 0.93]	0.83 [0.78 0.87]	-3.55 [-5.30 -2.02]	48.89%
	Bacc	0.84 [0.78 0.89]	0.75 [0.70 0.80]	-3.71 [-5.68 -2.00]	55.60%
Case 2	*AUC*	0.91 [0.88 0.94]	0.83 [0.79 0.87]	-3.26 [-4.91 -1.83]	36.75%
	Bacc	0.85 [0.80 0.90]	0.76 [0.71 0.81]	-3.51 [-5.37 -1.93]	47.35%
Case 3	*AUC*	0.91 [0.88 0.94]	0.84 [0.80 0.88]	-2.49 [-3.86 -1.31]	4.35%
	Bacc	0.91 [0.88 0.94]	0.83 [0.79 0.87]	-2.86 [-4.37 -1.53]	20.06%

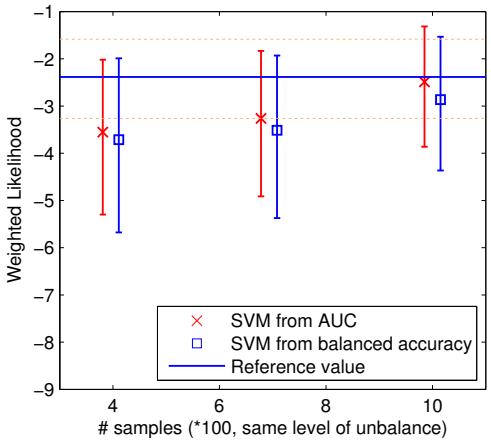


Figure 16. Weighted likelihood and its 95% confidence interval for different sizes of training samples.

selected based on AUC provide closer weighted likelihood to the reference value and smaller 95% confidence intervals than the ones based on balanced accuracy.

IV.E. Hip fracture risk assessment based on PSVM

When hyper-parameters of an SVM are selected, a PSVM-based hip fracture risk model can be constructed based on the predetermined SVM. The PSVM[6,7] model proposed by Platt approximates the probability of belonging to one class for any sample $\mathbf{x}_i$ by:

$$P(y_i = +1|\mathbf{x_i}) \approx P_{A,B}(s(\mathbf{x}_i)) \equiv \frac{1}{1 + \exp(A \cdot s(\mathbf{x}_i) + B)} \tag{13}$$

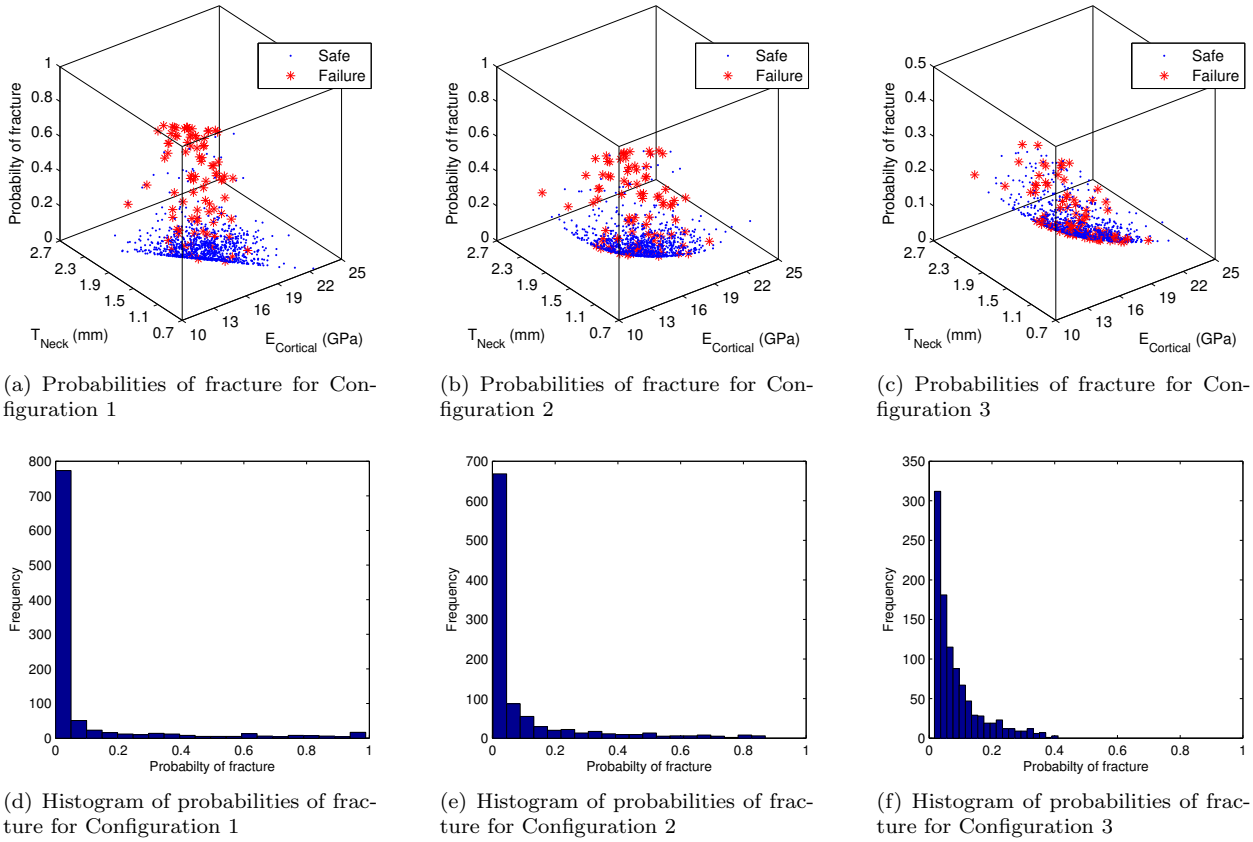
where $P(y_i = +1|\mathbf{x}_i)$ is the probability of being into the +1 class for $\mathbf{x}_i$, $s(\mathbf{x}_i)$ is the SVM output value as shown in Equation 3. It is worth mentioning that using $s(\mathbf{x})$ to calculate AUC will give the same results as using $P(y = +1|\mathbf{x})$, which is a monotonic function of $s(\mathbf{x})$. Parameters A and B are determined by solving the following maximum likelihood problem.

$$\min_{A,B} \; -\sum_{i=1}^{N} (t_i \log(P(y_i = +1|\mathbf{x}_i) + (1 - t_i) \log(1 - P(y_i = +1|\mathbf{x}_i))) \tag{14}$$

$$t_i = \begin{cases} 1, & \text{if } y_i = +1 \\ 0, & \text{if } y_i = -1 \end{cases}, i = 1, \ldots, N$$

where N is the number of samples and y_i is the actual class of sample $\mathbf{x}_i$.

For the 3 sample configurations in Figure 10, PSVM models are built based on the corresponding constructed SVMs. The probabilities of suffering from femur fracture and histograms of the probability are shown in Figure 17.



(a) Probabilities of fracture for Configuration 1

(b) Probabilities of fracture for Configuration 2

(c) Probabilities of fracture for Configuration 3

(d) Histogram of probabilities of fracture for Configuration 1

(e) Histogram of probabilities of fracture for Configuration 2

(f) Histogram of probabilities of fracture for Configuration 3

Figure 17. Probabilities of fracture for 3 sample configurations from PSVM-based hip risk model.

From the histograms shown in Figure 17, the distribution and range of the fracture probabilities vary with the level of separability. As level of separability increases, the hip risk model has more confidence in predicting fractured samples with $P(y_i = -1|\mathbf{x_i}) > 0.8$ for those samples. In the most non-separable case (Configuration 3), the hip risk model is less confident since $P(y_i = -1|\mathbf{x_i}) < 0.5$ for even fractured samples.

V. Acknowledgement

The authors are thankful to Mykola M. Khrystenko for helping in the creation of the fully parameterized finite element model of a femur. The support of NIH through grant NIAMS 1R21AR060811-01 is gratefully acknowledged.

VI. Conclusion

This article compared three commonly used validation metrics for the selection of optimal SVM parameters in the case of non-separable and unbalanced data. A systematic study with different levels of separability and levels of unbalance as well as sizes of training samples, were presented. The data sets used were created from a finite element model for the prediction of hip fracture. The results show the advantage of the AUC metric, mostly for case with large degrees of unbalance and non-separability. The next steps of this study will involve higher dimensional problems along with the use of clinical data.

References

[1]Metz, C. E., "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, Vol. 8, No. 4, 1978, pp. 283 – 298.

[2]Fawcett, T., "An introduction to ROC analysis," *Pattern recognition letters*, Vol. 27, No. 8, 2006, pp. 861–874.

[3]Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M., "The balanced accuracy and its posterior distribution," *Pattern Recognition (ICPR), 2010 20th International Conference on*, IEEE, 2010, pp. 3121–3124.

[4]Rijsbergen, C. J. V., *Information Retrieval*, Butterworth-Heinemann, Newton, MA, USA, 2nd ed., 1979.

[5]Matthews, B. W. et al., "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." *Biochimica et biophysica acta*, Vol. 405, No. 2, 1975, pp. 442.

[6]Platt, J. et al., "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, Vol. 10, No. 3, 1999, pp. 61–74.

[7]Basudhar, A. and Missoum, S., "Reliability Assessment using Probabilistic Support Vector Machines (PSVMs)," *International Journal of Reliability and Safety*, To appear in 2013.

[8]Burges, C. J., "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, Vol. 2, No. 2, 1998, pp. 121–167.

[9]Cristianini, N. and Shawe-Taylor, J., *An introduction to support vector machines and other kernel-based learning methods*, Cambridge university press, 2000.

[10]Basudhar, A., Missoum, S., and Harrison Sanchez, A., "Limit state function identification using Support Vector Machines for discontinuous responses and disjoint failure domains," *Probabilistic Engineering Mechanics*, Vol. 23, No. 1, 2008, pp. 1–11.

[11]Yang, Z. R., *Machine learning approaches to bioinformatics*, Vol. 4, World Scientific Publishing Company Incorporated, 2010.

[12]Tay, F. E. and Cao, L., "Application of support vector machines in financial time series forecasting," *Omega*, Vol. 29, No. 4, 2001, pp. 309–317.

[13]Basudhar, A. and Missoum, S., "An improved adaptive sampling scheme for the construction of explicit boundaries," *Structural and Multidisciplinary Optimization*, Vol. 42, No. 4, 2010, pp. 517–529.

[14]Konig, I., Malley, J., Pajevic, S., Weimar, C., Diener, H., and Ziegler, A., "Tutorial in Biostatistics Patient-centered prognosis using learning machines," 2005.

[15]Jack, L. and Nandi, A., "Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms," *Mechanical systems and signal processing*, Vol. 16, No. 2, 2002, pp. 373–390.

[16]Osuna, E., Freund, R., and Girosi, F., "Support vector machines: Training and applications," 1997.

[17]Vapnik, V., *The nature of statistical learning theory*, springer, 1999.

[18]Chang, C.-C. and Lin, C.-J., "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, 2011, pp. 1–27.

[19]McLachlan, G. J., Do, K.-A., and Ambroise, C., *Analyzing microarray gene expression data*, Vol. 422, Wiley-Interscience, 2004.

[20]Kohavi, R. et al., "A study of cross-validation and bootstrap for accuracy estimation and model selection," *International joint Conference on artificial intelligence*, Vol. 14, Lawrence Erlbaum Associates Ltd, 1995, pp. 1137–1145.

[21]Swets, J. A., *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers.*, Lawrence Erlbaum Associates, Inc, 1996.

[22]Goldberg, D. E. and Holland, J. H., "Genetic algorithms and machine learning," *Machine Learning*, Vol. 3, No. 2, 1988, pp. 95–99.

[23]Efron, B. and Tibshirani, R., "Improvements on cross-validation: the 632+ bootstrap method," *Journal of the American Statistical Association*, Vol. 92, No. 438, 1997, pp. 548–560.

[24]Varian, H., "Bootstrap tutorial," *Mathematica Journal*, Vol. 9, No. 4, 2005, pp. 768–775.

[25]ANSYS, A., "Programmer's Guide. ANSYS," *Inc. March*, 2002.

[26]Hölzer, A., Schröder, C., Woiczinski, M., Sadoghi, P., Scharpf, A., Heimkes, B., and Jansson, V., "Subject-specific finite element simulation of the human femur considering inhomogeneous material properties: A straightforward method and convergence study," *Computer methods and programs in biomedicine*, 2012.

[27]Bergmann, G., Deuretzbacher, G., Heller, M., Graichen, F., Rohlmann, A., Strauss, J., and Duda, G., "Hip contact forces and gait patterns from routine activities," *Journal of Biomechanics*, Vol. 34, No. 7, 2001, pp. 859 – 871.

[28]Bayraktar, H., Morgan, E., Niebur, G., Morris, G., Wong, E., and Keaveny, T., "Comparison of the elastic and yield properties of human femoral trabecular and cortical bone tissue," *Journal of biomechanics*, Vol. 37, No. 1, 2004, pp. 27–35.

[29]Doblaré, M., García, J., et al., "On the modelling bone tissue fracture and healing of the bone tissue," *Acta Científica Venezolana*, Vol. 54, No. 1, 2003, pp. 58–75.

[30]Grassi, L., Schileo, E., Taddei, F., Zani, L., Juszczyk, M., Cristofolini, L., and Viceconti, M., "Accuracy of finite element predictions in sideways load configurations for the proximal human femur," *Journal of Biomechanics*, Vol. 45, No. 2, 2012, pp. 394 – 399.