Accurate and Reliable Recommender System for Chronic Disease Diagnosis

Asmaa S. Hussein, Wail M. Omar, Xue Li
School of IT and Electrical Engineering
The University of Queensland
Queensland, Australia
asmaa.hussein@uqconnect.edu.au, w.omar@bnr-education.ca, xueli@itee.uq.edu.au

Modafar Ati
College of Engineering and Computer Science
Abu Dhabi University
Al Ain, United Arab Emirates
modafar.ati@adu.ac.ae

***Abstract*-With the rapid growth of chronic disease cases around the world, healthcare support systems like recommender systems play a major role in controlling the disease, through providing accurate and trustworthy disease risk diagnosis prediction and acknowledgement of disease risk status, that assists healthcare providers to have 24/7 remote patient monitoring system and assist patients to have 24 hour access to the medical care. Providing an accurate real-time recommendation for medical data is a challenge according to its complexity represented by unbalance, large, noisy and/or missing data. The Chronic Disease Diagnosis (CDD) recommender system expectation is to give a high accuracy and reliable disease risk prediction. This paper presents a CDD recommender system model using multiple decision tree classification algorithms. Decision tree algorithms are applied to achieve high accuracy disease risk predictive model. Historical patients' medical data from the Middle East is used to train the model. Determining the relevant features through Attribute Selection method is used to reduce data generation and improve the predictive model performance. Merging patients' lab and home test readings is considered to leverage the diagnosis fidelity. Diabetes diagnosis case study is designed through this research as experiment to show the feasibility of our model.**

Keyword-E-Health; Remote Chronic Disease Diagnosis prediction; Diabetes healthcare management system; Decision Tree; Random Forest.

I. INTRODUCTION

In recent years, a pattern of chronic diseases has started to emerge in the Middle East similar to the rest of the world [1]. These diseases appear in the form of increases in obesity, heart diseases, and diabetes (both types 1 and 2). This growth in chronic illnesses along with the increase of inactive lifestyle in the Middle East has imposed great pressure on healthcare providers, especially when trying to ensure a structured patient follow-up to be achieved after each therapeutic change.

The rapid increase in ICT (Information and Communications Technology) development has opened a new era for researchers to develop a number of E-Health applications that are starting to play a major role in improving healthcare services. Recommender systems [2] have been emerged and being increasingly used by many applications, like E-Commerce and E-Health [3] according to their feasibility in automatically extracting useful information and predicting and recommending appropriate results to consumers. Therefore, this research applies a recommender system to manage the chronic disease, which minimizes both the risk of such a disease as well as the cost that is associated with it.

The most popular technique of recommender systems is collaborative filtering, which utilizes users' experiences or histories to generate predictions of the unknown preferences [4]. For example, in a system recommending books there usually are two sets, a set of users *U* (e.g., the readers) and a set of items *I* (e.g., the books), and a utility function r between the two sets (e.g., the books rating by users). In its most common formulation, the recommendation problem is reduced to the problem of estimating ratings for the items that have not been seen by a user, selecting for each user *u* □ *U,* the item *i* □ *I* that maximizes the defined user's utility.

In our case the set of users, *U* represents patients, while the set of items *I* is the patients' medical record attributes represented by disease risk factors and diagnosis. The relationship between the two sets, instead of being a user preference rating, is the risk factors readings and an ordinal set of rating the disease risk diagnosis (represented by 0-10 scale). This rating is determined by physicians based on various factors such as patient's vital signs and lab tests.

The hypothesis is that, if a patient's chronic disease risk is predicted with high accuracy, we will expect the improvement of patients' health conditions, lifestyle adjustments along with reducing the healthcare services costs.

A. The Research Problem

Accuracy is the essential factor of recommender systems [5]. In our research, accuracy reflects how the recommendation is valuable and correspondent with the patient's physiologic state and complexity of disease in order to be valuable to the users who are looking for trusty recommendation. To achieve that, recommender systems should minimize false positive and false negative errors. For example, a specific disease risk diagnosis is not useful for a patient (false), which is recommended by the system (positive). While a specific diagnosis is useful for a patient (false), but it is not recommended (negative).

Our purpose of developing the recommender system for CDD is to provide efficient recommendations benefit users remotely. To improve the accuracy of recommendations for CDD; we consider: the nature and complexity of medical data such as different types of data, large number of parameter, and missing and noisy data.

For this consideration, as known, the patients' states of health and disease risk are monitored through symptoms, vital signs, laboratory test results, etc. Those results and readings are generated by different sources like different data entry-medical workers, medical home sensors, etc., which are not corresponding in structure or quality. Thus, medical data can have features of different types and may contain various types of errors (missing data and noisy data) that occur for a variety of reasons for example: erroneous attribute values during the insertion process, values cannot be recorded when the data is collected, a doctor may not order all applicable tests while diagnosing a patient [6], and some personal data is ignored by users because of privacy concerns. Dealing with such data is a challenge to get accurate and reliable recommendations as it can have a significant effect on the recommendations quality and accuracy.

Our contribution is to provide more accuracy and reliable recommendations in order to assist patients controlling their chronic disease and assist healthcare providers to have 24/7 remote patient monitoring system.

B. *Decision Tree Classification Approach*

In this work, the decision tree classification approach [7] has been adopted within a recommender system to improve the prediction accuracy. Decision tree is a very popular data mining method for classification and regression, where decision tree Collaborative Filtering models act as classifier to classify tasks [8].

Different decision tree classification models such as J48[9], Decision Stump[10], REP Tree[11] and Random Forest [12] have been constructed to predict the chronic disease risk. RF model has proven its feasibility in providing accurate predictions over the medical data, while the other decision tree classification models have been used for prediction performance comparisons.

C. *Random Forest*

RF algorithm [13, 14] was developed by Brieman in 2001. RF is an ensemble classifier that consists of many decision trees created by using bootstrap samples of the training data and random feature selection at each node to grow each tree; it outputs the class that is the mode of the class's output by individual trees. In this way, an RF ensemble classifier performs better than a single tree from classification performance point of view [12]. And as each node is split using the best among a subset of predictors randomly chosen at that node, RF performs very well compared with many other classifiers and is robust against over-fitting. Furthermore, RF can be used to estimate missing values [15]. Generally, RF is a powerful statistical classifier that has some advantages compared to other statistical classifiers; see [16] for RF advantages.

The remaining of the paper is structured as related work in section two. CDD recommender system is in section three, followed by CDD scenario and methodology. Diabetes diagnosis experiments, outcomes, and results are demonstrated and discussed in section five. Conclusion and future work are presented in section six.

II. RELATED WORK

This section demonstrates the current researches in the area of recommender system and its uses with healthcare applications, and the RF method used in disease diagnosis field. Recommender systems have been used by many researchers in different areas such as: movies, healthcare, E-commerce, etc. [17]. With focus on healthcare applications, Davis *et al.* [18] proposed a Collaborative Assessment and Recommendation Engine (CARE), which relies on patient's medical history using only ICD-9-CM codes without considering other information such as lab tests, etc. CARE predicted each patient's future disease risks based on their own medical history and that of similar patients using ICD-9-CM codes. The experimental results demonstrated that CARE performed well at capturing the future diseases and facilitating discussion about early testing and prevention. Sapon *et al*. [19] used supervised learning algorithms of Artificial Neural Network for diabetes prediction. The network was trained using the data of 250 diabetes patients. Scaled Conjugate Gradient algorithm with value of R=0.88026, produced the best performance in the prediction of diabetes compared to the other algorithms.

Accuracy of the recommender system is vital in many applications. Therefore, different Collaborative Filtering (CF) recommendation approaches have been used to improve the reliability, performance and accuracy of the recommender system. Noh *et al.* [20] used multiple imputation-based CF approach for recommendation system to improve the accuracy in prediction performance by solving the incomplete data problem of CF algorithm. The approach replaces each missing value with m>1 acceptable values from their predictive distribution and converts an incomplete dataset into m complete dataset and each dataset is used for analysis. Overall estimate is then obtained by combing these m estimates. The prediction power was improved by adopting CF algorithm on the complete dataset after imputation is made.

The RF is one of the decision tree classification methods that prove its feasibility to be used with healthcare applications. Özçift [12] presented a resampling strategy based RF ensemble classifier to improve diagnosis of cardiac arrhythmia. The resultant accuracy of the classifier was found to be 90.0%, and the results of experiments demonstrated the efficiency of random sampling strategy in training RF ensemble classification algorithm. Ko *et al*. [21] used RF to demonstrating an efficient white blood cell (WBC) image classification method. WBC was classified into five different categories that are necessary for accurate disease diagnosis. The experimental results showed that using the random forest with dynamic features could indeed improve the classification performance.

As demonstrated by the above surveys, recommender systems prove their usability in predicting and recommending the appropriate results to consumers. In this paper, our challenges are to provide a real-time and quality recommendation for a large, missing and lab tests merged with home tests medical data.

 ISBN: 978-1-61208-243-1

As such, because of the power and efficiency of the RF classification technique [16], it has been implemented in the development of a recommendation system for CDD. Such a system is presented as a case study for the model developed in this work.

III. Chronic Disease Diagnosis Recommender System

CDD recommender system is represented by prediction and recommendation. It depends on a set of patients' health history to train and build a model that is able to predict and recommend disease risk and disease risk status for the future cases. Those predictions and recommendations are approved by physicians. As shown in Fig. 1, the CDD recommender system requires input information to produce recommendation and predicted items. In this work, diabetes is used as a case study.

The input information that is needed to build up the predictive model of CDD for diabetes and make prediction for undiagnosed patients is categorized into:

Training data: a bulk of historical medical records of previous diabetic patients (935 records) has been collected from hospitals in Oman. The collected data records are represented by a number of attributes, values and doctors' assessment (diagnosis) for each case. Oman Health Authority based on international standards defines these factors. All of these records have been tested and diagnosed by doctors in Oman hospitals. A scale from 1 to 10 has been used for the diagnosis to give more opportunities for better classification. The doctor diagnosis scale is based on: 1–2 represents excellent, 3–4 good, 5–6 fair but needs control, 7–8 bad with bad control, and 9–10 represents a critical condition. Fig. 2 demonstrates a sample of the collected data.

Demographic data of active patient: refers to the user's profile such as: name, age, level of education, type of end user device and type of connectivity.

Medical dataset of active patient: formed by two types of data. The first type is known as home-tests such as blood sugar level, blood pressure, and weight. The other type involves results obtained in the laboratory, which is usually carried out over a long time interval, depending on the condition of the patient. Fig. 3 shows sample of the diabetes medical dataset of an undiagnosed patient case.

The output of the system is:

Prediction and recommendation: prediction is expressed as a numerical value that represents the disease risk diagnosis of an active patient and to the future cases. This prediction value should be within the same scale as the diagnosis values provided in the training dataset. Recommendation is expressed as disease risk status acknowledgment, which the active patient is seeking for.

IV. CDD Recommender System Scenario

The scenario consists of number of major activities, which starts by building the predictive model of recommender system, as shown in Fig. 4 and illustrated in the following sections. Multiple decision tree algorithms are adopted to achieve a high accurate predictive model.

A. *Building the Predictive Model*

The predictive model is built as follows:

1) *Data Preprocessing*

Filtering the data might be necessarily accomplished to avoid the creation of ambiguous or inappropriate models and improve the learning model performance [22]. In our system, the diabetes dataset is filtered by determining the relevant features through InfoGainAttributeEval Attribute Selection method, furthermore, the data is also transformed to a form appropriate the classification. InfoGainAttributeEval has ranked the relevant and significance features sequentially as follows: LDLC, Cholesterol, Triglyceride, FBS, HBA1C, BMI, HDLC, DBP, Age, SBP and Sex. As seen, the result shows that LDLC attribute, Low Density Lipoprotein Cholesterol, has a significant and highest effect on the diabetes diagnosis. The scale of diagnose is increased in conjunction with the increase of LDLC level. The patients with worse condition of diabetes (class diagnose 10) exhibit the highest readings of LDLC and vice versa.

2) *RF Algorithm*

The RF algorithm is utilized to analyze and segment the patients' records in training data into distinguished and related groups of classes based on the seen diagnosis scales. These classes are used to generate a predictive model that is able to predict the class of diagnosis of an active patient.

The workflow of RF algorithm starts by classifying the training datasets that is provided by the healthcare provider. *Kth* tree is constructed by sampling with replacement. At each node, *mth* variables are selected at random out of the total number of variables *M*, and the split is the best split on these *mth* variables. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the *kth* tree. The out-of-bag data (oob) or bagging [23] is used to get an unbiased estimate of the classification error as trees are added to the forest.

It is also used to enhance accuracy through using random features and to get estimates of variable importance [14]. After each tree is built, each case left out in the construction of the *kth* tree is put down the *kth* tree to get a classification. In this way, a test set classification is obtained for each case in about one-third of the trees. At the end of the run, take *j* to be the class that received the most votes every time case *n*

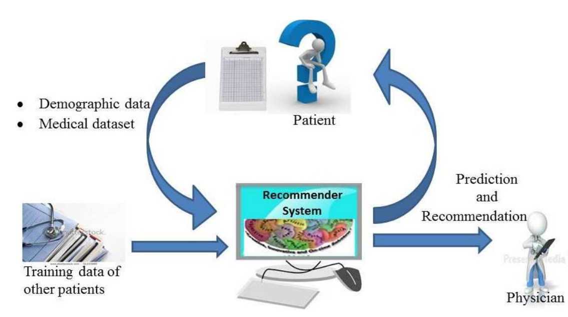

Figure 1: CDD Recommender System

PatientID	Age	Sex	BMI	SBP	DBP	FBS	HBA1C	Cholesterol	Triglyceride	HDLC	LDLC	(Class Lab Diagnosi
99901	40	2	23.2	130	80	186	6.6	206	92	51	136	6
99902	44	1	27.3	150	80	155	6.5	159	110	71	132	5
99903	51	2	22.1	140	80	188	5.5	169	84	55	98	5
99904	58	1	16.6	130	90	190	15.7	183	150	44	183	7
99905	38	2	28.7	130	90	144	10.7	228	111	58	148	7
99906	38	2	28.7	130	90	127	9.1	228	111	58	148	6
99907	45	2	26.3	184	108	142	6.2	221	124	55	142	7
99908	48	1	24	140	80	416	6.8	129	168	37	159	8
99909	54	1	19.5	160	110	151	15.3	171	146	29	113	7
99910	64	1	23	130	90	396	19	314	147	41	242	10

Figure 2: "10" sample records of total "935" diabetes training dataset

PatientID	Age	Sex	BMI	SBP	DBP	FBS	HBA1C	Cholesterol	Triglyceride	HDLC	LDLC	(Class label) Diagnosis
31477	69	1	26.4	120	70	166	7.4	158	119	39	97	?

Figure 3: Sample dataset of an undiagnosed active patient

was out of bag. In this case, the class of diagnosis is predicted by aggregating the predictions of the *kth* trees (the majority votes).

B. *New Data of Active Patient*

As shown in Fig. 4, the healthcare provider initiates a request to the request repository for collecting the home medical tests readings of an active patient. The request is sent to the patient-home medical sensor to collect those readings of the patient such as: blood sugar level, blood pressure, weight, etc. At the same time of sending the request, the healthcare provider provides the laboratory tests readings of the same patient (patient medical record). Those lab readings are merged with the home readings as a complete dataset in the profile of active patient (as shown in Fig. 3). This complete dataset has the same attributes of the training dataset, but the class diagnosis variable has no value and it is unknown.

C. *Prediction*

The complete dataset of an active patient is uploaded to the generated RF predictive model to predict the class of diagnosis and disease condition of patient. The dataset is analyzed and run over the tree model to finding the similarities between the current patient's dataset and the training dataset from other patients [24].

D. *Alert Message*

Once the diagnosis scale is produced, it is sent to healthcare providers. Healthcare providers would examine the accuracy of the prediction and approve/correct them. The patient profile is then updated by adding the approved diagnosis to the medical record of patient. Finally, an assessment and acknowledgement alert message of the predicted disease diagnosis and recommended disease risk status is sent to the patient, and the updated patient's medical record is stored in the training dataset repository. The latter action is performed in order to add an extra element of data to the training dataset, which as a result, would help in improving the performance of CDD recommendations for future cases.

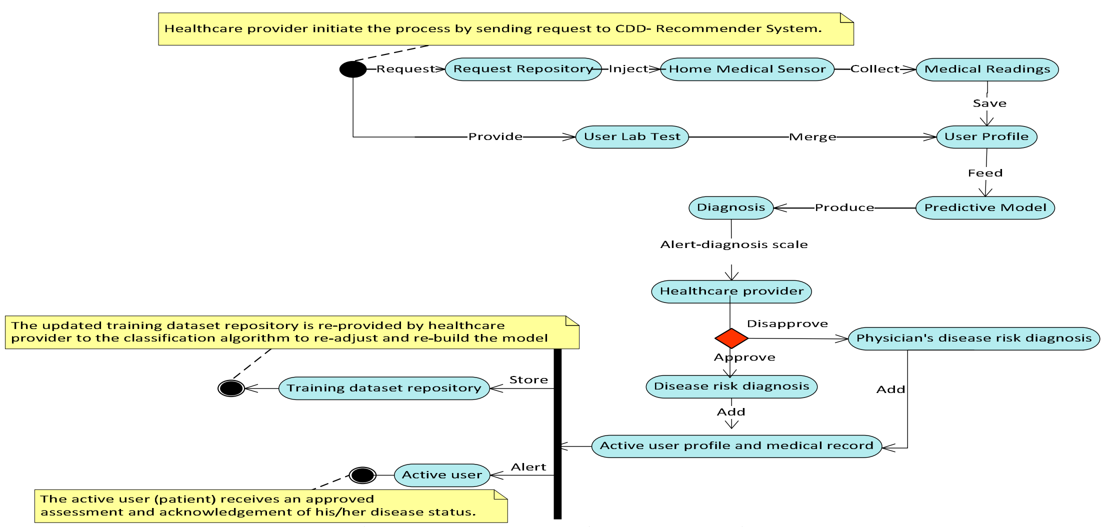

Figure 4: CDD Recommender System Scenario

 ISBN: 978-1-61208-243-1

V. OUTCOMES

In this section, we present experimental results of the proposed predictive model and compare the prediction performance to other decision tree models like: J48, Decision Stump and REP Tree. The dataset used for this experiment is diabetes data set collected from hospitals in Oman. The collected data represented by 935 patients' records, "13" attributes including the doctors' diagnosis scale, and values of attributes for each case.

A. Metrics

Metrics such as MAE [7], Precision and Recall [15] are used to evaluate and compare the prediction performance of algorithms used. The accuracy of the system is high as the MAE of the prediction system is low. Precision and Recall should be maximized for well-performed prediction system.

The MAE is given by:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \qquad (1)$$

where MAE is an average of the absolute errors $ei = fi - yi$, fi is the predicted class (diagnosis) and yi is the actual class.

$$\text{Precision} = \frac{true\ positive}{true\ positive + false\ positive} \qquad (2)$$

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive+false negetive}} \qquad (3)$$

B. Experimental results

Our experimental in building the predictive model for CDD recommender system using decision tree algorithms, has shown the following:

- RF has been built of 10 trees. Out of bag estimated error is 0.2877.
- RF has demonstrated 99.7% of correctly classified cases.
- Since we do not have a large dataset, we conducted a 10-fold cross validation [25]. Cross validation has been used with: J48, Decision Stump, and REP Tree algorithms. However, RF does not need cross validation method [14] because the test set error is estimated internally during the run.
- RF improves the accuracy by handling the missing data internally through the training and testing phase. In our case study, the dataset is complete. However, we randomly put missing values along the training data, and re-build the model with the aforementioned algorithms. The results show that our RF model still outperformed all other models.

The comparison results of predictions performance for different models show:

The model built using RF has less MAE and high Precision and Recall results compared with the other models results (as shown in Fig. 5). That shows the high quality and performance of our recommendation system for CDD built based on RF.

C. Evaluation and Results

To test our system, evaluate the performance and determine the efficiency of the algorithm used, 20 random samples of real patients' medical records have been tested and applied on the RF learning model and the other trees models. The results show that the RF model outperformed all other models (as seen in Fig. 6).

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented an effective and reliable recommender system approach for CDD using data mining technique. The main contributions of this study are summarized as follows:

- The utilized method is expected to generate effective disease risk prediction for chronic disease patients.
- An ensemble trees classification algorithm is adopted. It handles the shortcomings of the medical data and generates a robust learning model that is able to provide more accurate predictions for online users.
- An Attribute Selection method, that determines the relevant features of medical data, is applied.
- The laboratory and daily-home test data of chronic disease are merged to accomplish more accurate results.
- Reliable prediction and recommendation results approved by physicians are sent to patients as alert messages.

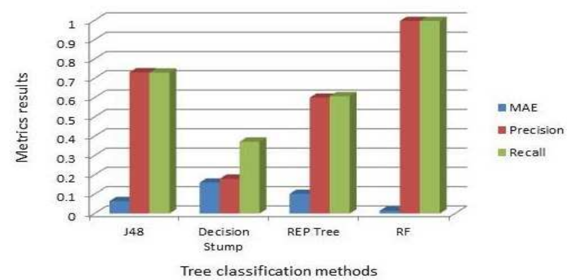

Figure 5: Metrics results of evaluating the prediction performance of J48, Decision Stump, REP Tree and RF

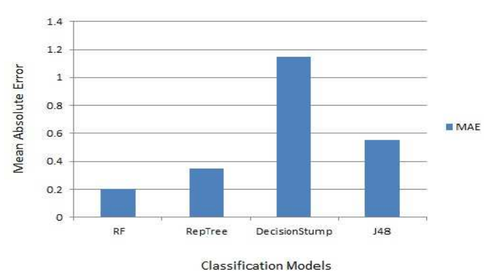

Figure 6: MAE values of four tree classification models tested on "20" samples of patients medical records

Further study on testing different predication and recommendation methodologies will be considered to improve the system accuracy. In addition, more chronic disease case studies will be tested with the generated model. Moreover, the prediction result will be considered as one of the patient external features during our future work of extending the CDD recommender system. The extended work aims to provide medical advices and treatments recommendation along with the disease risk status.

REFERENCES

[1] J.A. Al-Lawatia and J. Tuomilehto, "Diabetes risk score in Oman: A tool to identify prevalent type 2 diabetes among Arabs of the Middle East," Journal of Diabetes and Clinical Practice, vol. 77, pp. 438-444, 2007.

[2] L. Duan, W. N. Street, and E. Xu, "Healthcare information systems: data mining methods in the creation of a clinical recommender system," Enterprise Information Systems, vol. 5, pp. 169-181, 2011/05/01 2011.

[3] B. M. Sarwar, G. Karypis , J. Konstan , and J. Riedl, "Recommender Systems for Large-Scale E-Commerce: Scalable Neighborhood Formation Using Clustering," presented at the The Fifth International Conference on Computer and Information Technology (ICCIT 2002), 2002.

[4] D. Sun, Z. Luo, and F. Zhang, "A novel approach for collaborative filtering to alleviate the new item cold-start problem," presented at the Communications and Information Technologies (ISCIT), 2011 11th International Symposium on, Hangzhou, 2011.

[5] G. Pitsilis and L. Marshall, "Trust as a Key to Improving Recommendation Systems

Trust Management." vol. 3477, P. Herrmann, et al., Eds., ed: Springer Berlin / Heidelberg, 2005, pp. 421-432.

[6] B. M. Marlin, "Missing Data Problems in Machine Learning," Doctor of Philosophy, Computer Science, University of Toronto, 2008.

[7] N. S. N. Gayatri, A. V. Reddy, and R. Chitra, "Performance Analysis of Datamining Algorithms for Software Quality Prediction," presented at the Proceedings of the 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 2009.

[8] X. Su and T. M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," Advances in Artificial Intelligence, vol. 2009, 2009.

[9] H. Kaur and G. P. S. Raghava, "A neural-network based method for prediction of γ-turns in proteins from multiple sequence alignment," Protein Science, vol. 12, pp. 923-929, 2003.

[10] T. Kudo and Y. Matsumoto, "A Boosting Algorithm for Classification of Semi-Structured Text," in Proceedings of EMNLP 2004, 2004, pp. 301-308.

[11] J. Park, H. R. Tyan, and C. C. J. Kuo, "Internet Traffic Classification for Scalable QOS Provision," in Multimedia and Expo, 2006 IEEE International Conference on, 2006, pp. 1221-1224.

[12] A. Özçift, "Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis," Computers in Biology and Medicine, vol. 41, pp. 265-271, 2011.

[13] L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5-32, 2001.

[14] L. Breiman, and A. Cutler. Random Forests. Available: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm #workings. Accessed on August 2012.

[15] Y. Qi , J. Klein-seetharaman , and Z. Bar-joseph, "Random Forest Similarity for Protein-Protein Interaction Prediction," Pac Symp Biocomput, vol. 2005, pp. 531--542, 2005.

[16] D. R. Cutler, T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. C. Gibson, and J. J. Lawler, "Random Forest For Classification In Ecology " Ecology, vol. 88, p. 10, 2007 2007.

[17] P. Pattaraintakorn, G. Zaverucha, and N. Cercone, "Web Based Health Recommender System Using Rough Sets, Survival Analysis and Rule-Based Expert Systems Rough Sets, Fuzzy Sets, Data Mining and Granular Computing." vol. 4482, A. An, et al., Eds., ed: Springer Berlin / Heidelberg, 2007, pp. 491-499.

[18] D. A. Davis, N. V. Chawla, N. A. Christakis, and A.-L. Barabási, "Time to CARE: a collaborative engine for practical disease prediction," Data Mining and Knowledge Discovery, vol. 20, pp. 388-415, 2010.

[19] M. Sapon, K. Ismail, S. Zainudin, and C. Ping, "Diabetes Prediction with Supervised Learning Algorithms of Artificial Neural Network," presented at the International Conference on Software and Computer Applications, Kathmandu, Nepal, 2011.

[20] Noh, H.-ju, Kwak, M.-jung, Han, and I. goo, "Handling Incomplete Data Problem in Collaborative Filtering System," Journal of intelligent information systems, vol. 9, pp. 51-63, 2003.

[21] B. C. Ko, et al., "Cell image classification based on ensemble features and random forest," Electronics Letters, vol. 47, pp. 638-639, 2011.

[22] S. B.Patil, and Y.S.Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network," European Journal of Scientific Research, vol. 31, pp. 642-656, 2009.

[23] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, pp. 123-140, 1996.

[24] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti, "Similarity-based Classification: Concepts and Algorithms," Machine Learning Research, vol. 10, pp. 747-776, 2009.

[25] M. Saeedmanesh, T. Izadi, and E. Ahva, "HDM: A Hybrid Data Mining Technique for Stock Exchange Prediction," presented at the Proceedings of the International MultiConference of Engineers and Computer Scientists 2010, Hong Kong, 2010.