ABSTRACT

Hypertension disease is one of the common chronic diseases to be found in human beings. It is always better to detect the disease at an earlier stage of life to prevent further harmful effects and to devise proper treatment. In this study, the hypertension is taken into consideration for early prediction and diagnosis of the disease by using the data mining technique to analyze the data. In healthcare organizations the volume of data is more. To get knowledge from those data we need an efficient technique. Data mining is used for the purpose of discovering knowledge from vast amount of database. To classify the stage of hypertension, classification technique which is one of the data mining technique is used. The data is collected from 149 household of students from Public Health College in Jazan University. The research work is done with WEKA open source software under Windows7 environment. An experimental study is carried out using data mining techniques such as J48 and Random Forest tree. The data records are classified as five categories such as normal, elevated, Hypertensive crisis, stage 1 and stage 2. As a result, the performance is evaluated for both classification techniques and their accuracy compared through confusion matrix. It has been concluded that J48 gives better accuracy than the Random Forest tree technique.

Keywords: Hypertension, Data Mining, Classification, Decision Tree.

I. INTRODUCTION

Hypertension is one of the conditions which leads heavy dangerous health issues when the blood vessels have persistently raised pressure. Blood pressure is formed by the force of blood pushing against the walls of blood vessels (arteries) because it is pumped by the center. The higher the pressure the harder the center have gotten to pump. Hypertension may be a major disease and may be a significant explanation for death everywhere in the planet. The relevant researches show that the cardiovascular disease is an important risk causing hypertension\textsuperscript{1,2}. As defined by the National High Blood Pressure Education Program (NHPEP), hypertension can be summarized as shown in Table 1. Hypertension, also known as high or raised blood pressure, is a condition during which the blood vessels have persistently raised pressure. As defined by the National High Blood Pressure Education Program (NHPEP), hypertension can be summarized as shown in Table 1\textsuperscript{1,2,3}.

Table 1. Stages of Blood Pressure

<table>
<thead>
<tr>
<th>Classification of blood pressure for adults aged 18 and older (NHPEP, 2002) Category</th>
<th>Systolic (mm Hg)</th>
<th>Diastolic (mm Hg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>&lt;120</td>
<td>&lt;80</td>
</tr>
<tr>
<td>Elevated</td>
<td>&lt;130</td>
<td>&lt;85</td>
</tr>
<tr>
<td>Stage 1</td>
<td>130-139</td>
<td>85-89</td>
</tr>
<tr>
<td>Stage 2</td>
<td>140-179</td>
<td>90-99</td>
</tr>
</tbody>
</table>
Hypertension disease may be a significant ill health, and patients might not be ready to recognize this disease for years. As a result, it is going to damage the patient's kidney, heart and veins. Hence, early diagnosis and hypertension treatment is extremely important. The rationale for this importance is that the damage caused by hypertension on organs, and high treatment costs and loss of labor that occur as a result. A serious challenge facing health care institutions is that the supply of quality services at reasonable prices. Poor clinical choices will end in devastating consequences and the worth of clinical tests should be reduced by Hospitals. This will be achieved by using computer-based information or decision support systems. Many hospital information systems are designed to support patient billing, inventory management and generation of straightforward statistics. Some hospitals use decision support systems but are largely limited within the sense that they cannot answer complex queries. However, they cannot answer complex queries like “Given patient records, predict the probability of patients getting hypertension. Several times, clinical decisions are often made supported doctors’ intuition and knowledge without counting on the knowledge rich data hidden within the database. This practice results in unwanted biases, errors and unwarranted medical costs which affects the standard of service provided to patients. The anticipated system is aimed toward reducing medical errors, enhance patient safety, decrease unwanted practice disparity, and improve patient outcome. This research work is presents data processing as a viable tool for generating a knowledge rich environment which may help to significantly improve the standard of clinical decisions. Data processing combines applied math analysis, machine learning and knowledge technology to extract hidden patterns and relationships from huge databases. During this work J48 and Random Forest tree algorithm is used to form a model with predictive capabilities.

1.1 Problem Statement
This research focus on the performance analysis of different data mining techniques applied on hypertensive data set to classify the stage of hypertension.

1.2 Objectives
1.2.1 General Objective
The aim of this study is to develop a predictive model using J48 and Random Forest tree classifier.

1.2.2 Specific Objectives
a. To implement data mining techniques such as J48 and Random Forest tree to predict hypertension stage from the hypertension dataset.

b. To compare the performance of classifiers to identify which classifier predict the disease correctly with high accuracy in terms of confusion matrix and help the physicians to predict the stage of the disease and take decision at proper time.

II. REVIEW OF LITERATURE
M. Tureet.al, 2005 conducted study on Comparing classification techniques for predicting essential hypertension. In this study, performance of classification techniques is compared to predict the danger of hyperpiesia disease. They performed retrospective analysis in 694 subjects (452 patients and 242 controls). They compared performances of three decision trees, four statistical algorithms, and two neural networks. They found neural networks performed better performance than others.

Huang et.al, 2011, they constructed a case-based classifier for hypertension detection. In their work, a classification approach was constructed with the use of hybrid of case-based reasoning and genetic algorithms for hypertension detection using anthropometric body surface scanning data. The obtained result reveals the connection between a subject’s 3D scanning data and hypertension disease.

The GA is adopted to work out the acceptable feature weights for CBR. The proposed approaches were experimented and compared with a daily CBR and other widely used approaches including neural nets and decision trees. The experiment showed that applying GA to work out the acceptable weights in CBR may be a feasible approach to improving the effectiveness of case matching of hypertension disease. They also

www.turkjphysiotherrehabil.org
demonstrated the different weighted CBR approach that presents better classification accuracy over the results obtained from other approaches also.

Zheng et.al, 2012 developed Intelligent ZHENG Classification of Hypertension Depending on ML-kNN and Information Fusion. In this paper, ML-kNN—a multilabel learning model—is used because the classification model for hypertension. Feature-level information fusion is additionally used for further utilization of all information. Their results shown from experiment gives that the ML-kNN can model the hypertension's ZHENG differentiation well. Information fusion helps improve models' performance.

Aljumah and Siddiqui.et.al, 2014, they computed the probability and prediction of hypertension using data mining techniques and concluded that smoking cessation is the best intervention followed by exercise, diet, weight and drug for the hypertension intervention in Saudi Arabia. Hence, all hypertension patients are unambiguously advised to stop smoking.

Kokyeret.al, 2015, they created hypertension database belonging to patients who arrived at hospital in different times which includes: age, sex, body mass index, HDL, LDL, triglyceride, uric acid, smoking and whether that person has hypertension or not; and the data were analyzed through Decision Table and Random Forest algorithms, which are data mining classification algorithms. In this way, a system ready to predict whether or not hypertension patient candidates are hypertension was developed.

III. METHODOLOGY

3.1 Data Collection

A simple pre-coded questionnaire will be developed, and data is collected from 149 household of students from Public Health College in Jazan University. The data includes clinical and their demographic background. There are totally 149 instances in hypertension dataset. In these instances, 61 instances are normal, 16 instances are elevated, 11 instances are hypertensive crisis, and 37 instances are stage 1 and 24 instances are stage 2. In our research work we have taken 9 attributes which will be used to classify the data. The hypertensive data set is given below table 1.

<table>
<thead>
<tr>
<th>SN</th>
<th>Attribute Name</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>Numeric</td>
</tr>
<tr>
<td>2</td>
<td>Sex</td>
<td>Male, Female</td>
</tr>
<tr>
<td>3</td>
<td>Chest pain</td>
<td>T,F</td>
</tr>
<tr>
<td>4</td>
<td>Height</td>
<td>Numeric</td>
</tr>
<tr>
<td>5</td>
<td>Weight</td>
<td>Numeric</td>
</tr>
<tr>
<td>6</td>
<td>Systolic blood Pressure</td>
<td>Numeric</td>
</tr>
<tr>
<td>7</td>
<td>Diastolic blood Pressure</td>
<td>Numeric</td>
</tr>
<tr>
<td>8</td>
<td>Fasting blood</td>
<td>Numeric</td>
</tr>
<tr>
<td>9</td>
<td>Class</td>
<td>Normal, Elevated, Hypertensive, Stage2, Stage1</td>
</tr>
</tbody>
</table>

3.2 Proposed System
a Preprocessing
Data preprocessing is the main technique in data processing that involves transforming data into a clear format. When the Real-world data is collected then there is usually incomplete, inconsistent, and in some places, there will be a lacking in certain behaviors or trends and is probably going to contain many errors. Data preprocessing may be a proven method of resolving such issues. There are many tasks in Data preprocessing prepares raw data for further processing. These tasks include data cleaning, integration, transformation and data reduction. The data is collected from the 149 household of students from Public Health College in Jazan University. The collected data was checked for the presence of error in data entry including misspellings and missing data. In our work, we have used Replace with missing values filter to fill the missing values to make the data complete.

b Classification
Classification is one among the data mining Technique. it is accustomed classify the data supported similarity of instances. There are two forms of learning. One is supervised and another is unsupervised learning. it is a supervised learning, during which predefined training data is out there. most popular data processing classification techniques are decision trees and neural networks and so on.

d Decision Tree
One of the most important classification techniques in data mining is Decision Tree. It is tree-like graph. Testing each attribute represented as internal node and each branch represents an outcome of the test, and the leaf node represents classes. It is a graphical representation of possible solutions, based on these solutions, optimum course of action is carried out. In this research, we have used two decision tree classifiers such as Random Forest Tree and J48 to classify the hypertension data set. The Algorithm of J48 and random forest tree is given below.

3.2.1 J48 Algorithm
J48 is a tree-based learning approach. It is developed by Ross Quinlan which is predicated on iterative dichotomiser (ID3) algorithm. J48 uses divide-and-conquer algorithm to separate a root node into a subset of two partitions till leaf node (target node) occur in tree. Given a group T of total instances the subsequent steps are wont to construct the tree structure.

Step 1: If all the instances in T belong to the same group class or T is having fewer instances, than the tree is leaf labeled with the most frequent class in T.
Step 2: If step 1 doesn't occur then select a test supported one attribute with a minimum of two or greater possible outcomes. Then consider this test as a root node of the tree with one branch of every outcome of the test, partition T into corresponding T1, T2, T3........, according to the result for every respective cases, and therefore the same could also be applied in recursive thanks to each sub node.

Step 3: Information gain and default gain ratio are ranked using two heuristic criteria by algorithm J48.

3.2.2 Random Forest Tree
Random forest Tree is a collective learning method for enhancing the uses of classification, regression, and other tasks. The algorithm operates by constructing a mess of decision trees at training time and it gives the output
as the category with the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.13,14

IV. EXPERIMENTS WITH WEKA

In this study, we used the Waikato Environment for Knowledge Analysis (Weka)12. It is a comprehensive suite of Java class libraries that implement many algorithms for data mining clustering, classification, regression, analysis of results. WEKA can be downloaded from the website12.

4.1 Performance Measures of Classifier

In our experiment data is supplied to classifier of J48 Algorithm and Random Forest tree to classify the data. The classifier’s performance is evaluated through Confusion Matrix.

A Confusion Matrix

It is used for measuring the performance of classifiers. In the confusion matrix, sum of the diagonal elements are called correctly classified instances and others are called incorrectly classified instances.

B Accuracy

Accuracy is defined as the ratio between correctly classified instances and total number of instances in the dataset. Where correctly classified instances are termed as True positive and True Negative others are called False Positive and False Negative.

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

C Error Rate

The Misclassification error rate is calculated by the following formula

\[
\text{Misclassification error rate} = 1 - \text{Accuracy}
\]

V. RESULTS AND ANALYSIS

There are totally 149 records and 9 attributes in the hypertension dataset. These records are classified into 5 classes such as Normal, Elevated, Hypertensive crisis, stage 1 and stage 2. Normal – 61 instances, Elevated - 16 instances, Hypertensive crisis- 11 instances, stage 2 – 37 instances, stage 1 – 24 instances. The following figure 3 and 4 represents the output of J48 and Random Forest tree Algorithm.
The following Figure 1 represents the output of Random Forest Tree Algorithm.

The following Table 2 represents confusion matrix of J48 Algorithm.

<table>
<thead>
<tr>
<th>Target Class</th>
<th>Normal</th>
<th>Elevated</th>
<th>Hypertensive Crisis</th>
<th>Stage1</th>
<th>Stage2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>58</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Elevated</td>
<td>3</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

www.turkjphysiotherrehabil.org
In J48 classifier, the correctly identified instances are 130 and incorrectly identified instances are 19. The following Table 3 represents confusion matrix of Random Forest Algorithm.

### Table 3. Confusion matrix of Random Forest Algorithm

<table>
<thead>
<tr>
<th>Target Class</th>
<th>Normal</th>
<th>Elevated</th>
<th>Hypertensive Crisis</th>
<th>Stage 1</th>
<th>Stage 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>57</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Elevated</td>
<td>5</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Hypertensive Crisis</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Stage 1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>Stage 2</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>32</td>
</tr>
</tbody>
</table>

In Random Forest tree classifier, the correctly identified instances are 120 and incorrectly identified instances are 29. The following table 4 shows the accuracy, time taken to build the model and error rate of J48 and Random Forest tree Algorithm.

### Table 4. Accuracy, Error Rate and Time taken to Build Model

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Time Taken(Secs)</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>87.25%</td>
<td>0</td>
<td>12.75%</td>
</tr>
<tr>
<td>Random Forest Tree</td>
<td>80.54%</td>
<td>0.03</td>
<td>19.46%</td>
</tr>
</tbody>
</table>

Table 4 shows that the accuracy of J48 (87.25%), accuracy of Random Forest tree (80.54%). J48 takes 0 seconds to build the model, Random Forest Tree takes 0.03 seconds to build the model. The error rate of Random Forest tree is 19.46% and the error rate of J48 is 12.75%. While comparing J48 with other classifier, J48 is giving highest accuracy (87.25%), minimum error rate (12.75%) and minimum time (0 seconds) than Random Forest Tree algorithms.

The following chart 1 shows the accuracy, error rate and time taken to build model of classifiers.

![Performance Analysis](chart1.png)
In this chart 1, X axis represent the classifier and Y axis represent the accuracy, error rate and time. It shows that the accuracy of J48 classifier is 87.25%, time to build the model is 0 seconds and error rate is 12.75% which is best than Random Forest tree algorithms.

VI. CONCLUSION AND FUTURE SCOPE

Diagnosis of disease is an incredibly challenging task in the field of health care. Various data mining techniques have proven to be extremely helpful in decision making. In our work, we have used data cleaning task to fill up the missing values and we have applied J48and Random Forest Tree data mining classification techniques which are used to classify the hypertensive disease. The performances of classifiers are evaluated through the confusion matrix in terms of accuracy and error rate. The J48 Algorithm gives 87.25% which is providing better accuracy than Random Forest tree accuracy and J48 Algorithm gives very minimum error rate (12.75%) and time (0 Seconds) than Random Forest tree. As a future work the same technique is used to apply for other disease datasets such as heart disease, Lung cancer and so on.

Acknowledgement

The authors would like to thank their family members, colleagues and friends for their great support to complete this research. They would also like to thank the University for providing an environment to carry out this research.

Conflict of Interest

There was no conflict of interests between the authors.

REFERENCES


16. Jiawei Han, KamberMicheline., Datamining: Concepts and Techniques, Morgan Kaufmann Publisher,2009.