International journal of science, engineering, and information technology
Volume 05, Issue 02, July 2021
Journal homepage: https://journal.trunojoyo.ac.id/ijseit

Feature Selection and K-nearest Neighbor for Diagnosis Cow Disease

Yeni Kustiyahningsih[a], Mula'ab[b], Qori A. Nurhamidin[b], Jaka Purnama[c]

[a] Information System, University of Trunojoyo Madura, Indonesia
[b] Informatics Engineering, University of Trunojoyo Madura, Indonesia
[c] Industrial Engineering University of 17 Agustus 1945 Surabaya, Indonesia

ABSTRACT

The large number of cattle population that exists can increase the potential for developing cow disease. Lack of knowledge about various kinds of cattle diseases and their handling solutions is one of the causes of decreasing cow productivity. The aim of this research is to classify cattle disease quickly and accurately to assist cattle breeders in accelerating detection and handling of cattle disease. This study uses K-Nearest Neighbour (KNN) classification method with the F-Score feature selection. The KNN method is used for disease classification based on the distance between training data and test data, while F-Score feature selection is used to reduce the attribute dimensions in order to obtain the relevant attributes. The data set used was data on cattle disease in Madura with a total of 350 data consisting of 21 features and 7 classes. Data were broken down using K-fold Cross Validation using k = 5. Based on the test results, the best accuracy was obtained with the number of features = 18 and KNN (k = 3) which resulted in an accuracy of 94.28571, a recall of 0.942857 and a precision of 0.942857.

Keywords: K-Nearest Neighbor, F-Score, Feature Selection, Cow Disease, multi-class.

Article History
Received 24 February 21 Received in revised form 27 June 21 Accepted 26 July 21

1. Introduction

Cows are a very potential livestock commodity in Bangkalan Madura. The support for general grazing land covering 19,025 hectares and the potential for fodder crops of 54,550 hectares contributed to the development of production for cattle breeders amounting to 663,290 [1]. The need for high beef must be balanced with good quality beef. The quality and safety of beef has several criteria, one of which is safe or does not contain germs [2]. Handling of livestock health is examining sick cattle through examinations and changing the changes that occur in livestock with visible symptoms so that they can be taken from the disease [3]. The lack of knowledge of cattle breeders regarding the various diseases that attack livestock as well as solutions for handling cow disease is one of the reasons for the health management process in cattle [4]. There are several types of existing cattle disease, namely intestinal worms, dystocia and others [3,4]. In the study of cow disease, there are 21 symptoms, namely: fever, dull hair, uneasiness, cough, ear scabs, itching, runny nose, paralysis, thinness, limping, runny nose, weakness, hair loss, skin disorders, difficulty breathing, miscarriage, ulcers, bleeding wounds, decreased appetite, bloody stools. Many diseases in cattle cause farmers difficulty in determining detection. Classification is a process approach used to classify data or disease symptoms based on certain categories [5].

The algorithm used for grouping or classifying cattle disease is the K-Nearest Neighbor (KNN) algorithm. KNN is an algorithm that is easy and flexible in giving problems in using the distance approach [5]. The contribution of this study is the classification of SAPI disease with multi-class feature selection using the f-score-KNN method and z-score for classification.

KNN has been widely used in various fields, including disease diagnosis, determining new student admissions, e-learning recommendations and others. The advantage of the KNN method is that it is superior to noisy training data and is effective for large training data [6]. Research on the performance of the KNN algorithm with Naive Bayes, J48 and the Support Vector Machine to determine the position in the building, with 41 public space locations on the UKDW campus shows that the KNN algorithm is better than using the Naive Bayes algorithm, J48 and Support Vector Machine [6]. The KNN is one of a supervised machine learning algorithm used for classification of objects based on learning data, calculates the k of nearest neighbours in the feature, and a sample of a specific category [7,8, 9]. This algorithm involves several main factors: distance measurement, K-value selection and so on [10,11]. The aim of this study was to classify diseases based on feature selection to obtain high accuracy. Data that has been selected for features will be carried out in the normalization process.

* *Corresponding author*
E-mail address: ykustiyahningsih@trunojoyo.ac.id

E-ISSN 2548-4214

Normalization is the process of scaling attribute values over a certain range. The weight value used is more stable and affects the level of accuracy [12]. Several normalization methods are z-score, min-max, and decimal scale. Z-score is the method with the highest accuracy value [13]. Based on previous research, KNN has not been used to classify cows with the characteristics of Pamekasan cattle, and it is implemented by feature selection and normalization with z scores. Select the features that are used to reduce or eliminate the features or symptoms of Cow disease that are less relevant. The study with the title of applying the feature selection method to improve the diagnosis of breast cancer showed that the performance of the C4.5 and Naïve Bayes algorithms improved after using the f-score feature selection [14,15,16]. Therefore, the classification of bovine disease in this study uses the f-score for feature selection, the z-score for normalization, and the KNN for determining the disease diagnosis.

2. Feature Selection

Feature selection is a preprocessing stage that is used to remove features or terms that are less relevant to a data or document [5]. Feature selection reduces features or terms that are less relevant and has no effect on modeling or classification. Feature selection is divided into three categories namely filter models, wrapper models and embedded models [5]. The f-score feature selection is a feature selection model that is included in the filter feature selection model [17]. The technique used in the f-score measures the discrimination of two sets of real numbers. This feature selection can evaluate the features individually. With the training vector X_k, k = 1…, m, if the number of positive and negative n + and n- respectively. Then the f-score of i feature is defined by equation (1):

$$\mathrm{Fi}=\frac{\left(x_i^{(+)}-xi\right)^2+\left(x_i^{(-)}-xi\right)^2}{\frac{1}{n_{+}-1}\sum_{k-1}^{n+}\left(x_{ki}^{(+)}-xi^{(+)}\right)^2+\frac{1}{n_{-}-1}7\sum_{k-1}^{n-}\left(x_{ki}^{(-)}-xi^{(-)}\right)^2} \qquad (1)$$

With :

x_i = Average of features to - *i*,

$x_i^{(+)}, x_i^{(-)}$ = Positive and negative dataset

$x_{ki}^{(+)}, x_k^{(+)}$ = The i feature of k-positive case and k-negative case

The numerator shows discrimination between positive and negative sets and the denominator shows the features in the two sets [5]. The f-score feature selection works as a multi-class. In the Cattle disease dataset, there are seven classes. After selecting the f-score feature, the feature that has a greater threshold value is selected and the rest will be reprocessed for other classes. Figure1. This is the feature selection work process flow. The process of this flowchart starts with inputting disease data, calculating the mean and f-score value, input threshold, if the f-score is less than the threshold value, it will remove the feature.

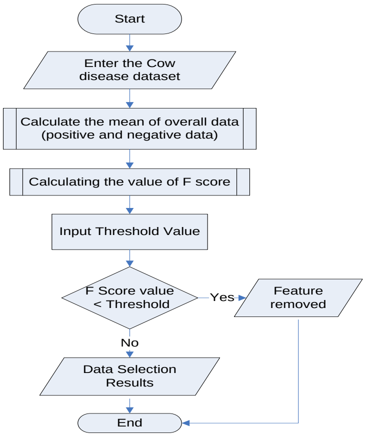

Start
Enter the Cow disease dataset
Calculate the mean of overall data (positive and negative data)
Calculating the value of F score
Input Threshold Value
F Score value < Threshold
Yes
Feature removed
No
Data Selection Results
End

Figure 1.F-Score Feature Selection Flowchart

2.1. K-Nearest Neighbour (KNN)

K-Nearest Neighbor is a supervised learning algorithm, where the results of new query instances are classified based on majority of categories on KNN. The most arising class will be the class resulting from the classification [18]. K-Nearest Neighbor facilitates the training data modeling process until it is needed to classify test data samples. The training data sample is described by a numerical item. When the sample test data is unknown, K-Nearest Neighbor will look for the k training sample closest to the test data sample. In this study, distance measurements will be carried out use Euclidean Distance. The Euclidean Distance formula is presented in equation (2) [12]. The following are steps for calculating KNN in this study are Determining the value of k, Calculating the distance between the test data and the training data, Sorting the distance from smallest to largest, Taking as much data as the nearest k and choosing the major value

$$d_{(xi,xj)} = \sqrt{\textstyle\sum_{r=1}^{n}(x_{ir} - x_{jr})^2} \qquad (2)$$

With :

d(xi,xj) = Euclidean Distance

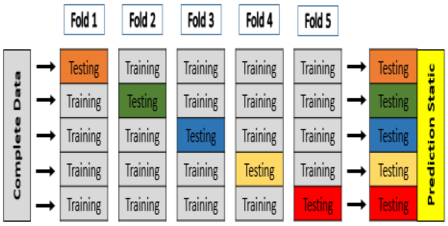
n = Data dimension

x_i = Test Data

x_j = Training Data

2.2. Cross Validation

Cross Validation is a method that can be used for system testing [19]. Cross Validation processes data by dividing the data used into two parts. The first part is used as training data and the second is used as testing data. In general, the k value test was carried out 5 times [20]. K-fold cross validation can be seen in Figure 2



Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	
Testing	Training	Training	Training	Training	Testing
Training	Testing	Training	Training	Training	Testing
Training	Training	Testing	Training	Training	Testing
Training	Training	Training	Testing	Training	Testing
Training	Training	Training	Training	Testing	Testing

Figure2. *Cross Validation* Model

3. System Design

The system design model is the stages of the data processing process to the output in the form of a disease diagnosis. The system design model is shown in Figure 3.

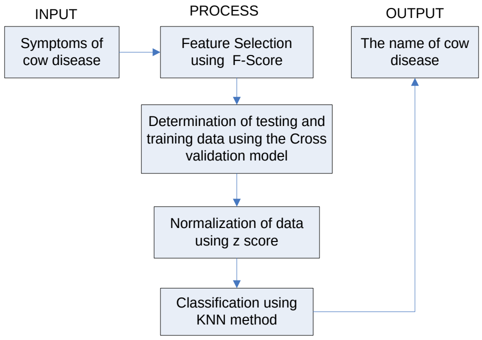
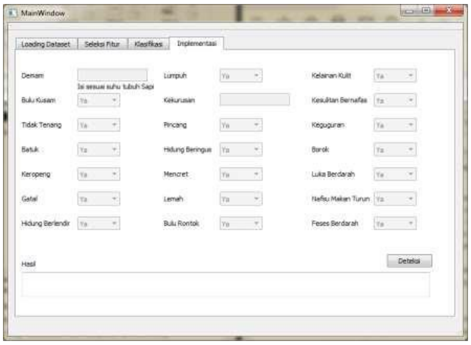

INPUT
PROCESS
OUTPUT
Symptoms of cow disease
Feature Selection using F-Score
The name of cow disease
Determination of testing and training data using the Cross validation model
Normalization of data using z score
Classification using KNN method

Figure 3. System Architecture

Base on Figure 3. The system architecture consists of three stages, namely input, in the form of symptoms of cow disease. The process consists of feature selection using the F-score method, then testing the system data with k-fold cross validation, by determining the training data and test data. After this process, normalization is carried out using the Z-Score method, and determining the output in the form of the name of the disease using KNN algorithm. The data needed in this study are cow disease data and cow disease symptoms which are used as training data and test data. Data on the name of cow disease are Bovine Ephemeral Fever (BFE), Cacingan, Scabies, Malignant Catarrhal Fever, Infectious Bovine Rhinotracheitis (IBR), Miasis, Septicemia epizootica. Data on symptoms or features of cow disease are Fever (F1), Dull Hair (F2), Uneasy (F3), Cough (F4), Ear Scab (F15), Itching (F6), Nasal Mucus (F7), Lame (F8), Thinness (F9), Limp (F10), Runny Nose (F11), Diarrhea (F12), Weakness (F13), Hair Loss (F14), Skin Disorders (F15), Difficulty Breathing (F16), Miscarriage (F17), Ulcers (F18), Blood Sores (F19), Decreased Appetite / Anorexia (F20), Bloody Stool (F21). The process of classification of cow disease is shown in Figure 4. Feature Selection Form. Based on the figure, there are 3 stages, namely loading a dataset of bovine symptoms and diseases, then selecting features, the classification process and implementation based on the best accuracy of the disease classification process.



Demam:
Isi sesuai suhu tubuh Sapi
Buku Kusam: Ya
Tidak Tenang: Ya
Batuk: Ya
Keropeng: Ya
Gatal: Ya
Hidung Berlendir: Ya
Lumpuh: Ya
Kekurusan:
Pincang: Ya
Hidung Beringus: Ya
Mencret: Ya
Lemah: Ya
Bulu Rontok: Ya
Kelainan Kulit: Ya
Kesulitan Bernafas: Ya
Keguguran: Ya
Borok: Ya
Luka Berdarah: Ya
Nafsu Makan Turun: Ya
Feses Berdarah: Ya
Hasil

Figure 4. Form for Classification Process

4. *Testing and Analysis*

Cow disease data used is from District Livestock Service Office Bangkalan. The data obtained total are 350 data, consisting of 21 features as symptoms of disease and 7 classes as name of the disease. Furthermore, the F-Score feature selection is carried out by calculating the relevant or irrelevant features, in this process it is obtained from the ranking of each feature. Data that has been selected for features is divided using K-fold Cross Validation into 5 parts consisting of 1 part of the test data and 4 other parts as training data and applied to each data segment. Furthermore, training data and test data are carried out by normalizing the Z-Score so that the range of values between features is not too far away. Furthermore, KNN calculation is carried out between the training data and the test data, so that the label is obtained from the test data. After completion, it is necessary to evaluate using the Confusion Matrix which results in accuracy, recall and precision. In this study, experiments were carried out without feature selection (Trial Scenario 1) and with feature selection (Trial Scenario 2).

4.1. *Trial Scenario 1*

This test was carried out using K-Nearest Neighbor with different k values, namely (3, 5, 7, 9, 11) and 21 features. The trial using without this feature selection is shown in Table 1.

Table 1. Trial Scenario Results 1

K	21 Features		
	Accuracy	*Recall*	*Precision*
3	14.512	27.576	27.714
5	13.528	27.316	27.429
7	14.319	25.818	25.429
9	12.925	25.798	25.714
11	14.674	28.571	28.571
Average	13.992	27.016	26.971

Table I. Shows that the highest accuracy without feature selection is at K-fold Cross Validation located at K-Nearest Neighbor (k = 11), which is 14.67393, a recall of 28.57142 and a precision of 26.97143.

4.2. *Trial Scenario 2*

This test uses K-Nearest Neighbor with a selection of f-score features and number of features 20,19,18,15,13,10,7,5 with different k values, namely 3, 5, 7, 9, 11. Table 2 using 20 features selection. Trials using this feature selection are shown in Table 2 toTable 10.

Table 2. Trial Scenario Results 2

	20 features		
K	Accuracy	Recall	Precision
3	81.710	75.654	74.286
5	85.862	85.504	84.571
7	87.814	87.863	86.571
9	89.887	87.186	86.000
11	90.354	88.004	87.429
Average	87.125	87.125	83.771

Table 2. Shows the results of 20 feature selection trial, with the highest accuracy at K = 11. The accuracy result for K-11 is 90,354, recall is 88,004 and precision is 87,429.

Table 3. Trial Scenario Results 3

	19 Features		
K	Accuracy	Recall	Precision
3	82.034	76.895	75.143
5	85.461	85.387	84.286
7	91.108	89.063	88.857
9	92.177	91.077	90.857
11	91.618	90.120	90.000
Average	88.480	86.508	85.829

Table 3. Shows the results of 19 feature selection trial, with the highest accuracy at K = 9. The accuracy result for K-9 is 91.177, recall is 91.077 and precision is 90.857.

Table 4. Trial Scenario Results 4

	18 Features		
K	Accuracy	*Recall*	*Precision*
3	91.714	0.917	0.917
5	90.612	90.977	90.857
7	92.112	90.989	91.143
9	91.867	90.035	90.000
11	91.430	89.413	89.143
Average	91.547	72.466	72.412

Table 4. Shows the results of 18 feature selection trial, with the highest accuracy at K = 7. The accuracy result for K-7 is 92.112, recall is 90.989 and precision is 91.143.

Table 5. Trial Scenario Results 5

	15 Features		
K	**Accuracy**	***Recall***	***Precision***
3	74.447	70.876	68.571
5	76.347	73.258	71.143
7	74.337	73.039	71.143
9	68.559	72.868	71.143
11	67.851	70.032	68.286
Average	72.308	72.014	70.057

Table 5. Shows the results of 15 features selection trial, with the highest accuracy at K = 3. The accuracy result for K-3 is 74.447, recall is 70.876 and precision is 68.571.

Table 6. Trial Scenario Results 6

	13 Features		
K	**Accuracy**	***Recall***	***Precision***
3	63.556	63.398	62.571
5	63.081	61.574	60.571
7	57.238	60.706	59.143
9	51.300	59.875	58.286
11	50.847	58.695	57.429
Avera	57.204	60.850	59.600

Table 6. Shows the results of 13 features selection trial, with the highest accuracy at K = 3. The accuracy result for K-3 is 63.556, recall is 63.398 and precision is 62.571.

Table 7. Trial Scenario Results 7

	10 Features		
K	**Accuracy**	***Recall***	***Precision***
3	51.906	59.396	58.000
5	64.483	60.348	58.571
7	45.181	58.385	45.181
9	38.945	57.103	55.143
11	37.938	56.173	54.000
Average	47.691	58.281	54.179

Table 7. Shows the results of 10 features selection trial, with the highest accuracy at K = 5. The accuracy result for K-5 is 64.483, recall is 60.348 and precision is 58.571

Table 8. Trial Scenario Results 8

	7 Features		
K	**Accuracy**	***Recall***	***Precision***
3	34.990	48.512	46.857
5	34.604	46.925	45.143
7	25.415	40.956	40.571
9	23.693	38.569	38.000
11	22.420	37.379	36.571
Average	28.225	42.468	41.429

Table 8. Shows the results of 7 features selection trial, with the highest accuracy at K = 3. The accuracy result for K-3 is 34.990, recall is 48.512 and precision is 46.857

Table 9. Trial Scenario Results 9

	5 Features		
K	**Accuracy**	***Recall***	***Precision***
3	32.758	42.455	40.000
5	35.450	43.407	40.571
7	32.271	42.836	40.000
9	24.896	40.448	37.429
11	23.602	39.258	36.000
Average	29.795	41.681	38.800

Table 9. Shows the results of 5 features selection trial, with the highest accuracy at K = 5. The accuracy result for K-5 is 35.450, recall is 43.407 and precision is 40.571

Table 10. Trial Scenario Results 10

Number of	**Experiment Results**		
Features (K)	**Accuracy**	***Recall***	***Precision***
20	90.354	88.004	87.429
19	92.177	91.077	90.857
18	92.112	90.989	91.143
15	74.447	70.876	68.571
13	63.556	63.398	62.571
10	64.483	60.348	58.571
7	34.990	48.512	46.857
5	35.450	43.407	40.571

Table 10. Show the results of the feature selection trial were 20,19,18,15,13,10,7,5, the highest accuracy was at K-Nearest Neighbor (k = 19) which was 92,177, recall was 91,077 and precision was 92,177.

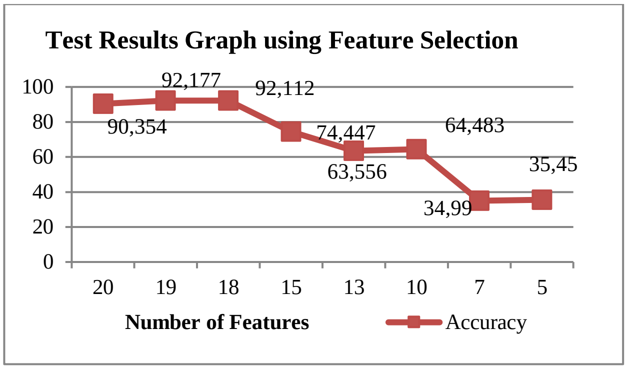


Figure 7. Graph of Feature selection accuracy results

Based on Figure 7. shows that the feature selection with K = 19 has the highest accuracy, namely 92.177, so that this rule is used as a classification of cow disease

Table 11. Comparison Trial Results

Scenario	Experiment Results		
	Accuracy	***Recall***	***Precision***
No Feature	14.674	28.571	28.571
Feature Selection	92.177	91.077	90.857

Table 11 is the comparison result of the test scenario 1 and test scenario 2 using the K-fold Cross Validation (k = 5). In the 1st trial scenario, the application of the K-Nearest Neighbor model without using feature selection resulted in an accuracy of 14.67393, a recall of 28.57142 and a precision of 28.57142. Whereas in the second trial of the K-Nearest Neighbor classification model using feature selection that produces the highest accuracy with the number of features = 19 and K-Nearest Neighbor (k = 5) which is equal to 92.17703, recall of 91.07744 and precision of 90.85714

5. Conclusion

Based on the research results, it can be concluded that the comparisons without feature selection and using feature selection are higher than feature selection. The selection of features with the highest accuracy is found in feature = 19, K-Nearest Neighbor (k = 5) and K-fold Cross Validation (k = 5), so the rules for this feature selection can be used to determine the diagnosis of cow disease.

References

[1] M. M. Jannan, H. Supriyono 2018 Android-Based Decision Support System for Cattle Disease Jurnal Emitor, vol. 18, no. 02, pp. 8-13.

[2] S. Harwati, 2014 Efforts to Provide Healthy and Quality Beef, Bangka Belitung Regency : Dinas Pertanian, Perkebunan dan Peternakan.

[3] L. G. Sri Astiti, 2010 Management of the Prevention and Control of Cow Disease, West Nusa Tenggara: Kementrian Pertanian,

[4] S. S. Emanuel, L. Schoonman and C. J. Daborn, 2010 Knowledge and Attitude Towards among Animal Health and Livestock Keepers in Arusha an Tanga, Tanzania," Tanzania Journal of Health Research, pp. 272-277.

[5] P.L.Venjakob, R.Staufenbiel, W.Heuwieser, S.Borchardt, 2021 Association between serum calcium dynamics around parturition and common postpartum diseases in dairy cows Journal of Dairy Science Volume 104, Issue 2, Pages 2243-2253

[6] J. Tong, S. Alelyani and H. Liu, "Feature Selection for Classification: A Review," Sch, 2013.

[7] Y. Lukito and A. R. Chrismanto, 2015 Comparison of Classification Methods for Indoor Positioning Systems Jurnal Teknik Informatika dan Sistem Informasi, vol. 1, no. 2, pp. 123-131.

[8] Fan Cunjia, Wang Yousheng, Bian Hang. 2015 An Improved KNN Text Classification Algorithm[J]. Foreign Electronic Measurement Technology, 12: 39-43.

[9] Hilal Arslan, Hasan Arslan, 2021 A new COVID-19 detection method from human genome sequences using CpG island features and KNN classifier, Engineering Science and Technology, an International Journal, https://doi.org/10.1016/j.jestch.2020.12.026

[10] Z. F. Ma, H. Tian, Z. C. Liu, Z. w. Zhang, 2020 A new incomplete pattern belief classification method with multiple estimations based on KNN, Applied Soft Computing, Volume 90, 106175, https://doi.org/10.1016/j.asoc.2020.106175

[11] Jin li Zhang, HailongYou, RenxuJia 2020 Reliability hazard characterization of wafer-level spatial metrology parameters based on LOF-KNN method Author links open overlay". Microelectronics Reliability, Volume 107, https://doi.org/10.1016/j.microrel.2020.113599

[12] Z. Chen, L. J. Zhou, X. Da Li, J. N. Zhang, W. J. Huo, 2020 The Lao Text Classification Method Based on KNN" Procedia Computer Science 166 523–528. DOI: 10.1016/j.procs.2020.02.05

[13] D. A. Nasution, H. H. Khotimah and N. Camidah, 2019 Comparison of Normalized Data for Wine Classification using the K-NN Algorithm," CESS (Journal of Computer Engineering System and Science), vol. 4, no. 1, pp. 78-82.

[14] D. Valentina and R. C. Wihandika, 2019 Toddler Fingerprint Recognition Using Zone Based Linear Binary Pattern and Extreme Learning Machine Method," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 3, no. 2, pp. 1851-1859.

[15] E. S. Wahyuni 2016 Application of the Feature Selection method to improve the results of Breast Cancer Diagnosis," Jurnal simetris, vol. 7, no. 1, pp. 284-294

[16] C. Saranya and G. Manikandan, 2013 A Study on Normalization Techniques for Privacy Preserving Data Mining," International Journal of Engineering and Technology (IJET), vol. 5, no. 3, pp. 2701-2704.

[17] D. L. Al Shalabi and D. Z. Shaaban, 2006 Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix," proceedings of the International Conference on Dependability of Computer System, pp. 207-214,

[18] D. Kusnianingtyas, B. A. Rahardian, D. P. Mahardika, A. Kartika and D. Angraeni K., 2017 Decision Support System for Beef Cattle Disease Diagnosis Using K-Nearest Neighbor (K-NN)," Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), vol. 4, no. 2, pp. 122-126.

[19] H. Leidiyana, 2013 Application of the K-Nearest Neighbor Algorithm for determining the risk of motorized vehicle ownership credit," Jurnal Penelitian Ilmu Komputer, System Embedded & Logic, vol. 1, no. 1, pp. 65-76.

[20] J. Y. Sari, R. A. Saputra 2017 Finger Vein Introduction Using Local Line Binary Pattern and Learning Vector Quantization, "ULTIMA Computing , vol. IX, no. 2, pp. 52-57.