DOI: http://dx.doi.org/10.26483/ijarcs.v9i2.5856

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

LIVER CANCER PREDICTION FOR TYPE-II DIABETES USING CLASSIFICATION ALGORITHM

J. Kumara Kumar
Assistant Professor, Computer Science and Engineering,
Pondicherry Engineering College, Puducherry, India

S. Agilan
M. Tech Student, Computer Science Dept.
Pondicherry Engineering College, Puducherry, India

Abstract: In recent years, type II diabetes with liver cancer became a serious disease that threatens the health and mind of human. Efficient predictive modelling is required for medical researchers and practitioners. To develop a prediction model using data mining technique for type II diabetes patients with liver cancer within 6 years of diagnosis. Data has been collectedfrom the NHIRD (National Health Insurance Research Database). That selected patients who were newly diagnosed with type II diabetes. In this data 2060 cases were founded and assigned them to a case group (diagnose patients with liver cancer) and control group (diagnosed patients without liver cancer). In This proposal a liver cancer prediction for type II diabetes predictive model based on random forest which aims at analysing some readily available indicator (age, liver diseases, Alcoholic fatty liver diseases, hyperlipidaemia, etc.)using this the risk factor were identified, then chi-square test was conducted on each independent variable to make a differentiate between patients with liver cancer and patients without liver cancer. The dataset were randomly divided into two groups (training group and testing group). The training group contain of 70% of dataset (1442 cases) where the prediction model was done using training dataset. The remaining 30% of dataset is assigned to the test group for model validation. Random forest algorithm uses multiple decision trees to train the samples, and integrates weight of each tree to get the final results. The validation result shows that the random forest algorithm can greatly reduce the problem of modelling error of the single decision tree, and it can effectively predict the impact of these readily available indicators on the risk liver cancer for diabetes patients. Additionally, to get better prediction accuracy in random forest model than using the Artificial Neural Network (ANN), AdaBoost and Logistic Regression algorithm.

Keywords: Artificial Neural Network (ANN), AdaBoost, LogisticRegression, Random forest algorithm.

I. INTRODUCTION

Liver cancer is the sixth most common cancer world-wide and is the third leading cause of cancer related to death [1,2]. The main reason for liver cancer is alcohol usage [3] and a high occurrence of hepatitis b and hepatitis c, which added to chronic cirrhosis and hepatocellular carcinoma [4]. the another study shows non-alcoholic fatty liver diseases is common in type-II diabetes patients [5] and several liver diseases like alcoholic fatty liver diseases, non- alcoholic fatty liver diseases and cirrhosis may also increase the risk of liver cancer development [6–8].

Although these risk factor have been investigated adequately, that more papers reporting a negative and positive association between cancer and diabetes. The pathogenic contrivances underlying the relationship between cancer and diabetes were explained [9]. The us studies have indicated that type II diabetes at the risk of cancer with those without diabetes [10-12]. To diabetes patients the liver and pancreas are exposed to high involvement of insulin, it has the high probability of liver cancer may increase [13].In recent years, using predictive classification in medical diagnosis has received a strong boost owing to earnest research activity in this field in recent times. And majority of papers published deal with the goal of improving accuracy. For example, Karol Grudzinski used the KNN algorithm (k=22) to obtain the highest accuracy of the model is 75.5% [14]. The neural network achieved an accuracy of 75.4% whereas the Bayesian approach achieved 79.5% accuracy in reference [15]. Allah Verdi proposed a hybrid neural network (artificial neural networks (ANN) and fuzzy neural network (FNN) model), the precision of this method to get the value of 84.24% [16]. Although the accuracy of the model has been improving, it is obvious that the research methods and the improved algorithms are the indisputable single classifier. However, the ensemble classifier is better than the single classifier in many cases. Also, these predictors in the model are not directly visible and need to be measured by certain medical equipment, which increases the cost of the patient's diagnosis. Besides, the datasets contain a lot of readily available indicators (such as sex, age, alcoholic cirrhosis, non-alcoholic cirrhosis, alcoholic hepatitis, viral hepatitis, other types of chronic hepatitis,), using these indicators to predict diabetes patients that can greatly reduce the cost of diabetes liver cancer prediction so the system may choose these external readily available indicators to judge the impact on diabetes and try to prevent it in the bud. Random forest is an ensemble classifier composed of multiple decision trees, which has the advantages of high accuracy and good robustness [17]. Therefore, the present study uses random forest as the basic classifier.

On the basis of the prediction model, the main concept is to develop an application to enable physicians to detect the probability of liver cancer in future 6-year period.

II. PROPOSED MODEL

2.1. Data source

Data were taken from the NHIRD of Taiwan, The NHIRD encompasses all medical privileges data of almost 23.72 million people, including over 99% of the Taiwan population [18]. This study used the LHIRD (Longitudinal Health Insurance Database) 2010, which covers the health insurance data of 2 million people in 6 year time period [19].

This paper chose recently determined patients to have compose type II diabetes patients (from 2000 to 2003 who did not have a history of cancer (n = 65,871) [19]. A while later, this model utilizes encoded singular recognizable proof information to perform information linkage with the disease registry database to distinguish whether the patients had been determined to have liver growth(International Classification of Diseases, (ICD-O-3 = C22.0 and C22.1))between 2001 and 2009.In fig.1, it is found 515 diabetes patients who established liver cancer within 6 years after diabetes diagnosis. Those studies [20-21] have reported that the ratio of the test group to the control group should not be more than 1:3 ratio; using other ratios it may lead to a biased comparison.

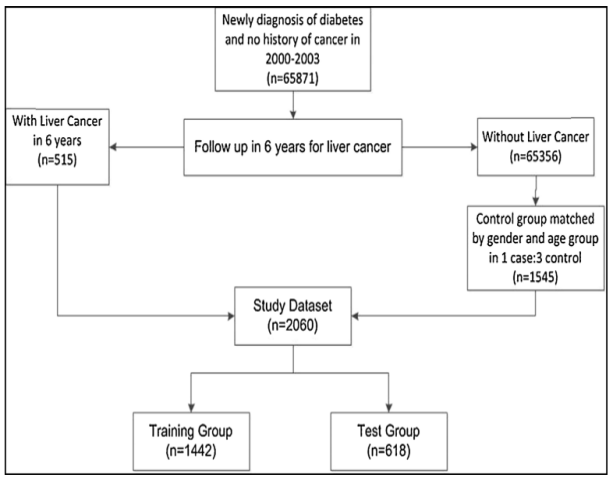

Fig.1 Research Flow

2.2 Random Forest model

The random forest algorithm, proposed by Dr.Breiman in 2001, has been to a great degree effective as a universally useful grouping and relapse technique. The approach, which joins separate randomized decision trees and aggregates their predictions by averaging which has shown excellent performance in settings where the quantity of factors is significantly bigger than the quantity of observations [22].. It is an algorithm based on statistical learning theory, which uses Bootstrap randomized re-sampling way to extract multiple versions of the sample sets from the original training datasets, then building a decision tree model for each sample set, the final combined all the results of the decision trees to predict the results of classification by the established voting mechanism. The detailed process is shown in Fig.2.

2.3 Data Pre-Processing

The data should be carefully collected, integrated and prepared for analysis. In this study, the model applied the techniques of data pre-processing to improve the quality of the mining results and the efficiency of the mining process. The raw dataset is provided by NHIRD of Taiwan. Which has 2060 cases The raw data was randomly categorized into two groups (training group and test group); the training group consist of 1442 cases (70% of dataset). The prediction model was developed based on training dataset. The remaining 30%cases is assigned as test group. The 70/30 percentage rule was applied on the basis of some studies such as by Antonio Mucherino [23] and CogNova Technologies [24], and each tester consists of 10 features including age group, gender, alcoholic cirrhosis, other cirrhosis, alcoholic hepatitis, viral hepatitis, etc . In this datasets, it is easy to judge whether or not the tester has liver cancer using sequential mining optimization algorithm, they have indicate that 70% of data is sufficient for developing random forest model and remaining data can be used for validation, if the dataset is small 90% is used as training set and remaining 10 fold validation is used.

This model is also tested the SVM,ANN(artificial neural network)and Logistic regression on our data in this study and employed WEKA to devise this models. The decision tree included sex, alcoholic cirrhosis, cirrhosis, viral hepatitis, chronic hepatitis, alcoholic fatty liver disease, hyperlipidaemia, and age as parameters, and the decision tree algorithm was used to construct random forest model. By contrast, to devise the ANN model, it is used in the sequential minimal optimization algorithm, and included as factors sex, alcoholic cirrhosis, cirrhosis, viral hepatitis, chronic hepatitis, alcoholic fatty liver disease, and other types of fatty liver disease,

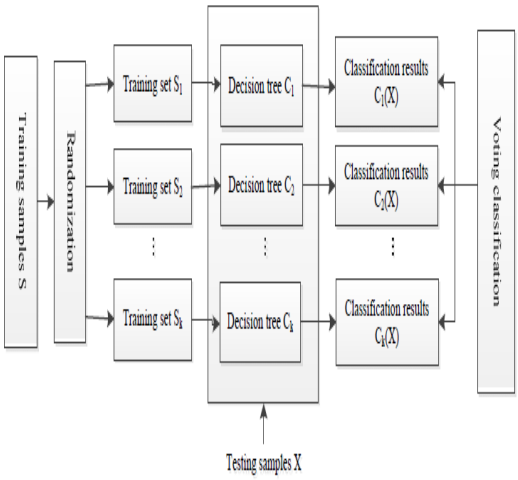

Fig. 2. Flow chart of the random forest algorithm

Hyperlipidaemia, and age. Additionally, the random forest model have some features of missing values in different degrees, including alcoholic cirrhosis, non-alcoholic cirrhosis, alcoholic hepatitis, viral hepatitis, etc. . However, our study are mainly used some readily available indicators to predict the risk of liver cancer for diabetes, it can conduct dimensionality reduction first. The key advantage of dimensionality reduction is to enhance the execution of the calculation, because the dimensionality reduction can separate the unrelated features and reduce the noise [25]. As the statistical result are shown in TABLEI.

Table I: Statistics data –relationship between factors and liver cancer.

Factors	***P*-value**
Gender	0.869
Age group	0.999
Alcoholic liver cirrhosis	0.018*
Liver cirrhosis	<0.001*
Alcoholic hepatitis	0.134
Viral hepatitis	<0.001*
Chronic hepatitis	<0.001*
Alcoholic fatty liver	0.125
Other fatty liver	0.248
Hyperlipidaemia	<0.001*

$*P$-value < 0.05, it means the factor have significant relationship with liver cancer.

In order to improve the accuracy of the model, the continuous features are often needed to be discretized[26]. Discretization involves two tasks: First, to determine the number of classification that it need; Second, to determine how to map continuous features values to these classification values. For the first sub-tasks, it can handle like this: after the continuous features values are sorted, divide them into n intervals by specifying the n-1 points. As for the second subtasks, it will map all the values in an interval to the corresponding classification value. Therefore, the discretization is to choose the number of split points and determine the point location problem. In order to facilitate the processing of the data, it will divide each feature into three parts and using low, medium and high represent these feature values, respectively. The next step is to determine the split point; there are three kinds of methods to determine the split point, namely: width discretization, frequency discretization and k-means discretization. After the experimentation, the performance of k-means discretization is the best [27].According to the centre point of each feature obtained by k-means discretization, the features of discretization are shown in TABLE II

Table II. Discretized features

Name	**Low**	**Medium**	**High**
Age	0	0.55~0.90	≥ 1
Liver cirrhosis	0	0.001~0.015*	≥1
Alcoholic hepatitis	0	0.01~0.134	≥1
Alcoholic fatty liver	0	0.001~0.134	≥1

III. EXPERIMENT DESIGN

After the data pre-processing, the next goal is to dig out the relationships between the various features and extract some useful patterns. Now, the main idea is to develop a risk of liver cancer for type II diabetes model to predict whether a person will develop liver cancer. The construction steps of the random forest mainly include generating a training set, choosing the splitting point, repeating construct the classification and regression tree and the voting. Detailed procedure is as follows:

Step1: using Bootstrap re-sampling techniques to generate k (In this paper, the k is 10) samples. Theoretically k samples cover 2/3 of the original datasets, and the rest of the data is called Out-Of-Bag (OOB), OOB can be used as test data [28].

Step2: using the k samples to generate k decision trees. At each node of each tree, that are randomly selected m features (m<M) in the M features, it is suggested starting with m = $\sqrt{M}$and then decreasing or increasing m until the minimum error for the OOB data set is obtained. Finally choose the best split according to the Gini criterion.

Gini criterion and prediction class labels are shown in the Eq. (1) and Eq. (2).

$$\text{Gini}(A_i) = 1 - \sum_{i=1}^{n} p_i^2$$

-→ Eq. (1)

Where $p_i$represents the probability of the i-th class instance; n is the number of classes; A_i represents the i-th feature.

$$C_D = argmax_c \frac{I}{K} \sum_{k=0}^{n} I \frac{n_{h_i}}{n_{h_i}}, (C)$$

-→ Eq. (2)

Where C_D represents prediction class labels; arg max$c$ represents a parameter to find the maximum score c; k represents the number of decision trees in a random forest; I(*) represents indicator function; n_{h_i}, C, represents the classification results of the decision tree for the c class; n hi represents the number of leaf nodes in the decision tree h_i.

Step3: according to the previous two steps to predict the test samples, and combined with the test results of each tree and determines the final result in accordance with majority rule voting mechanism.

In order to validate the effectiveness of the proposed methods, it utilizes another three algorithm, namely ANN model, Logistic regression algorithm and AdaBoost algorithm. Additionally, in order to further demonstrate the effectiveness of the method used in this study. This model is designed in a different set of contrast experiments. First, the data set was divided into four subsets (20%, 40%, 60%, and 80%)of the total data set, respectively), and each model was compared in each subset. The overall framework of model building is shown in Fig.3

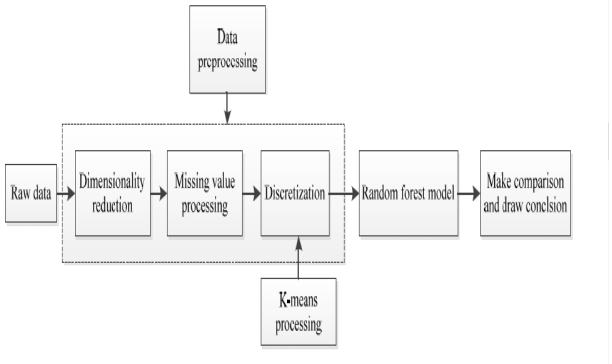

Fig.3 The framework of model building

IV. RESULTS ANALYSIS AND MODEL EVALUATION

According to the above experiment, it is easy to see that the root node of each tree in a random forest, including Alcoholic liver cirrhosis, Liver cirrhosis, Alcoholic hepatitis, A0lcoholic fatty liver and Age. It indicates that these are all important external indicators to determine whether they are suffering from liver cancer . The rules are as follows:

1) If Liver cirrhosis ≥0 and Alcoholic fatty liver ≥1, then diabetic.
2) If Alcoholic fatty liver ≥ 1 and Chronic hepatitis ≤ 1, then diabetic.

If the first rule is considered, the rule is supported by a previous study , which showed that Liver cirrhosis or Alcoholic fatty liver discriminate better the cases with diabetes from those without, as compared with Alcoholic hepatitis. Considering the second rule, recent study have shown a similar study which showed that Alcoholic fatty liver -to- Chronic hepatitis can be used to identify liver cancer for Type II diabetes.

In medical diagnosis accuracy, sensitivity and specificity are the common measures of performance metrics. Accuracy determines ability of the classifier to produce accurate disease diagnosis. Sensitivity measures the ability of the model to identify the occurrence of target class accurately. Specificity measures the ability of the model to separate the target class. The Accuracy, Sensitivity and Specificity are measured as follows [29].

-→Eq. (3)

-→Eq. (4)

-→Eq. (5)

Where the True Positives (TP) and True Negatives (TN) are correct classifications. A False Positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A False Negative (FN) occurs when the outcome is incorrectly predicted as no when it is actually yes.
The k-fold cross validation is a best measure for classifier performance [30]. Therefore, in our study, ituses the 10-fold cross validation method to evaluate the reliability of the model. According to the Eq. (3), the accuracy rate of the random forest model is 85.00%, which is calculated by the 10-fold cross validation method. In addition, this paper also carried out ANN, Logistic regression and AdaBoost algorithms to obtain accuracy of the model are as follows: 78.57%, 79.89% and 84.19%. Fig.4 presents the bar graph of accuracy for 4 models'

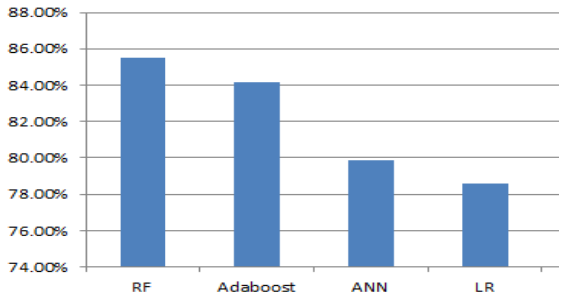

Fig. 4. Accuracy for different models

According to the Eq. (4), Eq. (5), it is found that the sensitivity reached a comparative high level, they achieved a sensitivity of 91.17%, 92.11%, 99.05% and 100% respectively. While the specificity achieved a quite low level. By observing the data and searching the literature, it is found that the data imbalance leads to this problem. And this is a direction for the next research. However, compared to ANN, Logistic regression and AdaBoost, the accuracy and sensitivity of random forest model are more satisfactory. It has a certain guiding significance for the early warning of diabetes whether having liver cancer or not. Taking into account the amount of data used in the experiments is relatively small, it is made 4 groups of comparative experiment to strengthen the persuasiveness of experiments. According to the characteristics of the data set, the model is being set up four subsets of different sizes, accounting for 20%, 40%, 60%, 80% of the total data set, respectively. With the above methods, the accuracy of each experimental group were shown in TABLE III..

Table III. Comparison with different scale data

Algorithms	**Group 1**	**Group 2**	**Group 3**	**Group 4**
Random Forest	60.00%	67.78%	80.07%	84.13%
ANN	66.65%	69.93%	78.64%	80.50%
Logistic regression	62.25%	66.73%	66.36%	72.94%
AdaBoost	66.70%	70.10%	79.06%	81.61%

In order to better observe the effect of random forest algorithm in each experimental group, and make its line graph. It is shown in Fig. 5.
In Fig. 6, it can be concluded that with the expansion of the data, the accuracy of the random forest model is constantly improved. Additionally, the accuracy of the random forest model is also constantly improving while the same amount of increased data in the similar proportion of cases. So the random forest model can effectively predict the liver cancer for type II diabetes patients in the case of a sufficient amount of data

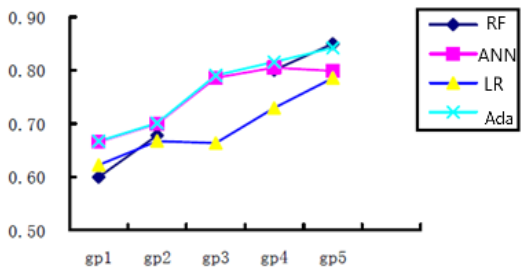

Fig. 5. The line graph of comparison with different scale data

.

V. CONCLUSION

The data mining technique has played a very important and decisive role in the medical industry. In this paper, it is to obtain some simple decision rules by establishing the random forest model, and which can make a simple prediction of whether having liver cancer or not for type II diabetes patients by these simple and readily available indicators. Additionally, these indicators are relatively easy to obtain and measured by physician, so they can greatly reduce the cost of diagnosis. By using these indicators to predict liver cancer for diabetes patients will have a certain practical significance.

In this paper, just use some readily available indicators to predict the risk of liver cancer for type II diabetes and there is no further study the impact of other indicators of illness, also not taken into account the impact of the tester itself suffering from other diseases on the prediction of diabetes. Expand other indicators to predict the risk of disease and update the perspective of data mining are the future direction of the prediction of the risk of liver cancer for type II diabetes patients.

VI. ACKNOWEDGEMENT

The authors gratefully acknowledge the authorities and Pondicherry Engineering College for the facilities offered to carry out this work.

VII. REFERENCE

[1] Jemal, Ahmedin, et al. "Global cancer statistics." CA: a cancer journal for clinicians 61.2 (2011):pp. 69-90.

[2] Singh, Gopal K factors b., Mohammad Siahpush, and Sean F. Altekruse. "Time trends in liver cancer mortality, incidence, and risk y unemployment level and race/ethnicity, United States, 1969–2011." Journal of community health 38, no. 5 (2013): 926-940.

[3] Alcohol Use and Cancer, 2014, Available from:http://www.cancer.org/cancer/cancercauses/dietandphysicalactivity/alcohol-use-and-cancer.

[4] Momin, Behnoosh, and Lisa Richardson. "An analysis of content in comprehensive cancer control plans that address chronic hepatitis B and C virus infections as major risk factors for liver cancer." Journal of community health 37, no. 4 (2012): 912-916.

[5] Schulz, P. O., Ferreira, F. G., Nascimento, M. D. F. A., Vieira, A., Ribeiro, M. A., David, A. I., &Szutan "Association of non-alcoholic fatty liver disease and liver cancer." World Journal of Gastroenterology: WJG 21, no. 3 (2015):p. 913.

[6] Vongsuvanh, Roslyn, David van der Poorten, and Jacob George. "Non-alcoholic fatty liver disease-related hepatocellular carcinoma: a sleeping tiger in the Asia Pacific." Hepatology international 7, no. 2 (2013): 823-832.

[7] Hashimoto, Etsuko, and KatsutoshiTokushige. "Hepatocellular carcinoma in non□alcoholic steatohepatitis: Growing evidence of an epidemic?." Hepatology Research 42, no. 1 (2012): pp. 1-14.

[8] Baffy, György, Elizabeth M. Brunt, and Stephen H. Caldwell. "Hepatocellular carcinoma in non-alcoholic fatty liver disease: an emerging menace." Journal of hepatology 56, no. 6 (2012):pp. 1384-1391.

[9] P.F.F. Vigneri, L. Sciacca, G. Pandini, R. Vigneri, Diabetes andcancer, Endocr. Relat. Cancer 16 (4) (2009) pp. 1103–1123

[10] Lowenfels, Albert B., and Patrick Maisonneuve. "Risk factors for pancreatic cancer." Journal of cellular biochemistry 95, no. 4 (2005):pp. 649-656.

[11] Chang, Chia□Hsuin, Jou□Wei Lin, Li□Chiu Wu, Mei□Shu Lai, Lee□Ming Chuang, and K. Arnold Chan. "Association of thiazolidinediones with liver cancer and colorectal cancer in type 2 diabetes mellitus." Hepatology 55, no. 5 (2012):pp. 1462-1472..

[12] Bosetti, Cristina, ValentinaRosato, DaniloBuniato, AntonellaZambon, Carlo La Vecchia, and Giovanni Corrao. "Cancer risk for patients using thiazolidinediones for type 2 diabetes: a meta-analysis." The oncologist 18, no. 2 (2013): pp.148-156.

[13] Printz, Carrie. "Diabetes associated with increased risk of liver cancer." Cancer 120, no. 9 (2014): 1288-1288.

[14] Grudziński, Karol. "Towards Heterogeneous Similarity Function Learning for the k-Nearest Neighbors Classification." In International Conference on Artificial Intelligence and Soft Computing, pp. 578-587. Springer, Berlin, Heidelberg, 2008.

[15] Bioch, Jan C., Onno Van Der Meer, and Rob Potharst. "Classification using Bayesian neural nets." In Neural Networks, 1996., IEEE International Conference on, vol. 3, pp. 1488-1493. IEEE, 1996.

[16] Kahramanli, Humar, and NovruzAllahverdi. "Design of a hybrid system for the diabetes and heart diseases." Expert systems with applications 35, no. 1-2 (2008):pp 82-89.

[17] Fawagreh, Khaled, M. M. Gaber, and EyadElyan. "Random forests: from early developments to recent advancements." Systems Science & Control Engineering: An Open Access Journal 2, no. 1 (2014): pp. 602-609.

[18] Cheng, C. L., Kao, Y. H. Y., Lin, S. J., Lee, C. H., & Lai, M. L. "Validation of the National Health Insurance Research Database with ischemic stroke cases in Taiwan." Pharmacoepidemiology and drug safety 20, no. 3 (2011):pp. 236-242.

[19] Rau, H. H., Hsu, C. Y., Lin, Y. A., Atique, S., Fuad, A., Wei, L. M., & Hsu, M. H. (2016) "Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network." computer methods and programs in biomedicine125 (2016):pp. 58-65

[20] N. Chawla, Data Mining for Imbalanced Datasets: AnOverview Data Mining and Knowledge Discovery Handbook,2010, pp. 875–886.

[21] Lee, Paul H. "Resampling methods improve the predictive power of modeling in class-imbalanced

datasets." International journal of environmental research and public health 11, no. 9 (2014):pp. 9776-9789.

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

[23] A.P.P. Mucherino, P.M. Pardalos, Data mining in agriculture. Vol. 34. Springer Science, 2009.

[24] Karaboga, Dervis, and CelalOzturk. "Neural networks training by artificial bee colony algorithm on pattern classification." Neural Network World 19, no. 3 (2009):pp. 279..

[25] M. G. Ahamad, A. Aljumah, and M. K. Siddiqui, "Application of data mining: Diabetes health care in young and old patients." Journal of King Saud University-Computer and Information Sciences 25, no. 2 (2013):pp. 127-136

[26] J. C. Han, J. C. Rodriguze, and M. Beheshti, "Diabetes data analysis and prediction model discovery using rapidminer." In Future Generation Communication and Networking, 2008. FGCN'08. Second International Conference on, vol. 3, pp. 96-99. IEEE, 2008.

[27] P.N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Pearson Addison Wesley, 2006

[28] Seyedhosseini, Mojtaba, and TolgaTasdizen. "Disjunctive normal random forests." Pattern Recognition 48, no. 3 (2015):pp. 976-983.

[29] Z. H. Zhou, Machine Learning, 1st ed., Tsinghua University Press, 2016,pp.28-36.

[30] N. Esfandiari, M. R. Babavalian and V. K. Tabar, "Knowledge discovery in medicine: Current issue and future trend." Expert Systems with Applications 41, no. 9 (2014):pp. 4434-4463.