Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients

Youn-Jung Son, RN, PhD1, Hong-Gee Kim, PhD2, Eung-Hee Kim, ME2, Sangsup Choi, PhD2, Soo-Kyoung Lee, RN, Doctoral Candidate2

1Department of Nursing, Soonchunhyang University, Cheonan; 2Biomedical Knowledge Engineering Laboratory, Seoul National University, Seoul, Korea

Objectives: Heart failure (HF) is a progressive syndrome that marks the end-stage of heart diseases, and it has a high mortality rate and significant cost burden. In particular, non-adherence of medication in HF patients may result in serious consequences such as hospital readmission and death. This study aims to identify predictors of medication adherence in HF patients. In this work, we applied a Support Vector Machine (SVM), a machine-learning method useful for data classification.

Methods: Data about medication adherence were collected from patients at a university hospital through self-reported questionnaire. The data included 11 variables of 76 patients with HF. Mathematical simulations were conducted in order to develop a SVM model for the identification of variables that would best predict medication adherence. To evaluate the robustness of the estimates made with the SVM models, leave-one-out cross-validation (LOOCV) was conducted on the data set.

Results: The two models that best classified medication adherence in the HF patients were: one with five predictors (gender, daily frequency of medication, medication knowledge, New York Heart Association [NYHA] functional class, spouse) and the other with seven predictors (age, education, monthly income, ejection fraction, Mini-Mental Status Examination-Korean [MMSE-K], medication knowledge, NYHA functional class). The highest detection accuracy was 77.63%. Conclusions: SVM modeling is a promising classification approach for predicting medication adherence in HF patients. This predictive model helps stratify the patients so that evidence-based decisions can be made and patients managed appropriately. Further, this approach should be further explored in other complex diseases using other common variables.

Keywords: Heart Failure, Medication, Patient Adherence, Support Vector Machine

I. Introduction

The support vector machine (SVM) is a relatively new classification or prediction method developed by Cortes and Vapnik [1] in the 1990s as a result of the collaboration between the statistical and the machine learning research community. SVM tries to classify cases by finding a separating boundary called hyperplane. The main advantage of the SVM is that it can, with relative ease, overcome ‘the high dimensionality problem’, i.e., the problem that arises when there is a large number of input variables relative to the number of available observations [2]. Also, because the SVM approach is data-driven and possible without theoretical framework, it
may have important discriminative power for classification, especially in cases where sample sizes are small [3]. This technique has recently been used to improve methods for detecting diseases in clinical settings [4,5]. Moreover, SVM has demonstrated high performance in solving classification problems in bioinformatics [6,7]. In this study, we use SVM to predict medication adherence of HF patients using common variables that can be obtained with relative ease.

Over the last several decades, HF has increased in both incidence and prevalence worldwide [8]. HF is a condition characterized by unpleasant symptoms, high mortality, recurrent hospitalization and significant cost burden [9,10]. HF is one of the many chronic conditions that require patients to adhere to a lifelong therapeutic regimen such as medication, exercise, diet, fluid restriction and daily weight measurements to achieve optimal outcomes [11]. Especially, patients with HF require multiple medications to decrease morbidity and mortality and to attain maximal therapeutic benefits [10,12]. Estimates of medication adherence rates in patients with HF range from 7% to 90%, depending on the definition and how adherence is measured [10]. Adherence rates decrease as patients stay longer with the condition. Medication nonadherence is thought to be the most common cause of HF exacerbation and subsequent hospital readmission in patients with HF. Because of the importance of medication adherence in managing HF, a full understanding of the factors associated with medication adherence in patients with HF is needed so that effective interventions to improve medication adherence can be developed [13].

In Western countries, studies related to application of SVM in cardiovascular patients have been continuously conducted [14,15]. On the other hand, in Korea, classification algorithm or pattern analysis was mainly focused on Bayesian and artificial neural networks in broad healthcare domain [16,17]. Up to date, SVM have not yet been studied for prediction of medication adherence in HF patients in Korea nor in other countries. Therefore, the aim of this study is to investigate the use of a SVM based classification model for determining the predictors of medication adherence in HF patients.

**II. Methods**

**1. Data and Data Preprocessing**

For the model building, Ninety four consecutive HF patients who visited at outpatient clinics from January to April 2010 participated in the study. As eighteen patients with 1 or more missing data points were excluded, 76 (21 men, 55 women) were used for SVM models. The mean age of the HF patients was 74.8 years (SD, 5.9). Table 1 shows the demographic and clinical characteristics of the patients with HF.

The study was approved by the ethics committee of the Soochunhyang University Hospital prior to the start of the data collection. The committee waived informed consent because of the non-interventional study design.

**2. Measurements**

The questionnaire was designed to yield information about

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Category</th>
<th>No. (%)</th>
<th>Mean ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Men</td>
<td>21 (27.6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Women</td>
<td>55 (72.4)</td>
<td></td>
</tr>
<tr>
<td>Age (yr)</td>
<td></td>
<td>74.8 ± 5.9</td>
<td></td>
</tr>
<tr>
<td>Spouse</td>
<td>No</td>
<td>39 (51.3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>37 (48.7)</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>Illiteracy</td>
<td>38 (51.3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Elementary school</td>
<td>23 (30.3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Middle school</td>
<td>6 (7.9)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>High school</td>
<td>7 (9.2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Above college</td>
<td>2 (2.6)</td>
<td></td>
</tr>
<tr>
<td>Monthly income</td>
<td>None</td>
<td>53 (69.7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,000-1,499</td>
<td>14 (18.4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,500-1,999</td>
<td>2 (2.6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>≥2,000</td>
<td>1 (1.3)</td>
<td></td>
</tr>
<tr>
<td>MMSE-K</td>
<td>21.5 ± 4.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration of heart failure diagnosis (mo)</td>
<td>32.1 ± 38.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ejection fraction (%)</td>
<td>57.8 ± 14.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NYHA</td>
<td>I</td>
<td>20 (26.3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>39 (51.3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>III/IV</td>
<td>17 (22.4)</td>
<td></td>
</tr>
<tr>
<td>Daily frequency of medication (per day)</td>
<td>26 (34.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>41 (53.9)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>9 (11.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>≥4</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Medication knowledge</td>
<td>12.8 ± 3.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

demographic characteristics such as age, gender, education, spouse and monthly income. To assess ability of cognition, we used Mini Mental Status Examination–Korean version (MMSE-K) by interviewing. Clinical data were obtained from medical records. New York Heart Association (NYHA) functional class was used as an index of functional impairment. The NYHA functional class identifies patients in one of four categories based on physical symptoms and activity restriction: Class I is no symptoms with ordinary activity and no limitations on activity; Class II is slight to moderate symptoms with normal activity and slight limitation of activity; Class III is moderate symptoms with less than normal activity and marked limitation of activity; Class IV is inability to carry out any physical activity without discomfort and symptoms may occur at rest. Ejection fraction (EF) was used as an index of cardiac function. EF was determined by left ventricular angiography or echocardiography.

We developed five self-report questions to assess patients’ medication knowledge. These items asked how well the patients knew the names, purposes, recommended doses, frequencies, and side effects of their medications. The 5 items with a 5-point Likert type scale was used. Internal consistency measured by Cronbach’s alpha was 0.83. A simple single item to measure medication adherence was used. The patient was categorized as either compliant or non-compliant.

3. Variable Selection
A total of 11 variables were used in the model building and analysis. The variables were selected because they either had shown to have an impact on medication adherence in HF patients in previous research [8-10,12,13] or were of potential clinical importance as indicated by a panel of experts. The 11 variables were gender, age, spouse, education, monthly income, and duration of HF diagnosis, daily frequency of medication, EF, MMSE-K, medication knowledge, and NYHA functional class.

4. Analytical Tool – SVM Basics
Among various data mining methods, SVM is well known for its discriminative power for classification, especially in the cases where sample sizes are small and a large number of features (variables) are involved (i.e., high-dimensional space). Kim et al. [18] demonstrated an outstanding performance of SVM on the classification for prognosis prediction of Class III malocclusion in comparison to a conventional statistical method, stepwise Guyon et al. [19] also showed that SVM performs much better than correlation based techniques in a gene selection problem for cancer classification with the feature selection method, recursive feature elimina-

5. Development of the SVM Based Models
In practice, however, it is frequently not possible to clearly separate the given data set because some of the data points in the two classes might fall into gray area that is not easy to separate linearly. As one of the solutions for this problem, data are mapped to a higher dimension such that the two classes could be separable in the higher dimension (called kernel function). Consider a training set of vector-label pairs \((x_i, y_i), i = 1, \ldots, l\) where \(x_i \in \mathbb{R}^n\) and \(y_i \in \{1, -1\}\) where \(x_i\) is a vector in an n-dimensional space. If the training set is linearly separable, there exist infinitely many hyperplanes \(f(x) = w^T x + b\) which correctly classify all vectors in the training set, i.e. \(\text{sign}(f(x_i)) = y_i\). For n-dimensional space, the hyperplane will be n-1 dimensional. The objective of SVM is to choose the optimal hyper-plane that clearly separates vectors in the training set into two groups, +1 or -1, and maximizes the distance (margin) between the hyperplane and the support vectors. In summary, the purpose of SVM can be regarded as the solution of the following optimization problem:

\[
\min_{w, b, \xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \quad \text{subject to } y_i (w^T \Theta(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0,
\]

where \(\xi_i \geq 0\), where the training data are mapped to a higher dimensional space by the function \(\Theta\), and C is a penalty parameter on the training error. Therefore, for any testing vector \(x\), the decision function (predictor) is \(f(x) = \text{sign}(w^T \Theta(x) + b)\).

In addition, as mentioned above, most training sets used in a variety of domains are not linearly separated so that it is hard to derive the optimal hyperplane correctly classifying vectors in two classes. To solve this non-linearity problem, several solutions called kernel functions have been proposed.
and adopted for SVM. A kernel function is written \( K(x_i, x_j) \equiv \langle \phi(x_i), \phi(x_j) \rangle \) and the most widely used four kernel functions are

- **Linear**: \( K(x_i, x_j) = x_i^T x_j \)
- **Polynomial**: \( K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \ \gamma > 0 \)
- **Radial basis function (RBF)**: \( K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \ \gamma > 0 \)
- **Sigmoid**: \( K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r). \)

where \( \gamma, r \) are kernel parameters.

### 6. SVM Simulation Method

We used LIBSVM [20], a freely available SVM software library, to generate the SVM models. To generate the data set for model training, we randomly selected non-cases to match the number of cases in the training data set. To meet the format requirement of SVM, the collected data are transformed as follows. Values of selected input features (equivalent to ‘input variables’ in regression analysis) were normalized to values from -1 to +1. Values of categorical variables such as Education are arbitrarily assigned to numbers between -1 and +1. For example, -1, -0.5, 0, 0.5, 1 represents illiteracy, elementary school graduates, middle school graduates, high school graduates, college graduates, respectively. Values of continuous variables were transformed into values between -1 and +1 by dividing them by an appropriate number. In the training data set, the last column of the input data was set to the known outcome, i.e., 1 for positive, -1 for negative. Different kernel functions, including linear, polynomial, sigmoid, and radial basis functions (RBF), were tested and selected for the models on the basis of performance.

### 7. Performance Evaluation

To evaluate the robustness of the estimates from the SVM models, the leave-one-out cross-validation (LOOCV) was performed. In LOOCV, one test sample is extracted from a total of \( n \) samples. This test sample is then used for calculating the classification accuracy of the remaining \( n-1 \) training samples, and this process is repeated \( n \) times. That is, LOOCV is a special case of \( k \)-fold cross-validation where \( k \) is the same as the number of samples in training data set \([3,18,21]\). To generate summary performance estimates, we used the statistics including sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the cross-validations.

Five widely used statistics were adopted to evaluate the performance of a model: sensitivity, specificity, PPV, NPV, and accuracy.

\[
\text{Sensitivity} = \frac{TP}{(TP + FN)}, \quad \text{Specificity} = \frac{TN}{(TN + FP)}
\]

\[
\text{PPV} = \frac{TP}{(TP + FP)}, \quad \text{NPV} = \frac{TN}{(TN + FN)}
\]

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + FN + FP + TN)}
\]

where \( TP, FP, TN, \) and \( FN \) refer to the number of true positives, false positives, true negatives, and false negatives statuses, respectively.

### III. Results

#### 1. Feature Selection

To build an optimal solution which identifies predictors of medication adherence in HF patients using our questionnaire survey data, we applied SVM for all possible feature-combinations from the dataset. We generated all possible combinations (\( 2^{11} - 1 = 2047 \)) of 11 features (gender, age, spouse, education, monthly income, duration of HF diagnosis, daily frequency of medication, ejection fraction and MMSE-K, medication knowledge, and NYHA functional class) and found out combinations that represent the highest classification accuracy.

Feature selection technique is one of the key issues in data mining for reducing classifier-building time as well as increasing the performance of classifiers. Fortunately, our sample data consists of 11 dimensional vectors which have relatively small dimensionality, therefore it allowed us to build classifiers for 2047 (\( 2^{11} - 1 \)) combinations of features and to evaluate their performance in a reasonable time.

#### 2. Cross-Validation Test

Consequently, using the four kernel functions, four models for partitioning people into two categories - medication adherence of HF patients and the others - were developed. The performance of each model was evaluated using LOOCV which randomly partitions the dataset into 76 equal size subsets having a sample as its element and uses each subset as a test dataset to verify the performance of each model which is derived by 75 remaining data subsets.

Table 2 depicts the comparison of the best accuracy of the four models (Linear, Polynomial, RBF, and Sigmoid). Since SVM with RBF showed better performance than with other kernel functions in the accuracy, hereafter we only report the performance of models with RBF.

The model showing the best accuracy consists of five features (gender, spouse, daily frequency of medication, medication knowledge, NYHA functional class) and seven features (age, education, monthly income, EF, MMSE-K, medication knowledge, NYHA functional class).

Table 3 shows the performance of the best model in detail.
with statistics of sensitivity, specificity, PPV, and NPV.

Table 4 shows that the two sets of medication adherence-related variables with the best classification performance: a set of five variables (gender, daily frequency of medication, medication knowledge, NYHA functional class, spouse) and seven variables (age, education, monthly income, EF, MMSE-K, medication knowledge, NYHA functional class). The best detection accuracy was 77.63%. The 2 variables, medication knowledge and NYHA functional class, are shared by the two models. Swets suggests the following guidelines for interpretation: 0.5-0.7, rather low accuracy; 0.7-0.9, moderate accuracies useful for some purposes; and >0.9, rather high accuracy [22]. Therefore, the classification accuracy of these five-variable and seven-variable models can be described as the level of fair performance.

IV. Discussion

A full understanding of the factors associated with medication adherence in patients with HF is needed so that effective interventions to improve medication adherence can be developed [3]. But, most studies of medication adherence in patients with HF have had a number of limitations that reduce the usefulness of their findings. One of these limitations is the failure to use multivariate analysis methods to study medication adherence. This study addressed this problem by using SVM in identifying predictors of medication adherence in HF patients.

Our application of SVM found models that achieved fair classification performance, with a leave-one-out cross validated accuracy of around 77.6%. The LOOCV performance is a realistic indicator of performance of a classifier on unseen data and is a widely used statistical technique. The LOOCV is used during the training of a classifier to prevent overfitting of a classifier on the training set. But the LOOCV is rarely adopted in large-scale applications since it is computationally expensive [23]. In further research, if there are many samples, we may have to consider other ways such as 10-fold cross validation, etc.

Some of the selected variables in HF patients showed unexpected tendencies that are contrary to common medical thoughts. For example, a longer duration of HF diagnosis might be associated with medication non-adherence. However, this variable was absent in the two models that showed the best accuracy. In future work, we will investigate the performance using the SVM algorithm as well as other features selection methods to identify a group of significant factors.

In general, factors known to influence medication adherence in HF include medication knowledge and NYHA functional class [24,25]. Each of the models showed a strong preference for these factors. Especially, medication knowledge as a single variable was a strong predictor. Medication knowledge is well known as risk factor of medication non-adherence in HF [24,25]. The results of this study suggest that patients with higher levels of knowledge utilized the health care system more often, perhaps have greater awareness, hence a stronger tendency of self care.

Although our sample size was large enough to demonstrate fair accuracy, a larger sample size and more heterogeneous sample may be needed to more thoroughly investigate predictors of medication adherence and to obtain results that can be generalized to large populations.

Our study has several limitations. The primary limitation of this study is its small sample size, which made it very difficult for any of the endpoints to achieve statistical significance. The second limitation was that we did not directly

Table 2. Comparison of the best accuracy for four kernel functions

<table>
<thead>
<tr>
<th></th>
<th>Linear</th>
<th>Polynomial</th>
<th>RBF</th>
<th>Sigmoid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best accuracy</td>
<td>73.438</td>
<td>67.188</td>
<td>77.632</td>
<td>64.063</td>
</tr>
<tr>
<td>RBF: radial basis function.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Detailed sensitivity, specificity, PPV, NPV and accuracy for the best model of RBF kernel function

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.776</td>
<td>0.776</td>
<td>0.778</td>
<td>0.816</td>
<td>0.776</td>
</tr>
</tbody>
</table>

PPV: positive predictive value, NPV: negative predictive value, RBF: radial basis function.

Table 4. Predictors of medication adherence in heart failure patients

<table>
<thead>
<tr>
<th>No.of features</th>
<th>Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Gender, daily frequency of medication, medication knowledge, NYHA functional class, spouse</td>
<td>77.632</td>
</tr>
<tr>
<td>7</td>
<td>Age, education, monthly income, EF, MMSE-K, medication knowledge, NYHA functional class</td>
<td>77.632</td>
</tr>
</tbody>
</table>

measure medication adherence. Third, the cross-sectional design does not allow any inference to be drawn with respect to the causal relationship among variables. Finally, our data were mainly based on patient information, and so, Berkson’s bias [26] may contribute to inaccurate or excessive approval. SVM has been widely used in various areas, such as recognition, reliability evaluation, bioinformatics and medicines, for survival time classification and assessment of the severity of many acute and chronic diseases [27]. However, to our knowledge, this is the first study to investigate predictors of medication adherence and non-adherence in Korean patients with HF by means of SVM.

This study only illustrates a potential use of the SVM technique. Further studies should be conducted where the discriminative power of our SVM is compared with that of commonly used logistic regression, Bayesian network and neural network models.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

An abstracts of this study was presented with poster in 2010 KOSMI Spring Conference.

References

5. Thurston RC, Matthews KA, Hernandez J, De La Torre F. Improving the performance of physiologic hot flash measures with support vector machines. Psychophysiol-
SVM Prediction of Medication Adherence for HF