Support Vector Feature Selection for Early Detection of Anastomosis Leakage from Bag-of-Words in Electronic Health Records

Cristina Soguero-Ruiz *Member, IEEE*, Kristian Hindberg, José Luis Rojo-Álvarez, *Senior Member, IEEE*, Stein Olav Skrøvseth, Fred Godtliebsen, Kim Mortensen, Arthur Revhaug, Rolv-Ole Lindsetmo, Knut Magne Augestad, Robert Jenssen *Member, IEEE*

***Abstract*—The free text in Electronic Health Records (EHRs) conveys a huge amount of clinical information about health state and patient history. Despite a rapidly growing literature on the use of machine learning techniques for extracting this information, little effort has been invested towards feature selection and the features' corresponding medical interpretation. In this work, we focus on the task of early detection of anastomosis leakage (AL), a severe complication after elective surgery for colorectal cancer (CRC) surgery, using free text extracted from EHRs. We use a Bag-of-Words model to investigate the potential for feature selection strategies. The purpose is earlier detection of AL and prediction of AL with data generated in the EHR before the actual complication occur. Due to the high dimensionality of the data, we derive feature selection strategies using the robust support vector machine linear maximum margin classifier, by investigating: (a) a simple statistical criterion (leave-one-out based test); (b) an intensive-computation statistical criterion (Bootstrap resampling); and (c) an advanced statistical criterion (kernel entropy). Results reveal a discriminatory power for early detection of complications after CRC (sensitivity 100%; specificity 72%). These results can be used to develop prediction models, based on EHR data, that can support surgeons and patients in the preoperative decision making phase.**

***Index Terms*—Electronic Health Record; Support Vector Machine; Feature Selection; Kernel Entropy; Bag-of-Words; Anastomosis Leakage; Colorectal Cancer; Early Detection.**

CSR and JLRA are with Dept. Teoría de la Señal y Comunicaciones, Universidad Rey Juan Carlos (URJC). 28943 Fuenlabrada, Madrid, Spain (email: {cristina.soguero,joseluis.rojo}@urjc.es). JLRA is Prometeo Researcher with Electric and Electronic Department, Universidad de las Fuerzas Armadas ESPE, Ecuador. KH and FG are with Dept. Mathematics and Statistics, University of Tromsø(UiT), Tromsø, Norway (email: {kristian.hindberg,fred.godtliebsen}@uit.no). SOS, KMA and RJ are with Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway (UNN), Norway (email: {stein.olav.skrovseth,Knut.Magne.Augestad}@telemed.no). SOS is with IBM T.J. Watson Research Center, Yorktown Heights, New York, USA. KM, AR, ROL and KMA are with Dept. of Gastrointestinal Surgery, UNN; Tromsø, Norway (email: {Kim.Erlend.Mortensen,arthur.revhaug,Rolv-Ole.Lindsetmo@unn.no}). KM, ROL and KMA are with Institute of Clinical Medicine, UiT, Tromsø, Norway. AR is with the Clinic for Surgery, Cancer and Women's Health, UNN, Tromsø, Norway. KMA is with Dept. of Colorectal Surgery, University Hospitals Case Medical Center, Cleveland, Ohio, USA. RJ is with Dept. of Physics and Technology, UiT, Tromsø, Norway (email: {robert.jenssen}@uit.no).

This work was partially supported by projects TEC2010-19263-TCM and TEC2013-48439-C4-1-R from Spanish Government, by the Prometeo Project of the Secretariat for Higher Education, Science, Technology and Innovation of the Republic of Ecuador and was performed as part of Tromsø Telemedicine Laboratory, funded by the Research Council of Norway grant no. 174934. CSR is supported by FPU grant AP2012-4225 from Spanish Government.

I. Introduction

Electronic Health Records (EHRs) are collections of health information in digital storage format, which can in theory be shared among systems to convey the relevant information of a patient [1]. EHRs have three levels of medical understanding, namely, data storage, information, and knowledge [2]. While technology seems to have successfully covered the data storage level, the others are currently intensive research tracks. A considerable amount of literature exists on extraction of knowledge from EHRs to support clinical decision making (see [3] and references therein). Specifically, analysis of the (unstructured) EHR free text may potentially extract a large amounts of information regarding patient health status and medical history, which may not be fully available in the structured data that are also available in EHRs [4], [5].

Machine learning methods have recently demonstrated great potential at free text analysis for decision support and medical information retrieval. Several such methods are based on the simple, but often powerful, Bag-of-Words (BoW) model. Wright et al. [5] used this model to identify relevant documents in EHRs pertaining to a user's query on progress notes in diabetes, and in [6], a system for automatic case identification was proposed for observational epidemiological studies. Using various levels of sophistication in the BoW model, the authors in [4] developed a framework for general-purpose automatic diagnosis in traditional Chinese medicine. Furthermore, the authors in [7] derived a semi-supervised Support Vector Machine (SVM), for automated identification of primary care records from the General Practice Research Database, with applications to retrieval of coronary angiogram and ovarian cancer diagnoses, and in [8] a comprehensive bag-of-concepts system was proposed for quantifying a patient's risk of mortality and complications. The interested reader can also see reference [9] for a recent review of natural language processing techniques for analysis of free text in EHRs, in addition to [10] for a review on extracting information from textual documents in EHRs, including the advances in the field from 1995 to 2008.

However, few studies have explored systematic *Feature Selection (FS) criteria* for machine learning-based applications using EHR data, or principled knowledge extraction from the machine learning engines. We focus on early detection of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361688, IEEE Journal of Biomedical and Health Informatics

Anastomosis Leakage (AL) using a BoW model extracted from an EHR. AL is one of the most common complications after colorectal cancer (CRC) surgery. CRC is the third most common cancer type and surgery is the only curative treatment [11]. However, standard elective colorectal resection is usually associated with a complication rate of 20-30%, which again have severe implications for the individual patient [12]. Indeed, AL is reported to occur in 5-15% of all patients who underwent colorectal cancer surgery and it is recognized as an important quality indicator in colorectal cancer surgery [13]. AL may be a lethal condition, therefore its early detection is vital [13], [14]. Authors in [15] showed that the risk of AL determined by surgeons' risk assessment appeared to have low predictive value. Early diagnosis and intervention can minimize systemic complications, but is hindered by current diagnostic methods that are non-specific and often uninformative [14]. A Colon Leakage Score was developed in [16] to predict the risk of AL based on information from the literature and experts opinions. Results showed that this score is a good predictor for AL, however, novel methods to identify and detect this complication at an early stage using EHR data are needed.

In this work, we propose several novel FS strategies that are capable of automatically identifying the relevant words, while permitting easy knowledge extraction from the system. This work was based upon a patient database (QUAKE, quality control of surgical performance with unstructured EHR data) which was extracted from the Department of Gastrointestinal Surgery at the University Hospital of North Norway.

A vast general literature exists on FS, see Sec. II-D for examples. Of particular interest is FS based on the weights obtained by a maximum margin SVM *linear* classifier, which we pursue in this exposition. There are several reasons for this: (i) the robustness of the linear SVM in high-dimensional and noisy low sample size problems; and (ii) the one-to-one relationship between the weights of the linear classifier and the features (words), which enables interpretation of the features. The latter is a significant advantage when compared to classifiers such as Gaussian maximum likelihood or artificial neural networks [17], [18], where the direct connection to the features are lost. The previous literature on SVM-based FS is to a large degree concentrated on the Recursive Feature Elimination (RFE) method [19], which has been shown to compare very favorably to many of the classical FS methods. RFE puts a threshold on the amplitudes of the weights obtained by the SVM. Hence, the user must either pre-specify the number of features to obtain, or, alternatively, engage in a computationally demanding cross-validation procedure, whereby features are eliminated recursively, thus requiring numerous SVM re-training procedures on subsets of features of decreasing size. This may be very time consuming, even for small sample sizes.

As novel alternatives to the RFE, we propose innovative FS methodology in order to avoid numerous SVM re-training procedures. Our present work introduces statistically principled FS methods, capable of working on the linear classifier weight amplitudes in an easy way with extremely high dimensional input spaces. The proposed methods require no pre-specification of the number of features to obtain, and are based on three different criteria (see Sec. II-D for details):

a A simple statistical criterion (leave-one-out (LOO) based test);
b An intensive-computation statistical criterion (Bootstrap resampling);
c An advanced statistical criterion (kernel entropy).

After adjusting for imbalanced classes, which is a well-known challenge in medical classification applications [6], [20], the proposed FS strategies are shown to significantly improve the detection of AL. Also, the results provide useful knowledge of the relevant words (without need of their pre-selection by clinicians) and their temporal evolution.

The paper is structured as follows. Section II presents the proposed methodology for FS, including the theoretical fundamentals of three benchmarked classifiers (SVM, Fisher criterion, and Naive Bayes), as well as the statistical description of the problem in terms of plug-in estimators for the distributions of the weights and figures of merit, and statistical considerations about training . In Section III, the clinical complication (i.e., AL) and the database are described. Section IV is devoted to analyze the performance of the classifiers in combination with the FS strategies, as well as the interpretation of selected words. Finally, Section V summarizes the discussion and conclusions of the work.

II. Distribution-based FS

In this section, we first discuss performance measures in binary classification. We then review the binary SVM for classifying data $\mathbf{x}$ into one of two classes (see Sec. III for the definition of $\mathbf{x}$ when represented by the BoW model). For completeness, we also mention Fisher discriminant analysis and the Naive Bayes methods, as alternatives to the SVM, and we also discuss the issue of imbalanced classes. Finally, we derive in detail the three proposed FS strategies.

A. Performance Measures

Performance measures in classification problems are essential to evaluate the quality of learned methods, as well as for free parameter tuning. Performance measures in binary classification problem may be constructed based on the confusion matrix (CM), as follows,

		Real diagnosis	
		Positive	Negative
Predictive diagnosis	Positive	TP	FP
	Negative	FN	TN

where TP and TN denote true positives and true negatives, and FP and FN denote false positives and false negatives, respectively [21]. The performance measures we consider are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361688, IEEE Journal of Biomedical and Health Informatics

$$P_e = \frac{FP+FN}{TP+FN+FP+TN} \quad (1)$$

$$Se = \frac{TP}{TP+FN} \quad (2)$$

$$Sp = \frac{TN}{TN+FP} \quad (3)$$

$$BER = \frac{1}{2}(Se+Sp) \quad (4)$$

where P_e is the error probability, Se and Sp are sensitivity and specificity, and BER denotes balanced error rate.

B. The SVM Linear Classifier

The data model for a general linear classifier is given by $y = \langle \mathbf{x}, \mathbf{w} \rangle + b$, where $\mathbf{x}$ is the input (column) vector, $\mathbf{w}$ is the weight vector, b is the bias term, and y is the classification output. We focus on the linear SVM classifier (see e.g. [22], [23]), integrating in the same classification procedure regularization such that model complexity is controlled, and the minimization of an upper bound of the generalization error. These theoretical properties make the SVM an attractive approach for several medical data problems.

Denote $D = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^n$ as a labeled training data set, where $\mathbf{x}^{(i)} \in \mathbb{R}^N$ and $y^{(i)} \in \{-1, +1\}$. The SVM classification algorithm seeks the separating hyperplane with the largest margin between the two classes. The hyperplane that optimally separates the data is the one minimizing $\|\mathbf{w}\|^2$, as well as the classification losses in terms of slack variables $\xi^{(i)}$. Consider the ν-SVM, introduced by Schölkopf et al. [24], in which we have to solve

$$\min_{\mathbf{w}, \{\xi^{(i)}\}, b, \rho} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + \nu\rho + \frac{1}{n}\sum_{i=1}^n \xi^{(i)} \right\} \quad (5)$$

subject to:

$$y^{(i)}(\langle \mathbf{x}^{(i)}, \mathbf{w} \rangle + b) \geq \rho - \xi^{(i)} \qquad \forall i = 1, \ldots, n \quad (6)$$

$$\rho \geq 0, \quad \xi^{(i)} \geq 0 \qquad \forall i = 1, \ldots, n \quad (7)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. The variable ρ adds another degree of freedom to the margin, and the margin size linearly increases with ρ. The parameter $\nu \in [0, 1]$ acts as an upper bound on the fraction of margin errors, and it is also a lower bound on the fraction of support vectors.

The decision function for any test vector $\mathbf{x}_*$ is given by

$$f(\mathbf{x}_*) = sgn\left(\sum_{i=1}^n y^{(i)} \alpha^{(i)} \langle \mathbf{x}^{(i)}, \mathbf{x}_* \rangle + b \right) \quad (8)$$

where $\alpha^{(i)}$ are Lagrange multipliers corresponding to constraints in Eq. (6). The so-called support vectors (SVs) are those training samples $\mathbf{x}^{(i)}$ with corresponding Lagrange multipliers $\alpha^{(i)} \neq 0$. The bias term b is calculated by using the *unbounded* Lagrange multipliers as $b = 1/k \sum_{i=1}^k (y^{(i)} - \langle \mathbf{x}^{(i)}, \mathbf{w} \rangle)$, where k is the number of unbounded Lagrange multipliers ($0 \leqslant \alpha^{(i)} < 1$) and $\mathbf{w} = \sum_{i=1}^n y^{(i)} \alpha^{(i)} \mathbf{x}^{(i)}$. The use of the ν-SVM algorithm allows us a compact implementation of the free parameter search strategy for the linear SVM classifier.

Two additional and well-known linear classifiers can be used for benchmarking the performance of the classification engine in our system. These are briefly described next.

1) Fisher Criterion: The goal of Fisher's discriminant analysis (FDA) in the two-class problem [25] is to find an optimally discriminating linear projection $\langle \mathbf{w}, \mathbf{x} \rangle$, by simultaneously maximizing the between-class scatter and minimizing the within-class scatter on the projected output given by the cost function

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}} \quad (9)$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ denote the between class scatter matrix and the within-class scatter matrix in the original space. These are defined by $\mathbf{S}_B = \sum_{c=1}^2 n_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top$ and $\mathbf{S}_W = \sum_{c=1}^2 \sum_{i \epsilon C_c} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_c)(\mathbf{x}^{(i)} - \boldsymbol{\mu}_c)^\top$, respectively. Here, n_c is the number of samples in class C_c, with $c = 1, 2$. Furthermore, $\boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{i \epsilon C_c} \mathbf{x}^{(i)}$ and $\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}^{(i)}$. In order to classify the projected points $\langle \mathbf{w}, \mathbf{x}^{(i)} \rangle$, $i = 1, \ldots, n$, a threshold akin to the bias term b has to be determined. There is no general rule for finding this threshold, but a common choice is the average between the class-conditional means.

2) Naive Bayes: The Naive Bayes (NB) classifier [26] estimates the class-conditional probability density functions assuming conditionally independent features, i.e.,

$$p(\mathbf{x}|y = c) = \prod_{m=1}^N p(x_m|y = c) \quad (10)$$

where $\mathbf{x} = (x_1, \ldots, x_N)$ is the feature vector and $c = 1, 2$ denotes the class.

The model is called *naive* since we do not expect the features to be independent, even conditional on the class label. Despite this, classifiers based on NB have been successful in many applications, sometimes giving competitive results with respect to other more sophisticated methods [27], [28].

C. Handling Imbalanced Classes

Medical classification problems are frequently imbalanced. For example, in the binary case, the number of patients in the positive class may be substantially smaller than the number of patients in the negative class. Several previous machine learning studies have shown that balanced classes in the training data set provide improved overall classification performances (see e.g. [29] and references therein). Common strategies to balance the classes include undersampling, i.e., removing samples from the majority class, at the risk of information loss, or oversampling the minority class has also been studied, at the risk of overfitting. We employ an undersampling strategy, see Sec. IV, for details.

D. Proposed FS Strategies

FS is defined as a series of actions to choose a subset of features that are relevant, while holding or improving the learning method. The problem of FS is well known in the machine learning literature (for a review, see [19], [30], [31]). Three different types of FS are common in the classification

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361688, IEEE Journal of Biomedical and Health Informatics

literature [30]. First, filter methods select features as a preprocessing step performed independently of the classifier. Second, wrapper methods evaluate the performance of the classifier based on subsets of features. An third, embedded methods integrate feature selection and classifier performance into the training procedure of the classifier [19], [30]. Examples of previous FS methods range from feature-ranking techniques based on correlation, to sensitivity analysis [32], and to maximum margin criteria [19], [33]. FS in text documents have focused on criteria such as the document frequency, the term frequency, mutual information, information gain, odds ratio, χ^2 statistic, and term strength, to name a few [34]–[36].

We propose FS strategies based on the weights of the linear SVM, by investigating: (a) a simple statistical criterion (leave-one-out test); (b) an intensive-computation statistical criterion (Bootstrap resampling test); and (c) an advanced statistical criterion (kernel entropy based threshold), as explained below. For the m-th feature with $m = 1, \ldots, N$, a given linear classifier yields a weight w_m, whose statistical distribution can be approximated with different empirical resampling criteria, denoted as $\hat{f}_{w_m}(w_m)$. Note that in a BoW problem, each feature corresponds to a word. In the following, features are initially sorted in descending order in terms of relative frequency in the BoW.

1) LOO-based Test: The LOO cross-validation method has been shown to give an almost unbiased estimator of the generalization properties of statistical learning models [37]. The concept can be used for estimating the probability density function (pdf) for each feature m.

We create a matrix of weights $\mathbf{W}$ with n rows and N columns, where n is the number of instances and N is the number of features. Each row of $\mathbf{W}$ is a weight vector corresponding to the linear SVM solution by using LOO cross-validation. The LOO technique partitions the original data set into n subsets, one for validation and the remaining $n-1$ for training. This process is repeated n times, setting apart for evaluation each of the n subsets just once, hence yielding $\mathbf{W}$.

The estimated Confidence Interval (CI_m) is built for each w_m, which has all the LOO estimations for the m-th feature, to obtain if this feature is relevant. Then, CI_m is used to perform a hypothesis test on the m-th feature, with $H_0 : 0 \in CI_m$ (feature m is irrelevant for the model) vs alternative hypothesis $H_1 : 0 \notin CI_m$ (feature m is relevant for the model).

2) Bootstrap Resampling-based Test: Bootstrap resampling methods [38] are very useful approaches for nonparametric estimation of the distribution of statistical magnitudes. We propose a Bootstrap resampling scheme (see Fig. 1) for building a statistical test for FS, as follows. We use $\mathbf{W}$ to provide a statistical description of the noise assuming its variance is globally dependent on the weight magnitude, and locally constant for weights with similar magnitude. Under these conditions, for each feature m with $m = 1, \ldots, N$, a local window of δ radius encompassing the 2δ nearest features is considered to build the set of weights given by $R_m = \{w^{(i)}_{m-\delta}, \ldots, w^{(i)}_{m-1}, w^{(i)}_{m+1}, \ldots, w^{(i)}_{m+\delta}\}^n_{i=1}$. Hence, R_m represents a noisy set of weights, with low (still non-null) probability of including representative weights. Third, For each m-th feature, the set $S_m = \{w^{(i)}_m\}^n_{i=1}$ represents the weight set to be tested for significance.

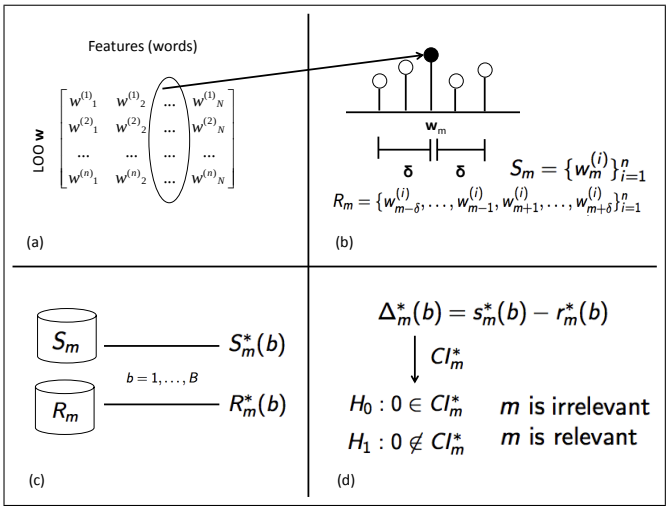


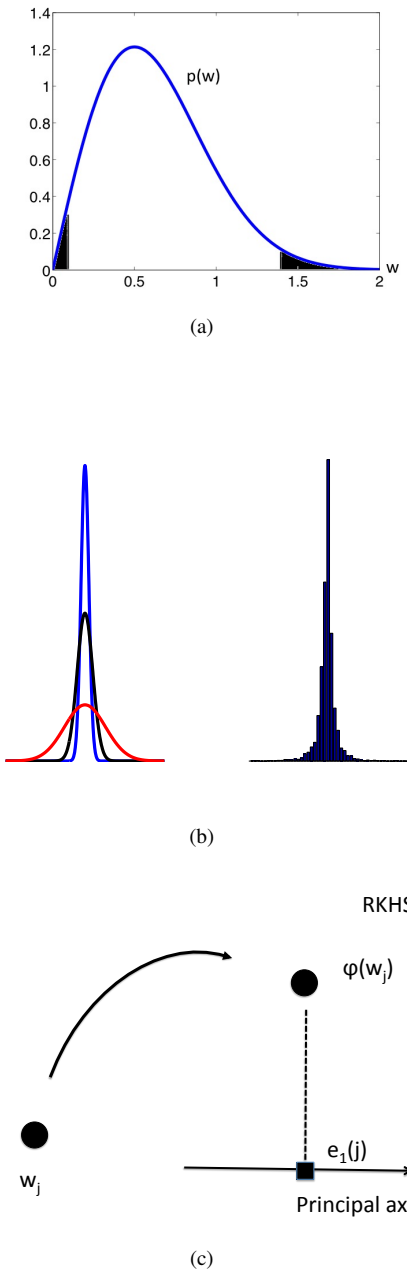
Fig. 1. Schema of the proposed Bootstrap Resampling-based Test: (a) Matrix of weights $\mathbf{W}$; (b) 2δ nearest features represents a noisy set of weights R_m, whereas S_m is the weight set to be tested for significance; (c) Bootstrap resamples; and (d) Bootstrap replication and hypothesis test.

The weight sets R_m and S_m are used to estimate the marginal distribution of noisy and potentially relevant weights for the m-th input feature, respectively, by constructing Bootstrap resamples. A *Bootstrap resample* is a new set obtained from sampling with replacement the elements of the original set (R_m and S_m in our case), providing resamples R^*_m and S^*_m, respectively. The resampling process is repeated B times, with b indexing the resampling number ($b = 1, \ldots, B$). Thus, the b-th resamples $S^*_m(b)$ and $R^*_m(b)$ contain $2\delta n$ and n elements of S_m and R_m, respectively, appearing zero, one, or several times. A *Bootstrap replication* of an estimator is constrained to the elements in the Bootstrap resample. The Bootstrap replication of the statistics of interest is $\Delta^*_m(b) = s^*_m(b) - r^*_m(b)$, where $s^*_m(b)$ and $r^*_m(b)$ are elements, randomly chosen, from $S^*_m(b)$ and $R^*_m(b)$, respectively. The B Bootstrap replications for each feature m allow us to estimate the Confidence Interval (CI^*_m) for the statistics Δ^*_m. Then, CI^*_m is used to perform a hypothesis test on the m-th feature, with $H_0 : 0 \in CI^*_m$ (feature m is irrelevant for the model) vs alternative hypothesis $H_1 : 0 \notin CI^*_m$ (feature m is relevant for the model). Note that we only sample one pair of $s^*_m(b)$ and $r^*_m(b)$ for each b, producing one $\Delta^*_m(b)$ for each b, and that the process results in a feature being found relevant if it has a large absolute value compared to the mostly noise weights that have mostly smaller absolute weights.

3) Kernel Entropy Inference Test: The basic idea behind the proposed kernel entropy inference test for feature selection, is to select those features that correspond to the high entropy part of a probability density function (pdf), describing a random variable considered to generate the features. The high entropy part of a pdf represents the most informative part, and is associated with the tails of the pdf. Fig. 2 (a) illustrates a pdf, where the sum of the areas represented by the black regions represent the tail probability.

In order to achieve the entropy-based feature selection, we concentrate on Renyi's second order entropy [39] for a random

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361688, IEEE Journal of Biomedical and Health Informatics



(a)

(b)

(c)

Fig. 2. Kernel Entropy Inference Test: (a) The tail probability refers to the sum of the areas corresponding to the black regions under the probability density function $p(w)$. (b) Illustration of the role of the bandwidth, σ, in KDE. A large bandwidth (red) provides more smoothing compared to a small bandwidth (blue). (c) KECA is related to principal components in a RKHS corresponding to the positive semi-definite kernel function used in KDE.

variable w, given by

$$H(p) = -\log V(p), \quad V(p) = \int p^2(w')dw' \qquad (11)$$

where $p(w)$ is the probability density function of w. The reason for this choice is that this measure is easily estimated using the modern technique known as kernel entropy component analysis [40] (KECA). KECA estimates the entropy using a kernel density estimator (KDE),

$$\hat{p}(w) = \frac{1}{N}\sum_{m=1}^{N} k_\sigma(w, w_m) \qquad (12)$$

Here, w_m, $m = 1, \ldots, N$, are realizations of $\mathbf{w}$ and the kernel function provides a smoothing of the histogram, where the bandwidth parameter σ governs the amount of smoothing. A common choice of kernel, which we also pursue in this paper, is $k_\sigma(w, w_m) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(w-w_m)^2}$. Figure 2(b) illustrates the role of σ. A relatively big σ will tend to produce a too smooth density estimate and vice versa. Note that w is in this approach considered a one-dimensional random variable, and in that case reliable data-driven (automated) procedures exist for the selection of σ, meaning that a different σ is computed for different samples (data sets), see Section IV-B for details. Furthermore, in the current exposition, the elements in the SVM weight vector $\mathbf{w}$ represent the samples w_m of the random variable w. Based on one particular such $\mathbf{w}$, the left panel in Fig. 2 (b) (best viewed in color) shows the KDE based on an automated bandwidth selection procedure (blue), corresponding to the most narrow function shape. The broadest function (red) shows a Gaussian best fit. The right panel shows the histogram for $\mathbf{w}$ indicating that the KDE performs better than the Gaussian model. In addition, the middle function (black) shows a kernel density estimate where we have manually doubled the selected σ. Note how the function becomes more smooth, in this case deviating more from the peaky shape.

When inserting Eq. (12) into Eq. (11), the KECA estimator for the Renyi entropy becomes $\hat{V}(p) = \frac{1}{N^2}\sum_{m=1}^{N}\left[\sqrt{\lambda_m}\mathbf{e}_m^\top \mathbf{1}\right]^2$. Here, λ_m and $\mathbf{e}_m$ are eigenvalues and eigenvectors of the so-called kernel matrix $\mathbf{K}$ where $K_{t,j} = k_\sigma(w_t, w_j)$ and $\mathbf{1}$ is a vector of ones. We have experienced robust estimates of $V(p)$ using only the top component (eigenvalue), such that in our case $\hat{V}(p) = \left[\sqrt{\lambda_1}\mathbf{e}_1^\top \mathbf{1}\right]^2$ (leaving out eigenvectors may be considered a de-noising process).

There is a one-to-one relationship between the elements in the vector $\mathbf{e}_1$ and the features stored in the SVM vector $\mathbf{w}$, and we use this in the feature selection. Since the kernel function is positive semidefinite, it computes an inner-product in a reproducing kernel Hilbert space (RKHS) [41]. That is, $w \mapsto \phi(w)$ such that the RKHS inner-product is $k_\sigma(w_t, w_j) = \langle \phi(w_t), \phi(w_j) \rangle$. It is furthermore known, that in RKHS, the projection of the jth point $\phi(w_j)$ equals $\mathbf{e}_1(j)$, i.e. the jth element of the eigenvector $\mathbf{e}_1$. This is the RKHS principal component corresponding to $\phi(w_j)$. Hence, the feature w_j corresponds to the element $\mathbf{e}_1(j)$ for $j = 1, \ldots, N$. This is illustrated in Fig. 2 (c).

The kernel entropy FS idea is the following. The tails of $p(w)$ contribute the most to the entropy of the random variable w and the features corresponding to the tail are represented by the smallest principal components in the RKHS (i.e, the smallest principal components contribute the most to $\hat{V}(p)$). In the FS, we fix a tail probability, for example to the value 0.05, and select those features that correspond to the tail by identifying the corresponding smallest principal components

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361688, IEEE Journal of Biomedical and Health Informatics

(elements of $\mathbf{e}_1$). Note that the number of selected features by this proposed procedure is not pre-specified, but it depends on the chosen tail probability.

III. Database and Preprocessing

A. Database Description

The database used in the current study consisted of unstructured Norwegian text extracted from the EHR used at the Department of Gastrointestinal Surgery at the University Hospital of North Norway. All documents related to both inpatient and outpatient visits between 2004-2012 were extracted. The most frequent document types that were extracted were nurses notes, journal notes, outpatient notes, radiology reports, referrals, discharge letters and admission notes. A clinician (author KEM) manually reviewed the EHR of 402 patients admitted for CRC surgery in 2006-2011, and 31 patients with AL were identified. The negative class consisted of the 371 remaining patients.

A BoW model was subsequently built, by counting all unique words appearing in the database. There were a total of 65328 unique words in the database. Hence, the database is represented as $D = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{n}$ where each $\mathbf{x}^{(i)}$, representing the i-th patient, is 65328-dimensional. For compact notation, we collect the data samples $\mathbf{x}^{(i)}$, $i = 1, \ldots, n$, in the matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. In the linear SVM $y = \langle \mathbf{w}, \mathbf{x} \rangle + b$, each element, or feature, in $\mathbf{x}$ hence corresponded to the number of appearances in a EHR for a given patient of one of the unique words.

B. Preprocessing

Initially all words were transformed to lowercase and all grammatical symbols were removed. Furthermore, all numbers and stop words were filtered out. Apart from that, advanced natural language processing procedures, such as combining words with identical meanings or corrections of obvious misspellings, were not considered in this work.

These unique words represent the ”bag” in the BoW model by only keeping those words appearing at least a certain number of times, the bag cardinality was reduced, assuming that e.g. misspelled words appear relatively infrequently. As a first preprocessing approach, several word thresholds were evaluated in terms of confusion matrix values. Thus we obtained improved results with a threshold of 10, meaning that only words appearing at least 10 times were included in the BoW. This threshold actually reduced the dimensionality of the vectors x_i from 65328 to 13188. Of course, enforcing a threshold may lead to information loss. Note that enforcing a too high threshold may lead to information loss. In the remainder of the paper, the data set consists of the resulting 13188 words.

In previous general-purpose text classification studies using SVM [42]–[45], normalization has been suggested for preprocessing. Normalization may be obtained in several ways. Term frequency - inverse document frequency (TF-IDF) representation [46] is a common method. Here we considered this and other normalization strategies, such as standardization to mean zero and unit variance. Alternatively, feature vectors may be normalized to equal (Euclidean) length. In our study, such normalizations did not influence the results much, and they were not pursued further.

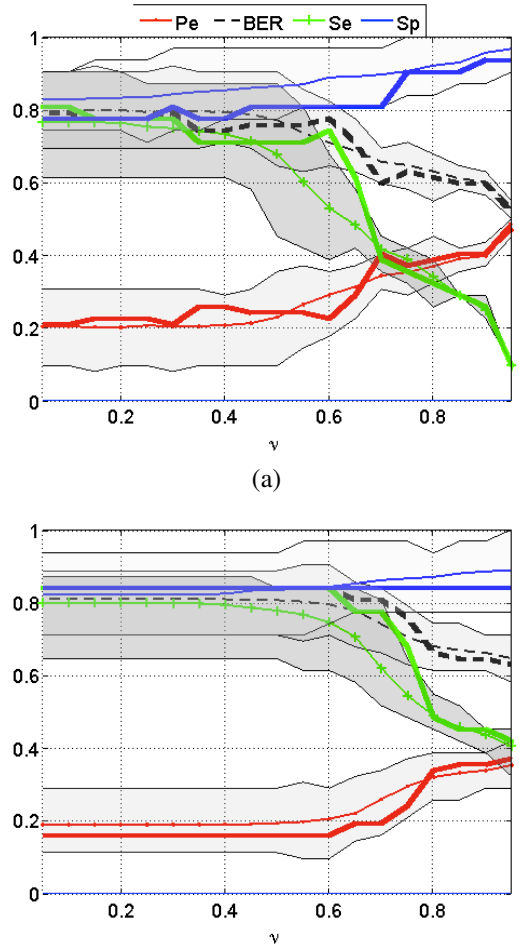


Fig. 3. Free parameter (ν) tuning in terms of several figures of merit (Pe, Se, Sp and BER) for random (thick) and resampling (fine and filled, CI shaded) downsampling, evaluated for $\mathbf{X}$ and $\mathbf{X}^{bin}$ in (a) and (b), respectively.

Finally, the feature set can be represented on a binary basis, by the presence or absence of each word, so that the influence of high frequency words that do not necessarily exhibit discriminatory power is reduced. This binary dataset is denoted by $\mathbf{X}^{bin}$.

IV. Experiments

This experiments section starts by analyzing and discussing the tuning of the free parameter ν in the SVM, and then comparing the SVM AL classification performance on the dataset $\mathbf{X}$ without FS, with those of FDA and NB. We subsequently analyze in detail the effect of the proposed FS strategies, and show that results improve significantly. Finally, a temporal analysis explores the viability of early detection of AL by means of the BoW model.

A. Parameter Tuning

The linear ν-SVM algorithm requires the tuning of a single free parameter $\nu \in (0, 1)$, which has to be tuned. This

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361688, IEEE Journal of Biomedical and Health Informatics

parameter must be tuned based on the available training set. We adopted a LOO strategy for the tuning of ν, ensuring that the parameter tuning was always based on out-of-sample performance. For completeness, we evaluated several different performance measures, namely, Pe, Se, Sp, and BER (see Sec. II-A).

The training set was constructed using an undersampling strategy in order to enforce balanced classes (see Sec. II-C). Towards that end, a *random subset* (31 samples) of the negative class was selected, together with the 31 positive samples in the database. This random subset was used for the tuning of ν. The results, one for each performance measure, are shown in Fig. 3, indicated by the thick line (see figure text for further explanation). Observe that the best performance was obtained for a relatively wide range of smaller values of ν, independently of the figure of merit used. We computed CM for $\nu \in [0.05, 0.4]$ (not shown) finding that the error rates were basically the same over this range of ν. In the end, we decided to use $\nu = 0.05$ in subsequent experiments (see below). The reason for this choice was that ν represents an upper bound on the fraction of margin errors and a lower bound of the fraction of support vectors relative to the total number of training examples. As few support vectors as possible, while maintaining performance, is in general considered a positive property of any SVM method.

In order to analyze the appropriateness of the particular random subset used here, in a statistical sense, we extracted further 50 random resamples (with replacement) from the negative class. Figure. 3 shows the mean performance (fine line, see figure text) and the confidence interval (filled tube) for each of the figures of merit. It is important to note that the results corresponding to the initial random sample lie well within the confidence interval, and may therefore be considered representative for the negative class.

The test or generalization performance of the SVM received special attention in this work. The key element when evaluating the generalization ability is to keep the training and the testing process independent, as far as possible in the given database. For this purpose, the overall data set was divided in two parts, one part in which there is a balance between positive and negative instances (balanced part), and a second part consisting of the remaining negative instances. The generalization ability was measured by a two-stage process, whereby LOO cross-validation was first invoked on the balanced part (i.e. by re-training the SVM D times on near balanced classes, where D is the number of samples when classes are balanced), and then it was combined with the test results obtained on the remaining negative instances when applying the SVM classifier based on the balanced set. Table I shows CM for the SVM using $\nu = 0.05$, together with the FDA and NB methods (using the same generalization procedure), for both feature spaces $\mathbf{X}$ and $\mathbf{X}^{bin}$ (SVM only). First of all, the table shows that FDA and NB performances are clearly lower to those of the SVM. Interestingly, for the SVM, results were better on $\mathbf{X}^{bin}$ compared to $\mathbf{X}$. We used a Paired Bootstrap resampling test as proposed in [47] to establish statistical significance of the different performances across methods, obtaining that $\mathbf{X}^{bin}$ performs better than the other ones.

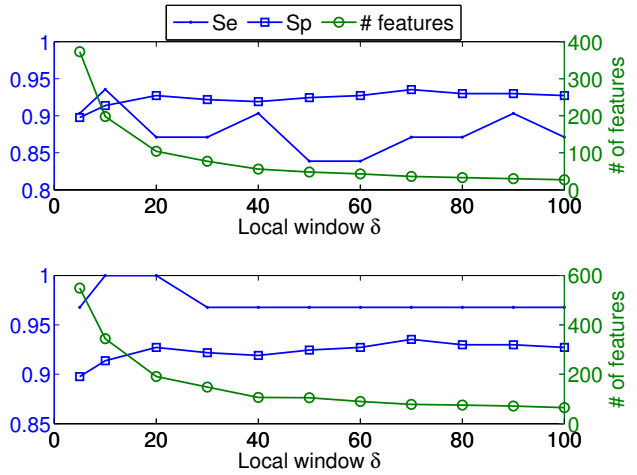


Fig. 4. Number of features, Se and Sp depend on the size of a local window δ of neighbor weights for $\mathbf{X}$ (upper panel) and $\mathbf{X}^{bin}$ (lower panel).

TABLE I
Performance for ν-SVM, FDA, and NB classifiers.

	ν-SVM, $\boldsymbol{X}$	ν-SVM, $\boldsymbol{X}^{bin}$	FDA, $\boldsymbol{X}$	NB, $\boldsymbol{X}$
CM	$\begin{bmatrix} 25 & 56 \\ 6 & 315 \end{bmatrix}$	$\begin{bmatrix} 26 & 52 \\ 5 & 319 \end{bmatrix}$	$\begin{bmatrix} 15 & 208 \\ 16 & 163 \end{bmatrix}$	$\begin{bmatrix} 10 & 28 \\ 21 & 343 \end{bmatrix}$
Se	81%	84%	48%	32%
Sp	85%	86%	42%	92%

When using FDA, it is well-known that the inherent matrix inversion is problematic when the number of features, i.e. the dimensionality, is greater than the number of samples. For that reason, we forced the dimensionality of the feature vectors to be less than 402, which was the number of samples, by considering the 350 most frequently occurring words. A problem when using NB, is that some of the most infrequent words, or features, are not appearing in both classes. In order to avoid this problem, we kept only those features appearing in both classes.

In the testing phase, the classes were imbalanced. For this reason, we also display the performance, or generalization ability, of the SVM in terms of Se and Sp in Table I. The SVM results on $\mathbf{X}^{bin}$ also stand out with respect to Se measures.

B. FS Experiments

In this section, we turn our attention to the analysis of the proposed FS strategies, namely, a simple statistical criterion (LOO based test), an intensive-computation statistical criterion (Bootstrap resampling), and an advanced statistical knowledge criterion (kernel entropy). The core idea is the following: 1) a subset of relevant features was selected by one of the proposed algorithms, and 2) the linear ν-SVM classifier was retrained with the selected features and used to classify test samples. As shown below, the performance of the classifier *increases* as a result of the FS.

First, we provide a brief discussion on the selection of free parameters in the FS algorithms. For Bootstrap resampling, the free parameter corresponds to the size of a local window

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361688, IEEE Journal of Biomedical and Health Informatics

TABLE II
FS criteria analysis. CM, Se, Sp and number of selected features obtained by LOO based test, Bootstrap resampling, and kernel entropy criterion (Keca) for $\mathbf{X}$ (upper) and $\mathbf{X}^{bin}$ (lower) inputs spaces.

	All	LOO	Boot ($\delta = 10$)	Keca
CM	$\begin{bmatrix}25 & 56\\ 6 & 315\end{bmatrix}$	$\begin{bmatrix}28 & 52\\ 3 & 319\end{bmatrix}$	$\begin{bmatrix}31 & 39\\ 0 & 332\end{bmatrix}$	$\begin{bmatrix}25 & 55\\ 6 & 316\end{bmatrix}$
Se	81%	90%	100%	81%
Sp	85%	86%	89%	85%
♯ features	13188	6896	196	212
CM	$\begin{bmatrix}26 & 52\\ 5 & 319\end{bmatrix}$	$\begin{bmatrix}29 & 52\\ 2 & 319\end{bmatrix}$	$\begin{bmatrix}31 & 42\\ 0 & 329\end{bmatrix}$	$\begin{bmatrix}31 & 45\\ 0 & 326\end{bmatrix}$
Se	84%	94%	100%	100%
Sp	86%	86%	89%	88%
♯ features	13188	8073	292	189

TABLE III
Proposed FS benchmarked with RFE for non-binary ($\mathbf{X}$, upper) and binary ($\mathbf{X}^{bin}$, lower) input feature spaces.

	LOO	RFE	Boot	RFE	Keca	RFE
Se	90%	87%	100%	87%	81%	80%
Sp	86 %	86%	89%	82%	85%	82%
♯ features	6896	6896	196	196	212	212
Se	94%	90%	100%	100%	100%	100%
Sp	86%	86%	89%	88%	88%	88%
♯ features	8073	8073	292	292	189	189

δ of neighbor weights. Small δ values provide higher number of selected features, whereas the opposite is true for larger δ values. This is illustrated in Fig. 4 where Se and Sp results are shown based on Bootstrap FS retraining over a range of δ values. Results suggest that good performance was obtained when considering $\delta = 10$, which is the value used in subsequent experiments.

Kernel entropy component FS requires the selection of the tail probability and the kernel size (σ value). We have experienced that a tail probability of 0.05 provides good results. Furthermore, since w is a one dimensional random variable, Silverman's rule [48] for kernel size selection is known to be reliable, and for that reason we used that criterion in the remainder. With this approach, kernel size is obtained as follows: $\sigma = 1.06std(\mathbf{w})N^{-1/5}$, where std is the standard deviation and $\mathbf{w}$ is the weight vector obtained for each dataset.

Table II shows the benefit of FS in terms of CM, Se, Sp and the number of selected features obtained by the LOO based test, the Bootstrap resampling, and the kernel entropy criterion for both databases $\mathbf{X}$ and $\mathbf{X}^{bin}$. The power of the proposed FS methods can be observed by noting that all of them *improve* Se and Sp measures. Furthermore, these improvements were obtained by using *far fewer features*, compared to the original dimensionality of the data.

Results suggest that the best performance is obtained with the Bootstrap resampling approach for both $\mathbf{X}$ (Se 100%, Sp 89%) and $\mathbf{X}^{bin}$ (Se 100%, Sp 89%). The number of features selected to obtain these results were 196 for $\mathbf{X}$ and 292 for $\mathbf{X}^{bin}$.

For completeness, we compared the proposed feature selection strategies with the RFE method [19]. Results obtained using the proposed FS methods and RFE are shown in Table III, using for clarity the same number of features for RFE as the number of features selected by the proposed methods, respectively. Recall that RFE requires the training of multiple classifiers on subsets of features of decreasing size, and for this reason, it does not trivially provide the optimum number of features to be selected. We also implemented the RFE cross-validation procedure (requiring up to 12 hours run-time on a standard research-purpose laptop for one data set) obtaining results which were very similar to those displayed in Table III. This shows that the proposed FS methods may extract useful information from the EHRs, similarly or better when compared to the RFE, however, it is based on statistical criteria requiring no pre-specification of the number of features to be selected, nor any computationally demanding cross-validation.

C. Early AL Detection Experiments

We further explored the *early detection* of AL. Towards that end, we created several additional databases. The databases $\mathbf{X}_{op}$ and $\mathbf{X}_{op}^{bin}$ represented the BoW based on all journal notes up to and including the day of surgery. At this point in time, none of the patients who eventually experienced AL, had developed the condition.

Furthermore, the BoW databases $\mathbf{X}_{+4}$ and $\mathbf{X}_{+4}^{bin}$ were created, where "$+4$" indicates that this BoW is based on all journal notes up to and including postoperative day four.

Table IV shows CM, Se, Sp, and the number of selected features for all the considered databases. The Area Under the Curve was also explored, but similar results were obtained.

Note that discriminatory power is revealed, even for $\mathbf{X}_{op}$ and $\mathbf{X}_{op}^{bin}$. In particular, for $\mathbf{X}_{op}^{bin}$, the results show that given that the patient will eventually experience AL, our FS method detects that in 100% of the cases. On the other hand, given that the patient does not eventually experience AL, our FS method correctly reveals that in about 72% of the cases. This means that the FS approach advocated in this paper, has capacity for detecting AL patients at an early stage. Note also that the number of features selected in order to achieve these results is dramatically lower than the cardinality of the input feature space. As one would expect, the discriminatory power in the data increases with time.

D. Interpretation of Selected Words

One of the major advantages of training a linear SVM on the EHR is that each weight in the weight vector $\mathbf{w}$ corresponds to a particular word in the BoW database, enabling knowledge extraction by analyzing the weights. In this section, we present those words that correspond to the dominant SVM weights, and interpret the words in the context of AL detection.

We focus on the databases $\mathbf{X}_{op}^{bin}$ and $\mathbf{X}^{bin}$ due to the promising AL detection results presented in the previous section. These databases contained only positive elements (binary numbers), such that a positive weight corresponded directly with the positive class (AL) and a negative weight was associated with the negative class, since the classification into the positive or negative class is based on the sign of $\mathbf{w}^\top\mathbf{x}$.

Consider Fig. 5, which shows the selected (with the Bootstrap method) SVM weights obtained for the database $\mathbf{X}_{op}^{bin}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361688, IEEE Journal of Biomedical and Health Informatics

TABLE IV

Temporal analysis (CM and number of features) for different data time slots: up to and including day of surgery; four days after surgery or until patients leave the hospital, for non-binary and binary input feature spaces.

FS	$\mathbf{X}_{op}$	$\mathbf{X}_{+4}$	$\mathbf{X}$	$\mathbf{X}_{op}^{bin}$	$\mathbf{X}_{+4}^{bin}$	$\mathbf{X}^{bin}$
All	$\begin{bmatrix}19 & 186\\ 12 & 185\end{bmatrix}$	$\begin{bmatrix}17 & 126\\ 14 & 245\end{bmatrix}$	$\begin{bmatrix}25 & 56\\ 6 & 315\end{bmatrix}$	$\begin{bmatrix}20 & 145\\ 11 & 226\end{bmatrix}$	$\begin{bmatrix}22 & 112\\ 9 & 259\end{bmatrix}$	$\begin{bmatrix}26 & 52\\ 5 & 319\end{bmatrix}$
Se	61%	55%	81%	65%	71%	84%
Sp	50%	66%	85%	61%	70%	86%
♯ features	5409	6858	13188	5409	6858	13188
LOO	$\begin{bmatrix}28 & 193\\ 3 & 178\end{bmatrix}$	$\begin{bmatrix}25 & 118\\ 6 & 253\end{bmatrix}$	$\begin{bmatrix}28 & 52\\ 3 & 319\end{bmatrix}$	$\begin{bmatrix}29 & 131\\ 2 & 240\end{bmatrix}$	$\begin{bmatrix}30 & 93\\ 1 & 278\end{bmatrix}$	$\begin{bmatrix}29 & 52\\ 2 & 319\end{bmatrix}$
Se	90%	81%	90%	94%	97%	94%
Sp	48%	68%	86%	65%	75%	86%
♯ features	2840	3912	6896	2991	3992	8073
Boot	$\begin{bmatrix}30 & 196\\ 1 & 175\end{bmatrix}$	$\begin{bmatrix}30 & 130\\ 1 & 241\end{bmatrix}$	$\begin{bmatrix}31 & 39\\ 0 & 332\end{bmatrix}$	$\begin{bmatrix}31 & 105\\ 0 & 266\end{bmatrix}$	$\begin{bmatrix}31 & 82\\ 0 & 289\end{bmatrix}$	$\begin{bmatrix}31 & 42\\ 0 & 329\end{bmatrix}$
Se	97%	97%	100%	100%	100%	100%
Sp	47%	65%	89%	72%	78%	89%
♯ features	107	102	196	120	142	292
Keca (5%)	$\begin{bmatrix}29 & 181\\ 2 & 190\end{bmatrix}$	$\begin{bmatrix}30 & 146\\ 1 & 225\end{bmatrix}$	$\begin{bmatrix}25 & 55\\ 6 & 316\end{bmatrix}$	$\begin{bmatrix}29 & 125\\ 2 & 246\end{bmatrix}$	$\begin{bmatrix}31 & 85\\ 0 & 286\end{bmatrix}$	$\begin{bmatrix}31 & 45\\ 0 & 326\end{bmatrix}$
Se	94%	97%	81%	94%	100%	100%
Sp	51%	61%	85%	66%	77%	88%
♯ features	90	110	212	86	106	189

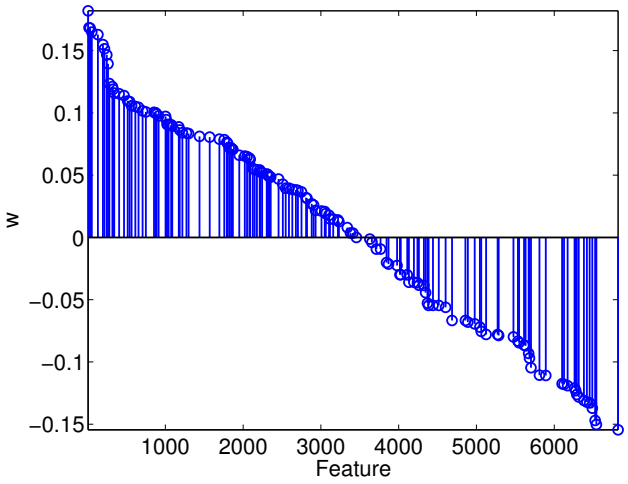


Fig. 5. Selected weights in decreasing order after Bootstrap FS for $\mathbf{X}_{op}^{bin}$.

TABLE V

Words associated with selected (Bootstrap) SVM positive weights corresponding to $\mathbf{X}_{op}^{bin}$ (first column) and $\mathbf{X}_{op}$ (second column).

$\mathbf{X}_{op}^{bin}$	$\mathbf{X}^{bin}$
anastomosis	anastomosis leakage
shaved	anastomosis
easy	re-operated
relieving	re-operation
low	butt
localized	insufficiency
air	saline
info	anterior
up	vatan
anterior	colorectal
peripheral	some
far	drainage
anesthesia	sigmoidostomi
evt	suture
stapler	stapler
loop ileostomy	furix
coloanal	localized
daily	rectal stump
end	atelectasis
locations	drainage
recti-diagnostics	anastomosis
transition	leakage
thought	rectum resection
irregular	irritated
leads	tired
indentation	step
anastomosis leakage	rectally
scope	rise

Those weights with the largest positive values correlate the most with the positive class, and vice versa for the negative class. Table V (first column) shows the words corresponding to some of the largest positive weights (in order) for $\mathbf{X}_{op}^{bin}$. These were the words which SVM associates with the positive class, i.e. the class of patients experiencing postoperative AL. For surgeons (authors ROL, KEM, KMA, AR), the appearance of several of these words in association with AL seemed reasonable from a clinical perspective. Some examples are presented below.

Tumors located in the lower part of the rectum are known to increase the risk of AL, and are removed by the surgical procedure known as *low anterior resection*. Similar reasons may explain the appearance of the word *anterior* in Table IV (first column). The word *air* may be an indicator of a leakage, since the presence of air outside of the bowel will be due to a leakage. A diverting *loop iliostomy* will be performed in patients with the highest risk of anastomotic leakage (low rectal cancer with *coloanal* anastomosis and after neoadjuvant treatment with irradiation).

Regarding the words associated by the SVM to the negative class, we provide some examples in the following words: *amputation*, *abdominoperineal*, *endcolostomy*, *proximal*, *sonor percussion sound*. One of the words is *amputation*. This word simply refers to the removal of the whole rectum and anus in order to remove a distal rectal tumor oncologically safe. In that

case the problem of AL is completely avoided and the patient will have a permanent endcolostomy. Abdominoperineal amputation is the name of the operation. Patients with *proximal* (means located in the upper part of rectum) rectal cancer do not need deep pelvic surgery and are thereby less exposed to anastomotic leakage. The expression *sonor percussion* sound is used when the physician describes the normal sounds that appear before the operation, when he/she carefully knocks on a finger placed on the patients chest in order to detect pleural fluid collections or abnormal air distribution in the chest. AL is often followed by lung and heart complications.

We also analyzed the selected (Bootstrap) SVM weights corresponding to the databases $\mathbf{X}_{+4}^{bin}$ and $\mathbf{X}^{bin}$. The distribution of positive and negative weights change for these databases, compared to $\mathbf{X}_{op}^{bin}$. We focus here on the words corresponding to $\mathbf{X}^{bin}$. Table V (second column) shows the words corresponding to some of the largest positive weights (in order) for $\mathbf{X}^{bin}$. Several of the words from Table V in the first column reappear in the second column. However, there are differences. For example, the weight associated with the word *anastomosis leakage* is now the largest of all the weights. Furthermore, words like *re-operation* and *re-operated* are also associated with large weights.

This analysis shows that the selected words, obtained by the proposed FS strategies based on the BoW model, may be reasonably interpreted in the medical context of AL. Future work may consider highlighting words of particular medical relevance when training decision support systems, or flag certain selected words as indicators of the AL complication.

V. Discussions

In this paper, we demonstrate that the clinical narrative contains relevant information for early detection of AL following surgery for colorectal cancer. The discriminatory power in the data is based on a simple BoW model, where classification and feature selection is based on a linear ν-SVM.

Results show that both binary and non-binary approaches have discriminatory power. A binary input space yields a sensitivity of 100% and specificity of 72%, while performance worsens when using a non-binary input space, to 97% and 47% respectively. The number of relevant features is also lower in the latter case. In multidisciplinary studies like the present one, validation by clinicians is highly necessary in order to extract correct and useful knowledge. The set of words shown in Table V was therefore validated by a group of surgeons who concluded that several words (in bold) appear to have relevance for identification of patients with increased risk of anastomotic leakage after colorectal surgery.

The study has some limitations. In particular, the number of cases is low, and hence prone to over-fitting, such that external validation of the results would be desirable. A manual annotation process as used here is likely to provide accurate labels, but is very time consuming. By using automated phenotyping [49] of the EHR, one can effectively gather larger cohorts at the loss of some accuracy. The extracted text does not contain all information about the patient, and notably the surgery notification form is unavailable. Thus there is much information missing about patients preoperative status, which could be important additional indicators of subsequent complications and could improve accuracy.

In studies of risk assessment models there is always the concern that the signal may be censored when a clinician spots a pattern leading to a complication and takes appropriate action to avoid the complication [50]. This would result in a significant number of cases where the pattern leading to the adverse event is present but not the event itself as that was successfully averted, effectively constituting mislabeled cases. This might be a concern in the current study, and would, if the classifier is good, lead to a decreased specificity. In the current paper we used a BoW model, which is arguably the simplest possible model for text processing. Nevertheless we demonstrated potential for feature selection for improving the AL detection.

In future work we will incorporate more advanced NLP tools to build models that may be more robust to erroneous inputs such as misspelled or accidentally omitted words, and unusual inputs such as words or structures that have not been encountered by the classifier before. However, most available methods are designed for English language, and not directly applicable for Norwegian clinical language. For English, a common practice is to use the Unified Medical Language System (UMLS) [51], which enables a consistent representation of clinical language, to which no Norwegian counterpart exists to our knowledge. These issues represent a challenge for future work. As the present study only includes the written narrative, other information about the patient such as sex, age, tumor status and other structured data are not included. Structured data that are available but not used in the current study include prior diagnosis codes, procedure codes and test results. Future studies will combine these with the presented approach to obtain a more complete risk assessment status of the patient. The results of this paper indicate that the clinical narrative can be used as a basis for a clinical decision support system. A complete system must consider all available information as outlined above, and should be designed both in collaboration with clinicians and EHR providers to provide a streamlined, usable and useful system integrated in the patient care. There are significant technological and operational challenges, but establishing robust and methodologically sound methods is a vital first step in the process to design such systems and integrating them in the surgical workflow.

This innovate study describes the development of an early computerized warning system that, when fully developed, will be a useful supplement for the clinician to be alerted at an early stage and act promptly to avoid potentially lethal postoperative outcomes. It is important that the information provided by the system is actionable on the part of the physician, in that there is an option to change the course of action for patients with increased risk. In the case where the risk is evident prior to the index surgery, potential courses of action can be to postpone the surgery until all known risk factors are corrected or to protect the anastomosis by a diverting stoma or avoid any anastomosis by giving the patient a permanent stoma in the first place. Additionally, the patient can be involved in the preoperative decision-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361688, IEEE Journal of Biomedical and Health Informatics

making and sign an informed consent form based on a better understanding of the preoperative risk-assessment for AL. In the case of increased risk postoperatively, potential actions in the case of alarm signals indicating an anastomotic leakage would be emergency CT scans, diagnostic laparoscopy, or laparotomy. The latter two are resource demanding and not without potential complications. It is therefore beneficial to have additional computerized algorithms as described in this paper, in addition to sound clinical judgment.

VI. Conclusions

We have shown that there is information in the clinical narrative that can be used to predict anastomosis leakage after colorectal surgery. Thus, the text can be a piece of the input to a clinical information system that warns clinicians of the potential for complications in individual patients. To the best of our knowledge, the FS methods proposed in this paper are novel contributions in this field. Experimental results corroborate the feasibility and sustainability of the proposed framework, although future work could further enhance results to support early diagnosis decisions.

References

[1] J. Rodrigues (Ed.), *Health Information Systems: Concepts, Methodologies, Tools, and Applications*, IGI Global, Hershey New York, 2010.

[2] S. Garde, P. Knaup, E.J.S. Hovenga, and S. Heard, "Towards semantic interoperability for electronic health records: domain knowledge governance for openEHR archetypes," *Method Inform Med*, vol. 46, no. 3, pp. 332–43, 2007.

[3] J.M. Buckley, S.B. Coopey, J.Sharko, F. Polubriaginof, B. Drohan, A.K. Belli, E.M. H. Kim, J.E. Garber, B.L. Smith, M.A. Gadd, M.C. Specht, C.A. Roche, T.M. Gudewicz, and K.S. Hughes, "The feasibility of using natural language processing to extract clinical information from breast pathology reports," *J Pathol Inform*, vol. 3, no. 23, pp. 1–7, 2012.

[4] Y. Wang, Z. Yua, Y. Jiangb, Y. Liuc, L. Chena, and Y. Liua, "A framework and its empirical study of automatic diagnosis of traditional chinese medicine utilizing raw free-text clinical records," *J Biomed Inform*, vol. 45, no. 2, pp. 210–223, 2012.

[5] A. Wright, A.B. McCoy, S. Henkin, A. Kale, and D.F. Sittig, "Use of a support vector machine for categorizing free-text notes: assessment of accuracy across two institutions," *J Am Med Inform Assoc*, vol. 20, no. 5, pp. 887–890, 2013.

[6] Z. Afzal, M. Engelkes, K.M. Verhamme, H.M. Janssens, M.C. Sturkenboom, J.A. Kors, and M.J. Schuemie, "Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases," *Pharmacoepidemiol Drug Saf*, vol. 22, no. 8, pp. 826–33, 2013.

[7] Z. Wang, A.D. Shah, A.R Tate, S. Denaxas, J. Shawe-Taylor, and H. Hemingway, "Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning," *PLoS One*, vol. 7, no. 1, pp. 1–9, 2012.

[8] R. Cobb, S. Puri, D. Zhe Wang, T. Baslanti, and A. Bihorac, "Knowledge extraction and outcome prediction using medical notes," in *ICML workshop on Role of Machine Learning in Transforming Healthcare*, Atlanta, Georgia, USA, 2013.

[9] P.M. Nadkarni, L. Ohno-Machado, and W.W. Chapman, "Natural language processing: an introduction," *J Am Med Inform Assoc*, vol. 18, pp. 544–51, 2011.

[10] S.M. Meystre, G.K. Savova, K.C. Kipper-Schule, and J.F. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research," *Year Med Inform*, pp. 128–44, 2008.

[11] I.K. Larsen, *Cancer in Norway 2011 - Cancer incidence, mortality, survival and prevalence in Norway*, Cancer Registry of Norway, Oslo: Cancer Registry of Norway, 2013.

[12] H. Kehlet, "Fast-track colorectal surgery," *The Lancet*, vol. 371, no. 9615, pp. 791–3, 2008.

[13] H.S. Snijders, D. Henneman, N.L. van Leersum, M. Ten Berge, M. Fiocco, T.M. Karsten, K. Havenga, T. Wiggers, J.W. Dekker, R.A. Tollenaar, and M.W. Wouters, "Anastomotic leakage as an outcome measure for quality of colorectal cancer surgery," *BMJ Qual Saf*, vol. 22, no. 9, pp. 759–67, 2013.

[14] N.A. Hirst, J.P. Tiernan, P.A. Millner, and D.G. Jayne, "Systematic review of methods to predict and detect anastomotic leakage in colorectal surgery," *Colorectal Dis*, 2013.

[15] A. Karliczek, N.J. Harlaar, C.J. Zeebregts, T. Wiggers, P.C. Baas, and G.M. van Dam, "Surgeons lack predictive accuracy for anastomotic leakage in gastrointestinal surgery," *International Journal of Colorectal Disease*, vol. 24, no. 5, pp. 569–76, 2009.

[16] J.W Dekker, G.J Liefers, J.C de Mol van Otterloo, H Putter, and R.A Tollenaar, "Predicting the risk of anastomotic leakage in left-sided colorectal surgery using a colon leakage score," *J Surg Res*, vol. 166, no. 1, pp. e27 – e34, 2011.

[17] G. Hughes, "On the mean accuracy of statistical pattern recognition," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 55–63, 1968.

[18] K. Fukunaga, "Effect of sample size in classifier design," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 8, pp. 873–85, 1989.

[19] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach Learn*, vol. 46, no. 1, pp. 389–422, 2002.

[20] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Med Inform Decis Mak*, vol. 11, no. 51, pp. 1–13, 2011.

[21] J. Peacock and P. Peacock, *Oxford Handbook of Medical Statistics*, Oxford University Press Print, Oxford, UK, 2010.

[22] V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, Inc., New York, 1998.

[23] B. Schölkopf and A.J. Smola, *Learning with kernels*, MIT Press, Cambridge, MA, 2002.

[24] B. Schölkopf, A.J. Smola, R.C. Williamson, and P.L. Bartlett, "New support vector algorithms," *Neural Comput*, vol. 12, no. 5, pp. 1207–45, 2000.

[25] R.A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann Eugen*, vol. 7, pp. 179–88, 1936.

[26] T. Mitchell, *Machine Learning*, McGraw-Hill, Boston, MA, 1997.

[27] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Mach Learn*, vol. 129, pp. 103–30, 1997.

[28] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach Learn*, vol. 29, pp. 131–63, 1997.

[29] H. He and E.A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–84, 2009.

[30] R. Kohavi and G.H. John, "Wrappers for feature subset selection," *Artif Intell*, vol. 97, pp. 273–324, 1997.

[31] I. Guyon, S. Gunn, M. Nikravesh, and L.A. Zadeh (Ed.), *Feature extraction: foundations and applications*, Springer, Heidelberg, 2006.

[32] P. Pavlidis, J. Cai, J. Wetson, and W.N. Grundy, "Gene functional analysis from heterogeneous data," in *Proc. of the 5th IC on Computation Molecular Biology*, Montreal, QC, Canada, 2000, pp. 242–8.

[33] J. Brank, M. Grobelnik, N. Mili-Frayling, and D. Mladeni, "Feature selection using support vector machines," in *Proc. of the 3rd Int. Conf. on Data Mining Methods and Databases for Engineering, Finance, and Other Fields*, Bologna, Italy, 2002, pp. 84–89.

[34] Y. Yang and J.O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. of the 14th ICML97*, San Francisco, CA, USA, 1997, pp. 412–320.

[35] D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and Naive Bayes," in *Proc. of the 16th ICML99*, Bled, Slovenia, 1999, pp. 258–67.

[36] G. Forman, "An experimental study of feature selection metrics for text categorization," *J Mach Learn Res*, vol. 3, pp. 1289–305, 2003.

[37] G.C. Cawley and N.L.C Talbot, "Fast exact leave-one-out cross-validation of sparse least-squares support vector machines," *Neural Networks*, vol. 17, pp. 1467–75, 2004.

[38] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann Stat*, vol. 7, no. 1, pp. 1–26, 1979.

[39] A. Renyi, "On measures of entropy and information," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1960, pp. 547–561.

[40] R. Jenssen, "Kernel entropy component analysis," *IEEE Trans Pattern Anal Mach Intell*, vol. 33, no. 5, pp. 847–860, 2010.

[41] J. Shawe-Taylor and N. Cristianini (Ed.), *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, UK, 2004.

[42] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Machine Learning: ECML-98, Lecture Notes in Computer Science*, Chemnitz, Germany, 1998, vol. 1398, pp. 137–42.
[43] H. Drucker, D. Wu, and V.N. Vapnik, "Support vector machines for spam categorization," *IEEE Trans Neural Netw*, vol. 10, no. 5, pp. 1048–54, 1999.
[44] E. Leopold and J. Kindermann, "Text categorization with support vector machines. How to represent texts in input space?," *Mach learn*, vol. 46, pp. 423–44, 2002.
[45] A.B.A. Graf, A.J. Smola, and S. Borer, "Classification in a normalized feature space using support vector machines," *IEEE Trans Neural Netw Learn Syst*, vol. 14, no. 3, pp. 597–605, 2003.
[46] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–23, 1988.
[47] C. Soguero-Ruiz, F.J. Gimeno-Blanes, I. Mora-Jiménez, M.P. Martínez-Ruiz, and J.L. Rojo-Álvarez, "On the differential benchmarking of promotional efficiency with machine learning modeling (i): Principles and statistical comparison," *Expert Syst Appl*, vol. 39, no. 17, pp. 12772–83, 2012.
[48] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall CRC, London, UK, 1986.
[49] C. Shivade, P. Raghava, E. Fosler-Lussier, P.J. Embi, N. Elhadad, S.B. Johnson, and A.M. Lai, "A review of approaches to identifying patient phenotype cohorts using electronic health records," *J Am Med Inform Assoc*, vol. 21, no. 2, pp. 221–30, 2014.
[50] C. Paxton, A. Niculasci-Mizi, and S. Saria, "Developing predictive models using electronic medical records: Challenges and pitfalls," *Am. Med. Informatics Assoc. Annu. Symp*, , no. 2, pp. 1109–15, 2013.
[51] National Library of Medicine (US), *UMLS Reference Manual*, 2009.

Cristina Soguero-Ruiz received the Telecommunication Engineering Degree and B.Sc. Degree at Business Administration ad Management from the Rey Juan Carlos University, Spain, in 2011, and the M.Sc degree in Biomedical Engineering at Rey Juan Carlos University in 2012. Currently, she is attending a PhD program in Telecommunication at Rey Juan Carlos University. Her main research interests include statistical learning theory, digital signal processing, with application to eHealth, Electronic Health Records and marketing.

Kristian Hindberg holds an MSc/PhD in physics/mathematical statistics from the University of Troms (UiT), Norway. He also holds an Msc in Space Studies from the International Space University in Strasbourg, France. He worked as a scientist at the Norwegian Defence Research Establishment from 2004 to 2005. He is currently working as a postdoctoral fellow at UiT on projects connected to Troms Telemedicine Laboratory

José Luis Rojo-Álvarez (SenM12) received the Telecommunication Engineering Degree from the University of Vigo, Spain, in 1996, and the Ph.D. in Telecommunications from the Polytechnical University of Madrid, Spain, in 2000. Since 2006, he has been an Associate Professor at the Department of Signal Theory and Communications, Rey Juan Carlos University, Madrid, Spain. He is the coauthor of more than 60 indexed papers and more than 150 national and international conference communications. His current research interests include statistical learning theory, digital signal processing, complex system modeling, and Electronic Health Recording, with applications to digital communications and to cardiac signal and image processing.

Stein Olav Skrøvseth holds a PhD and MSc in physics from the Norwegian University of Science and Technology in Trondheim, Norway. In 2007-2008 he was a postdoctoral fellow at the University of Sydney, Australia, and has since 2009 been a research scientist at the Norwegian Centre for Integrated Care and Telemedicine. He is currently a visiting scientist at IBM TJ Watson research laboratory in Yorktown Heights, New York.

Knut Magne Augestad is a postdoctoral research fellow at Department of Colorectal Surgery, University Hospitals Case Medical Center. Until the summer 2013 I was practicing surgeon and research leader working at the University of Hospital North Norway. My research focuses on surgical quality assessment, decision analytic modeling, predictive analytics and telemedicine. He is the principal investigator in a Norwegian surgical quality assessment project, aiming to integrate big data analyses with continuous surgical quality surveillance. In particular he is working to predict major complications in surgery, using data extracted from the electronic patient record.

Robert Jenssen is an Associate Professor and board member at the Department of Physics and Technology at the University of Tromsø (UiT), Norway, and a Research Professor (20%) at the Norwegian Center for Integrated Care and Telemedicine. He has held guest researcher positions at the Technical University of Denmark (DTU Compute), 2012/2013, at the Technical University of Berlin, 2008/2009 and at the University of Florida, 2002/2003, March/April 2004. In his research, he has focused on information theoretic machine learning, with strong connections to Mercer kernel methods and to spectral clustering and dimensionality reduction methods. Jenssen's paper "Kernel Entropy Component Analysis" was the Featured Paper of the May 2010 issue of IEEE Transactions on Pattern Analysis and Machine Intelligence, and the paper "Kernel Entropy Component Analysis for Remote Sensing Image Clustering" won the IEEE Geoscience and Remote Sensing Society Letters Prize Paper Award, 2013. He also won the "Honorable Mention for the 2003 Pattern Recognition Journal Best Paper Award", the "2005 IEEE ICASSP Outstanding Student Paper Award" and the "2007 UiT Young Investigator Award." Jenssen served on the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee 2006-2009, and is currently an Associate Editor of Pattern Recognition.