A Comparison of C4.5 and K-Nearest Neighbor Algorithm on Classification of Disk Hernia and Spondylolisthesis in Vertebral Column

Irma Handayani1,* and Suprapto2

1Faculty of Science & Technology, Informatics, University of Technology Yogyakarta, Sleman, Indonesia
2FMIPA, Department of Computer Science and Electronics, Gadjah Mada University, Yogyakarta, Sleman, Indonesia

Email: 1 irma.handayani@staff.uty.ac.id, 2 suprapto@ugm.ac.id

Correspondence Author Email: irma.handayani@staff.uty.ac.id

Abstract—Good spinal health is needed to carry out daily activities. Trauma to the vertebral column can affect the spinal cord's ability to send and receive messages from the brain to the body's sensory and motor control systems. Disk hernia and spondylolisthesis are examples of pathology of the vertebral column. Research on pathology or damage to bones and joints of the skeletal system is rare. Whereas the classification system can be used by radiologists as a "second opinion" so that it can improve productivity and diagnosis consistency from that radiologist. This study compared the accuracy values of the C4.5 and K-NN algorithms in the classification of herniated disc disease and spondylolisthesis as well as a comparison of the speed of time in the classification process. Tests were carried out using data from 310 patients with normal conditions (100 patients), herniated disks (60 patients), and spondylolisthesis (150 patients). The results showed that the accuracy of the C4.5 classifier was 89% and the K-NN classifier was 85%. The average time needed to classify the C4.5 classifier is 0.00912297 seconds and the K-NN classifier is 0.000212303 seconds.

Keywords: Vertebral Column; C4.5 Algorithm; K-NN Algorithm; Disk Hernia; Spondylolisthesis

1. INTRODUCTION

Vertebral column or spinal sequence is a flexible structure formed by a number of bones called vertebra or vertebrae. The vertebral column supports the body’s physical structure and nervous system, enabling movement and sensation. Pathology of the spine can lead to debilitating outcomes on quality of life. The vertebral column (spine) defines the animal subphylum Vertebrata, or vertebrates, of the phylum Chordata. In humans, it is composed of 33 vertebrae that include 7 cervical, 12 thoracic, 5 lumbar, 5 sacral, and 4 coccygeal. Along with the skull, ribs, and sternum, these vertebrae make up the axial skeletal system [1]. Good spinal health is needed to carry out daily activities. Spinal health problems can interfere with daily activities and carrying out daily activities in the wrong way can also interfere with spinal health.

Spinal Cord Injury (SPI) or spinal injury can cause permanent disruption in body function. Spinal injuries are divided into two, namely injuries that only affect some of the movement functions and those that affect almost all of the motion functions. Spinal injuries are generally caused by driving accidents, falls, acts of violence, sports injuries, and diseases such as cancer, arthritis, and osteoporosis. Among complications of spinal cord injury, there are those related to the nervous and osteoarticulat system, such as spasticity, contractures, loss of joint range of motion, and osteoporosis, the result of neurological damage and disuse of the affected limbs [2]. Hernia Disc and Spondylolisthesis are also examples of diseases that can occur in the spine. Disk Hernia is the abnormal rupture of the inner core of the intervertebral disks are flexible and cushion-like discs found in the vertebral column (spine). These disks provide shock absorption effects to the vertebral column. The spine is composed of 24 bones, referred to as vertebrae. The vertebrae are so structured to protect the spinal cord, which houses critical nerves connecting the brain to the rest of the body. Each intervertebral disk is composed of two important organic matters, which are known as the annulus fibrous and nucleus pulposus. The nucleus pulposus is hard and supple and while the annulus fibrous is a jelly-like matter. A number of factors can cause the jelly-like matter (inner core) to rupture and push against the outer core, resulting in lower back pain. In severe cases, the rupture may so pressurize the outer core such that spinal nerves are affected; leading to the impairment of the nervous activities, especially in the lower body (thighs, legs, etc.) [3]. Meanwhile, spondylolisthesis is a slipping of the vertebra, however, the neural arch is still intact. Spondylolisthesis may compromise the stability of spine and increase axial load that may stress surrounding structures [4]. Under normal circumstances the vertebrae are arranged to collide with each other in a straight line.

Such problems can cause pain in the lower back and neck which worsens with the age and may require surgery and other medication. Timely and accurate diagnosis of pathologies is an important task in bioinformatics. Hence, data mining is a way that aims to find patterns automatically or semi-automatically from existing data in databases or other data sources that are used to solve a problem through various process rules. Machine learning can play a key role in the field of bioinformatics. In recent years machine learning has aided the practitioners in medical diagnosis, thus facilitating significant improvements in clinical decision support systems. The machine learning models are applied both in disease diagnosis and in prognosis. The models predict a specific label based on the input data [5].
Data mining can be defined as the process of extracting valid, previously unknown and actionable information from large data sets. The purpose of the data mining is to use the extracted information to make crucial business decisions [6].

Several data mining techniques, research related to the C4.5 and K-NN algorithms have been carried out by [7]. In this study, we want to predict the cause of bad credit by paying attention to the data entered in the process of borrowing money. The method to be used to predict bad credit is a comparison between K-NN algorithm method and C4.5 algorithm method. The accuracy of the C4.5 algorithm scored better than the K-NN algorithm of 61.64% while the accuracy of K-NN was 45.21%, it can be concluded that the C4.5 algorithm is more accurate for determining bad credit. A study by [8] This study will analyze the accuracy of the K-Nearest Neighbor and Naive Bayes algorithms for the classification of breast cancer. So that patients with existing parameters can be predicted which are malignant and benign breast cancer. The test results using K-fold (k-10) cross validation, followed by confusion matrix of 455 data consisting of 284 data on cases of benign cancer, 171 data on malignant cancer cases, on the K-NN classifier were able to correctly classify 441 data with an accuracy rate of 0.97%, while the Naive Bayes classifier was able to correctly classify 428 data with an accuracy rate of 0.94%. A study by [9] The position of digital media user skills is categorized by the k-Nearest Neighbor technique data mining method. The objective of this research using the k-Nearest Neighbor technique data mining method was carried out by compiling categories that refer to attributions of the 10 competencies of Japelidi digital media user skills. The data collection of Yogyakarta residents' competence was carried out in the June-July 2020 period. The research results provided information that the competence of using digital media for Yogyakarta residents in the June-July 2020 period was in the Consuming Function in the first quadrant. A study by [10] The research method used was to prepare data from the breast cancer dataset, conduct training and testing upon the data, then perform a comparative analysis. The research target is to produce the best algorithm in classifying breast cancer so that patients with existing parameters can be predicted which ones are malignant and benign breast cancer. By making comparisons, this method produces 95.79% for K-Nearest Neighbor and 93.39% for Naive Bayes. A study by [11] The purpose of this research is to classify trends in freight transport violations based on violation data in the UPPKB. The expected result of this research is to be able to find out the pattern of classification trends for freight vehicle disturbances based on the results of the C.45 algorithm decision tree, so that the research results can be used as a reference in making decisions and making policies. The results of this study indicate that the accuracy performance in data mining tests for the classification of trends in freight vehicle disturbances with 10 fold cross validation linear sampling produces an accuracy of 86.31% +/- 1.23% (micro average: 86.31%), shuffled sampling produces an accuracy of 86.34% +/- 0.67% (micro average: 86.34%) and stratified sampling produces an accuracy of 86.34% +/- 0.67% (micro average: 86.34%).

The C4.5 algorithm is an algorithm that has a high accuracy value as seen in related studies and the K-NN algorithm is an algorithm that is robust against training data that has a lot of noise.

Data types greatly affect the performance and accuracy of an algorithm. The best algorithm for one data type is not necessarily good for another type. It is even possible that a good algorithm will be very bad for other data types.

This study compared the C4.5 and K-NN algorithm methods in the classification of herniated disc disease and spondylolisthesis in the vertebral column. The data are organized into two but related classifications. The first data classifies patients as one of the categories, namely Normal (100 patients), Hernia Disk (60 patients), and Spondylolisthesis (150 patients). The second one, the Hernia Disk and Spondylolisthesis categories are combined into one category which is labeled as “abnormal”. Thus, the second task classified patients as one of two categories: Normal (100 patients) or Abnormal (210 patients).

2. RESEARCH METHODOLOGY

2.1 Theoretical Framework

2.1.1 Classification

Classification is the process of finding a set of models/functions that explain and differentiate data into certain classes, with the aim of using the model in determining the class of an object whose class is unknown. There are two classification processes which consist of: the learning/training process is building a model using training data, where each record in the training data is analyzed based on its attribute values to obtain a model. Then the testing process is to test the testing data using the model that has been obtained from the training process [12].

2.1.2 C4.5 Algorithm

The C4.5 algorithm is one of the algorithms contained in data mining classification and is a model or function that explains or differentiates concepts or data classes with the aim of estimating an unknown class of an object. One of the data mining classification techniques is the C4.5 algorithm. The C4.5 algorithm can be used to create decision trees and is an algorithm specifically for supervised learning. C4.5 algorithm can be used to classify data that has numeric, continuous, and categorical attributes [13].
The components that make up the C4.5 algorithm in the form of a decision tree are:

### 2.1.1 Entropy

Entropy is a probability distribution in information theory and is adopted into the C4.5 algorithm to measure the degree of homogeneity of the class distribution of a set (data set). As an illustration, the higher the entropy level of a data set, the more homogeneous the class distribution in the data set. Calculation of entropy is shown in equations (1) and (2).

\[
\text{Info}(S_1, \ldots, S_m) = -\sum_{i=1}^{m} P_i \log_2(P_i)
\]

Where \( P_i = \frac{s_i}{S} \) is the probability from sample which has \( C_i \) class.

\[
E(A) = \sum_{i=1}^{V} \frac{s_{ij} + \cdots + s_{mj}}{s} \times I(S_1, \ldots, S_m)
\]

\[
\frac{s_{ij} + \cdots + s_{mj}}{S} \text{ is the weight of the subset } j \text{ and the number of samples in the subset (which has the value } a_j \text{ of } A \text{) divided by the total number of samples in } S.
\]

### 2.1.2 Information Gain

After dividing the data set based on an attribute into smaller subsets, the entropy of the data will change. This entropy change can be used to determine whether the data distribution that has been done is good or not. This change in entropy is called information gain in the C4.5 algorithm. This information gain is measured by calculating the difference between the entropy of the data set before and after the splitting is done. The best distribution will result in the smallest entropy subset, thereby having the greatest information gain [14]. To choose an attribute as the root, it is based on the highest gain value of the existing attributes and can be shown by equation (2).

\[
\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i)
\]

Where \( S \) is the case set, \( A \) is the computed attribute, \( n \) is the number of assignments to attribute \( A \), \( |S_i| \) is the number of cases on the I partition, and \( |S| \) is the number of cases in \( S \).

### 2.1.3 K-NN Algorithm

The K-NN algorithm is a method for classifying objects based on learning data that are closest to the object according to the number of nearest neighbors or the value of \( k \). The K-NN algorithm was first introduced in the 1950s with the concept of training data for the classification process. The data training process is based on a comparison of the training data and the tested data. The data being trained is data with a number of \( n \) attributes. Each attribute represents a data dimension that becomes a data mining pattern (pattern). The advantages of K-NN are that it is tough against training data which has a lot of noise and is effective if the amount of training data is large [15].

### 2.1.4 Numerical Attributes Similarity

In the numerical attribute there is a calculation of the distance (the distance between two objects) which can be done using the Euclidean distance, Manhattan distance, and Minkowski distance calculations. In this study the authors used Euclidean distance to calculate the distance between two objects with nominal attributes. Euclidean distance is the calculation of the distance from 2 points in Euclidean Space [16]. Near or far neighbors are calculated based on the Euclidean distance indicated by equation (4).

\[
dist(X, Y) = \sqrt{\sum_{k=1}^{n} (X_k - Y_k)^2}
\]

Where \( d \) is the distance of the test data to training data, \( X_k \) is the \( k \) attribute value of the test data \( x \), with \( k=1, 2, \ldots, n \), and \( Y_k \) is the \( k \) attribute value of the training data \( y \), with \( k=1, 2, \ldots, n \). After the distance or dissimilarity (d) is calculated then converted into similarity (s) with an interval between 0 to 1 (\( s \in [0,1] \)) calculated using equation (5).

\[
s = \frac{1}{1+d}
\]

### 2.1.5 K-Fold Cross Validation

Cross validation is a technique for assessing or validating the accuracy of a model built based on a particular dataset. One of the cross-validation methods is k-fold cross validation [17]. Fold amount the standard for predicting error rates from data is to use 10-fold cross validation. Cross validation is used in order to find the best parameters of one model. Use of k-fold cross validation to eliminate bias in the data. Training and testing were carried out \( k \) times [18]. This was done by testing the amount of error in the testing data. In cross validation, data is divided into
k samples of the same size. From the k subset of data used will be used in k-1 sample training data and 1 remaining sample for testing data. This is often called k-fold validation. For example, there are 10 subsets of data, 9 subsets will be used for training and the remaining 1 subset for testing. This is done for all possibilities. There are 10 times training where each training data have 9 subsets of data for training and 1 subset of data for testing. Then, the average error is calculated (the error mean). If there are 3 models, then each model was tested 10 times in each combination of training-testing subset and in every run an error will be found for each model. The model that gives the average smallest error is the best method.

2.1.6 Confusion Matrix

The Confusion Matrix is a technique for evaluating the results of model testing [19]. Confusion Matrix shows the result of identification between the amount of correct prediction data and the number of incorrect predictive data compared to the facts produced. Table 1 shows the Confusion Matrix.

Table 1. Confusion Matrix

<table>
<thead>
<tr>
<th>Actual</th>
<th>Prediction</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Positive</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

With,

a: a lot of data predicted by the system with the correct results is indicated healthy, the doctor states indicated healthy.
b: a lot of data predicted by the system with wrong results is indicated by malaria, the doctor said indicated healthy.
c: a lot of data predicted by the system with true results is indicated wrong, the doctor stated indicated malaria.
d: a lot of data predicted by the system with the correct results is indicated malaria, the doctor stated indicated malaria.

There are several terms based on Table 1.

True Positive (TP) is positive data correctly indicated on the model. Calculation TP values can be calculated using equation 6.

\[ TP = \frac{d}{c+d} \] (6)

False Positive (FP) is positive data incorrectly indicated on the model. Calculation FP values can be calculated using equation 7.

\[ FP = \frac{b}{a+b} \] (7)

True Negative (TN) is negative data that is correctly indicated in the model. Calculation TN value can be calculated using equation 8.

\[ TN = \frac{a}{a+b} \] (8)

False Negative (FN) is negative data that is incorrectly indicated in the model. Calculation FN values can be calculated using equation 9.

\[ FN = \frac{c}{c+d} \] (9)

2.1.7 Accuracy Measurement

The accuracy of the classification results on the given test data is the percentage of test data that is correctly classified by the classification model that has been made [20]. Testing the level of accuracy means to find the percentage of accuracy in the process of classifying the testing data that is being examined. The level of accuracy is calculated using equation (10).

\[ Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN} \] (10)

Where a: the result of a positive classification with the actual class is positive, b: the result of a negative classification with the actual class is positive, c: the result of a positive classification with the actual class is negative, and d: the result of a negative classification with the actual class is positive.

2.2 Research Stages

The methodology used in this research was divided into several stages such as those shown in Figure 1.
2.2.1 Literature Study and Problem Analysis

In the initial stage, it is done by searching and studying library materials both from journals, digital libraries, papers, books, e-books, internet sites or scientific works that can support the process writing. This stage is carried out to obtain information related to data mining, classification, C4.5 and K-NN algorithm. Information obtained from observing problems related to factors needed to be used in this study and observing related studies data classification using the C4.5 and K-NN method.

2.2.2 Data Collection

The next stage is to prepare training and testing data taken from the Vertebral Column dataset from UCI (University of California, Irvine) Machine Learning Repository. Vertebral column dataset is a collection of Biomedical data set built by Dr. Henrique da Mota during a medical residence period in the Group of Applied Research in Orthopedics (GARO). The data have been organized in two different but related classification tasks. The first task consists in classifying patients as belonging to one out of three categories: Normal (100 patients), Disk Hernia (60 patients), or Spondylolisthesis (150 patients). For the second task, the categories Disk Hernia and Spondylolisthesis were merged into a single category labelled as ‘abnormal’. Thus, the second task consists in classifying patients as belonging to one out of two categories: Normal (100 patients) or Abnormal (210 patients).

2.2.3 Implementation and Testing

Based on the data that has been obtained and also various references that have been complete, the steps the following is implementation and testing. The results of the system design are outlined in the form program implementation which produces writing program code to get the test data results. The test carried out is testing the classification accuracy produced by the system with using the C4.5 and K-NN algorithm. Accuracy measurements are carried out using the k-fold cross method validation. In addition to measuring accuracy, calculation of the length of the classification process is also carried out test data that has been prepared. This is done to analyze the C4.5 and K-NN algorithm inside classifying disk hernia and spondylolisthesis in the vertebral column.

2.3 System Description

The system is designed to be able to classify Vertebral Column data using the C4.5 and K-NN algorithms. The disease is then divided into 3 classes, namely disc herniation, spondylolisthesis, and normal class. The process applied to the system is divided into 3 stages, including pre-processing, classifier design, and post-processing stages. The system flow is shown in Figure 2.

Figure 1. Research Methodology Flow Diagram

Figure 2. System Flow

i. The pre-processing stage is the stage that starts with the data collection process. The data collected is then grouped based on the influence on each class. After that, the data is normalized where the data is entered into
the appropriate class. After normalization, the data is then divided into 2 parts, namely training data and test data by dividing 70% training data and 30% test data randomly.

ii. After the pre-processing stage is complete, the data is then entered into each classifier as knowledge. The classifier then learns from the data that has been entered and evaluated. If some of the specified attributes have not been trained, then the system training process will be repeated with different structures and functions.

iii. The third stage is the post-processing stage where the classification results are displayed in a form that is easier to understand. The system will display whether the patient is normal or has vertebral column disease and the type of disease suffered whether it is a herniated disk or spondylolisthesis.

3. RESULT AND DISCUSSION

3.1 Disease Data Compilation Process

The data used as training data and test data are data regarding herniated disc disease and spondylolisthesis. There are 310 training data consisting of 150 data on spondylolisthesis cases, 60 data on cases of herniated disc disease, and 100 data on normal cases.

3.2 System Requirements Analysis

The system designed is a system used to classify existing patient data into several classes, namely normal classes and abnormal classes. For abnormal classes themselves are divided into 2 subclasses, namely disk hernia class for disk hernia patients and spondylolisthesis class for patients with spondylolisthesis. The class division is used based on the value of each owned by each patient, namely pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, and degree spondylolisthesis.

3.3 Input Data

Input data is data that will be used as input to the system. This input data will then be processed using the C.45 and K-NN classification methods to determine the patients’ class. The data used include: Pelvic incidence (PI) as the angle between the perpendicular line between the sacral plate and the line connecting the midpoint of the sacral plate to the bicoxofemoral axis; PI is a specific and constant value for each patient; Pelvic tilt (PT) is the orientation of the pelvis that connects the femur with the rest of the body. PT can be moved forwards, backwards, or in other directions; Lumbar lordosis angle (LA) is a characteristic feature of the human spine and is a reference to whether a patient's posture is good or bad; Sacral slope (SS) is the slope that occurs between the sacral plate and the horizontal plate; Pelvic radius (PR) is a value that influences the development of large lumbar lordosis; and Degree spondylolisthesis (DG) is often also called the grade of spondylolisthesis, which is a measurement that states how much of the patient's body that slips forward exceeds the lower body.

3.4 The C4.5 Classification Process

The C4.5 classification process is carried out by tracing the position of the test data based on the tree that has been built based on the previous training data. The process of forming a tree is carried out in the following stages:

a. Select attribute as root
   In choosing an attribute as the root, it is based on the highest gain value of the existing attributes. The gain value of each attribute is calculated using equation (3).

b. Create a branch for each value
   After getting the root in the previous stage, namely GS, the next step is to make a branch of the decision tree to find a solution or look for the next node. GS has 3 attributes, namely low, normal, and high.

c. Split cases in a branch
   The process of dividing cases into this branch is carried out in the same way as the initial process, but the GS value is used as the root so that the entropy value used is the GS entropy value.

d. Repeat the same process on all branches
   The gain calculation process will continue until the tree finds a solution or until a leaf is found, at which point the calculation cannot be continued.

3.4.1 Data Normalization and Attribute Class Division

Data normalization is done to balance the data if the data has a wide range of values and makes it easier for the classification process [21]. The division of attribute classes used for classification using the C4.5 algorithm is divided into low, normal, and high. In addition to the division of attribute classes, the data used for calculations is shown in Table 2.

<table>
<thead>
<tr>
<th>No.</th>
<th>PI</th>
<th>PT</th>
<th>LA</th>
<th>SS</th>
<th>PR</th>
<th>DS</th>
<th>Diagnosis</th>
<th>Type of Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>53.94</td>
<td>9.31</td>
<td>43.1</td>
<td>44.64</td>
<td>124.4</td>
<td>25.08</td>
<td>AB</td>
<td>SL</td>
</tr>
</tbody>
</table>

Irma Handayani, Copyright © 2023, MIB, Page 1508
Submitted: 18/06/2023; Accepted: 31/07/2023; Published: 31/07/2023
The classification process is a process where the input data will be classified by the system into the appropriate class based on the tree that has been formed before. Test data that has been normalized is entered into the system for classification. The data that has been input will then carry out a search or sequence to which branch of the tree to arrive at a solution.

### 3.5 K-NN Classification Process

K-NN classification process is done by comparing the similarities between test data and training data which have been owned by the system. If the similarity of the case value in the training data compared to the test data is greater, then it will be collected as a solution. Data collected as a set of solution is as much as the value of k, so the case with k similarity value as much as k will be used as the solution set. Class diagnosis that has the most frequency will be taken and displayed as a solution by the system. Examples of cases in training data are shown in Table 4.

### Table 4. Training Data

<table>
<thead>
<tr>
<th>No.</th>
<th>PI</th>
<th>PT</th>
<th>LA</th>
<th>SS</th>
<th>PR</th>
<th>DS</th>
<th>Diagnosis</th>
<th>Type of Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>53.94</td>
<td>9.31</td>
<td>43.1</td>
<td>44.64</td>
<td>124.4</td>
<td>25.08</td>
<td>AB</td>
<td>SL</td>
</tr>
<tr>
<td>2</td>
<td>84.97</td>
<td>33.02</td>
<td>60.86</td>
<td>51.95</td>
<td>125.66</td>
<td>74.33</td>
<td>AB</td>
<td>SL</td>
</tr>
<tr>
<td>3</td>
<td>89.01</td>
<td>62.94</td>
<td>111.48</td>
<td>6.06</td>
<td>NO</td>
<td>NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>85.35</td>
<td>124.42</td>
<td>67.02</td>
<td>70.76</td>
<td>AB</td>
<td>SL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>45.37</td>
<td>30.52</td>
<td>122.34</td>
<td>2.29</td>
<td>NO</td>
<td>NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>50.75</td>
<td>48.01</td>
<td>93.56</td>
<td>56.13</td>
<td>AB</td>
<td>SL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>77.11</td>
<td>64.64</td>
<td>112.15</td>
<td>70.76</td>
<td>AB</td>
<td>SL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>77.41</td>
<td>40.75</td>
<td>118.45</td>
<td>93.56</td>
<td>AB</td>
<td>SL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>74.38</td>
<td>84.97</td>
<td>143.56</td>
<td>56.13</td>
<td>AB</td>
<td>SL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>50.91</td>
<td>118.15</td>
<td>118.45</td>
<td>57.66</td>
<td>NO</td>
<td>NO</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The process of classifying test data is divided into several steps, namely the process of calculating similarity, the process sorting highest similarity, and the process of determining solutions as a result of classification that is shown in Table 5.

### Table 5. Test Data Sample

<table>
<thead>
<tr>
<th>No.</th>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PI</td>
<td>55.29</td>
</tr>
<tr>
<td>2</td>
<td>PT</td>
<td>20.44</td>
</tr>
<tr>
<td>3</td>
<td>LA</td>
<td>34</td>
</tr>
<tr>
<td>4</td>
<td>SS</td>
<td>34.85</td>
</tr>
</tbody>
</table>
3.5.1 The Process of Sorting the Highest Similarity

After the calculation of the similarity between the test data and all the training data has been completed, the next step is to sort the similarity results from the highest to the lowest and then take the greatest similarity as much as k, in this example we will use k = 5. The results of the 5 training data with the highest similarity are shown in Table 6.

<table>
<thead>
<tr>
<th>No.</th>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>PR</td>
<td>115.88</td>
</tr>
<tr>
<td>6</td>
<td>DS</td>
<td>3.56</td>
</tr>
</tbody>
</table>

Table 6. Training Data with the Highest Similarity

3.6 System Interface

To enable interaction between the user and the system, an intermediary called the system interface is needed. At the system interface, users can enter data and perform analysis using the C4.5 algorithm and the K-NN algorithm.

3.6.1 Analysis Page

The analysis page is the page where the user enters the data he wants to classify. On the analysis page, users can choose an algorithm to classify the data they have. The available options for classifying are the C4.5 algorithm and the K-NN algorithm. The analysis page display is shown in Figure 3.

![Analysis Page Display](image-url)

Figure 3. Analysis Page Display

3.6.2 Form C4.5

Form C4.5 is a form for displaying the results of data mining analysis using the C4.5 algorithm with previously prepared training data. Besides being able to display the results of the analysis, form C4.5 also allows the user to see the calculations performed by the system. The appearance of form C4.5 is shown in Figure 4.

![Display Form C4.5](image-url)

Figure 4. Display Form C4.5

3.6.3 Form K-NN

The K-NN form is a form that displays the results of data mining analysis using the K-NN algorithm. In this form the user can see the results of the classification of data entered based on the previous training data and see system calculations for the data entered using the K-NN algorithm. The K-NN form is shown in Figure 5.
3.7 Testing

3.7.1 System Testing

Tests conducted on the system is a k-fold cross validation test with k-10 with previously randomized data with details of 150 patients with spondylolisthesis, 100 data normal patients and 60 data for patients suffering from disk hernia. Then the randomized data is shared to 10 fold with each fold containing 31 pieces of data. Division of data into fold shown in Table 7.

<table>
<thead>
<tr>
<th>Fold</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-31</td>
</tr>
<tr>
<td>2</td>
<td>32-62</td>
</tr>
<tr>
<td>3</td>
<td>63-93</td>
</tr>
<tr>
<td>4</td>
<td>94-124</td>
</tr>
<tr>
<td>5</td>
<td>125-155</td>
</tr>
<tr>
<td>6</td>
<td>156-186</td>
</tr>
<tr>
<td>7</td>
<td>187-217</td>
</tr>
<tr>
<td>8</td>
<td>218-248</td>
</tr>
<tr>
<td>9</td>
<td>249-279</td>
</tr>
<tr>
<td>10</td>
<td>280-310</td>
</tr>
</tbody>
</table>

3.7.2 C4.5 Classifier Testing

Testing is carried out by entering test data one by one into the system and then recording the results of the classification and the running time needed by the system to carry out the classification. Accuracy calculation is done by using equation (6). Based on the test results on the C4.5 classifier, a system accuracy rate of 89% was obtained and the details of the test results are shown in Table 8.

<table>
<thead>
<tr>
<th>No</th>
<th>Type of Disease</th>
<th>Amount of Data</th>
<th>Accuracy</th>
<th>Average Running Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Spondylolisthesis</td>
<td>150</td>
<td>96%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Normal</td>
<td>100</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Disk Hernia</td>
<td>60</td>
<td>86%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>310</td>
<td>89%</td>
<td>0.00912297</td>
</tr>
</tbody>
</table>

3.7.3 K-NN Classifier Testing

The test is carried out by entering the test data one by one and recording the results to calculate the accuracy based on equation (4) to get a total accuracy of 83%. Details of the test results are shown in Table 9.

<table>
<thead>
<tr>
<th>No</th>
<th>Type of Disease</th>
<th>Amount of Data</th>
<th>Accuracy</th>
<th>Average Running Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Spondylolisthesis</td>
<td>150</td>
<td>87%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Normal</td>
<td>100</td>
<td>78%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Disk Hernia</td>
<td>60</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>310</td>
<td>83%</td>
<td>0.000212303</td>
</tr>
</tbody>
</table>

3.7.4 Comparison of C4.5 and K-NN Classifier

At this stage, a comparison of the accuracy and time required for the system to classify the data used will be displayed. The test data used was 93 test data from a total of 310 data obtained. The recapitulation of the test results is shown in Table 10.
4. CONCLUSION

Based on the research that has been done, the test results using k-fold (k = 10) cross validation followed by a confusion matrix on 310 data consisting of 60 data from herniated disk patients, 150 data from spondylolisthesis patients, and 100 normal patient data have an accuracy of 83% for the classifier K-NN and 89% for classifier C4.5. The C4.5 method gets a higher accuracy value which caused by initiating rules that allow certain patient’s data to enter the appropriate class, while K-NN with a value of K = 5 requires decisions from the 5 closest neighbors which may not be neighbor that has the highest closeness value but based on the frequency of the number of classes in the solution set. The average time needed to classify the C4.5 classifier is 0.00912297 seconds and the K-NN classifier is 0.000212303 seconds. This is because the K-NN method only performs ordinary mathematical calculations to carry out classifications, while C4.5 requires more time to search the test data on the decision trees that are built. Thus, it is better to use the C4.5 algorithm to classify Hernia Disk disease and Spondylolisthesis because the level of accuracy is higher, but for running time in the classification process the K-NN algorithm has a faster classification time.

REFERENCES


