Pattern Classification of Foot Diseases using Decision Tree

Jungkyu Choi¹, Keun-Hwan Jeon², Yonggwan Won³, and Jung-Ja Kim⁴,⁵,*

¹Department of Healthcare Engineering, Graduate school
Chonbuk National University
664-14, Deokjin-dong 1, Deokjin-ku, Jeonju, Jeolabuk-do, 561-756
Republic of Korea
chaosjoozak@nate.com

²Department of Occupational therapy
Kunjang College
608-8, Doamri, Sungsan-myun, Gunsan, Jeollabuk-do
Republic of Korea
khjeon@kunjang.ac.kr

³School of Electronics and Computer Engineering
Chonnam National University
77 Yongbong-ro, Buk-gu, Gwangju, 500-757
Republic of Korea
ykwon@chonnom.ac.kr

⁴Division of Biomedical Engineering
Chonbuk National University
664-14, Deokjin-dong 1, Deokjin-ku, Jeonju, Jeolabuk-do, 561-756
Republic of Korea

⁵Research Center of Healthcare & Welfare Instrument for the Aged
Chonbuk National University
664-14, Deokjin-dong 1, Deokjin-ku, Jeonju, Jeolabuk-do, 561-756
Republic of Korea
jungjakim@jbnu.ac.kr

Abstract: - The purpose of this study was to investigate significant knowledge from developing the prediction model for pattern classification of foot diseases using decision tree based on clinical data stored in the Foot Clinic. Sample data of 80 patients diagnosed with a single disease in 1267 patients. Dependent variable was composed of 8 diseases. Independent variables were selected with 28 variables in the whole 37 attributes closely related to disease. The whole data was divided into training data and test data, and prediction rate was verified by C5.0, C&RT, QUEST and CHAID algorithm. As the result, the C5.0, the top of prediction rate, is applied for final analysis of disease category. The importance factors related closely with 8 disease were RCSP, right pelvis elevation and pelvis rotation condition. In the case of both pes planus and left pes planus, it were shown same result that left RCSP was higher than right RCSP. However, the left pes planus had possibility to accompany the body unbalance by pelvis rotation and pelvis elevation difference. Pes cavus was always shown if any feet had above 1° for RCSP, and gastro-soleus muscle tightness was appeared if there was pelvis elevation difference with similar RCSP on both foot. In conclusion, we were able to conclude that factor of each node, including RCSP, which formed decision-tree, major diagnosis for distinction lower limbs disease and obtain the 8 diagnosis rules.

Key-Words: - Biomechanical analysis, Custom-made insole, Pes cavus, Foot pressure, EMG

1 Introduction
The foot is the most important role in the standing posture and bipedalism of the human body [1-2]. When a human walks for 1 km, there is about a 15 t weight increase on the foot, only 5 % of the entire surface of the body, and the pressure that occurs by weight or the push-off exercise causes stress or soft tissue strain [3-4]. In this condition, there are close connections between form of the abnormal foot and
cause of the various diseases in lower limbs as biomechanical concept.

Generally, the foot shape is classified into three types such as pes planus, normal foot, and pes cavus by using examinations like visual inspection, radiographic examination, resting calcaneal stance position (RCSP) and footprint examination [5]. Pes planus is form of the foot touched excessively on the floor in the midfoot by the lower medial longitudinal arch (MLA) abnormally and caused by the loose spring ligament. However, most of cases are idiopathic [6]. Pes cavus is a medical condition, in which the height of MLA is raised, and accepted to be rigid structurally, and runs a complicated deformity, to cause the equinus of the foot, or the varus of the hindfoot. As a consequence, the contact area of the foot on the floor is narrowed, while the ankle or heel is tilted out exteriorly [7]. This foot condition is known to be commonly caused by the Achilles’ tendon constricted, Charcot-Marie-Tooth, polio or high-heel shoes for woman. These abnormal foot shapes have a bad influence on balance of the knee, pelvis and spine besides the ankle and the unstable body cause excessive and unnatural movement in the joints [8].

In recent years, the various field including medical area have a large-scale data by development of information technology. The result analyzed on the data is applied index or information extracted on the basis of experience. Datamining, which is highly favoured as the technology of information modeling on account of search for useful pattern and rule systematically and automatically in this data, is method to find out new and meaningful knowledge with utilizing pattern recognition technology, statistics technique or mathematic algorithm [9-10]. Decision-tree algorithm, which is one of representative technique in the datamining, is an analysis method to predict or classify a couple of subgroup from interested object group by observing relationship and modeling regulation in data, and non-parametric method that linearity, normality and homoscedasticity supposition are unnecessary [11]. The decision-tree has advantage to understand and explain analysis process easily compared with neural network, discriminant analysis and regression analysis because it show result of learning as tree structure shape [12]. The typical algorithm to generate decision-making are C5.0 based on ID3 algorithm, C&RT, Quest and CHAID [13-14]. This datamining methods are utilized in several medical field such as research, diagnosis, hospital management and customer management in the present and used for analysis of clinical data as useful tool, especially [15-16]. According to previous studies, the evolutionary-driven support vector machine was used to anticipate stage of hepatic fibrosis that determine hardness degree of liver or operation in chronic hepatitis C, and the logistic regression and neural network are utilized to extract attribute and perform learning based on widespread clinical data of acute myocardial infarction for forecast short-term relapse mortality of ST-segment elevation myocardial infarction (SEMI) patients. Through this study, the model to foresee short-term mortality of SEMI patients is suggested [17-18]. In addition, age, associated disease, pathology scale, course of hospitalization, respiratory failure and congestive heart failure were came out to be danger factors on death of pneumonia by using the decision-tree model for analysis of death factor on pneumonia patients [19]. More intelligent and exquisite analysis is necessary because most existing studies just figured out simple correlation by only quantitative analysis based on statistical method, despite it is very important in disease to understand various interconnection of several parameters measured with reference to disease on analysis of medical data. Valuable knowledge found by data mining increase accuracy of diagnosis and treatment about patients [20].

Accordingly, the purpose of this study was to find out significant knowledge from developing prediction model that distinguished category of lower limbs disease and deducting relevant rule based on clinical data of the Foot Clinic.

2 Decision-Tree

A decision tree is a model of decisions and a special form of tree structure [21]. Analysis of decision-tree goes through several steps like growing the tree, pruning, interpretation and prediction [22]. Generally, the structure of decision-tree is composed of node and branch. The top of node is root node, the end of node is terminal node, and the middle of node is internal node. The direction of progress for decision-tree starts from the root node to the terminal node through the internal node. The way from the root node to any terminal node is only one, and this is expressed as rule [23]. Algorithm used for this study were C5.0, C&RT, CHAID and QUEST, and they have different construction process in splitting criterion, stopping rule and pruning.
The C5.0 construct a simple model and use entropy concept of information theory for accurate classification. The entropy is to show quantitatively mixed degree of different patterns. It is higher if there is more data of different class in subtree responding any node of decision-tree, or, conversely, it is lower if data is composed of a single class. The C5.0 form simple decision-tree model by selecting attribute for lowest entropy from the root node and expanding model [14]. The C&RT is algorithm to select predictor by using decrement of variance as splitting criterion, and perform binary split. The binary split measure impurity by the Gini index, Twoing index, entropy index and variance. In categorical target variable, the Gini index and Twoing index have many case to form similar tree structure. Generally, the Gini index is preferred because the Twoing index need a lot of time for calculation if the number of category on target variable is too much [24]. The QUEST carry out the binary split by chi-squared test statistics or p-value of F-statistics as splitting criterion in nominal target variable [25]. The CHAID perform multiway split to repeat selection, separation or merger of predictor by using the chi-squared test statistics or p-value of F-statistics. In case of that target variable is categorical, the likelihood ratio chi-squared statistics is used, too [26].

### 3 Method

#### 3.1 Subjects

The first medical examination clinical data of total 1285 patients in the Foot Clinic were utilized in this study. The data was composed of parameters of 37 attribute measured and diagnosed by a podiatrist besides gender and age. Sample data of 80 patients diagnosed with a singular disease in data of 1267 patients except data of 18 patients with missing value was used for analysis.

#### 3.2 Variables

Dependent variable in this study was single disease and consisted of total 8 items such as (L)Pes cavus (left pes cavus), (L)Pes planus (left pes planus), (R)Pes cavus (right pes cavus), (R)Pes planus (right pes planus), Gastro-soleus m. (gastro-soleus muscle tightness), Pes cavus, Pes planus and Scoliosis (Table 1).

Independent variables were selected with 28 variables of the whole 37 attributes closely related to disease through opinion of a podiatrist (Table 2).

#### 3.3 Study Procedure

In this study, combination of variables that explained effectively 8 diseases, nominal dependent variable, were analyzed by insert the 28 independent variables in algorithm at the same time. Data analysis was performed by IBM SPSS Modeler 14.2 (SPSS Inc., Chicago, IL, USA). It was efficient to make a number of predictive model from one data and carry out comparison analysis for producing ideal model. [27] The whole data, therefore, was divided into training data (70%) and test data (30%) by using partition node. Prediction rate was verified by analysis node after model of each data was developed by C5.0, C&RT, QUEST and CHAID algorithm of datamining (Fig. 1). As follow the

![Fig. 1. Study procedure](image)

**Table 1. Dependent variable**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathologies</td>
<td></td>
</tr>
<tr>
<td>(L) Pes cavus</td>
<td></td>
</tr>
<tr>
<td>(L) Pes planus</td>
<td></td>
</tr>
<tr>
<td>(R) Pes cavus</td>
<td></td>
</tr>
<tr>
<td>(R) Pes planus</td>
<td></td>
</tr>
<tr>
<td>Gastro-Soleus m.</td>
<td></td>
</tr>
<tr>
<td>Pes cavus</td>
<td></td>
</tr>
<tr>
<td>Pes planus</td>
<td></td>
</tr>
<tr>
<td>Scoliosis</td>
<td></td>
</tr>
<tr>
<td>Variables</td>
<td>Type</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>Sex</td>
<td>Nominal</td>
</tr>
<tr>
<td>(R) TIBIA_TMA</td>
<td>Numeric</td>
</tr>
<tr>
<td>(L) TIBIA_TMA</td>
<td>Numeric</td>
</tr>
<tr>
<td>Hip_ExtRot</td>
<td>Nominal</td>
</tr>
<tr>
<td>Hip_IntRot</td>
<td>Nominal</td>
</tr>
<tr>
<td>(L) STJ_Inversion</td>
<td>Numeric</td>
</tr>
<tr>
<td>(R) STJ_Inversion</td>
<td>Numeric</td>
</tr>
<tr>
<td>(L) STJ_Eversion</td>
<td>Numeric</td>
</tr>
<tr>
<td>(R) STJ_Eversion</td>
<td>Numeric</td>
</tr>
<tr>
<td>(L) STJ_ROM</td>
<td>Numeric</td>
</tr>
<tr>
<td>(R) STJ_ROM</td>
<td>Numeric</td>
</tr>
<tr>
<td>(L) FFtoRF</td>
<td>Numeric</td>
</tr>
<tr>
<td>(R) FFtoRF</td>
<td>Numeric</td>
</tr>
<tr>
<td>(L) RCSP</td>
<td>Numeric</td>
</tr>
<tr>
<td>(R) RCSP</td>
<td>Numeric</td>
</tr>
<tr>
<td>Pelvis_Tilting</td>
<td>Nominal</td>
</tr>
<tr>
<td>Pelvis_Rot</td>
<td>Nominal</td>
</tr>
<tr>
<td>(L) Pelvis_Trend</td>
<td>Numeric</td>
</tr>
<tr>
<td>(R) Pelvis_Trend</td>
<td>Numeric</td>
</tr>
<tr>
<td>(L) Pelvis_Eleva</td>
<td>Numeric</td>
</tr>
<tr>
<td>(R) Pelvis_Eleva</td>
<td>Numeric</td>
</tr>
</tbody>
</table>

L: Left, R: Right, Mi: Mild, Mo: Moderate, B: Bad, S: Severe, A: Anterior tilting, Lt: Lateral rotation
Fig. 2. The result of C5.0 decision-tree model
4 Result

Data of 80 patients (Male: 31, Female: 49) diagnosed 8 disease in the whole 1267 patients data was used for experiment. Four algorithm of decision-tree were used to form models, and the C5.0 algorithm, the top of prediction rate, is applied for final analysis of disease category (Fig. 2). The measured prediction rate of C5.0 is Correct: 94.64 % and Wrong: 5.36% in the training data, and Correct: 80.77 % and Wrong: 19.23 % in the test data. The prediction rate of C&RT is Correct: 69.64 % and Wrong: 30.36 % in the training data, and Correct: 66.67 % and Wrong: 33.33 % in the test data. The prediction rate of QUEST is Correct: 50 % and Wrong: 50 % in the training data, and Correct: 41.67 % and Wrong: 58.33 % in test data. The prediction rate of CHAID is Correct: 91.07 % and Wrong: 8.93 % in the training data, and Correct 68.42 % and Wrong: 31.58 % in test data (Table 3).

As a result of analysis on disease by using C5.0 algorithm, significant prediction importance factor was shown in the order: (L)RCSP (0.58), (R)RCSP (0.21), (R)PELVIS_Elva (0.16), PELVIS_Rot (0.05). In addition, total 8 rules were verified about training data (70%) of the whole data : 1) If ‘(R)RCSP’ is below 1°, ‘(L)RCSP’ is below -2°, ‘PELVIS_Rot’ is left-lateral rotation (L Lt-R_N) and ‘(R)PELVIS_Elva’ is below 2°, then ‘(L)Pes planus’, 2) If ‘(R)RCSP’ is below 1°, ‘(L)RCSP’ is below -2° and ‘PEVIS_Rot’ is normal, then ‘Pes planus’, 3) If ‘(R)RCSP’ below -2° and ‘(L)RCSP’ is above -2° ~ below 1°, then ‘(R)Pes cavus’, 4) ‘(R)RCSP’ is below 1° and ‘(L)RCSP’ is above 1°, then ‘(L)Pes cavus’, 5) ‘(R)RCSP’ is above 1° and ‘(L)RCSP’ is below 1°, then ‘(R)Pes cavus’, 6) ‘(R)RCSP’ is above 1° and ‘(L)RCSP’ is above 1°, then ‘Pes cavus’, 7) ‘(R)RCSP’ and ‘(L)RCSP’ are above -2° ~ below 1° and ‘(R)PELVIS_Elva’ is below 1°, then ‘Gastro-Soleus m.’, 8) ‘(R)RCSP’ is below 1°, ‘(L)RCSP’ is below -2°, ‘PELVIS_Rot’ is left-lateral rotation (L Lt-R_N) and ‘(R)PELVIS_Elva’ is above 2° or ‘(R)RCSP’ and ‘(L)RCSP’ are above -2° ~ below 1° and ‘(R)PELVIS_Elva’ is above 1°, then ‘Scoliosis’ (Fig. 4).

5 Conclusion

In this study, the object was to develop prediction model that classified category of lower limbs disease and draw diagnosis rule from learning data measured on patient based on foot clinical data by utilizing the decision-tree algorithms.

The sample data of 80 patients diagnosed with 8 disease for the first medical examination clinical data of 1285 patients in the Foot Clinic was used for analysis. Dependent variable was consisted of total 8 disease and independent variables were selected
with 28 variables. The data was divided into training data (70%) and test data (30%) for producing ideal model, and prediction rate was confirmed for comparison after model of each data was developed by C5.0, C&RT, QUEST and CHAID of decision-tree. As the result of prediction rate, C5.0 algorithm, the top of prediction rate, Correct: 94.64 % and Wrong: 5.36% in the training data, and Correct: 80.77 % and Wrong: 19.23 % in the test data, was utilized for final pattern analysis of disease.

As the result, the importance factors related closely with 8 disease were RCSP, right pelvis elevation and pelvis rotation condition. In addition, we could find out useful information from discovered rules. In the case of pes planus and left pes planus, it were shown same result that left RCSP was higher than right RCSP. Left pes planus, especially, had to possibility to accompany the body unbalance by pelvis rotation and pelvis elevation difference. Pes cavus was always shown if any feet had above 1° for RCSP, and gastro-soleus muscle tightness was appeared if there was pelvis elevation difference with similar RCSP on both feet. In conclusion, we were able to conclude that factor of each node, including RCSP, which formed decision-tree, major diagnosis for distinction lower limbs disease and obtain the 8 diagnosis rules.

This study shows a similar result to previous studies classified shape of the foot according to RCSP. [29], [30] A clue, especially, which the body unbalance would come together, was figured out if there was gastro-soleus muscle tightness or pes planus on one side. In the study, the error rate of prediction rate on model of test data was relatively high due to extract 80 patients diagnosed with 8 diseases from the whole data, and predicted to decrease if decision-tree model learn more data. On the basis of result in this study, datamining analysis for the whole data will proceed from now on.

Acknowledgments
This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science and Technology (NRF-2012R1A1B3003952, NRF-2013R1A2A2A04016782).

References:


