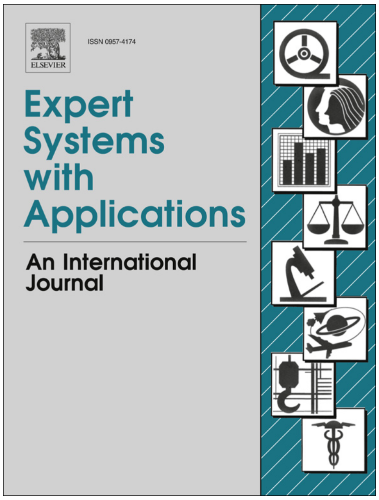

Predictive modeling of hospital readmissions using metaheuristics and data mining

Bichen Zheng, Jinghe Zhang, Sang Won Yoon, Sarah S. Lam, Mohammad Khasawneh, Srikanth Poranki

PII: S0957-4174(15)00308-5
DOI: http://dx.doi.org/10.1016/j.eswa.2015.04.066
Reference: ESWA 10013

To appear in: *Expert Systems with Applications*

Received Date: 14 August 2014
Revised Date: 21 April 2015
Accepted Date: 26 April 2015

Please cite this article as: Zheng, B., Zhang, J., Yoon, S.W., Lam, S.S., Khasawneh, M., Poranki, S., Predictive modeling of hospital readmissions using metaheuristics and data mining, *Expert Systems with Applications* (2015), doi: http://dx.doi.org/10.1016/j.eswa.2015.04.066

Predictive modeling of hospital readmissions using metaheuristics and data mining

Bichen Zheng[a], Jinghe Zhang[a], Sang Won Yoon[a,*], Sarah S. Lam[a], Mohammad Khasawneh[a], Srikanth Poranki[b]

[a] *Department of Systems Science and Industrial Engineering*
State University of New York at Binghamton
Binghamton, New York 13902
[b] *United Health Services Hospitals*
Binghamton, New York 13903

Abstract

This research studies the risk prediction of hospital readmissions using metaheuristic and data mining approaches. This is a critical issue in the U.S. healthcare system because a large percentage of preventable hospital readmissions derive from a low quality of care during patients' stays in the hospital as well as poor arrangement of the discharge process. To reduce the number of hospital readmissions, the Centers for Medicare and Medicaid Services has launched a readmission penalty program in which hospitals receive reduced reimbursement for high readmission rates for Medicare beneficiaries. In the current practice, patient readmission risk is widely assessed by evaluating a LACE score including length of stay (L), acuity level of admission (A), comorbidity condition (C), and use of emergency rooms (E). However, the LACE threshold classifying high- and low-risk readmitted patients is set up by clinic practitioners based on specific circumstances and experiences. This research proposed various data mining approaches to identify the risk group of a particular patient, including neural network model, random forest (RF) algorithm, and the hybrid model of swarm intelligence heuristic and support vector machine (SVM). The proposed neural network algorithm, the RF and

[*] Corresponding author. Tel.: +1 607 777 5935; fax: +1 607 777 4094.
Email addresses: bzheng6@binghamton.edu (Bichen Zheng), jzhang51@binghamton.edu (Jinghe Zhang), yoons@binghamton.edu (Sang Won Yoon), sarahlam@binghamton.edu (Sarah S. Lam), mkhasawn@binghamton.edu (Mohammad Khasawneh), Srikanth_Poranki@uhs.org (Srikanth Poranki)

the SVM classifiers are used to model patients' characteristics, such as their ages, insurance payers, medication risks, etc. Experiments are conducted to compare the performance of the proposed models with previous research. Experimental results indicate that the proposed prediction SVM model with particle swarm parameter tuning outperforms other algorithms and achieves 78.4% on overall prediction accuracy, 97.3% on sensitivity. The high sensitivity shows its strength in correctly identifying readmitted patients. The outcome of this research will help reduce overall hospital readmission rates and allow hospitals to utilize their resources more efficiently to enhance interventions for high-risk patients.

Keywords: Neural networks, Support vector machine, Particle swarm optimization, Hospital readmission, Risk prediction

1. Introduction

Healthcare has become one of the largest industries globally, and as such, it consumes a large amount of resources. In recent years hospital readmission has become a major topic of discussion in the U.S. healthcare system due to significant unnecessary costs associated with it. In 2004 about one-fifth of the Medicare beneficiaries were readmitted to hospitals within 30 days of discharge. It was estimated that the unplanned readmission of Medicare patients cost $17.4 billion (Jencks et al., 2009). Many of the preventable readmissions were related to low quality of care during patient stays in the hospital, as well as to poor arrangement of the discharge process (Malnick et al., 2008). Hospital readmission rate is thus recognized as a quality indicator of inpatient care for which effective, preventative interventions can be implemented (Hasan et al., 2010). The Centers for Medicare and Medicaid Services (CMS) has launched a readmission payment reduction program in which hospitals are financially penalized when Medicare patients are rehospitalized within 30 days of discharge (Centers for Medicare and Medicaid Services, 2012a). Thus, it is advantageous for hospitals to reduce their readmission rates by using effective and efficient interventions during patient stays and the discharge process. Currently, the finalized readmission penalty program focuses on acute myocardial infarction (AMI), heart failure (HF), and pneumonia (PN) since the readmissions from these diagnoses are more common, expensive, and preventable (Centers for Medicare and Medicaid Services, 2012c; QualityNet, 2012). Various interventions are implemented to reduce readmission rates, including enhanced education for patients during the discharge process, medication reconciliation, follow-ups, etc. (Koehler et al., 2009).

Considering that healthcare resources (including physicians, nurses, and other medical resources) are very costly and limited, it is impractical and inappropriate for hospitals to provide equal efforts and interventions for all patients. Therefore, a prediction model that can be used to identify high-risk patients in advance could greatly benefit healthcare providers by enabling them to target resources on risky patients and, by extension, reduce the overall readmission rate (Centers for Medicare and Medicaid Services, 2012b). Once a particular patient is identified as high-risk, intensive interventions can be made to prevent a potential readmission. To corroborate this, one study found that tele-monitoring high-risk patients and corresponding private health plans enabled a 15% reduction in readmissions at a home healthcare facility (Minott, 2008).

However, the process of identifying patients who are very likely to be readmitted within 30 days of discharge is very difficult based on clinical expertise. This is due to the complex causes of readmission, such as a patient's health condition, quality of inpatient care and social determinants. Therefore, the objective of this research is to model the readmission patterns appropriately to predict the likelihood of readmission accurately. To describe the implicit patterns that lead to readmission and non-readmission, there are two clusters of approaches: analytical modeling and data mining. Since the readmission patterns, i.e. the relationship between predictors and dependent variables, are unknown, it is impractical to build an analytical model for accurate pattern description. However, historical data provides good evidence of those implicit patterns. Consequently, researchers proposed the concept and various algorithms of data mining and machine learning to capture hidden patterns from data.

Risk assessment models have been proposed to address the readmission problem for patients with various conditions such as general medicine patients and stroke and heart failure patients. In this research, the readmission rate of HF patients in a community hospital is studied. The majority of past research in hospital readmission used cohort study, logistic regression, and scoring systems to address the problem (Ross et al., 2008). In general, existing risk-prediction models of hospital readmission perform poorly, according to the review research conducted by the Department of Veterans Affairs in 2011 (Kansagara et al., 2011).

In this study, classification models that use neural networks, random forest (RF) and support vector machines (SVM) are proposed to predict the readmission risk of a particular HF patient. The remainder of this paper is structured as follows: Section 2 discusses the related literature in risk prediction modeling, especially those applied to assess patients' readmission risks. Proposed methodologies are described in Section 3; in Section 4, experiments are conducted to train and test those classification models, and the result analyses are discussed to compare the quality of those classifiers. Finally, the summary of this research and future work are addressed in Section 5.

2. Literature Review

Risk-prediction models are broadly implemented in clinical and medical fields to support diagnostic decision-making. These include risk-prediction models for the risk assessment of breast cancer, type 2 diabetes, cardiovascular disease, and mortality for critically-ill hospitalized adults, as well as many others (Lindstrom and Tuomilehto, 2003; Siontis et al., 2012). There are two types of risk-prediction models regarding breast cancer: identifying the risk that a patient will develop breast cancer over a certain time period and estimating the probability that a breast cancer-related gene mutation will occur in an individual (Claus et al., 1994; Parmigiani et al., 1998). According to the risk-assessment results, a high-risk patient will be referred to intensive interventions and attention (e.g. screening and counseling) to prevent potential breast cancer. These models can help reduce the mortality rate among high-risk patients and can control the cost and complications for low-risk patients (Domchek et al., 2003). Noticeably, data-driven machine learning algorithms have been introduced into various medical decision-support domains, including cancer diagnosis (Nahar et al., 2012; Mukti and Ahmed, 2013; Zheng et al., 2014), cardiovascular abnormality detections (Sufi and Khalil, 2011), risk prediction (Siontis et al., 2012), etc. Data-oriented risk-prediction models have become effective tools that help medical decision-making and offer a number of benefits to both healthcare providers and patients. Also,

As an important quality indicator of healthcare services, the high hospital readmission rate has attracted increasing attention and effort from the government, healthcare institutions, insurance payers and patients. Risk-prediction models can help prevent avoidable readmissions and eventually reduce the overall readmission rate. Various methodologies and techniques have been used to develop risk-prediction models for hospital readmission. A brief overview of the previous studies in readmission prediction is presented in Table 1.

Among the proposed methods, a cohort study and statistical models such as logistic regression and Cox proportional hazards regression, are the most common methods to identify risk factors. After those are used, weighted scoring systems are developed to measure the readmission risk of patients based on significant risk factors (Hasan et al., 2010; Whitlock et al., 2011). Cohort studies are commonly used in clinical areas in which groups are tracked from risk factor (exposure) to disease (outcome) in order to identify the correlation between them. As a longitudinal study, the exposure-disease association is determined with a higher quality and less bias. However, in a cohort study, it is expensive and difficult to achieve a high degree of similarity in the control group and to compensate for class imbalance in real cases (Grimes and Schulz, 2002; The Himmelfarb Health Sciences Library, 2011). In addition, logistic regression is a popular classification approach, especially when the outcome is binary. As one of the risk-prediction models, the LACE score has already been implemented in some hospitals. It is developed to predict unplanned readmission and mortality based on a prospective cohort study. This index considers four independent variables, including length of stay (L), acuity level of admission (A), comorbidity condition (C), and use of emergency rooms (E). A LACE score has been developed to evaluate and assess the patient readmission risk based on the LACE index assuming the linear relationship among the four variables. For instance, a LACE score can be obtained by summing up the values of those four variables (van Walraven et al., 2010). A threshold is set up to determine patient readmission risk based on clinics' specific circumstances to classify patients into different risk groups.

The risk-prediction models listed above employ different variables and target different diagnosis-related groups (DRGs). However, considering the ease of implementation and the difficulty of collecting medical and healthcare data, a model with fewer variables is more applicable. In general, those models are not capable of providing an accurate prediction of the readmission risk of a particular patient. Some perform poorly with an accuracy of less than 50%, and very few of them can predict correctly in over 70% of cases (Kansagara et al., 2011). Additionally, most models use statistical approaches, including logistic regression and Cox proportional hazards regression.

Therefore, other approaches for risk prediction, such as data mining and

Table 1: Research map to match between research areas and approaches in the context of patient readmission predictions

	Condition	Sample Size	Attributes	Main Methodology	Readmission Length
(Anderson and Steinberg, 1985)	All except End Stage Renal Disease	270,226	10	Logistic regression	60 days
(Smith et al., 1985)	All	1,506	5	Multivariate analysis	90 days
(Holloway et al., 1990)	Veteran	2,970	14	Logistic regression	30 days
(Boult et al., 1993)	$\geq$ 70 yrs	2,176	8	Logistic regression	4 years
(Thomas, 1996)	12 conditions	1,163-14,590	4	Logistic regression	15/30/60/90 days
(Philbin and DiSalvo, 1999)	CHF	42,731	12	Logistic regression	30 days
(Krumholz et al., 2000)	HF	2,176	4	Cox proportional hazard model	6 months
(Morrissey et al., 2003)	All	1,219	–	Logistic regression	12 months
(Billings et al., 2006)	All	–	21	Logistic regression	12 months
(Bottle et al., 2006)	Emergency admission	2,895,234	12	Logistic regression	12 months
(Halfon et al., 2006)	All	131,809	$\geq$ 3	Poisson regression	30 days
(Novotny and Anderson, 2008)	All	1,077	8	Probability of repeated admission	41 days
(Silverstein et al., 2008)	All	29,292	16	Logistic regression	30 days
(Howell et al., 2009)	Chronic disease	3,129	8	Logistic regression	12 months
(Amarasingham et al., 2010)	HF	1,372	29	Logistic regression	30 days

(Hasan et al., 2010)	All	10,946	18	Logistic regression	30 days
(van Walraven et al., 2010)	All	1,004,812	4	Logistic regression	30 days
(Allaudeen et al., 2011)	≥ 65	159	8	Probability of repeated admission	30 days
(Hammill et al., 2011)	HF	24,163	36	Generalized linear regression	30 days
(Grafa et al., 2012)	$\geq$ 75 and discharged from ED	345	6 and 5	Identification of Senior At Risk (ISAR)/Triage Risk Stratification Tool (TRST)	1/3/6/12 months
(Kociol et al., 2012)	ST-segment elevation myocardial infarction (STEMI)	5,745	–	Logistic regression	30 days
(Kramer et al., 2012)	Intensive care unit	229,375	27	Logistic regression	30 days
(Dharmarajan et al., 2013)	HF, AMI, and PN	1,330,157 /548,834/ 1,168,624	3	Logistic regression and Cox proportional hazard model	30 days
(Garrison et al., 2013)	Family medicine patients	276	11	Logistic regression	30 days

machine learning, can be utilized to achieve a better performance. Jeejeebhoy et al. (2015) utilized the logistic regression model to predict patient readmission risks after 30 days of discharge with nutritional assessment. Another logistic regression model was developed with elastic net regularization to extract patient features automatically and predict the patient readmission risk (Tran et al., 2014). Their main contribution is that the prediction accuracy was maintained with feature reductions, which is important for the large volume of healthcare data. Braga et al. (2014) utilized a support vector machine, decision trees and naive Bayes models to predict patients' readmission into intensive care units. In this research, an oversampling method was used to tackle the issues of imbalanced patient readmission data sets. It showed that the best accuracy can be achieved by the naive Bayes model with a precision of 98.91%. Classification is an important branch of data mining that learns the relationships between attributes and targets in data sets. There are two potential risk groups in this research, high and low readmission risk, which makes this a traditional binary classification problem. To address this problem, instances from both classes are used in the training process in order to identify the characteristics of each class or to find the hyperplane that can separate the two classes. According to the classification results, hospitals can identify high-risk patients and focus their resources to reduce readmissions.

3. Proposed Methodology

3.1. Data Description

A data set derived from medical records is used to study the implicit regularities in hospital readmissions of HF patients. There is a total of 1,641 instances, and 316 of them are readmitted to hospitals within 30 days of discharge. In this dataset, there are nine attributes as presented in Table 2. Various nominal attribute values are indexed by the numeric values for predictive model preferences. For instance, MS-DRG codes are composed of heart failure and shock (HFS) with major comorbidity conditions (MCC), HFS with comorbidity conditions (CC) and HFS without MCC/CC. More detailed attribute distributions and statistical analysis are further shown in Figure 1. It is noted that 21.63% of the total records come from readmitted patients, which indicates the imbalance property of the data set. The data set is separated randomly into training and testing set to validate the proposed methodologies.

Table 2: An overview of input attributes

Attribute Name	Type	Value/Range
Patient Age	Ordinal	[19, 101]
Length of Stay (L)	Ordinal	[0,7] days
Admission Acuity (A)	Nominal	Acute /Non-acute
Comorbidity Index Score (C)	Ordinal	[0,5]
Use of ED (E)	Ordinal	[0,4]
Gender	Nominal	Male/Female
Patient Readmission Risk	Nominal	High-risk/low-risk
MS-DRG Code	Nominal	291-HFS with MCC 292-HFS with CC 293-HFS without CC/MCC
Insurance Payer	Nominal	Commercial Indemnity Insurance Free Government HMO - Medicare HMO/PHSP Medicaid HMO/PHSP Other Medicaid Medicare No Fault Non-profit Indemnity Insurance Workers Compensation

3.2. Data Preprocessing

The objective of this research is to predict the readmission risk of a particular patient by identifying the right categories. Since there are two class labels, low-risk and high-risk, this is a binary classification problem. Due to the low prevalence of HF readmissions, which is about 20%, this is a class imbalance problem. Literally, class imbalance refers to the fact that different categories are not represented equally (Chawla, 2010). Class imbalance is a common issue in clinical and medical fields, since the prevalence of certain diagnosis/disease is very low in the population (Mazurowski et al., 2008). For example, the incidence rate of breast cancer is 0.124% per year according to the statistics of the National Cancer Institute (National Cancer Institute, 2012). In such cases, the classification models will focus on the negative class, which is not the interest group of research. Therefore, the classification models for those problems have to compensate for the impact

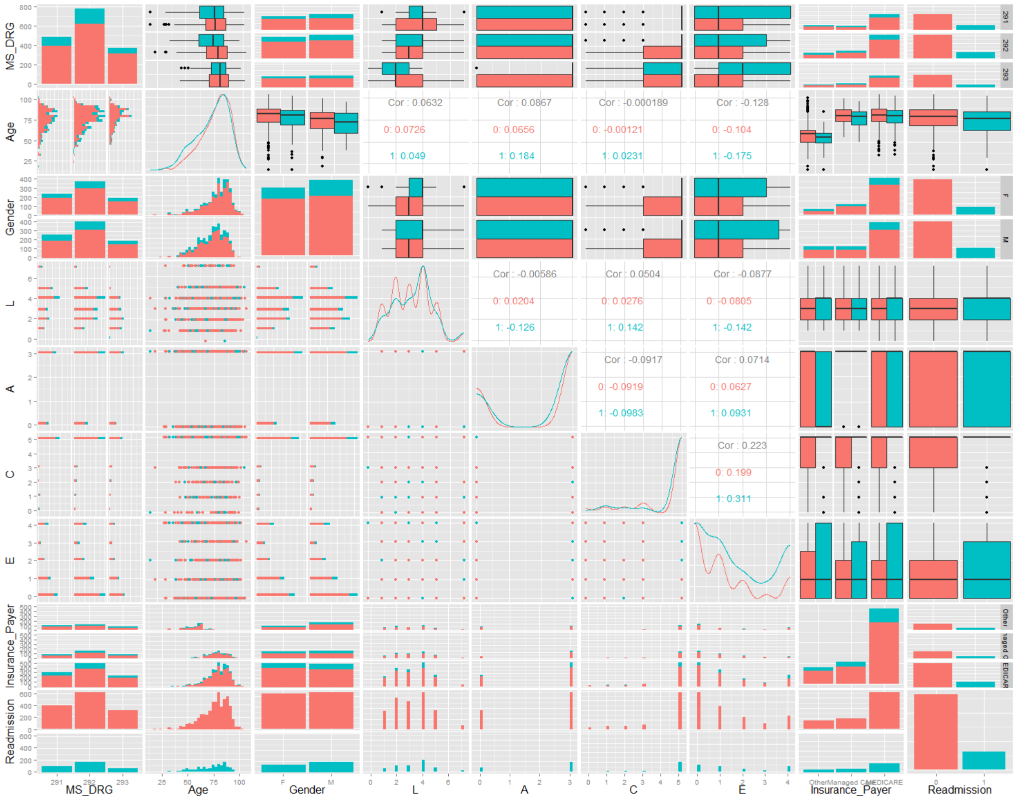

Figure 1: Descriptive statistics of raw data derived from medical records

of imbalanced datasets.

Generally, there are two compensation strategies that are used to obtain balanced classes in data mining: over-sampling and under-sampling. Over-sampling creates more input patterns from the minority class, whereas under-sampling removes some input patterns from the majority class (Chawla, 2010). In this study, a random over-sampling technique is implemented in which observations from the underrepresented class are randomly sampled and replicated to create balance between the two classes. Under-sampling, due to its inferior performance and the possibility of losing important patterns, is not implemented in this research (Mazurowski et al., 2008).

3.3. Radial Basis Function Neural Networks

An artificial neural network (ANN) is a popular approach in data mining that can be utilized to perform classification, clustering and function approximation. ANNs have been studied for many years, and a variety of networks and learning algorithms have been proposed to solve different problems. As a type of supervised learning, classification is performed under the guidance of targets, which indicates the class label of an instance. The ANN does not require a linear relationship between the independent attributes and the target, and allows for a combination of multiple training algorithms and easy generalization (Tu, 1996). Basically, an ANN consists of an input layer, hidden layer(s) and an output layer. Each layer has a number of neurons connected with other units that indicate the learned correlations, and the transfer function at the output layer transforms the input signal from antecedent units to obtain the output of the network.

A radial basis function neural network (RBFNN) is developed based on interpolation theory and consists of three layers: an input layer, a hidden layer and an output layer (Lu and Saratchandran, 1998). In the hidden layer, a unit takes the input vectors and performs a nonlinear computation based on a RBF for which a Gaussian function is commonly applied (Haykin, 2008). Since the impact of an input is related to its distance from a particular center, the RBFNN is locally tuned. From the hidden layer to the output layer, a weight matrix is trained under the guidance of the targets. Therefore, RBFNN is considered a combination of unsupervised learning and supervised learning.

Back-propagation (BP) is a very popular training algorithm in neural networks since it can be used to solve problems in various domains. A backpropagation neural network (BPNN) consists of an input layer, at least one hidden layer and an output layer. Most of the implemented BPNNs have one to two hidden layers. The training of the BPNN includes a forward process and a backward process. In the feed-forward process, the function signals are computed through the network, and the error is computed at the output layer by comparing the network output with the target. Weights are updated in the backward process (Haykin, 2008). The iterative training process continues to minimize an error function until the stopping criteria are met.

3.4. Random Forest

Random forest algorithm, one of the recent data mining algorithms for classification and regression, has received tremendous attention from industrial and academic researchers because of its simplicity and ensemble learning characteristics (Breiman, 2001). Ensemble learning classifiers leverage the advantages of each incorporated weak algorithm to produce a stronger classification accuracy. Random forecast algorithms introduce the random sampling process with replacements in the model. Unlike transitional single-decision tree models, the best classifier at each node is identified by a random subset of all the predictors (Liaw and Wiener, 2002). In this study, the patient readmission risk is predicted by the majority votes from all the decision trees in the random forest. Although the random forest may require more computational resources, such as storage spaces, random forest algorithms have demonstrated their strengths in the prediction accuracy, over-fitting avoidances and scalability, which is preferred by practitioners and researchers in data exploitations (Verikas et al., 2011).

3.5. Particle Swarm Optimization Based SVM

SVM is a statistical learning method proposed by (Vapnik, 1995). Here, the inputs are denoted as $\mathbf{X}$ and the output as $\mathbf{Y}$. Given a set of training data $(\mathbf{x}_i, y_i)$, where i=1, $\cdots$, N, $\mathbf{x}_i \in \mathbf{R}^d$, $y_i \in \{-1, 1\}$. Suppose there are some hyperplanes that separate the data points with different class labels. The hyperplane H is defined as $\mathbf{wx}+b = 0$ and the perpendicular distance between the hyperplane and the origin is $\frac{|b|}{\|\mathbf{w}\|}$ when $\mathbf{w}$ is normal to H (Burges, 1998). For a binary classification problem, the data points in the negative class satisfy $\mathbf{wx} + b \leq -1$, while those inthe positive class satisfy $\mathbf{wx} + b \geq 1$. Accordingly, there are two hyperplanes defined as: $H_1 : \mathbf{wx} + b = -1$ and $H_2 : \mathbf{wx} + b = 1$ (Tan et al., 2005). Those data points lying on Hyperplanes H_1 and H_2 are support vectors, circled in Figure 2 and 3. The objective of training an SVM classifier is to find a hyperplane that can classify inputs into correct classes with a maximum margin between H_1 and H_2 for better generalization (Wu et al., 2007). The geometric illustration of SVM is presented in Figure 2. Since the training of SVMs is based on support vectors, i.e. a subset of training data, the model complexity is greatly reduced and the generalizability of learning machines is improved.

Since the margin equals $2/\|\mathbf{w}\|$, the maximization problem is converted to the minimization of $\|\mathbf{w}\|/2$, and it is subjected to the constraint: $y_i \left(\mathbf{wx}_i + b\right) - 1 \geq 0$ (Burges, 1998). Lagrangian multipliers α_i, $i = 1, \cdots, N$, are implemented to solve this problem by transforming the constrained optimization

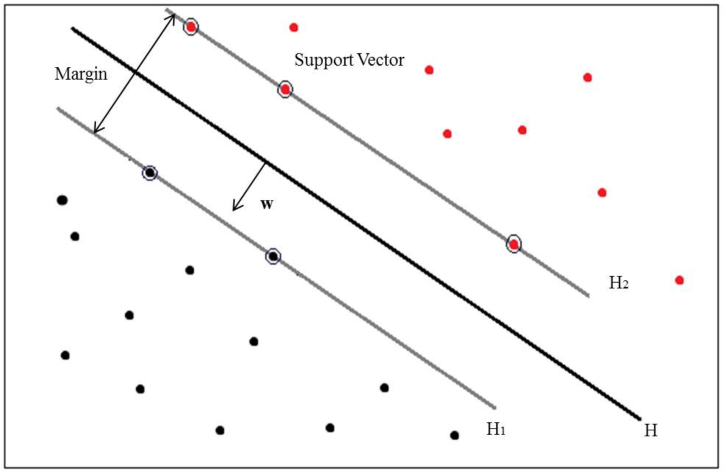

Figure 2: SVM Hyperplanes in Binary Classification

into an unconstrained optimization (Tan et al., 2005), defined as

$$L_p = \frac{1}{2}\left\|\mathbf{w}\right\|^2 - \sum_{i=1}^{N} \alpha_i y_i \left(\mathbf{w}\mathbf{x}_i + b - 1\right), \alpha_i \geq 0. \tag{1}$$

For nonlinearly separable problems, kernel method is implemented to transform it to a linearly separable problem. When there is a set of nonlinearly separable data points in input space O, a kernel function $K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \varphi\left(\mathbf{x}_i\right)^T \cdot \varphi\left(\mathbf{x}_j\right)$ can transform $\mathbf{x}$ from input space O to feature space Z in order to make the data points linearly separable (Sanchez A., 2003). The objective of using kernel method is also graphically illustrated in Figure 3 with a binary classification problem. There are three kernel functions implemented in this study, which are presented in Table 3.

Table 3: Three Types of Kernel Functions

Kernel Type	Function
Linear Function	$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathbf{x}_i^T \cdot \mathbf{x}_j$
Polynomial Function	$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(a\mathbf{x}_i^T \mathbf{x}_j + c\right)^m, a \geq 0.$
Radial Basis Function (RBF)	$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = e^{-\beta\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}$

However, it is difficult to find a hyperplane that can separate data points completely and correctly in some problems. Such a separation may result in

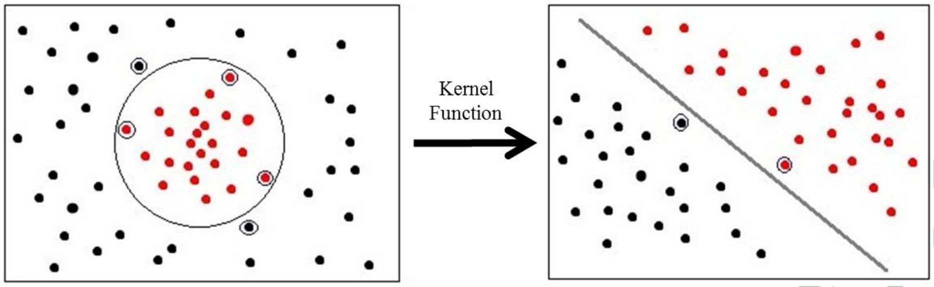

Figure 3: Kernel Function for Nonlinearly Separable Classification

a very complex hyperplane and reduce the generalizability of classifiers. To compensate for these issues, a soft margin is introduced into the model to allow some degree of violation (Haykin, 2008).

Although the SVM learning algorithm generally has a good performance and robust statistical foundation, the quality of an SVM classifier is largely influenced by the parameters of kernel functions (Jin et al., 2010). An exhaustive search to find the most appropriate parameter is too time-consuming, and so, heuristic and meta-heuristic approaches have been implemented to solve this problem. In this study, a swarm intelligence technique, Particle Swarm Optimization (PSO), is implemented for the parameter search in order to build an optimal classifier. The PSO is an efficient evolutionary optimization algorithm derived from the social behavior of flocks. In the food searching process, an individual's movement is attracted to the best solution found by itself as well as the best solution found by its neighborhood (Bratton and Kennedy, 2007). Individuals in the population are represented by particles, and each particle i consists of two components: position $\mathbf{p}_i$ and velocity $\mathbf{v}_i$. The population updates at each iteration by adjusting the position and the velocity of each particle (Esmaeili and Mozayani, 2009). The updates on velocity and position are defined in Equations (2) and (3) respectively:

$$\mathbf{v}_i\left(t+1\right) = \omega \mathbf{v}_i\left(t\right) + c_1 r_1 \left(\mathbf{p}_{ibest} - \mathbf{p}_i\right) + c_2 r_2 \left(\mathbf{p}_{gbest} - \mathbf{p}_i\right), \qquad (2)$$

$$\mathbf{p}_i\left(t+1\right) = \mathbf{p}_i\left(t\right) + \mathbf{v}_i\left(t+1\right), \qquad (3)$$

where t and ω represent iteration and inertia weight respectively, c_1 and c_2 are acceleration factors, and r_1 and r_2 are random numbers between 0 and 1 in order to take stochastics into consideration. Here $\mathbf{p}_{ibest}$ and $\mathbf{p}_{gbest}$ represent the best solution found by particle i and the best solution found by population respectively. The quality of a solution is measured through a fitness function defined specifically by a particular problem. Since the objective of this readmission prediction problem is to develop an SVM classifier that clusters patients into the correct classes, the fitness function is derived from a performance metric, the overall accuracy of which is illustrated in the following section. Therefore, the PSO is implemented to search for the solution that can optimize the value of the fitness function. The training process of this PSO-SVM algorithm is summarized in Table 4.

Table 4: The PSO-SVM Algorithm

```
Step 1   Set parameters for PSO:
           population size $S$,
           number of iterations $\leftarrow T$,
           fitness function $\leftarrow f$
Step 2   Initialize PSO:
           randomize $\mathbf{p}_{i1}$ and $\mathbf{v}_{i1}$
Step 3   Iterate PSO-SVM:
           for $t \leftarrow 1$ to $T$:
             construct SVMs with parameter $\mathbf{p}_{it}$
             evaluate performance of SVMs and output $f\left(\mathbf{p}_{it}\right)$
             if $f\left(\mathbf{p}_{it}\right) \geq f\left(\mathbf{p}_{ibest}\right)$
               $\mathbf{p}_{ibest} \leftarrow \mathbf{p}_{it}$
               if $f\left(\mathbf{p}_{ibest}\right) \geq f\left(\mathbf{p}_{gbest}\right)$
                   $\mathbf{p}_{gbest} \leftarrow \mathbf{p}_{ibest}$
             update $\mathbf{v}_{it}$ and $\mathbf{p}_{it}$
             until
               no improvement for more than $k$ iterations
Step 4   Return
           $\mathbf{p}_{gbest}$ and $f\left(\mathbf{p}_{gbest}\right)$
```

In this study, the position represents the spread of the RBF in SVM. At first, the position of each particle is initialized randomly, and the velocity of each particle is set to be 0. Also, it is predetermined that a possible position is between 0 and 10 in order to narrow the searching space and to accelerate the searching process. Both position and velocity are updated through iteration until the stopping criteria are met.

In binary classification, instances from both high- and low-risk groups are used to find the optimal hyperplane in order to classify patients correctly. However, in one-class classification, only high- or low-risk patients are adopted to find the classification hyperplane. Therefore, the other class of patients is identified as outliers since they are abnormal compared to the instances in the target class.

4. Experimental Results and Analysis

4.1. Performance Metrics

A classifier is evaluated by its complexity, required storage, training time, generalization, etc. Since the dataset for training is not very large, the storage and computation time required for training will not be considered as important measures to evaluate the model. The possible outcomes of a classification task can be interpreted as one of four categories:

1. True positive (TP): correctly classified as positive
2. False positive (FP): incorrectly classified as positive
3. True negative (TN): correctly classified as negative
4. False negative (FN): incorrectly classified as negative

The implemented performance metrics are accuracy, sensitivity and specificity. A positive pattern refers to a readmitted patient, whereas a negative pattern refers to a non-readmitted patient. Accuracy is the rate of correct classification and it is defined as

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{4}$$

Sensitivity, also known as recall, indicates the ability of a classifier to identify positive patterns (Seliya et al., 2009). It is defined as

$$Sensitivity = \frac{TP}{TP + FN}. \tag{5}$$

Specificity indicates the ability of a classifier to identify negative patterns and is defined as

$$Specificity = \frac{TN}{TN + FP}. \tag{6}$$

4.2. Training Result Analysis

In the section, the training process and results for proposed methodologies are discussed. Table 5 summarizes training accuracy for radial basis function neural networks, random forest, and PSO-SVM with radial basis kernel function. Detailed training process is discussed in the following subsequent sections.

Table 5: Training Performance of Proposed Methodologies

Training Model	Accuracy (%)
RBFNN	56.1
RF	87.6
PSO-SVM with RBF	83.8

4.2.1. Radial Basis Function Neural Networks

In the proposed RBFNN model, the parameters that have to be determined are the maximum number of neurons in the hidden layer as well as the spread of the Gaussian function in the hidden layer. Experiments are conducted using different numbers of hidden neurons and different spread values to search for a good RBFNN model. During the training process, hidden neurons are added progressively until it reaches the predefined maximum number of hidden units. Accordingly, training terminates when it reaches the maximum error allowed or the predetermined maximum hidden units. To control the complexity of this model, the maximum number of hidden neurons is initially set as 50. When the number of hidden neurons is 50, the performance of the classifier improves at the beginning but then starts to degrade, and finally it converges as the spread parameter decreases. Since the performance is not satisfactory, a classifier with more hidden neurons is developed and tested. When the maximum number of hidden neurons is set at 150, better performance is achieved when the spread parameter is around 0.01. Additionally, when the maximum number of hidden neurons increases to larger values (e.g. 500), the network performance stops improving and becomes stabilized when the number of hidden neurons reaches around 120. Here the performance refers to the value of mean squared error (MSE). When the MSE decreases, the performance improves.

Therefore, to balance the performance of the model and its complexity, the selected parameter set of the RBFNN model is the optimal parameter setting for the RBFNN after oversampling is 120 (maximum number of hidden neurons) and 0.01 (spread).

4.2.2. Random Forest Model Training

Random forest models provide a randomness layer based on the traditional classification and regression tree. The split at each node can be the best split among a subset of all variables. One of the classification trees in the random forest model is illustrated in Figure 4. At each node in the classification tree, the decision separator is selected from seven predictors since it gives the highest training accuracy as shown in Figure 5.

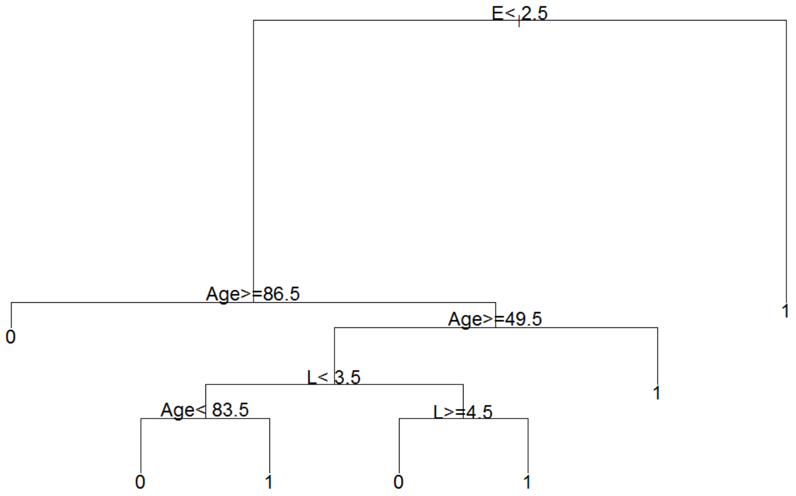

Figure 4: A sample of classification tree in the random forest model

4.2.3. PSO-SVM Training

SVMs with three kernel functions (linear, polynomial and radial basis functions) are built and trained in programing language R. The training and testing results have been summarized in Table 6. Since there are not many possible parameter values for the linear and polynomial kernels and the low prediction accuracy, the PSO is not implemented for parameter tuning of these two kernel functions. Instead, the exhaustive search is used.

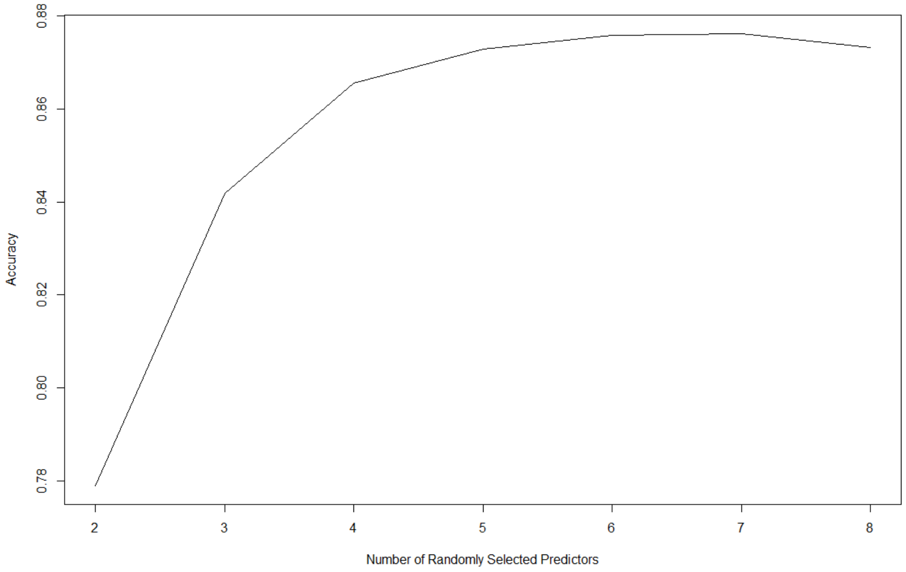

Figure 5: Accuracy comparisons over number of features at node level of random forest

The PSO is also developed in programming language R to search for the optimal parameter p (spread) for the RBF-based SVMs. In the experiment, the inertia weight ω and acceleration factors c_1 and c_2 are 2, 1 and 2 respectively. Figure 6 presents the best solutions of the population in each iteration during the parameter tuning process after over-sampling. The fitness function of the PSO can be defined as the average accuracy of the five folds cross-validation with parameter p_{it}:

$$f(\mathbf{p}_{it}) = \frac{1}{5}\left[\sum_{k=1}^{5}\left(\frac{TP+TN}{TP+FP+TN+FN}\right)_k\right]\Bigg|_{p=p_{it}}. \tag{7}$$

Given Equation (7), a larger value of the fitness function indicates a better classifier. At the beginning, the best solutions found by population make the corresponding SVMs perform poorly. However, when the search continues for more generations, the solutions converge, and the parameter tuning process terminates after no improvement for 10 iterations. The best spreads found by PSO is $\sigma = 4.9007$ for the SVM with radial basis function after over-sampling.

In addition, Table 6 presents the testing results of those SVMs. The SVMs

Table 6: Experimental results of SVM classifiers with various kernel functions

SVM Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)
Linear*	51.0	3.9	97.9
Polynomial*	50.8	1.3	100.0
RBF-based*	83.8	93.6	22.2
Linear**	50.6	3.0	98.9
Polynomial**	52.7	51.9	55.7
RBF-based**	69.5	78.7	35.7

* Training

** Testing

with RBF-based functions have better classification abilities than the SVMs with linear and polynomial functions. The performance of SVMs with linear and polynomial functions does not improve after over-sampling. However, the testing performance of RBF-based SVM is greatly improved, especially with the classification of high-risk patients. Therefore, of all the SVM classifiers, the RBF-based SVM trained after over-sampling demonstrates the best performance.

4.3. Proposed Methodology Experimental Testing Result Comparisons

Considering the generalization ability of those classifiers, the testing performance is more important than the training performance. As the detailed prediction testing results shown in Table 7 and Figure 7, the proposed methods tend to have better generalization results when trained with the data after over-sampling. Moreover, the PSO-SVM models demonstrate more accurate classification performances than other models, especially on high-risk patients, which are the group of interest under study. The high sensitivity indicates the proposed PSO-SVM's capacity in correctly identifying readmitted patients, with high accuracy compared to non-readmitted patient readmission. It also shows the implementation potentials for practitioners since the misclassification of non-readmitted patients has less impact and fewer consequences than readmitted patient misclassification.

5. Conclusions and Future Work

In this study, data mining and evolutionary algorithms are implemented to develop accurate readmission prediction models. Data mining algorithms

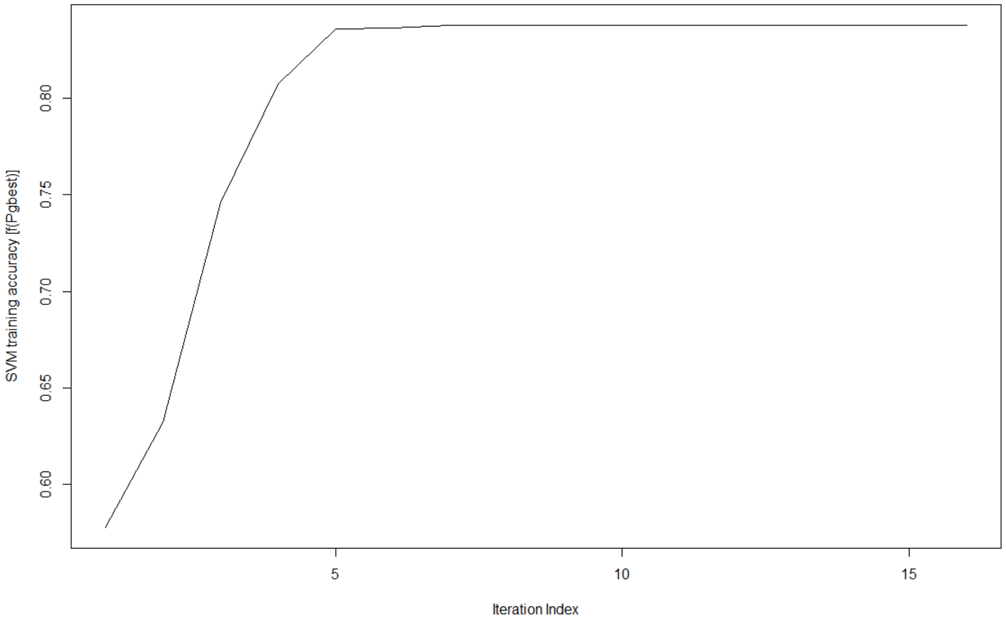

Figure 6: Patient readmission prediction accuracy comparisons

are used to explore a classifier to distinguish potential readmitted and non-readmitted patients. Evolutionary algorithms leverage their advantages in parameter optimizations to further improve the prediction accuracy with fine-tuning the parameters. The historical data of HF patients is utilized to learn the implicit patterns in order to correctly classify patients into the low-risk and high-risk groups. The imbalanced classes in the training dataset are likely to affect the performance of classifiers since some models may only focus on the majority class. However, the research interest of this study is the minority class, i.e. high-risk patients. Therefore, to compensate for the impact of class imbalance, random over-sampling is used as a data preprocessing technique to make the two classes balanced for training.

The proposed PSO-SVM classifier is a popular data mining technique based on robust statistical background. Given that some problems are not linearly separable, kernel function is incorporated into the SVM, and the instances are mapped into another higher dimensional space to make them linearly separable. In this study, three types of kernels are used: linear,

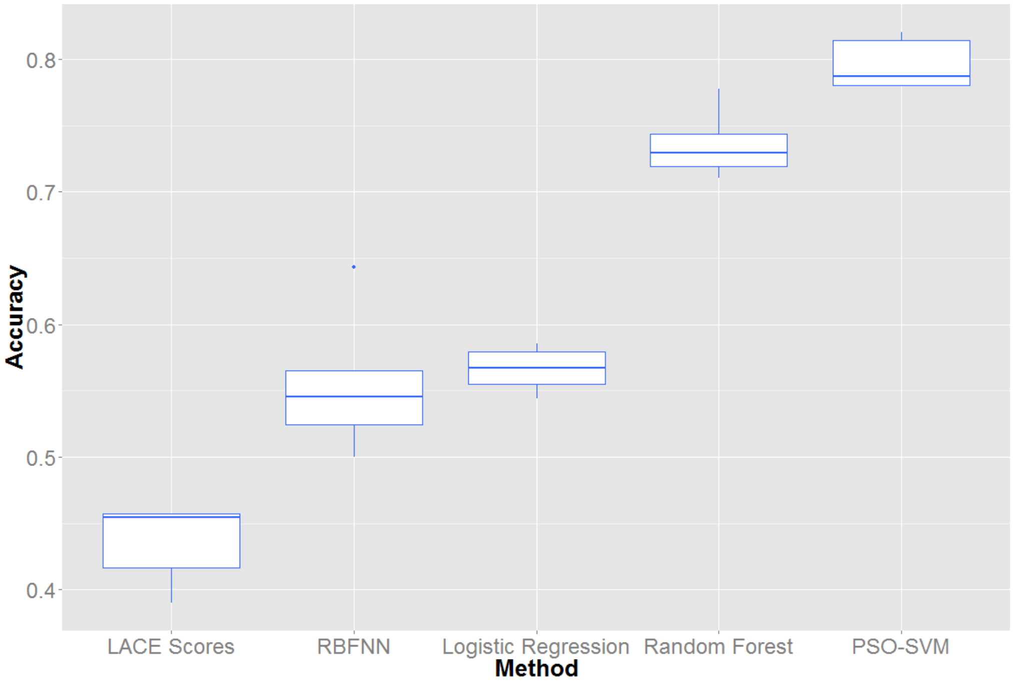

Figure 7: Prediction testing accuracy based on 10 repeated 5-fold cross validations

polynomial and radial basis function. The order in the polynomial function and the spread in the RBF are the two parameters that have to be determined. Since the possible values of the first parameter are very limited, an exhaustive search is implemented. However, the search space of the second parameter is very large, and the exhaustive search is infeasible, so instead, the metaheuristic method PSO is used. SVMs are trained with the data before and after over-sampling. The experimental results demonstrate that the RBF-based SVM has the best performance among all other SVMs in this research. Moreover, the RBF trained with data after over-sampling has better generalization ability and higher sensitivity in classifying readmitted patients than other models.

To compare the performance of the proposed classifiers with other readmission prediction models, random forest algorithms, RBFNN, LACE scores and logistic regression models are also tested on the same dataset. The experimental results indicate that the random forest and the RBF-based SVM using PSO for parameter tuning outperform the previous traditional meth-

Table 7: Prediction testing accuracy comparisons of proposed methodologies

SVM Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)
LACE Scores	43.5	51.8	21.8
RBFNN	54.6	56.1	49.3
Logistic Regression	57.9	60.5	49.3
Random Forest	74.4	87.4	30.7
PSO-SVM with RBF	78.4	97.3	8.6

ods for hospital readmission predictions. Also, these two models provide high sensitivity used to correctly predict readmitted patients, which are more desired by the healthcare practitioners. The PSO-SVM significantly improve the current patient readmission risk prediction accuracy.

Although the PSO-SVM outperforms other prediction models, the training and parameter tuning consumes tremendous computational resources and time. The processing training and parameter tuning need to be significantly reduced for tackling large-scale patient data records. One limitation of the PSO-SVM is that over-used parameter tuning may lead to future over-fitting in training process. A criterion to terminate particle swarm intelligence base parameter tuning may be developed for avoiding over-fitting issues. For compensation strategies that address the class imbalance, other over-sampling approaches, such as focused over-sampling and synthetic minority over-sampling technique (SMOTE), can be implemented to introduce more useful and representative data. Focused over-sampling replicates instances that are close to the boundary so that it can help reduce the likelihood of over-fitting to some extent. The SMOTE introduces new instances into the minority set, and the K-nearest neighbor method can be used to make the data more representative. Moreover, since the misclassification cost on a high-risk patient is much more than that on a low-risk patient, F-measure can be adjusted as an evaluation criterion for model selections. Additionally, the Markov decision process (MDP) can be studied to evaluate the effectiveness and to determine the optimal timing of certain interventions such as follow-ups. The proposed prediction models in this research can also be expanded to some other DRGs, such as AMI and PN, to enlarge the benefits.

References

Allaudeen, N., Schnipper, J. L., Orav, E. J., Wachter, R. M., and Vidyarthi, A. R. (2011). Inability of providers to predict unplanned readmissions. *Journal of General Internal Medicine*, 27(7).

Amarasingham, R., Moore, B. J., Tabak, Y. P., Drazner, M. H., Clark, C. A., Zhang, S., Reed, W. G., Swanson, T. S., Ma, Y., and Halm, E. A. (2010). An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Medical Care*, 48(11):981–988.

Anderson, G. F. and Steinberg, E. P. (1985). Predicting hospital readmissions in the medicare population. *Inquiry*, 22(3):pp. 251–258.

Billings, J., Dixon, J., Mijanovich, T., and Wennberg, D. (2006). Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *BMJ: British Medical Journal*, 333(7563):pp. 327–330.

Bottle, A., Aylin, P., and Majeed, A. (2006). Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. *JRSM*, 99(8):406–414.

Boult, C., Dowd, B., McCaffrey, D., Boult, L., Hernandez, R., and Krulewitch, H. (1993). Screening elders for risk of hospital admission. *Journal of the Am Geriatrics Society*, 41(8):811–817.

Braga, P., Portela, F., Santos, M. F., and Rua, F. (2014). Data mining models to predict patient's readmission in intensive care units.

Bratton, D. and Kennedy, J. (2007). Defining a standard for particle swarm optimization. In *Swarm Intelligence Symposium, 2007. SIS 2007. IEEE*, pages 120–127.

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

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Centers for Medicare and Medicaid Services (2012a). CMS set to penalize hospitals with high readmission rates. http://www.californiahealthline.org. Accessed on 2012-11-28.

Centers for Medicare and Medicaid Services (2012b). National medicare readmission findings: Recent data and trends.

Centers for Medicare and Medicaid Services (2012c). Readmissions reduction program. http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html. Accessed on 2012-09-20. Accessed: 2012-11-10.

Chawla, N. V. (2010). *Data Mining and Knowledge Discovery Handbook*, chapter Data mining for imbalanced datasets: an overview, pages 875–886. Springer.

Claus, E. B., Risch, N., and Thompson, W. D. (1994). Autosomal dominant inheritance of early-onset breast cancer. implications for risk prediction. *Cancer*, 73(3):643–651.

Dharmarajan, K., Hsieh, A. F., Lin, Z., Bueno, H., Ross, J. S., Horwitz, L. I., Barreto-Filho, J. A., Kim, N., Bernheim, S. M., Suter, L. G., Drye, E. E., and Krumholz, H. M. (2013). Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. *JAMA*, 309(4):355–363.

Domchek, S. M., Eisen, A., Calzone, K., Stopfer, J., Blackwood, A., and Weber, B. L. (2003). Application of breast cancer risk prediction models in clinical practice. *Journal of Clinical Oncology*, 21(4):593–601.

Esmaeili, A. and Mozayani, N. (2009). Adjusting the parameters of radial basis function networks using particle swarm optimization. In *Computational Intelligence for Measurement Systems and Applications, 2009. CIMSA '09. IEEE International Conference on*, pages 179–181.

Garrison, G. M., Mansukhani, M. P., and Bohn, B. (2013). Predictors of thirty-day readmission among hospitalized family medicine patients. *The Journal of the American Board of Family Medicine*, 26(1):71–77.

Grafa, C. E., Giannellia, S. V., Herrmanna, F. R., Sarasinb, F. P., Michela, J.-P., Zekrya, D., and Chevalleya, T. (2012). Identification of older patients at risk of unplanned readmission after discharge from the emergency department - comparison of two screening tools. *Swiss Medical Weekly*, 141:1–9.

Grimes, D. A. and Schulz, K. F. (2002). Cohort studies: marching towards outcomes. *The Lancet*, pages 341–345.

Halfon, P., Eggli, Y., Prêtre-Rohrbach, I., Meylan, D., Marazzi, A., and Burnand, B. (2006). Validation of the potentially avoidable hospital readmission rate as a routine indicator of the quality of hospital care. *Medical Care*, 44(11):pp. 972–981.

Hammill, B. G., Curtis, L. H., Fonarow, G. C., Heidenreich, P. A., Yancy, C. W., Peterson, E. D., and Hernandez, A. F. (2011). Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization. *Circulation: Cardiovascular Quality and Outcomes*, 4(1):60–67.

Hasan, O., Meltzer, D., Shaykevich, S., Bell, C., Kaboli, P., Auerbach, A., Wetterneck, T., Arora, V., Zhang, J., and Schnipper, J. (2010). Hospital readmission in general medicine patients: A prediction model. *Journal of General Internal Medicine*, 25:211–219.

Haykin, S. (2008). *Neural Networks and Learning Machines*. Prentice Hall, Upper Saddle River, NJ, 3 edition.

Holloway, J. J., Medendorp, S. V., and Bromberg, J. (1990). Risk factors for early readmission among veterans. *Health Services Research*, 25(1 Pt 2):213–237.

Howell, S., Coory, M., Martin, J., and Duckett, S. (2009). Using routine inpatient data to identify patients at risk of hospital readmission. *BMC Health Services Research*, 9:96.

Jeejeebhoy, K. N., Keller, H., Gramlich, L., Allard, J. P., Laporte, M., Duerksen, D. R., Payette, H., Bernier, P., Vesnaver, E., Davidson, B., Teterina, A., and Lou, W. (2015). Nutritional assessment: comparison of clinical assessment and objective variables for the prediction of length of hospital stay and readmission. *The American journal of clinical nutrition*.

Jencks, S. F., Williams, M. V., and Coleman, E. A. (2009). Rehospitalizations among patients in the medicare fee-for-service program. *New England Journal of Medicine*, 360(14):1418–1428.

Jin, G., Jin-ye, P., and Zhan, L. (2010). Application of improved pso-svm approach in image classification. In *Photonics and Optoelectronic (SOPO), 2010 Symposium on*, pages 1–4.

Kansagara, D., Englander, H., and Salanitro, A. (2011). Risk prediction models for hospital readmission: A systematic review. *JAMA*, 306(15):1688–1698.

Kociol, R. D., Lopes, R. D., Clare, R., Thomas, L., Mehta, R. H., Kaul, P., Pieper, K. S., Hochman, J. S., Weaver, W. D., Armstrong, P. W., Granger, C. B., and Patel, M. R. (2012). International variation in and factors associated with hospital readmission after myocardial infarction. *JAMA*, 307(1):66–74.

Koehler, B. E., Richter, K. M., Youngblood, L., Cohen, B. A., Prengler, I. D., Cheng, D., and Masica, A. L. (2009). Reduction of 30-day postdischarge hospital readmission or emergency department (ed) visit rates in high-risk elderly medical patients through delivery of a targeted care bundle. *Journal of Hospital Medicine*, 4(4):211–218.

Kramer, A., Higgins, T., and Zimmerman, J. (2012). Intensive care unit readmissions in u.s. hospitals: patient characteristics, risk factors, and outcomes. *Critical Care Medicine*, 40(1):3–10.

Krumholz, H. M., Chen, Y.-T., Wang, Y., Vaccarino, V., Radford, M. J., and Horwitz, R. I. (2000). Predictors of readmission among elderly survivors of admission with heart failure. *American Heart Journal*, 139(1):72 – 77.

Liaw, A. and Wiener, M. (2002). Classification and regression by random forest. *R news*, 2(3):18–22.

Lindstrom, J. and Tuomilehto, J. (2003). The diabetes risk score: A practical tool to predict type 2 diabetes risk. *Diabetes Care*, 26(3):725–731.

Lu, Y., S.-N. and Saratchandran, P. (1998). Performance evaluation of a sequential minimal radial basis function (RBF) neural network learning algorithm. 9(2):308–318.

Malnick, S., Balla, U., and Schattner, A. (2008). Early readmissions to the department of medicine as a screening tool for monitoring quality of care problems. *Medicine*, 87(5):294–300.

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., and Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2-3):427–436.

Minott, J. (2008). Reducing hospital readmissions. Technical report.

Morrissey, E., McElnay, J., Scott, M., and McConnell, B. (2003). Influence of drugs, demographics and medical history on hospital readmission of elderly patients. *Clinical Drug Investigation*, 23(2):119–128.

Mukti, M. Z. R. and Ahmed, F. (2013). Early detection of lung cancer risk using data mining. *Asian Pacific Journal of Cancer Prevention*, 14(1):595–598.

Nahar, J., Imam, T., Tickle, K. S., Ali, A. S., and Chen, Y.-P. P. (2012). Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer. *Expert Systems with Applications*, 39(16):12371–12377.

National Cancer Institute (2012). SEER stat fact sheets: Breast. http://seer.cancer.gov. Accessed on 2012-10-30.

Novotny, N. L. and Anderson, M. A. (2008). Prediction of early readmission in medical inpatients using the probability of repeated admission instrument. *Nursing Research*, 57(6):406–415.

Parmigiani, G., Berry, D. A., and Aguilar, O. (1998). Determining carrier probabilities for breast cancer susceptibility genes BRCA1 and BRCA2. *The American Journal of Human Genetics*, 62(1):145–158.

Philbin, E. F. and DiSalvo, T. G. (1999). Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. *Journal of the American College of Cardiology*, 33(6):1560 – 1566.

QualityNet (2012). Readmission measures overview: Publicly reporting risk-standardized, 30-day readmission measures for AMI, HF, PN, HWR, and THA/TKA. https://www.qualitynet.org/dcs. Accessed on 2012-11-09.

Ross, J., Mulvey, G., Stauffer, B., and et al (2008). Statistical models and patient predictors of readmission for heart failure: A systematic review. *Archives of Internal Medicine*, 168(13):1371–1386.

Sanchez A., V. D. (2003). Advanced support vector machines and kernel methods. *Neurocomputing*, 55(1–2):5–20.

Seliya, N., Khoshgoftaar, T., and Van Hulse, J. (2009). A study on the relationships of classifier performance metrics. In *Tools with Artificial Intelligence, 2009. ICTAI '09. 21st International Conference on*, pages 59–66.

Silverstein, M. D., Qin, H., Mercer, S. Q., Fong, J., and Haydar, Z. (2008). In *Proceedings (Baylor University. Medical Center)*, pages 363–72.

Siontis, G. C. M., Tzoulaki, I., Siontis, K. C., and Ioannidis, J. P. A. (2012). Comparisons of established risk prediction models for cardiovascular disease: systematic review. *BMJ*, 344.

Smith, D. M., Norton, J. A., and Mcdonald, C. J. (1985). Nonelective readmissions of medical patients. *Journal of Chronic Diseases*, 38(3):213 – 224.

Sufi, F. and Khalil, I. (2011). Diagnosis of cardiovascular abnormalities from compressed ecg: a data mining-based approach. *Information Technology in Biomedicine, IEEE Transactions on*, 15(1):33–39.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison Wesley, us edition.

The Himmelfarb Health Sciences Library (2011). Study design 101. http://www.gwumc.edu/library. Accessed on 2012-12-20.

Thomas, J. W. (1996). Does risk-adjusted readmission rate provide valid information on hospital quality? *Inquiry*, 33(3):pp. 258–270.

Tran, T., Luo, W., Phung, D., Gupta, S., Rana, S., Kennedy, R. L., Larkins, A., and Venkatesh, S. (2014). A framework for feature extraction from hospital medical data with applications in risk prediction. *BMC bioinformatics*, 15(1):6596.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11):1225–1231.

van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., Austin, P. C., and Forster, A. J. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Canadian Medical Association Journal*, 182(6):551–557.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Verikas, A., Gelzinis, A., and Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2):330–349.

Whitlock, T., Tignor, A., Webster, E., Repas, K., Conwell, D., Banks, P., and Wu, B. (2011). A scoring system to predict readmission of patients with acute pancreatitis to the hospital within thirty days of discharge. *Clinical Gastroenterology and Hepatology*, 9(5):175–80.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J., and Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14(1):1–37.

Zheng, B., Yoon, S. W., and Lam., S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Systems with Applications*, 41(4):1476–1482.

Highlights

- Risk predictions of hospital readmission for heart failure patients are investigated
- RBFNN, RF and PSO-SVM models are used to predict patient readmission risk
- The PSO-SVM achieves 78.4% on overall accuracy and 97.3% on sensitivity
- The PSO-SVM outperforms over traditional prediction models, including LACE scores