Application of K-Nearest Neighbor Algorithm on Classification of Disk Hernia and Spondylolisthesis in Vertebral Column

I Handayani*[1]

[1]Department of Informatics Engineering, Faculty of Information Technology and Electro, Universitas Teknologi Yogyakarta, Jl. Siliwangi, Yogyakarta 55285

E-mail: irma.handayani@staff.uty.ac.id[1]

Masuk: 20 Juni 2019, direvisi: 9 Agustus 2019, diterima: 24 Agustus 2019

Abstrak. Tulang punggung sebagai bagian dari tulang belakang yang mempunyai peran penting pada tubuh manusia. Trauma yang terjadi pada tulang punggung dapat mempengaruhi kemampuan sumsum tulang belakang untuk mengirim dan menerima pesan dari otak ke sistem tubuh yang mengendalikan sensorik dan motorik. Disk Hernia dan Spondylolisthesis merupakan contoh dari penyakit yang terjadi pada tulang punggung. Penelitian tentang klasifikasi penyakit atau kerusakan tulang dan sendi dari sistem kerangka masih jarang dilakukan. Padahal sistem klasifikasi tersebut dapat digunakan sebagai *second opinion* oleh para ahli radiologi sehingga dapat meningkatkan produktivitas dan konsistensi diagnosis dari ahli radiologi. Penelitian ini menggunakan *dataset* vertebral *column* yang memiliki tiga kelas (Disk Hernia, Spondylolisthesis, Normal) dan 310 *instance* yang terdapat pada UCI *Machine Learning*. Dalam penelitian ini menggunakan metode penerapan algoritma K-NN untuk klasifikasi penyakit Disk Hernia dan Spondylolisthesis pada tulang belakang. Data disusun dalam dua tugas klasifikasi yang berbeda namun terkait, yaitu kategori "normal" dan "abnormal". Algoritma K-NN melakukan pendekatan klasifikasi data dengan cara mengoptimalkan data contoh yang dapat dijadikan acuan sebagai data training untuk menghasilkan klasifikasi data tulang belakang berdasarkan proses belajar *(learning)*. Hasil penelitian menunjukkan bahwa akurasi dari *classifier* K-NN sebesar 83%. Rata-rata lama waktu yang dibutuhkan untuk melakukan klasifikasi *classifier* K-NN 0,000212303 detik.

Kata kunci: Algoritma K-NN; Disk Hernia; Spondylolisthesis; klasifikasi; tulang belakang

Abstract. Vertebral column as a part of backbone has important role in human body. Trauma in vertebral column can affect spinal cord capability to send and receive messages from brain to the body system that controls sensory and motoric movement. Disk hernia and spondylolisthesis are examples of pathologies on the vertebral column. Research about pathology or damage bones and joints of skeletal system classification is rare whereas the classification system can be used by radiologists as a second opinion so that can improve productivity and diagnosis consistency of the radiologists. This research used dataset Vertebral Column that has three classes (Disk Hernia, Spondylolisthesis and Normal) and instances in UCI Machine Learning. This research applied the K-NN algorithm for classification of disk hernia and spondylolisthesis in vertebral column. The data were then classified into two different but related classification tasks: "normal" and "abnormal". K-NN algorithm adopts the approach of data classification by optimizing sample data that can be used as a reference for

training data to produce vertebral column data classification based on the learning process. The results showed that the accuracy of K-NN classifier was 83%. The average length of time needed to classify the K-NN classifier was 0.000212303 seconds.

Keywords: K-NN algorithm; disk hernia; spondylolisthesis; classification; vertebral column

1. Introduction

Vertebral column or spinal sequence is a flexible structure formed by a number of bones called vertebra or vertebrae. Between every two segments of the vertebral column are cartilage pads. The length of the vertebral column in adults reaches 57 to 67 cm. In total there are 33 bone segments, 24 of which are apart and the remaining 9 segments join to form bone [1]. Spinal Cord Injury (SPI) or spinal injury can cause permanent disruption in the body function. In general the causes of SPI or spinal injuries are traffic accidents (50%), falls (25%), and sports-related injuries (10%), other than that due to violence and work accident. SPI due to trauma are estimated to occur in 30-40 per one million population per year, and around 8.000-10.000 sufferers each year, generally occurring in adolescents and adults [2]. Various kinds of trauma to the vertebral column are caused by many factors and causes malfunction. Disk hernia occurs when the disc is damaged due to several conditions, such as falls or accidents, the back strain when lifting or turning the back roughly, degeneration caused by disk aging, and spontaneous herniation that occurs without prior injury. When the discus breaks, the inner jelly-like center will bulge out through the soft outer fibrous tissue causing a bulge that presses the nerves around [3]. Under normal circumstances the vertebral column are arranged in a straight line. Spondylolisthesis is the condition of the spine where one vertebra is a shifted forward (*anterolisthesis*) or backward (*retrolisthesis*) [4].

Previous research by Kristy [5] used the decision tree (J48) method and bagging for the classification of disk hernia and spondylolisthesis. Decision tree method is easy to be presented or understood, but it is an unstable method and uncertain method that gives the same prediction when given a new case or test instance. Meanwhile, bagging is one of the ensemble methods that can be used to overcome this instability. Research on the classification of diseases or damages to bones and joints of the skeletal system is still rarely done because there is no database with quantitative numeric attributes that is able to describe the disease. Based on research [6], [7], [8], [9], [10] and [11], it is proven that the K-NN algorithm is good and results in high accuracy values for classification techniques data mining. This research applies the K-NN algorithm method for classification of disk hernia and spondylolisthesis in the vertebral column.

2. Theoretical Framework

2.1. Data Mining

Aggarwal [12] states that data mining is the process of collecting, cleaning, processing, analyzing and acquiring knowledge derived from data. Data mining is a terminology for bringing together the vast variety of data formats. Besides that, Berry and Linoff [13] define data mining as: "an automatic or semi–automatic exploration and analysis process of grear data in order to find important patterns and rules.". One of the main roles of data mining is classification.

2.2. Classification

Sulistyo [14] states that the classification comes from the Latin word "*classis*" which is the process of grouping, which means collecting the same object or entity and separating different objects or entities. In general it can be said that classification is the process of calculating existing data or is also called training data with new data or testing data. This process will generate possibilities in the testing data. Several algorithms can be used to calculate the classification process, one of which is K-NN algorithm. Classification is widely used to determine the decision according to new knowledge gained from past data processing using an algorithm. In the dataset classification, there is one destination attribute or it can also be called a label attribute. Attribute is what will be sought from the new data based on other attribute in the past. The number of attributes can affect the performance of an

algorithm. Some data mining classification techniques are proven to be good and resulting in high accuracy values, including the K-NN, Naïve Bayes and C4.5 algorithm.

2.3. K-Nearest Neighbor Algorithm

Larose [15] states that K-NN is a learning based algorithm where the dataset training is stored. So, the classification for the new record that is not classified is obtained by comparing the record that is most similar to the training set. Besides being used for classification, K-NN algorithm is also used for estimation and prediction. The steps of K-NN algorithm are:

a. Determining the parameter k (number of closest neighbors)
b. Calculating the distance (similarity) between all training records and new objects
c. Sorting data based on distance value from the smallest to the largest value
d. Retrieving data from a number of k values
e. Determining the most-frequent labels occuring in the k training records closest to the object

2.4. Numerical Attribute Similarity

In the numeric attribute, a calculation of the distance (distance between two objects) can be done using Euclidean distance, Manhattan distance and Minkowski distance calculation. In this research, the writer uses Euclidean distance to calculate the distance between two objects with nominal attributes. The neighbors proximity or distance is calculated based on Euclidean distance with Equation (1) [16].

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \quad (1)$$

Where:

$d(x,y)$: distance between data x and data y

x_k: attribute value to k from the test data (x), with k= 1, 2, …, n

y_k: attribute value to k from the training data (y), with k = 1, 2, …, n

After the distance or dissimilarity (d) is calculated then it is converted into similarity (s) with an interval between 0 and 1 ($s \in [0, 1]$) with Equation (2).

$$s = \frac{1}{1+d} \quad (2)$$

2.5. Confusion Matrix

Confusion Matrix is a table to evaluate the performance of the identification model. Confusion Matrix shows the result of identification between the number of correct prediction data and the number of incorrect predictive data compared to the facts produced. Table 1 shows the Confusion Matrix [17].

Table 1 Confusion Matrix

Actual	**Prediction**	
	Negative	**Positive**
Negative	a	b
Positive	c	d

Where,

a: the number of data predicted by the system with the true results that indicate healthy condition; the doctor states the patient to be healthy.

b: the number of data predicted by the system with false results that indicate malaria; the doctor states the patient to be healthy.

c: the number of data predicted by the system with true results and wrong indication, the doctor states that indication is malaria.

d: the number of data predicted by the system with the true and malaria indication, the doctor states the indication is malaria.

There are several terms based on Table 1.

- True Positive (TP) is positive data correctly indicated on the model. TP values can be calculated using Equation (3).

$$TP = \frac{d}{c+d} \tag{3}$$

- False Positive (FP) is positive data incorrectly indicated on the model. FP values can be calculated using Equation (4).

$$FP = \frac{b}{a+b} \tag{4}$$

- True Negative (TN) is negative data that is correctly indicated in the model. TN value can be calculated using Equation (5).

$$TN = \frac{a}{a+b} \tag{5}$$

- False Negative (FN) is negative data that is incorrectly indicated in the model. FN values can be calculated using Equation (6).

$$FN = \frac{c}{c+d} \tag{6}$$

2.6. Measurement Accuracy

Measurement accuracy is a step to prove the level of performance of an algorithm dataset used. In this research, confusion matrix is used as a performance measurement tool of classification algorithm. Confusion matrix is a calculation that compares datasets with the results of the classification in accordance with the actual data with the total number of data. The final result of this matrix is the level of accuracy in the units of percent (%). This level of accuracy will be used later as the researchers' reference to perform the classification algorithm. Confusion matrix contains information of the comparison between classification labels and actual labels. From Table 1 the level of accuracy from an algorithm model can be calculated using Equation (7) [18].

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

Where,

TP: True Positive, the number if positive data correctly classified by the system.
TN: True Negative, the number of negative data correctly classified by the system.
FN: False Negative, the number of negative data falsely classified by the system.
FP: False Positive, the number of positive data falsely classified by the system.

2.7. Requirements Analysis

The system designed is a system used to classify existing patient data into several classes, namely normal classes and abnormal classes. The abnormal classes 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 each value owned by each patient, namely *pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, and degree spondylolisthesis.*

2.8. Data Input

Input data is data that will be input to the system. These input data then will be processed using the K-NN classification method to determine the class of patients. The data include: *Pelvic Incidence* (PI) is the angle between the perpendicular line between the sacral plates and a line connecting the midpoint

of the sacral plate to the bicoxofemoral axis; PI is specific value and consistent for each patient; *Pelvic Tilt* (PT) is pelvic orientation which connects the femur with other body parts. PT can be moved forward, backward, as well as other directions; *Lumbar lordosis angle* (LA) is the characteristic of the human spine and the reference of human posture to see whether the posture is good or bad; *Sacral slope* (SS) is a slope located between the sacral plate and the horizontal plane; *Pelvic radius* (PR) is a value that affects the development of large lumbar lordosis; and *Degree spondylosisthesis* (DG) or grade of spondylolisthesis is a measurement that states the degree of how much of the lower body part is slipping forward.

2.9. System Description

The system is designed to be able to classify the vertebral column data using the K-NN algorithm. The disease is then divided into 3 classes, namely the disk hernia class, spondylolisthesis, and normal. The process applied to the system is divided into 3 stages, such as *pre-processing, design classifier,* and *post-processing stages*. The system flow is presented in Figure 1.

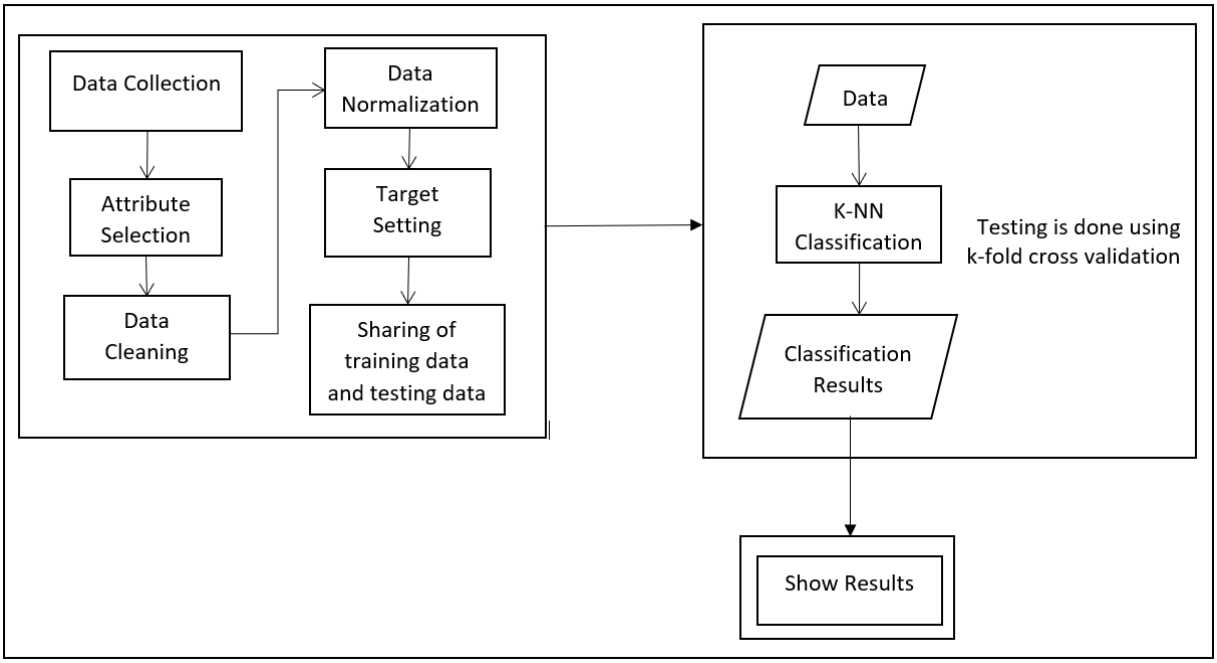

Figure 1. System Flow

The *pre-processing* stage is the stage that starts from the data collection process. The data collected later are grouped based on the influence on each class. After that, the normalized data are input into the appropriate class. After normalizing the data, the data are divided into 2 parts, namely training data and test data with the percentage of 70% training data and 30% randomized test data. After the *pre-processing* stage is completed, the data is then entered into the classifier as knowledge. The classifier then learns from the data that has been entered and evaluated. If the attribute has not been trained, then the system training process will be repeated with different structure and function. 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 suffers vertebral column disease, and the type of illness whether it is a disk hernia or spondylolisthesis.

3. Methodology

The methodology used in this research was divided into several stages as shown in Figure 2.

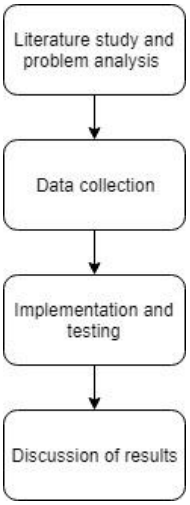

Figure 2. Research Methodology Flow Diagram

3.1. 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 of writing. This stage is carried out to obtain information related to data mining, classification and K-NN algorithm. Information is obtained from observing problems related to factors needed to be used in this study and observing related studies data classification using the K-NN method.

3.2. 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 dataset built by Dr. Henrique da Mota during a medical residence period in the Group of Applied Research in Orthopaedics (GARO). The data have been organized in two different but related classification tasks. The first task is to classify the patients into one out of three categories: Normal (100 patients), Disk Hernia (60 patients), or Spondylolisthesis (150 patients). For the second task, the categories of Disk Hernia and Spondylolisthesis are merged into a single category labelled as 'abnormal'. Thus, the second task is classifying patients into one out of two categories: Normal (100 patients) or Abnormal (210 patients).

3.3. K-Fold Cross Validation

Witten et al [19] state that cross validation is a simple form of statistical techniques. The number of standard fold to predict the error of the data is 10-fold cross validation. Cross validation is used in order to find the best parameters of one model [20]. This is done by testing the number of errors in the testing data. In cross validation, data is divided into in *k* samples of the same size. From the *k* subset of data used will be used *k-1* sample training data and 1 remaining sample for testing data. In cross validation, data is divided into *k* samples of the same size. From the k subset of data used, it will be used *k-1* sample as training data and 1 remaining sample for testing data. This is often called as 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 that consists of 9 a subset of data for training and 1 subset of data for testing. Then, the average error (error the mean) is calculated. If there are 3 models, then each model is tried 10 times in each combination of training-testing subset and every run will be found an error for each model. The model that gives the average with the smallest error is the best method.

3.4. Implementation and Testing

Based on the data that have been obtained and various references that have been completed, the following steps are implementation and testing. The results of the system design are outlined in the

form of implementation program 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 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 on test data that has been prepared. This is done to analyze the K-NN algorithm inside classifying disk hernia and spondylolisthesis in the vertebral column.

4. Result and Discussion

4.1. Disease Data Compilation Process

The data used as training data and test data are data about disk hernia and spondylolisthesis. There are 310 training data consisting of 150 cases of spondylolisthesis, 60 disk hernia case data, and 100 normal case data. Based on data obtained, the attributes used to do the classification are pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, and degree spondylolisthesis.

4.2. K-NN Classifier Process

K-NN classification process is done by comparing the similarities between test data with training data which have been owned by the system. If the similarity of the case value in the training data with the test data is greater, then it will be collected as a solution. The data collected as a set of solutions 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 2.

Table 2. Training Data Tables

No.	Pelvic Incidence (PI)	Pelvic Tilt (PT)	Lumbar Lordosis Angle (LA)	Scaral Slope (SS)	Pelvic Radius (PR)	Degree Spondylolisthesis (DS)
1	53.94	9.31	43.1	44.64	124.4	25.08
2	84.97	33.02	60.86	51.95	125.66	74.33
3	89.01	26.08	69.02	62.94	111.48	6.06
4	85.35	15.84	71.67	69.51	124.42	76.02
5	45.37	10.76	29.04	34.61	117.27	-10.68
…	…	…	…	…	…	…
214	40.75	1.84	50	38.91	139.25	0.67
215	47.81	10.69	54	37.12	125.39	-0.4
216	42.52	16.54	42	25.97	120.63	7.88
217	95.38	24.82	95.16	70.56	89.31	57.66

The process of classifying test data is divided into several steps, namely the process of calculating similarity, process sorting highest similarity, and the process of determining solutions as a result of classification, as shown in Table 3.

Table 3. Example of Test Data Tables

No.	Attribute	Value
1	PI	55,29
2	PT	20,44
3	LA	34
4	SS	34,85
5	PR	115,88
6	DS	3,56

4.3. The highest ranking process of similarity

After calculating the similarity between the test data and all the training data, the next step is sorting the results of similarity from highest to lowest, then the highest similarities of k is taken, in this example we will use k=5. Results of 5 training data with similarity the highest is shown in Table 4.

Table 4. Training Data Table with the Highest Similarity

No.	Training Data	Similarity	Diagnosis
1	28	0.183452	DH
2	126	0.10691	DH
3	182	0.095142	NO
4	6	0.094533	NO
5	128	0.081227	DH

4.4. 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 spondylolithesis, 100 data of 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. The division of data into fold is shown in Table 4.

Table 4. Division of Data into Fold

Fold	Data
1	1-31
2	32-62
3	63-93
4	94-124
5	125-155
6	156-186
7	187-217
8	218-248
9	249-279
10	280-310

4.5. Testing K-NN Classifier

The k-fold cross validation test is performed on the K-NN classifier by dividing the data in Table 4. The results using k=10 for k-fold cross validation are shown in Table 5.

Table 5. K-NN Classifier Test Results

No	Type of disease	Total data	Accuracy	Running time average (seconds)
1	Spondylolisthesis	150	87%	0,000212303
2	Normal	100	78%	
3	Disk Hernia	60	84%	
	Total	**310**	**83%**	

5. Conclusion

The test results using k-fold (k-10) cross validation, followed by confusion matrix of 310 data consisting of 60 data on disk hernia patients, 150 data in spondylolisthesis patients and 100 normal patient data have a K-NN classifier accuracy of 83%. This makes K-NN with k=5 needs decision from 5 closest neighbors which might not have the highest closeness but might have the most frequent class in the solution set. With 83% accuracy, K-NN method is proven to be good for data mining classification. The average speed of the classification process (running time) of the K-NN class is 0,000212303 seconds, which is considered to be very fast in carrying out the classification process because the K-NN method only does ordinary mathematical calculation to do the classification.

6. References

[1] E. C. Pearce, *Anatomi dan Fisiologi untuk Paramedis*. PT. Gramedia Pustaka Utama, Jakarta, 2012.

[2] J. Maja, "Diagnosis Dan Penatalaksanaan Cedera Servikal Medula Spinalis," *J. Biomedik*, vol. 5, no. 3, 2014.

[3] J. Jordan and K. Konstantinou, "Herniated Lumbar Disc," *Clin. Evid. (Online).*, vol. 9, no. June, pp. 34–44, 2016.

[4] K. A. Irianto, F. W. Hatmoko, and L. P. K, "Degenerative Spondylolisthesis : The preferable surgical technique," *Bali Med. J.*, vol. 7, no. 1, p. 215, 2018.

[5] M. A. Kristy, "Klasifikasi Penyakit pada Tulang Punggung Menggunakan Metode j48 dan Bagging," *Tesis, Dep. Ilmu Komput. FMIPA UGM, Yogyakarta*, 2013.

[6] B. G. Pratama, "Analisis Perbandingan Metode Pengukuran Jarak Pasangan Titik-Titik Ciri dan Metode Klasifikasi Terhadap Tiga Parameter Kantuk Pengemudi," *Tesis, Dep. Ilmu Komput. FMIPA UGM, Yogyakarta*, 2018.

[7] A. K. F. U. Harjoko, "KLASIFIKASI CITRA BATIK KAIN BESUREK DENGAN SPEED UP ROBUST FEATURES (SURF) DAN GRAY LEVEL CO-OCCURRENCE MATRIX (GLCM)No Title," *Tesis, Dep. Ilmu Komput. FMIPA UGM, Yogyakarta*, pp. 0–1, 2017.

[8] F. Kurniawan and Ivandari, "Komparasi Algoritma Data Mining Untuk Klasifikasi Penyakit Kanker Payudara," *J. Stmik*, vol. XII, no. 1, pp. 1–8, 2017.

[9] Indrayanti, D. Sugianti, and M. A. Al Karomi, "Optimasi Parameter K Pada Algoritma K-Nearest Neighbour Untuk Klasifikasi Penyakit Diabetes Mellitus," *Pros. SNATIF Ke-4 2017*, pp. 823–829, 2017.

[10] Mustakim and G. Oktaviani F, "Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa," *J. Sains, Teknol. dan Ind.*, vol. 13, no. 2, pp. 195–202, 2016.

[11] F. Agus, H. Hatta, Rahmania, and Mahyudin, "Pengklasifikasian Dokumen Berbahasa Arab Menggunakan K-Nearest Neighbor," *JSM (Jurnal SIFO Mikroskil)*, vol. 18, no. 1, pp. 43–56, 2017.

[12] C. C, Aggarwal, Data Mining: *The Textbook*, Switzerland, Springer, 2015.

[13] M. J. a. Berry and G. S. Linoff, *Data mining techniques,* Second Edition, Indianapolis, Wiley Publishing, 2004.

[14] B. Sulistyo, "Pengantar Ilmu Perpustakaan," PT. Gramedia Pustaka Utama, Jakarta, 1991.

[15] D. T. Larose, "Discovering An Introduction to Data Mining," *Discov. Knowl. Data*, 2005.
[16] T. M. Cover and P.E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 118–119, 1997.
[17] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," *Digit. Signal Process. A Rev. J.*, vol. 17, no. 4, pp. 694–701, 2007.
[18] F. Gorunescu, *Data Mining: Concept, Model and Techniques*. Heidelberg, Berlin: Springer, 2011.
[19] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems)*. Burlington: Elsevie, 2011.
[20] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning (2nd ed., web version)," *Math. Intell.*, pp. 369–370, 2008.