Journal of Physics: Conference Series

PAPER • OPEN ACCESS

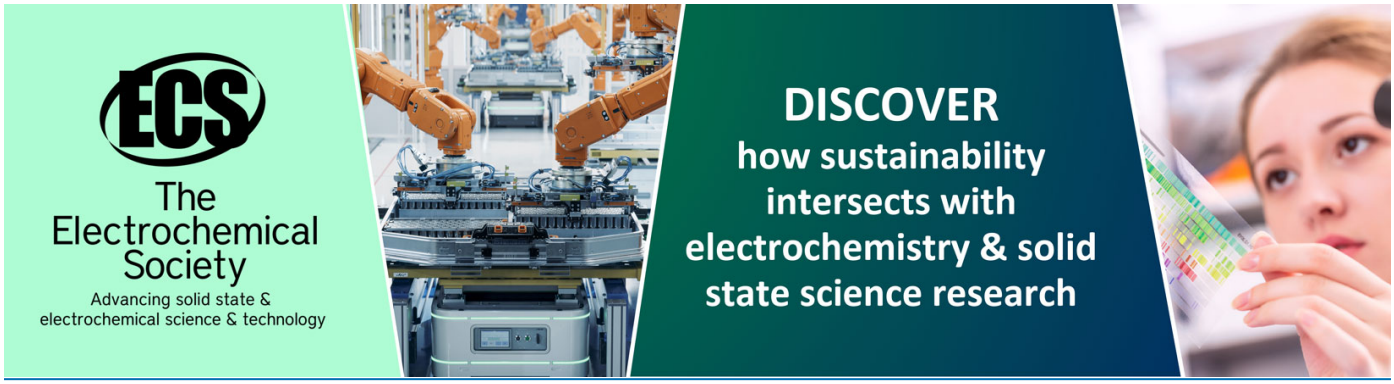
Classification of Infection Type Based on Leukocytes Examination Results Using K-Nearest Neighbor

To cite this article: S A N Suyanto *et al* 2020 *J. Phys.: Conf. Ser.* **1566** 012130

View the article online for updates and enhancements.

You may also like

- Unsupervised Method for Calculating Diameter and Number of Leukocyte Cells
Retno Supriyanti, Ahmad Haeromi, Yogi Ramadhani et al.
- The manifestation of optical centers in UV–Vis absorption and luminescence spectra of white blood human cells
Yu G Terent'yeva, V M Yashchuk, L A Zaika et al.
- Nanostructure characteristics of three types of platelet-rich fibrin biomaterial: a histological and immunohistochemical study
Thuy-Duong Nguyen-Thi, Bao-Song Nguyen-Tran, Thuan Dang-Cong et al.



This content was downloaded from IP address 3.144.252.201 on 08/05/2024 at 14:02

Classification of Infection Type Based on Leukocytes Examination Results Using K-Nearest Neighbor

S A N Suyanto[1], B Siregar[1*], E B Nababan[1], H A Fikri[1]

1 Department of Information Technology, Faculty of Computer Science and Information Technology, University of Sumatera Utara, Indonesia.

*Email: baihaqi@usu.ac.id

Abstract. Leukocytes are blood cells that contain nuclei, also called white blood cells. Leukocytes have a role in the cellular and humoral defiance of organisms against foreign substances. Laboratory tests of blood samples greatly influence the diagnosis of a disease. Manual blood tests do have a low price but still have some weaknesses such as the length of time needed will be longer, because health practitioners must examine them carefully to avoid misinformation. To help overcome these weaknesses, a classification of types of infections was carried out based on the results of leukocyte examination. Classification is a grouping of data where the data used has a label or target class. So that the algorithms for solving classification problems are categorized into supervised learning. The purpose of supervised learning is that label data or targets play a role as a 'supervisor' or 'teacher' who oversees the learning process in achieving a certain level of accuracy or precision. The algorithm used in this study is K-Nearest Neighbour. The data used in this study as many as 2,098 results of complete blood tests taken from one hospital in Medan. This study resulted in a classification accuracy of 92%.

1. Introduction

Laboratory tests of blood samples greatly affect the diagnosis of a disease. Based on the results of laboratory tests, such as those obtained from examination of leukocytes, it can be identified the possibility of disease that attacks the patient's body. The results of this examination can also identify the presence of infection in the body. Leukocytes are able to produce antibodies to support immune function. In addition, leukocytes also have the ability to diapedesis, namely the ability to penetrate capillary blood vessel walls and enter cells or tissues. Increases or decreases in leukocytes in the extreme can be indicated as an infection in the body. Leukocytes originate from the bone marrow and circulate throughout the bloodstream and are an important part of our immune system.

K-Nearest Neighbor (KNN) algorithm is one of the most widely used algorithms to determine classification. This algorithm works by grouping data based on the similarity or closeness that exists in the training data. The more data used for training, the greater the accuracy that can be generated from testing data because KNN works based on existing similarities. KNN stores all training data and almost all training data is needed during the testing period. KNN is done by finding the K group of objects in the training data that is closest (similar) to the object in new data or testing data. In other words, the purpose of the KNN algorithm is to classify new objects based on attributes and training data.

Research conducted by Dzikrulloh & Indriati (2017) uses KNN in the recruitment of prospective teachers and employees [1]. In this study four criteria were used namely the average GPA, academic test

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

results, results of general knowledge about science and technology, and interview test results. This research was successfully carried out by ranking so that the best results can be taken. The results of testing the effect of the best K value with several weight value criteria obtain an accuracy value of 94%. Research conducted by Hermawan & Agung (2017) on the application of sales data to predict sales based on item categories [2]. The data used in the two-year distance range produces an accuracy of 85.91%. Research conducted by Khamis et al. (2014) using the KNN algorithm can increase error reduction in patient diagnoses and reduce time to diagnose while still being able to improve efficiency and effectiveness in treatment with the accuracy of the results of this study was 75% [3]. Research conducted by Kataria & Singh (2013) has a success rate of almost 100% because the training data and test data used are still small in term and within easy reach [4]. Research conducted by Ndaumanu (2014) successfully analyzed the prediction of student resignation using KNN [5]. Research conducted by Johar et al. (2016) uses KNN and Simple Additive Weighting in decision making for selection of members of the Paskibraka reception [6]. Research conducted by Gunawan et al. (2019) is in identifying automatic plate number using KNN and works efficiently [7].

2. Material and methods

The data used in the study came from one hospital in Indonesia. From these data the patient's complete blood examination data was taken throughout 2017. After data collection, data analysis was performed according to the system requirements. Data analysis was performed using KNN. Total data used were obtained from 2,098 patient data. From these data there are five criteria used in the study. The five criteria for grouping are examination results, neutrophil examination reference values, eosinophil examination reference values, lymphocyte reference values, and monocyte examination reference values.

The method proposed to determine the classification of infection types consists of several stages. These stages start from the data collection that is data collection. Next, preprocessing data is carried out in the form of selecting to obtain the desired data section. Then cleaning is used to check for inconsistent data and fill in missing values. The labeling process is carried out to make a mark on the training data for machine learning. The next stage is determining the value of K, then calculating the distance between the testing data and the training data. At this stage, the distance of testing data is calculated with all training data. Then, sort the distance value based on K into the group that has the smallest Euclid value (sort the distance starts from the smallest value to the largest value). Next, collect classification categories based on K values and determine the best K value. The stages above can be seen in the general architecture in Figure 1.

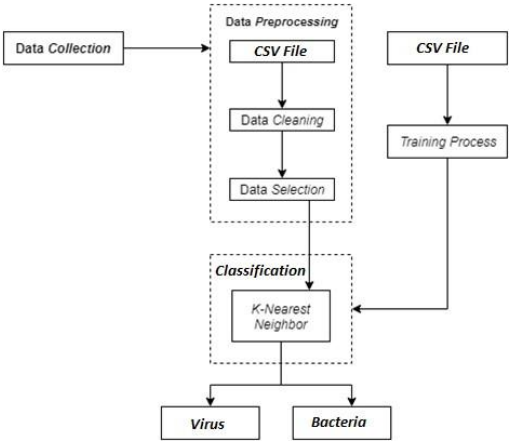


Figure 1. General Architecture.

In the system analysis stage, four types of attributes with quantitative scale are selected, namely the results of eosinophil examination, lymphocyte examination, monocyte examination, and neutrophil examination. Previously, the preprocessing process will be carried out on the raw data because not all data will be used in this study. The sample raw data obtained is shown in Table 1.

Table 1. Raw data samples before preprocessing.

laboratory_id	test_id	test_name	result	result_comment	reference_value
20681		Eritrosit	2.78		4.4 – 5.9
20681		MCH	21.9		27.0 – 31.0
20681		MCHC	31.3		32.0 – 36.0
20681		MCV	70.1		82.0 – 92.0
20681		Hematokrit	19.5		43 – 49
20681		Hemoglobin	6.1		14.0 – 17.0
20681		Leukosit	14.48		3.8 – 10.6
20681		Basofil	0.3		0 – 1
20681		Eosinofil	0.3		1 – 6
20681		Monosit A	0.27		0.2 – 0.4
20681		Trombosit	194		150 – 440
20681		Basofil A	0.04		0 – 0.1
20681		Eosinofil A	0.02		0 – 0.10
20681		RDW-CV	21.9		11.0 – 15.5
20681		Limfosit	10.3		20 – 40
20681		Monosit	4.5		2 – 8
20681		Neutrofil A	13.81		2.7 – 6.5
20681		Limfosit A	0.34		1.5 – 3.7
20681		PDW	8.4		9.6 – 15.2
20681		Neutrofil	83.6		50 – 70
20681		MPV	8.4		9.2 – 12.0
20681		RDW-SD	52.6		39 – 46

The first preprocessing process is the selection of data taken from a collection of operational databases. The selected data is stored in a separate file from the operational database. In this study, the data used are eosinophil, lymphocyte, monocyte and neutrophil attribute data, while other attribute data are ignored. Examples of data from the selection results can be seen in Table 2.

Table 2. Sample data selection results.

laboratory_id	test_id	test_name	result	result_comment	reference_value
20681		Eosinofil	0.3		1 – 6
20681		Limfosit	10.3		20 – 40
20681		Monosit	4.5		2 – 8
20681		Neutrofil	83.6		50 – 70
20682		Eosinofil	10.1		1 – 6
20682		Limfosit	36.2		20 – 40
20682		Monosit	0.8		2 – 8
20682		Neutrofil	24.5		50 – 70
20683		Eosinofil	13.2		1 – 6
20683		Limfosit	26.1		20 – 40
20683		Monosit	2.4		2 – 8
20683		Neutrofil	36.7		50 – 70

In the cleaning process, deletion of duplicate data is removed and the handling of missing value is carried out. An example of the data after the cleaning process can be seen in Table 3. The ‘laboratory_id’ attribute is changed to the ‘patient’ attribute and other attributes are replaced with each type of examination results that will be used in this study, namely eosinophils, lymphocytes, monocytes, and neutrophils.

Table 3. Data cleaning sample results.

Patient #	Eosinofil	Limfosit	Monosit	Neutrofil
1	0.3	10.3	4.5	83.6
2	10.1	36.2	0.8	24.5
3	13.2	26.1	2.4	36.7
4	0.1	15.6	14.2	72.1
5	6.3	12.6	1.8	48.2
6	0.1	5.1	14.4	80.2
7	1.3	23.3	9.4	78.1
8	2.0	41.5	21.0	34.9
9	1.8	44.7	8.2	42.8
10	0.1	4.5	9.8	59.5

The cleaning data is given a label of the type of infection that can then be used in the training data process. Data that has been labeled as a type of cause of infection can be seen in Table 4.

Table 4. Data samples that have been labeled type of infection.

Patient #	Eosinofil	Limfosit	Monosit	Neutrofil	Classification
1	0.3	10.3	4.5	83.6	Virus
2	10.1	36.2	0.8	24.5	Bacteria
3	13.2	26.1	2.4	36.7	Bacteria
4	0.1	15.6	14.2	72.1	Virus
5	6.3	12.6	1.8	48.2	Bacteria
6	0.1	5.1	14.4	80.2	Virus
7	1.3	23.3	9.4	78.1	Virus
8	2.0	41.5	21.0	34.9	Bacteria
9	1.8	44.7	8.2	42.8	Bacteria
10	0.1	4.5	9.8	59.5	Virus

The initial stage in the classification process using the KNN algorithm is the determination of the K value used. In this study, a K value of 5. was determined. At this stage the distance of testing data to be calculated was eosinophils = 14.9, lymphocytes = 20.6, monocytes = 7.8, and neutrophils = 65.6. Examples of testing data, for patients numbered 11, which will predict the type of cause of the infection with training data can be seen in Table 5.

Table 5. Samples of test data with training data are labeled.

Patient #	Eosinofil	Limfosit	Monosit	Neutrofil	Classification
1	0.3	10.3	4.5	83.6	Virus
2	10.1	36.2	0.8	24.5	Bacteria
3	13.2	26.1	2.4	36.7	Bacteria
4	0.1	15.6	14.2	72.1	Virus
5	6.3	12.6	1.8	48.2	Bacteria
6	0.1	5.1	14.4	80.2	Virus
7	1.3	23.3	9.4	78.1	Virus
8	2.0	41.5	21.0	34.9	Bacteria
9	1.8	44.7	8.2	42.8	Bacteria
10	0.1	4.5	9.8	59.5	Virus
11	14.9	20.6	7.8	65.6	?

The next job is to calculate the distance between the testing data with the training data that has been given a label. Examples of the results of calculating the distance between testing data and training data for patients numbered 11 can be seen in Table 6.

Table 6. Example calculation of distance between patient data.

Distance	Patient #									
	1	2	3	4	5	6	7	8	9	10
Patient #11	25.57	44.77	29.95	18.09	21.83	26.75	18.73	41.47	35.67	22.79

Calculations to get the value of this distance using formula 1.

$$\boldsymbol{Dxy} = \sqrt{\sum^{\boldsymbol{n}}_{\boldsymbol{i}=\boldsymbol{1}}(\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}})^{\boldsymbol{2}}} \quad (1)$$

with

D = distance
x = training data
y = testing data
n = the number of individual attributes
í = i[th] individual attribute

The value of the distance calculation results then sorted starting from the smallest distance to the largest distance that has been previously calculated, then grouping the data with its 5 closest neighbors (because the specified K is 5). It appears in Table 7 that the selected distance values are written in bold (distance to patients numbered 4, 7, 5, 10, and 1 in sequence).

Table 7. Distance and sample order of test data with a value of K = 5.

Patient #	Distance	Classification	Order
1	**25.57**	**Virus**	**5**
2	44.77	Bacteria	10
3	29.95	Bacteria	7
4	**18.09**	**Virus**	**1**
5	**21.83**	**Bacteria**	**3**
6	26.75	Virus	6
7	**18.73**	**Virus**	**2**
8	41.47	Bacteria	9
9	35.67	Bacteria	8
10	**22.79**	**Virus**	**4**

The results obtained from the training show that the classification of causes of infection that predominates and belongs to the closest neighbor group (K = 5) is infection caused by a virus, so testing data for patients numbered 11 with the results of eosinophil, lymphocyte, monocyte, and neutrophil examination results respectively worth 14.9, 20.6, 7.8, and 65.6. With the conclusion that the patient numbered 11 is indicated to suffer from an infection caused by a virus.

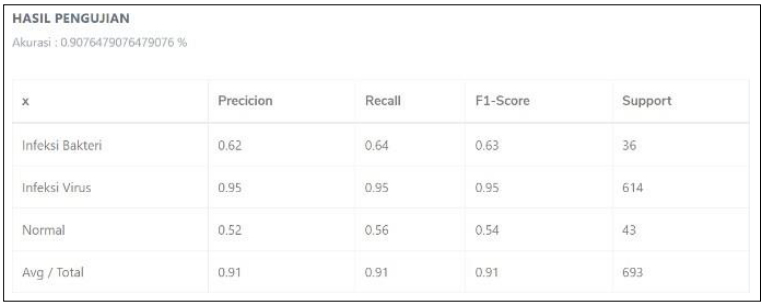
3. Result and discussion

This section will describe the results obtained from the implementation of the KNN algorithm (KNN) in the classification of infection based on the results of leukocyte examination. The testing process was carried out six times with different K values as shown in Table 8.

Table 8. Proportion of total data to be tested with different K values.

No.	Value of K	Number of Testing Data
1.	1	103
2.	3	103
3.	5	103
4.	7	103
5.	9	103

At each dataset test, a different K value will be given. From the results of this test we will get several different accuracy values. Examples of results from testing with a value of K = 1 get an accuracy of 90.7% which appears in the application dashboard built as in Figure 2. With an average value for precision, recall and f-score of 91%.


HASIL PENGUJIAN

Akurasi : 0.9076479076479076 %

x	Precicion	Recall	F1-Score	Support
Infeksi Bakteri	0.62	0.64	0.63	36
Infeksi Virus	0.95	0.95	0.95	614
Normal	0.52	0.56	0.54	43
Avg / Total	0.91	0.91	0.91	693

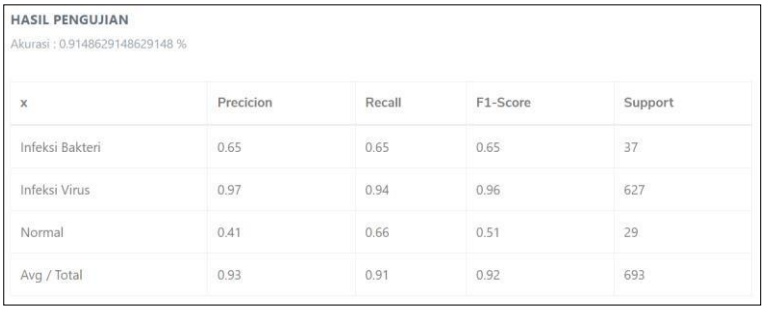
Figure 2. Parameter values for the test results with a value of K = 1.

The results of distance and sequence calculations for testing using sample test data with a value of K = 1 can be seen in Table 9. Selected distance values are marked in bold type, meaning that with a value of K = 1, the patient numbered 11 is indicated to have a viral infection (such as symptoms suffered by numbered patients 4).

Tabel 9. Distance and sample order of test data with a value of K = 1.

Patient #	Distance	Classification	Order
1	25.57	Virus	5
2	44.77	Bacteria	10
3	29.95	Bacteria	7
4	**18.09**	**Virus**	**1**
5	21.83	Bacteria	3
6	26.75	Virus	6
7	18.73	Virus	2
8	41.47	Bacteria	9
9	35.67	Bacteria	8
10	22.79	Virus	4

For K=3, an accuracy of 91.4% is obtained, as shown in the built application dashboard which can be seen in Figure 3.


HASIL PENGUJIAN

Akurasi : 0.9148629148629148 %

x	Precicion	Recall	F1-Score	Support
Infeksi Bakteri	0.65	0.65	0.65	37
Infeksi Virus	0.97	0.94	0.96	627
Normal	0.41	0.66	0.51	29
Avg / Total	0.93	0.91	0.92	693

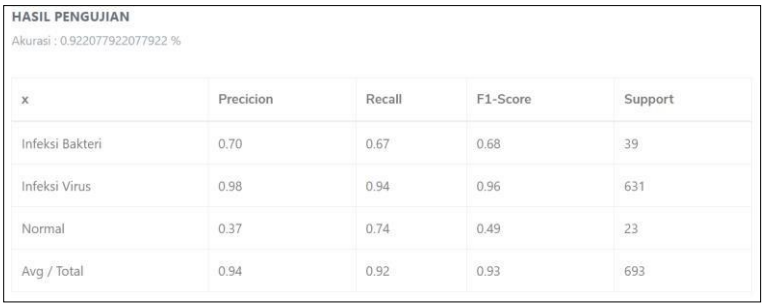
Figure 3. Parameter values for the test results with a value of K = 3.

From the test results, the average values for precision, recall, and f-score were 93%, 91%, and 92%, respectively. The results of distance and sequence calculations for testing using sample test data with a value of K = 3 can be seen in Table 10. Selected distance values are marked in bold type, meaning that with a value of K = 3, patients numbered 11 are indicated as more likely to suffer from infections caused by viruses rather than bacteria (such as symptoms suffered by patients numbered 4, 7, and 5 in order).

Table 10. Distance and sample order of test data with a value of K = 3.

Patient #	Distance	Classification	Order
1	25.57	Virus	5
2	44.77	Bacteria	10
3	29.95	Bacteria	7
4	**18.09**	**Virus**	**1**
5	**21.83**	**Bacteria**	**3**
6	26.75	Virus	6
7	**18.73**	**Virus**	**2**
8	41.47	Bacteria	9
9	35.67	Bacteria	8
10	22.79	Virus	4

For K=5, an accuracy of 92.2% is obtained, as shown in the built application dashboard which can be seen in Figure 4.


HASIL PENGUJIAN

Akurasi : 0.922077922077922 %

x	Precicion	Recall	F1-Score	Support
Infeksi Bakteri	0.70	0.67	0.68	39
Infeksi Virus	0.98	0.94	0.96	631
Normal	0.37	0.74	0.49	23
Avg / Total	0.94	0.92	0.93	693

Figure 4. Parameter values for the test results with a value of K = 5.

From the test results, the average values for precision, recall, and f-score were 94%, 92%, and 93%, respectively. The results of distance and sequence calculations for testing using sample data with a value of K = 5 can be seen in Table 7. Selected distance values are marked in bold type, meaning that with a value of K = 5, patients numbered 11 are indicated as more likely to suffer from infections caused by

viruses rather than bacteria (such as symptoms suffered by patients numbered 4, 7, 5, 10, and 1 in sequence).

For K=7, an accuracy of 91.9% is obtained, as shown in the built application dashboard which can be seen in Figure 5.

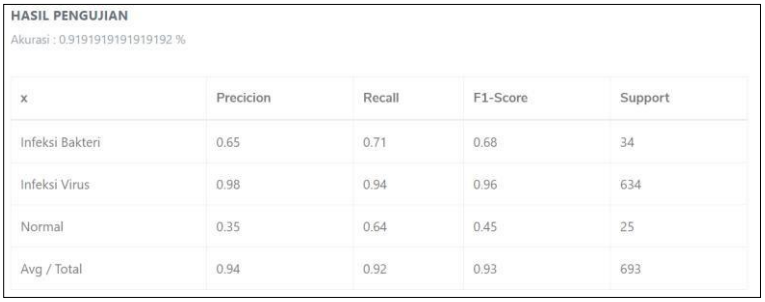


Figure 5. Parameter values for the test results with a value of K = 7.

From the test results, the average values for precision, recall, and f-score were 94%, 92%, and 93%, respectively. The results of distance and sequence calculations for testing using sample data with a value of K = 7 can be seen in Table 11. Selected distance values are marked in bold type, meaning that with a value of K = 7, patients numbered 11 are indicated as more likely to suffer from infections caused by viruses rather than bacteria (such as symptoms suffered by patients numbered 4, 7, 5, 10, 1, 6, and 3 in sequence).

Table 11. Distance and sample order of test data with a value of K = 7.

Patient #	Distance	Classification	Order
1	**25.57**	**Virus**	**5**
2	44.77	Bacteria	10
3	**29.95**	**Bacteria**	**7**
4	**18.09**	**Virus**	**1**
5	**21.83**	**Bacteria**	**3**
6	**26.75**	**Virus**	**6**
7	**18.73**	**Virus**	**2**
8	41.47	Bacteria	9
9	35.67	Bacteria	8
10	**22.79**	**Virus**	**4**

For K=9, an accuracy of 91.7% is obtained, as shown in the built application dashboard which can be seen in Figure 6.

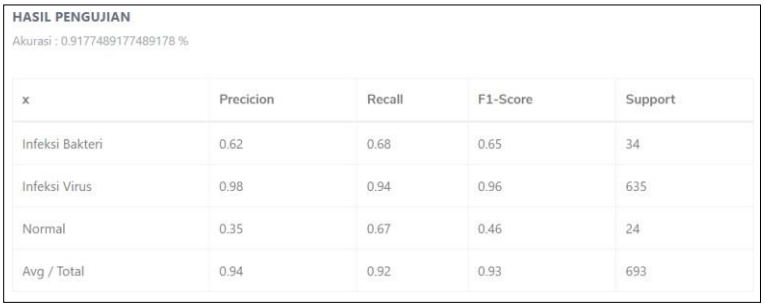


Figure 6. Parameter values for the test results with a value of K = 9.

From the test results, the average values for precision, recall, and f-score were 94%, 92%, and 93%, respectively. The results of distance and sequence calculations for testing using sample test data with a value of K = 9 can be seen in Table 12. Selected distance values are marked in bold type, meaning that with a value of K = 9, the patient numbered 11 is indicated to have an infection caused by a virus or bacteria (such as symptoms suffered by patients numbered 4, 7, 5, 10, 1, 6, 3, 9, and 8 respectively).

Table 12. Distance and sample order of test data with a value of K = 9.

Patient #	Distance	Classification	Order
1	**25.57**	**Virus**	**5**
2	44.77	Bacteria	10
3	**29.95**	**Bacteria**	**7**
4	**18.09**	**Virus**	**1**
5	**21.83**	**Bakteri**	**3**
6	**26.75**	**Virus**	**6**
7	**18.73**	**Virus**	**2**
8	**41.47**	**Bakteri**	**9**
9	**35.67**	**Bakteri**	**8**
10	**22.79**	**Virus**	**4**

After being tested with several different K values, we get a comparison of the test results shown in Figure 7.

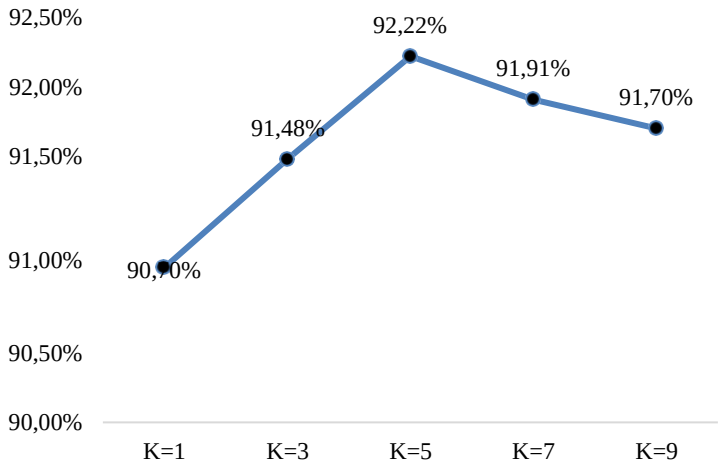


Figure 7. Comparison of the accuracy of the results of data testing with several different K values.

It appears that the highest accuracy value obtained from as many as five times testing in this study is at a value of K = 5, while the lowest accuracy is at a value of K = 1.

4. Conclusion

Based on the results of testing using the KNN algorithm, in the process of classifying the types of causes of infection based on the results of leukocyte examination, this study succeeded in classifying the results of examination of patients into types of infections caused by bacteria or viruses. By using the value of K = 5 that is obtained as many as 39 examination results entered into the classification of infections caused by bacteria, 631 examination results entered into the classification of infections caused by viruses, and 23 examination results entered into the normal classification.

References

[1] N. N. Dzikrulloh and B. D. Indriati, "Penerapan metode K-Nearest Neighbor dan metode Weighted Pı dalam penerimaan calon guru dan karyawan tata usaha baru berwawasan teknologi," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer,* pp. 378-385, 2017.

[2] F. Hermawan and H. Agung, "Implementasi metode K-Nearest Neighbor pada aplikasi data penjualan Multitek Mitra Sejati," *Jurnal Sains Dan Teknologi,* 2017.

[3] H. S. Khamis, K. W. Cheruiyot and S. KImani, "Application of k-Nearest Neighbor classification in medical data mining," *International Journal of Information and Communication Technology Resea* 2014.

[4] A. Kataria and A. Singh, "A review of data classification using K-Nearest Neighbor," *International Journal of Emerging Technology and Advanced Engineering,* 2013.

[5] R. Ndaumanu, "Analisis prediksi tingkat pengunduran diri mahasiswa dengan metode K- Nearest Neighbor," *Jurnal Teknologi Informasi (JATISI),* 2014.

[6] A. Johar, D. Yanosma and K. Anggriani, "Implementasi metode K-Nearest Neighbor dan Simple Add Weighting dalam pengambilan keputusan seleksi penerimaan anggota paskibraka," *Jurnal Pseudoc* 2016.

[7] D. Gunawan, W. Rohimah and R. F. Rahmat, "Automatic Number Plate Recognition for Indonesian License Plate by Using K-Nearest Neighbor Algorithm," in *IOP Conference Series: Materials Scie and Engineering*, 2019.