Application Of K-Nearest Neighbor (K-NN) Algorithm On Lung Disease Diagnosis Expert System

Olha Musa, A.Malik I Buna, Marlin Lasena, R. Rizal Isnanto, Suryono Suryono
0lh4mu54@gmail.com, mailqloex@gmail.com, marlinlasena@gmail.com, rizal_isnanto@yahoo.com, suryono@fisika.undip.ac.id
Academic of Management and Informatica of Computer Ichsan Gorontalo, Indonesia; Diponegoro University, Indonesia

ABSTRACT

Background: In this research diagnose of lung disease by using the K-NN method with some types of lung diseases that attack humans of various types. The lungs are very important organs for humans as part of important organs, it becomes one of the benchmarks for humans said to be physically fit in living their daily lives to do work activities. There are 12 (twelve) types of lung disease i.e. Pneumonia, Legionnaries, Pleural Efusion, Tuberculosis (TB), Pneumothorax, Asthma, Chronic Obstructive, Chronic Bronchitis, Emphysema, Lung Disease due to Work Environment Conditions (COPD), Silicosis, Asbestosis, if one of them attacks humans will cause decreased immunity and susceptible to other diseases. Method: The method of this research was designed of the application the K-Nearest Neighbor (K-NN) algorithm on an expert system for diagnose of lung disease in the RSUD. PROF. Dr. Aloei Saboe Gorontalo. From the 12 (twelve) types of disease, there are 8 (eight) types that currently attack the Gorontalo people. Result: The result of this research is a computerized expert system with the K-Nearest Neighbor (K-NN) method that produces allegations or the same diagnosis results with diagnoses performed by an expert (the doctor) at the Aloei Saboe Hospital in Gorontalo. Conclusion: The conclusions of this research is the decision-making process with information technology (computer applications) become easier, more effective and efficient in provided treatment for patients of lung disease which increasing every year.

Keywords: Lung Disease, K-NN method, Expert System, Application, Diagnosis

1. Background

As part of an important organ, the lungs become one of the benchmarks for humans to be said to be physically fit in living their daily lives to do work activities. Lung disease in humans is one of the many diseases that occur today. From year to year, patients with lung disease are increasing, in the last 3 (three) years from 2015 to 2017, patients with lung disease in hospitals. Prof. Dr. Aloei Saboe, Gorontalo City, out of 1030 outpatients and 4583 inpatients, the number of patients reaching 5613 became one of the reasons for this study which would later help diagnose patients with lung disease. There are 12 (twelve) types of lung diseases that attack humans in general if they cannot maintain good health, out of 12 (twelve) types of diseases there are 8 (eight) types of lung diseases that currently have suffered by the community gorontalo (RSUD. Prof. Dr. Aloei Saboe, Gorontalo City, 2017).

As for some types of lung diseases including Turbulence (TB), Asthma, Bronchitis, pneumonia, emphysema and lung cancer. The method used in this study is K-Nearest Neighboryang which is one method that is suitable for use in classifying lung disease. The advantages of KNN have several advantages, namely that it is resilient to noisy and effective data training when the training data is large [1]

The most common form of interstitial lung disease (ILD), is a different type of chronic, progressive, and fibrosis interstitial pneumonia with no known cause, especially in adults aged 1–3. The overall prognosis of IPF has remained poor for the past few years, with average survival ranging from 3 to 5 years and a 5-year survival rate ranging from 30% to 50% 4.5. There is no proven treatment other than lung transplants for IPF [2].

Nontuberculous mycobacteria (NTM) lung disease increases globally. Although the epidemiology of NTM etiology differs across regions, Mycobacterium avium complex (MAC) is the main cause of NTM lung disease in most countries, including mainland Japan. Okinawa is located in the southernmost region of Japan and is the only prefecture categorized as a subtropical region in Japan, therefore the possibility of epidemiology of etiology of NTM lung disease is different from mainland Japan [3]

Expert systems mimic the behavior of an expert in dealing with a problem. In the case of a patient who visits a doctor to check the health of a person who has a disorder, the doctor or health expert will examine and diagnose sometimes patients who want to seek treatment for too long waiting to know what disease the patient has.

This study focuses on the Application of the K-Nearest Neighbor (K-NN) Algorithm in the Expert System for Diagnosing Lung Disease according to its type. With the aim of the research to design the application of the application of the K-Nearest Neighbor (KNN) algorithm on the expert system of diagnosing lung disease in the RSUD. PROF. Dr. Aloei Saboe Gorontalo City.

In the title of the study "The Effect of Artificial Neural Network Models Combined with Six Tumor Markers in the Diagnosis of Lung Cancer". To evaluate the potential diagnosis of artificial neural network models (ANN) combined with six tumor markers in an additional diagnosis of lung cancer, to differentiate lung cancer from pulmonary benign disease, normal control, and gastrointestinal cancer. Carcino-embryonic antigen (CEA), gastrin, neuron-specific enolase (NSE), sialic acid (SA), Cu / Zn, Ca were measured by different experimental procedures in 117 lung cancer patients, 93 benign lung disease patients, 111 normal controls, 47 gastric cancer patients, 50 patients with colon cancer and 50 esophageal cancer patients, 19 basic information parameters surveyed among lung cancer, benign lung disease and normal controls, then developed and evaluated ANN model to differentiate lung cancer [4].

In the research title "Analysis of Lung Disease Using the K-Nearest Neighbor Algorithm at the Aloei Saboe Hospital in Gorontalo City". The lung disease analysis process using the K-Nearest neighbor algorithm, obtained the prediction results conducted by the K-Nearest Neighbor algorithm resulting in a fairly high accuracy reaching 91.90% thus able to detect lung diseases accurately and based on the precision values that reach 86.67%, the K-Nearest Neighbor algorithm is able to detect lung disease correctly [1]

In the research title "CBR framework with a choice of features based on improvements to the classification of lung cancer subtypes". Classification of molecular subtypes is a challenging field in the diagnosis of lung cancer. Although different methods have been proposed for biomarker selection, efficient discrimination between adenocarcinoma and squamous cell carcinoma in clinical practice presents several difficulties, especially when the latter is poorly differentiated. This is an increasingly important field, because certain treatments and other medical decisions are based on molecular and histological features. An urgent need for systems and a series of biomarkers that provide accurate diagnosis [5]

In the research title "The physiological and pathological role of congenital lymphoid cell tissue in the lungs" The lungs are important open organs and the main place of respiration. Many life-threatening diseases develop in the lungs, for example, pneumonia, asthma, chronic obstructive pulmonary disease (COPD), pulmonary fibrosis, and lung cancer. In the lungs, innate immunity serves as the front line in both anti-tumor and anti-tumor defense responses and is also important for mucosal homeostasis; thus, it plays an important role in this lung disease. Congenital lymphoid cells (ILCs), characterized by tight tissue culture and different functions in the mucosa, attract increased attention to innate immunity [6]

Distinguishing previous research with research conducted at this time is focused on the Application of the K-Nearest Neighbor (K-NN) Algorithm in the Expert System for Diagnosing Lung Disease according to its type, which results in a presumption or the same diagnosis with a diagnosis made by an expert (doctor). Moreover, would be developed on making software (information system for lung disease).

2. Methods

2.1 Classification of CBR

CBR classification describes the target problem using old experience, and finds a solution to the problem by taking similar cases stored on a case basis to the target problem where the base of the case is a specific knowledge base from experience. The case usually taken with learning techniques for CBR classifiers and the most common technique is K-NN. The process of finding common problems by CBR classifiers in fig. 2.1

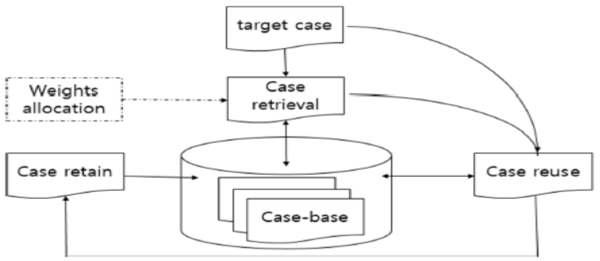


Fig 2.1 The process of finding common problems by CBR classifiers (Source: Amy M. Kwon., 2018)

The K-NN algorithm uses neighboring classification as the predictive value of the new test sample. Near or near neighbors are usually calculated based on distance traveled. The KNN method algorithm is very simple, works based on the shortest distance from the test sample to the training sample to determine the KNN.

After collecting KNN, then the majority of KNNs were taken to be predicted from the test sample. The data for the KNN algorithm consists of some multi-variate attributes.

$$d_i = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2}$$

Dengan:

x_1	= sampel data
x_2	= data uji
i	= variabel data
dist	= jarak
p	= dimensi data

Xi which will be used to classify Y. Data from K-NN can be on any size scale, from ordinal to nominal. The advantages of K-NN have some advantages, that it is resilient to training noisy and effective data if the training data is large. While the disadvantages of KNN are:

- KNN needs to determine the value of the parameter K (number of nearest neighbors)
- Learning based on distance is not clear about what type of distance to use and which attributes should be used to get the best results.

The search for similarities between the new case and the old case is done by matching the symptoms entered by the user according to the symptoms in the knowledge base. This retrieval process will use the K-Nearest Neighbor method.

CBR classification describes the target problem using old experience, and finds a solution to the problem by taking a similar case stored on the base case to the target problem where the basis of the case is the specific knowledge base from past experience. The case is usually taken by learning techniques for CBR classifiers, and the most common technique is K-NN. [7]

K-Nearest Neighbor (K-NN) known as one of the simplest nonparametric classifiers but in the high-dimensional accuracy settings, KNN is affected by interference features. In this research, K-Nearest Neighbor is important as a new approach to binary classification in high-dimensional problems [8]

The K-Nearest Neighbor algorithm (k-nearest neighbor or K-NN) is an algorithm for classifying objects based on learning data that is closest to the object. A special case where classification is predicted based on the closest learning data (in other words, k = 1) is called a nearest neighbor algorithm [9]

The purpose of this algorithm is to classify new objects based on the attributes and sample training. Classification does not use any model to matched and only based on memory. Given a test point, a number of K objects (training points) would found closest to the test point. Classification uses the most votes among classifications of K objects. The K-NN algorithm uses adjacency classification as the predictive value of the new test sample. Close or far neighbors usually calculated based on Eucledian distance. The KNN method algorithm is very simple, working on the shortest distance from the test sample to the training sample to determine the KNN.

2.2 Research Steps

Research on the Application of K-Nearest Neighbor (K-NN) Algorithm in the Expert System for Lung Disease Diagnosis in RSUD. Aloei Saboe Gorontalo, it has research steps that refer to the methodology of the K-Nearest Neighbors algorithm. The step of applying lung disease in the research procedure of fig. 2.2

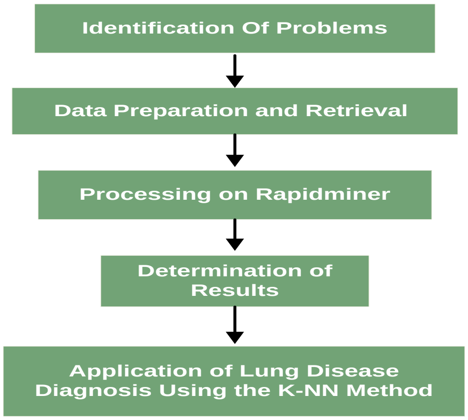

Identification Of Problems
Data Preparation and Retrieval
Processing on Rapidminer
Determination of Results
Application of Lung Disease Diagnosis Using the K-NN Method

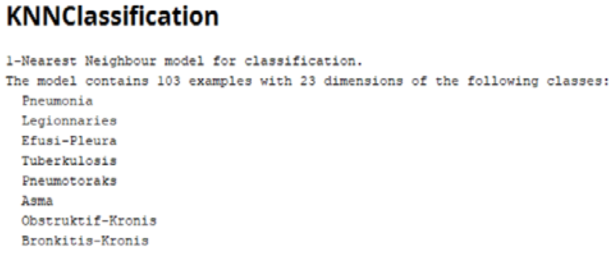
Fig. 2.2 Research procedure Application of Lung Disease Diagnosis by Using the K-Nearest Neighbors method

3. Finding and Discussion

Finding

The result of this research to get outside that achieved as folows:

K-NN clasification and dimention



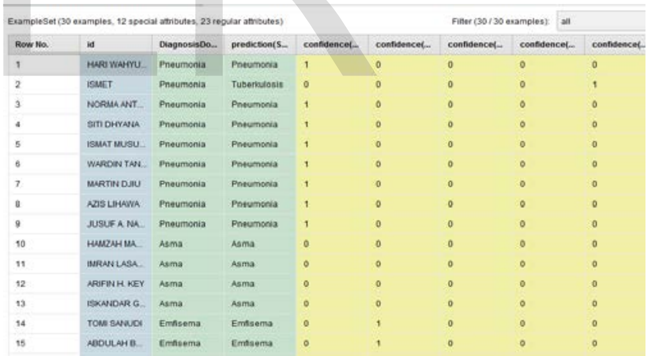
KNNClassification

```
1-Nearest Neighbour model for classification.
The model contains 103 examples with 23 dimensions of the following classes:
  Pneumonia
  Legionnaries
  Efusi-Pleura
  Tuberkulosis
  Pneumotoraks
  Asma
  Obstruktif-Kronis
  Bronkitis-Kronis
```

Apply of Data Model

It can be seen from the data model of the diagnosis of (test data) as many as 30 rows and predicted with sample data that has applied the KNN algorithm as many as 103 rows, then the prediction results are 1 wrong data and the rest are correct. Applying the Data Model in table 3.1

Tabel 3.1 Apply of Data Model



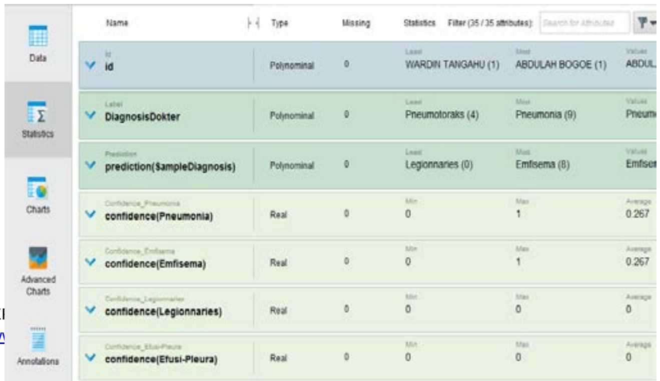
ExampleSet (30 examples, 12 special attributes, 23 regular attributes) Filter (30 / 30 examples): all

Row No.	id	DiagnosisDo...	prediction(S...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...
1	HARI WAHYU...	Pneumonia	Pneumonia	1	0	0	0	0
2	ISMET	Pneumonia	Tuberkulosis	0	0	0	0	1
3	NORMA ANT...	Pneumonia	Pneumonia	1	0	0	0	0
4	SITI DHYANA	Pneumonia	Pneumonia	1	0	0	0	0
5	ISMAT MUSU...	Pneumonia	Pneumonia	1	0	0	0	0
6	WARDIN TAN...	Pneumonia	Pneumonia	1	0	0	0	0
7	MARTIN DJIU	Pneumonia	Pneumonia	1	0	0	0	0
8	AZIS LIHAWA	Pneumonia	Pneumonia	1	0	0	0	0
9	JUSUF A. NA...	Pneumonia	Pneumonia	1	0	0	0	0
10	HAMZAH MA...	Asma	Asma	0	0	0	0	0
11	IMRAN LASA...	Asma	Asma	0	0	0	0	0
12	ARIFIN H. KEY	Asma	Asma	0	0	0	0	0
13	ISKANDAR G...	Asma	Asma	0	0	0	0	0
14	TOMI SANUDI	Emfisema	Emfisema	0	1	0	0	0
15	ABDULAH B...	Emfisema	Emfisema	0	1	0	0	0

Apply of Statistic Model

This table displays the average number of predictive data for each disease from the test data. Applying the statistical data model in table 3.2

Tabel 3.2 Apply of Statistic Model



Filter (35 / 35 attributes): Search for Attributes

Name	Type	Missing	Statistics		
Id id	Polynominal	0	Least WARDIN TANGAHU (1)	Most ABDULAH BOGOE (1)	Values ABDUL
Label DiagnosisDokter	Polynominal	0	Least Pneumotoraks (4)	Most Pneumonia (9)	Values Pneum
Prediction prediction(SampleDiagnosis)	Polynominal	0	Least Legionnaries (0)	Most Emfisema (8)	Values Emfise
Confidence_Pneumonia confidence(Pneumonia)	Real	0	Min 0	Max 1	Average 0.267
Confidence_Emfisema confidence(Emfisema)	Real	0	Min 0	Max 1	Average 0.267
Confidence_Legionnaries confidence(Legionnaries)	Real	0	Min 0	Max 0	Average 0
Confidence_Efusi-Pleura confidence(Efusi-Pleura)	Real	0	Min 0	Max 0	Average 0

Apply of Chart Model

The chart results by using the chart style Scatter model with x-Axis are prediction of disease and y-Axis is the test data to be entered. Applying the Chart data model in Figure 3.1

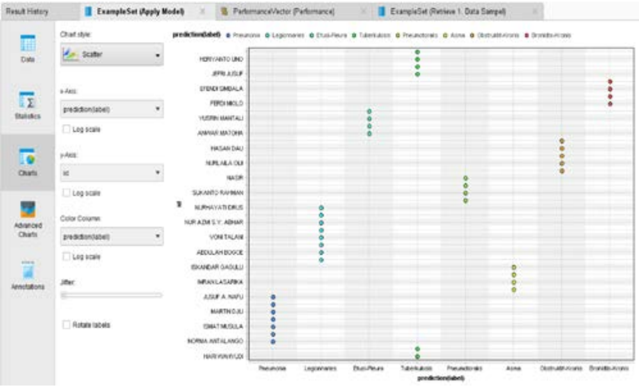


Chart style: Scatter
x-Axis: prediction(label)
y-Axis: id
Color Column: prediction(label)

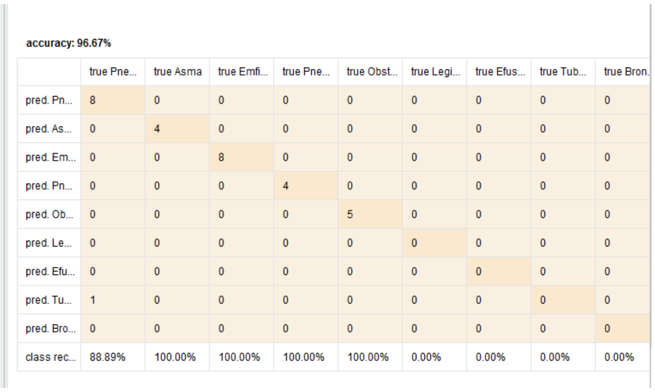
Fig. 3.1 Apply of Chart Model

Discussion

Performance Accuracy

Performance accuracy obtained from the test data is 96.97% of the total data of 30 rows.

Table 3.3 Performance Accuracy

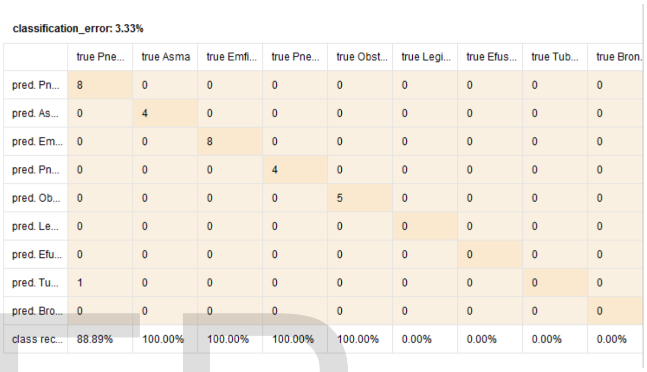

accuracy: 96.67%

	true Pne...	true Asma	true Emfi...	true Pne...	true Obst...	true Legi...	true Efus...	true Tub...	true Bron.
pred. Pn...	8	0	0	0	0	0	0	0	0
pred. As...	0	4	0	0	0	0	0	0	0
pred. Em...	0	0	8	0	0	0	0	0	0
pred. Pn...	0	0	0	4	0	0	0	0	0
pred. Ob...	0	0	0	0	5	0	0	0	0
pred. Le...	0	0	0	0	0	0	0	0	0
pred. Efu...	0	0	0	0	0	0	0	0	0
pred. Tu...	1	0	0	0	0	0	0	0	0
pred. Bro...	0	0	0	0	0	0	0	0	0
class rec...	88.89%	100.00%	100.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%

Performance Clasification Error

Performance Clasification Error which is obtained where the error is in the prediction of Pneumonia, which is supposed to be the disease, is Tuberculosis. Therefore, the accuracy rate was reduced by 3.33%. An example of displaying performance classification errors in table 3.4

Tabel 3.4 Performance Clasification Error


classification_error: 3.33%

	true Pne...	true Asma	true Emfi...	true Pne...	true Obst...	true Legi...	true Efus...	true Tub...	true Bron.
pred. Pn...	8	0	0	0	0	0	0	0	0
pred. As...	0	4	0	0	0	0	0	0	0
pred. Em...	0	0	8	0	0	0	0	0	0
pred. Pn...	0	0	0	4	0	0	0	0	0
pred. Ob...	0	0	0	0	5	0	0	0	0
pred. Le...	0	0	0	0	0	0	0	0	0
pred. Efu...	0	0	0	0	0	0	0	0	0
pred. Tu...	1	0	0	0	0	0	0	0	0
pred. Bro...	0	0	0	0	0	0	0	0	0
class rec...	88.89%	100.00%	100.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%

4 Conclusions and recommendations

Conclusion

Based on the results of testing on the prediction of lung disease using the K-Nearest Neighbors algorithm, conclusions are obtained as follows:

1. The results of the predictions carried out by the K-Nearest Neighbors algorithm produced a high accuracy of 96.97% and thus were able to detect accurately lung disease.
2. Based on the value of the preformance clasification error which only 3.33%, the K-Nearest Neighbor algorithm has a slight possibility of data errors.

So from the results achieved, this research should continue at the interface design stage in the form of software designed according to the K-Nearest Neighbors Algorithm modeling.

Suggestion

To hope that further research will add sample data to improve accuracy in predicting lung disease so that later it can be used as a reference for sample data to create an accurate and accurate lung disease expert system.

Acknowledgment

Thank you to the authors convey to all those who have helped in the process of writing this report, which include;

The Director General of Research and Development Strengthening who has been willing to provide Hiba University Cooperation Research (PKPT) until this research can carried out properly.

References

[1] Musa Olha, Alang "Analisis Penyakit PARU-PARU MENGGUNAKAN ALGORITMA," vol. 9, pp. 348–352, 2017.

[2] J. Park, J. Jung, S. H. Yoon, J. M. Goo, H. Hong, and J. Yoon, "Inspiratory Lung Expansion in Patients with Interstitial Lung Disease : CT Histogram Analyses," *Sci. Rep.*, no. October, pp. 1–13, 2018.

[3] H. Nagano, T. Kinjo, Y. Nei, S. Yamashiro, J. Fujita, and T. Kishaba, "Causative species of nontuberculous mycobacterial lung disease and comparative investigation on clinical features of Mycobacterium abscessus complex disease : A retrospective analysis for two major hospitals in a subtropical region of Japan," pp. 1–12, 2017.

[4] F. Feng, Y. Wu, Y. Wu, and G. Nie, "The Effect of Artificial Neural Network Model Combined with Six Tumor Markers in Auxiliary Diagnosis of Lung Cancer," pp. 2973–2980, 2012.

[5] J. Ramos-González, D. López-Sánchez, J. A. Castellanos-Garzón, J. F. de Paz, and J. M. Corchado, "A CBR framework with gradient boosting based feature selection for lung cancer subtype classification," *Comput. Biol. Med.*, vol. 86, pp. 98–106, 2017.

[6] H. Cheng, C. Jin, J. Wu, S. Zhu, Y. J. Liu, and J. Chen, "Erratum to: Guards at the gate: physiological and pathological roles of tissue-resident innate lymphoid cells in the lung (Protein & Cell, (2017), 8, 12, (878-895), 10.1007/s13238-017-0379-5)," *Protein Cell*, vol. 8, no. 12, p. 932, 2017.

[7] A. M. Kwon, "A rank weighted classification for plasma proteomic profiles based on case-based reasoning," *BMC Med. Inform. Decis. Mak.*, vol. 18, no. 1, pp. 1–10, 2018.

[8] H. R. Shahraki, S. Pourahmad, and N. Zare, ? "Important Neighbors : A Novel Approach to Binary Classification in High Dimensional Data," vol. 2017, 2017.

[9] Novita Mariana and dkk, "PENERAPAN ALGORITMA K-NN (nearest Neighbor) UNTUK DETEKSI PENYAKIT (KANKER SERVIKS) Novita Mariana, Rara Sriartati Redjeki, Jeffri Alfa Razaq Abstrak," vol. 7, no. 1, pp. 26–34, 2015.

[10] O. Musa, "Sistem Informasi Pemetaan Pendidikan Menggunakan Algoritma Data Mining," vol. 01, pp. 26–32, 2015.