An Intelligent Recommender System based on Short-term Risk Prediction for Heart Disease Patients

Raid Lafta, Ji Zhang, Xiaohui Tao, Yan Li and Vincent S. Tseng‡
Faculty of Health, Engineering and Sciences, University of Southern Queensland, Australia
‡Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan
{*RaidLuaibi.Lafta, ji.zhang, xtao, yan.li*}*@usq.edu.au,* ‡*vtseng@cs.nctu.edu.tw*

***Abstract*—In this paper, an intelligent recommender system is developed, which uses an innovative time series prediction algorithm to provide recommendations to heart disease patients in the tele-health environment. Based on analytics of each patient's medical tests in records, the system provides the patient with decision support for necessity of medical tests. The experimental results show that the proposed system yields satisfactory accuracy in recommendations. The system also offers a promising way for saving the workload for patients and healthcare practitioners in conducting daily medical tests. The research will help reduce the workload and cost in healthcare and help the healthcare industry transform from the traditional scenario to more a personalized paradigm in a tele-health environment.**

***Index Terms*—Intelligent system, Recommender system, Heart failure, Time series prediction**

I. Introduction

Most healthcare organizations such as hospitals and medical centers generate huge volume of data with semi-structural text, numbers and images [9]. These data contain wealthy information that may be used to support high-quality clinical decision-making support. Currently, the clinical decisions are usually made based on practitioners' experience with limited support from medical databases. Quite often this leads to undesirable biases, human errors and high medical costs and consequently, affects the quality of services provided to patients [7].

Recommender systems are powerful user-support tools providing users with useful suggestions by facilitating access to relevant items. These suggestions are in connection with various decision- making processes [22]. Many different techniques and algorithms such as collaborative filtering, association rules, content-based filtering have been used by recommender systems to generate recommendations [23].

In the last decades, much research efforts have been invested in the assessment of diseases risk in order to help make safe and effective decisions. Data mining techniques and statistical tools have been widely used for various diseases prediction [3,4, 5, 6, 7,25]. However, a challenge remains still in securing an effective analytic tool with high accuracy to help support personalized evidence-based decisions.

Heart disease is currently registering one of the highest death rates among non-infectious diseases, with a high associated cost in prevention and treatment. Great efforts have been taken to prevent heart disease using clinical decision support systems, for example, trying to predict heart disease at early stage [25]. In view of this, a heart disease prediction model for risk assessment is much desired to support high-quality and timely decision-making processes.

To tackle the challenge, we propose a heart disease prediction model in this work and incorporate it into an intelligent recommender system to conduct short-term risk assessment for heart failure patients. On the basis of assessment results, the system also provides recommendations to heart failure patients in relation to the necessity of medical tests taken on the following day. The research is conducted with an aim at providing evidence-based decision support to patients and healthcare practitioners and reducing their workload in medical checkups.

II. Related Work

Extensive research work has been carried out in data mining and analytic on medical data. The techniques used can be broadly classified as descriptive and predictive models. Descriptive models are used to identify patterns in data and the relationships between factors responsible for them [21]. Descriptive analytics involve methods such as clustering, summarizations, association rules, and sequence analysis [22].

On the other hand, predictive techniques focus on what will happen in the future. They can be divided into two categories: classification and prediction [23]. Classification, regression, and time series analysis are among the most important tasks of predictive data mining. There are various types of classification models including classification by decision tree induction, Bayesian classification, Neural Networks, Support Vector Machine (SVM), and classification based on association [23].

Various predictive techniques have been applied to different medical and healthcare problems. Genetic algorithm, logistic regression and decision tree have been used to predict the severity level of disease in patients [6, 10, 11, 12]. Prediction models have been designed to assess the risk of different diseases. In these works, laboratory measurements and symptoms were utilized to predict the disease risk. Different predictive data mining techniques such as Bayesian classifiers, decision

tree, logistic regression and back propagation neural network have been utilized by [4, 5, 7,13, 14] to predict diseases at early stage for patients. These techniques have been used in building effective predictive models to discover different diseases as early as possible in order to treat them effectively. Statistical analytic tools have been effectively used to estimate lifetime and long-term risk for patients with different diseases [15, 16, 17]. Clinical predictive models have also been developed based on measurements taken on patients in different timely basis, aiming to detect diseases at early stage and treat them at right time. The predictive survival models have been suggested by [18, 19, 20] to analyze patients' survival times and to help develop a superior treatment plan for the diseases. Statistical analytics such as supervised wavelet approximation coefficients and multivariable piecewise Poisson regression method have also been effectively utilized in these works. In all such studies, data in patients' medical profiles such as age, sex, blood pressure and blood sugar, etc. have played an important role and provided informative evidence in decision-making support.

III. Framework

In this study, we propose an intelligent recommender system equipped with a novel prediction algorithm to analyze the medical data of heart failure patients, assess the short-term risk of heart disease for the patients, and then provide them with recommendations based on the outcomes of prediction.

A. Data Preprocessing

Data pre-processing is an indispensable key step conducted on raw data to make them ready for the analytic tasks in question. Analytic tools could be misled and give wrong results if data have impurities such as missing or duplicate data. Therefore, it is necessary to preprocess the data before starting the data analytic process. In this phase, missing data problem, which is caused during the data collection or transmission, is resolved by filling the missing data with a global constant. Also, the noise records with incorrect readings are removed. Another important task of data preprocessing in this work is to extract the information for each individual patient from the original dataset for personalized data analysis and recommendations.

B. Time Series Recommendation Algorithm

The key component of the proposed system is the recommendation algorithm based on time series data analysis. The algorithm is developed to decide whether a given patient needs to take a medical measurement such as the heart rate test today based on a study of his/her measurement readings for the past *k* days. If the patient satisfies both of the following conditions for a measurement, a recommendation of "no test needed" will be generated, and the patient does not need to take the test on the following day for that measurement:

- *She (he) has taken the test for no less than p% of the past k days for this measurement (*$0 <= p <= 100$*), and*
- *All the readings of this measurement during the past k days are normal.*

The "no test needed" recommendation will be provided to the patient and stored into the backend database as a part of the patient?s historical records. If any of the conditions is not satisfied, a recommendation of "test required" will be generated and the patient is suggested to take the medical test on the following day. Again, the recommendation will be stored into the system as a historical record.

Overall, there are four parameters in the recommendation algorithm, i.e., the minimum (min) and maximum (max) of normal values for each measurement, setting up the boundary of healthy range; the length of the sliding time window k, and the minimum percentage (p) of days when medical test is conducted for the measurement in the past k days. The recommendation algorithm is presented as following.

Input : Patient's time series medical testing data (e.g., heart rates).
Output: $\mathbf{Rick} = [0|1]$ (0: low risk; 1: high risk);
$\mathbf{Recomm} = [0|1]$ (0: no test required; 1: test needed.)

1 let k be a limited number from the past days (the length of time window k);
2 let $p\%$ be a value between the range $(0 <= p <= 100)$;
3 let max and min be the boundary of healthy values in the test;
4 **foreach** *days for the patient* **do**
5 **if** *the patient has taken the test for no less than p% of the past k days* $(<= k - (p \times k))$ *and* $min <$ *measured value* $< max$ **then**
6 $\mathbf{Risk} = 0$;
7 **else**
8 $\mathbf{Risk} = 1$;
9 **end**
10 **if** $\mathbf{Risk} == 0$ **then**
11 $\mathbf{Recomm} = 0$;
12 **else**
13 $\mathbf{Recomm} = 1$;
14 **end**
15 **end**
16 return $\mathbf{Recomm}$.

Algorithm 1: Time Series Prediction Algorithm

The algorithm of our recommender system is presented in Algorithm 1. It evaluates the time series data collected for a patient on a continuous basis using a slide window with a length of k, which is the number of days in the past that the algorithm will look at in support of assessment evaluation for the following day. Two conditions are evaluated in the IF predicate from Line 5 to 14. The first one evaluates the percentage of the actual medical test that has been carried out in the past k days. If a test is skipped for a day, actual reading will be missing and as a result, the certainty and accuracy for risk assessment will drop for future days. Therefore, an upper bound is imposed in this condition on the total number of days when the medical testing is skipped in each sliding window. In addition to this bound, we also require that the readings of all the medical checkups conducted during the past k days are in the normal range for the measurement, as dictated by its corresponding minimum and maximum threshold values. Intuitively speaking, normal readings improve the confidence that the short-term risk is low, whereby a skip of the test

on this measurement can be recommended. If both the two conditions are satisfied, then the risk for skipping the physical test for this measurement is deemed low and accordingly a recommendation for skipping a test for the measurement can be made for the following day. Otherwise, the system will provide a recommendation urging patients to take medical test for the measurement on the following day and the reading of the measurement will be received and stored in the database. It's worthwhile mentioning that the above algorithm will be applied to a single measurement once at the time for risk assessment and recommendation.

C. Human Computer Interaction for the System

Our recommender system involves human computer interaction to receive input from human users concerning the values of the parameters that are used in the algorithm of our system. The recommendations generated by our system will be returned back to users through different channels and platforms including desktops, laptops and tablets to embrace the latest technological advancements for quick information dissemination. Besides returned back to the patients, the results can also be sent remotely to the practitioners such as doctors and nurses so that they can be informed and keep track of the physical checkups and overall health conditions of the patients.

IV. Evaluation Design

In this section, we will provide details regarding the design of our experimental evaluation including the dataset, performance metrics and the experimental platform.

We use a real-life dataset to test the practical applicability of the system we propose. A pilot study has been conducted on a group of heart failure patients and the resulting data were collected for their day-to-day medical readings of different measurements in a tele-health care environment. More specifically, the dataset contains the information of six patients with a total of 7,147 different records during a six-month period starting from May to November 2012. The dataset is by nature a time series and contains a set of measurements taken from the patients on different days. Each record contains a few numerical medical attributes including Ankles, Chest Pain, and Heart Rate (HR), Diastolic Blood Pressure (DBP), Mean Arterial Pressure (MAP), Systolic Blood Pressure (SBP), Oxygen Saturation (SO2), Blood Glucose (BG), and Weight (W).

This dataset is used as the ground truth result to test the performance of our recommendation system. The recommendations produced by our system will be compared with the actual readings of the measurements in the dataset to see how accurate our recommendations are.

We devise two performance metrics to evaluate the performance of the proposed system, namely *precision* and *workload saving*. Precision refers to the percentage of correctly recommended days against the total number of days that recommendations are provided, while workload saving refers to the percentage of the total number of days when recommendations are provided against the total number of days in the dataset. Mathematically, precision and workload saving are defined as follows:

$$Precision = \frac{NN}{NN + NA} \times 100\% \quad (1)$$

$$Saving = \frac{NN + NA}{|\mathcal{D}|} \times 100\% \quad (2)$$

Where NN denotes the number of days with correct recommendations, NA denotes the number of days with incorrect recommendations and $|\mathcal{D}|$ refers to the total number of days in the dataset.

V. Result Analysis

In this section, we discuss the results of the experimental evaluation conducted on our system. We focus on investigating the performance of our system from the perspective of different measurements and patients. Both precision and workload savings are employed to measure the performance of proposed system.

A. Performance under Different Measurements for all Patients

In this experiment, we evaluated the performance of the system when it is applied to different measurements for all patients based on $k = 5$. The algorithm was tested in four rounds for each patient with different medical tests (Heart rate, DBP, MAP and SO2). Figure 1 shows the detailed results we obtained for each patient. From the results, one may see that the algorithm yields recommendations with varying degree of precision, with the heart rate and SO2 measurements register the highest compared to others. This is because that there are intrinsic stronger correlations for the neighboring readings of heart rate and SO2 measurements so that the short-term prediction of risk becomes more accurate than other measurements.

By further aggregating the results for the measurements in Figure 1, Figure 2 demonstrates the averaged precision and workload savings for each patient. Generally speaking, the accuracy of the recommendations provided by the proposed system ranges from 75% to 100% across different patients. The system is capable of helping reduce on average 10% of workload for patients from their daily medical tests.

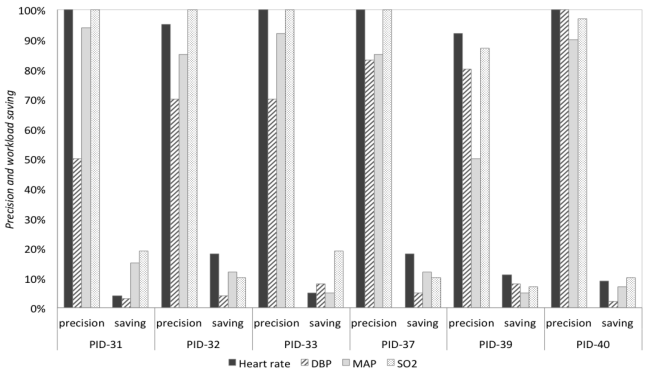

Fig. 1: The Precision and Workload Saving for Different Patients

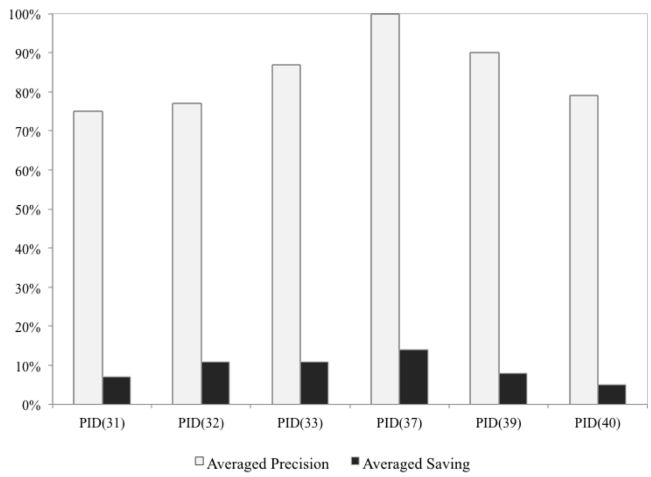

Fig. 2: Average Precision and Workload Saving for Each Patient

VI. Conclusions and Future Work

In this paper, we present an intelligent recommender system that predicts and assesses the short-term disease risk for heart failure patients. The system is developed aiming at improving the quality of clinical evidence-based decisions and helping reduce financial and timing cost taken by patients. A time series prediction algorithm is proposed to predict short-term risk for the heart failure patients. Based on the prediction result, the system provides a recommendation to the patient for necessity of taking a medical test. The work makes theoretical contribution by the time series prediction algorithm and applicable contribution by an intelligent system to improve the quality of health care services. As an ongoing project we envisage that there is much room available to further improve the system. In future work, we will further improve the predictive ability of the proposed algorithm in order to enhance the precision and workload saving performance. More comprehensive experiments will be conducted using larger, extensive datasets for evaluation.

References

[1] K. Ahmed, T. Jesmin, and M. Z. Rahman. Early Prevention and Detection of Skin Cancer Risk using Data Mining. In *International Journal of Computer Applications*, (0975C8887) Volume, 2013.

[2] C.-D. Chang, C.-C. Wang, and B. C. Jiang. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. In *Expert systems with applications*, vol. 38, no. 5, pp. 5507-5513, 2011.

[3] M. Farhadian, P. J. Lisboa, A. Moghimbeigi, J. Poorolajal, and H. Mahjub. Supervised Wavelet Method to Predict Patient Survival from Gene Expression Data. In *The Scientific World Journal*, vol. 2014, 2014.

[4] N.-C. Hsieh, L.-P. Hung, C.-C. Shih, H.-C. Keh, and C.-H. Chan. Intelligent postoperative morbidity prediction of heart disease using artificial intelligence techniques. In *Journal of medical systems*, vol. 36, no. 3, pp. 1809-1820, 2012.

[5] F. Huang, S. Wang, and C.-C. Chan. Predicting disease by using data mining based on healthcare information system. pp. 191-194, 2012.

[6] T. W. Joo, and S. B. Kim. Time series forecasting based on wavelet filtering. In *Expert Systems with Applications*, 2015.

[7] M. Kantardzic. In *Data mining: concepts, models, methods, and algorithms*, John Wiley and Sons, 2011.

[8] J.-K. Kim, J.-S. Lee, D.-K. Park, Y.-S. Lim, Y.-H. Lee, and E.-Y. Jung. Adaptive mining prediction model for content recommendation to coronary heart disease patients. In *Cluster Computing*, vol. 17, no. 3, pp. 881-891, 2014.

[9] H. C. Koh, and G. Tan. Data mining applications in healthcare In *Journal of healthcare information management*, vol. 19, no. 2, pp. 65, 2011.

[10] V. Krishnaiah, D. G. Narsimha, and D. N. S. Chandra Diagnosis of lung cancer prediction system using data mining classification techniques. In *International Journal of Computer Science and Information Technologies*, vol. 4, no. 1, pp. 39-45, 2013.

[11] M. Kurosaki, N. Hiramatsu, M. Sakamoto, Y. Suzuki, M. Iwasaki, A. Tamori, K. Matsuura, S. Kakinuma, F. Sugauchi, and N. Sakamoto. Data mining model using simple and readily available factors could identify patients at high risk for hepatocellular carcinoma in chronic hepatitis C In *Journal of hepatology*, vol. 56, no. 3, pp. 602-608, 2012.

[12] M. H. Lee, H. I. Yang, J. Liu, R. Batrla-Utermann, C. L. Jen, U. H. Iloeje, S. N. Lu, S. L. You, L. Y. Wang, and C. J. Chen. Prediction models of long-term Cirrhosis and hepatocellular carcinoma risk in chronic hepatitis B patients: Risk scores integrating host and virus profiles. In *Hepatology*, vol. 58, no. 2, pp. 546-554, 2013.

[13] D. M. Lloyd-Jones, E. P. Leip, M. G. Larson, R. B. d'Agostino, A. Beiser, P. W. Wilson, P. A. Wolf, and D. Levy Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. IN *Circulation*, vol. 113, no. 6, pp. 791-798, 2006.

[14] H.-Y. Lu, C.-Y. Huang, C.-T. Su, and C.-C. Lin. Predicting Rotator Cuff Tears Using Data Mining and Bayesian Likelihood Ratios. In *PloS one*, vol. 9, no. 4, pp. e94917, 2014.

[15] S. I. Nihtyanova, B. E. Schreiber, V. H. Ong, D. Rosenberg, P. Moinzadeh, J. G. Coghlan, A. U. Wells, and C. P. Denton. Prediction of Pulmonary Complications and Long-Term Survival in Systemic Sclerosis. In *Arthritis & rheumatology*, vol. 66, no. 6, pp. 1625-1635, 2014.

[16] L. Nobel, N. E. Mayo, J. Hanley, L. Nadeau, and S. S. Daskalopoulou. MyRisk_Stroke Calculator: A Personalized Stroke Risk Assessment Tool for the General Population. In *Journal of Clinical Neurology*, vol. 10, no. 1, pp. 1-9, 2014.

[17] S. Roessler, E. L. Long, A. Budhu, Y. Chen, X. Zhao, J. Ji, R. Walker, H. L. Jia, Q. H. Ye, and L. X. Qin. Integrative genomic identification of genes on 8p associated with hepatocellular carcinoma progression and patient survival. In *Gastroenterology*, vol. 142, no. 4, pp. 957-966. e12, 2012.

[18] M. Sabibullah, V. Shanmugasundaram and R. Priya. Diabetes patient's risk through soft computing model. In *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, vol. 2, no. 6, pp. 60-65, 2013.

[19] F. Siraj, and M. A. Abdoulha. Mining enrolment data using predictive and descriptive approaches. In *Knowledge-Oriented Applications in Data Mining*, pp. 53-72, 2007.

[20] S. Tuff—ry. *Data mining and statistics for decision making*. John Wiley & Sons, 2011.

[21] D.-Y. Yeh, C.-H. Cheng, and Y.-W. Chen. A predictive model for cerebrovascular disease using data mining. In *Expert Systems with Applications*, vol. 38, no. 7, pp. 8970-8977, 2011.

[22] F. Ricci, L. Rokach, and B. Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[23] S. Sivapalan, A. Sadeghian, H. Rahnama, and A. M. Madni. Recommender systems in e-commerce. In *World Automation Congress (WAC)*, pp. 179-184, 2014.