Image Classification of Tourist Attractions with K-Nearest Neighbor, Logistic Regression, Random Forest, and Support Vector Machine

Herry Sujaini

Department of Informatics, University of Tanjungpura, Jl. Prof.Hadari Nawawi, Pontianak, 78124, Indonesia
E-mail: hs@untan.ac.id

***Abstract*— K-Nearest Neighbor (KNN), Logistic regression (LR), Random Forest (RF), and Support Vector Machine (SVM) are four methods of identification. The methods are widely used in various research in data mining, especially classifications in recent years. We have used the four classification methods in the study to classify images of five natural attractions, namely Danau Toba (North Sumatra), Nusa Penida (Bali), Raja Ampat (West Papua), Tanah Lot (Bali), and Wakatobi (Southeast Sulawesi). Our research results have concluded that the Logistic Regression method's performance has the best performance in classifying natural images as done in this research. The LR method can classify images that other methods such as kNN, SVM, and RF cannot be correctly classified. However, SVM also shows good performance by only making one error in the classification results; it can even be corrected using the Linear Kernel. In general, it is shown that the LR method has the highest precision value of 100%, followed by the method of kNN and SVM with a precision of 91.9% and RF with a precision of 81.9%. Variations of the variables used in the experiment also determine each method's precision. Chebyshev Metric has the highest precision value in the kNN method, and Ridge Regularization has the highest precision value in the LR method. The number of best on the RF method is 11, and Linear Kernel is the Kernel that gets the best precision value on the SVM method.**

***Keywords*— image classification; K-Nearest Neighbor; logistic regression; random forest; support vector machine.**

I. INTRODUCTION

Image classification is the process of grouping all pixels in an image. These pixels are grouped so that they can be interpreted as specific properties. Pixels are determined to a specific class if they meet certain rules to suit the class. The status of a class is known or unknown. If class data can be accessed, then class data is recognized or unknown. The image classification technique is divided into two categories, namely parametric and non-parametric. Parametric techniques require distribution assumptions from the data used where the distribution of data needed is normal.

In contrast, non-parametric techniques do not require distribution assumptions, so the distribution of data is free. We can also categorize image classification techniques as supervised and unsupervised, or hard classifiers and soft classifiers. Depending on prior knowledge about the class, the technique is divided into two groups; classification techniques are monitored and not monitored. Cluster analysis is a form of learning pattern related to unsupervised learning, where the number of class patterns is not known beforehand. The clustering process divides data sets by grouping all pixels in the feature space into a cluster naturally.

The supervised method requires a training set, but the training set for each of these classes has not been recognized. One reason is the difficulty in determining the number of classes that are needed in the image, which reduces the challenge of finding which locations can be considered the most representative. This phenomenon encourages researchers in the field of pattern recognition to continue to produce algorithms capable of automatically pushing these groups of numbers [1].

Machine-based algorithm learning with non-parametric methods has received much attention from applications based on digital image processing in recent years. In this timeframe, the use of Random Forest and Support Vector Machine classification algorithms increased significantly. Articles that use MLC and ANN have fluctuated throughout the year but generally remain stable. In the last few years (2014, 2015, and 2017), several studies have used kNN. SVM and RF are not sensitive to noise or overtraining, which shows their ability to deal with unbalanced data [2].

Research on identification in an image has long been developed by distinguishing the texture of the image. Image texture can be characterized by density, regularity, uniformity, and roughness [3]. Because computers cannot recognize tastes like human vision, texture analysis is used to determine a digital image pattern. Texture analysis will

produce values from the characteristics or characteristics of the texture, which can then be processed by the computer for the classification process [4].

Feature information for each image is expressed as a vector containing feature elements, including contrast, energy, correlation, and homogeneity. The feature elements' results are used to classify the closest distance between the extracted image of the test leaf and the training leaves' extraction results. One classification method that is widely used is K-Nearest Neighbor [5]. In recent years, Logistic regression (LR), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) and Random Forest (RF) methods have been widely used in various researches in the field of data mining, especially classification [6]-[10].

Indonesia is a beautiful country, rich in diverse natural, tribal, and linguistic resources, and is a complete tourist destination that is a pity to miss. Indonesia is an archipelago that has many tourist attractions that showcase the beauty of natural sightings. If we talk about Indonesia, Bali has denied the same thing in the world of eyes. Nevertheless, besides Bali, it turns out that Indonesia also has a popular tour that is worldwide. The research we report in this article focuses on the image classification of five natural tourism objects, namely Danau Toba (North Sumatra), Nusa Penida (Bali), Raja Ampat (West Papua), Tanah Lot (Bali), and Wakatobi (Southeast Sulawesi). The purpose of this study was to compare four popular classification methods, namely kNN, LR, RF, and SVM, to the images of five significant natural tourism objects in Indonesia.

II. Material and Method

Figure 1 presents the steps of the research. The first step is collecting data in the form of an image file, and the data is then extracted from its features. The feature extraction results are then used interchangeably for four classifiers, namely kNN, LR, RF, and SVM. Each classification result is then tested and assessed so that it can be concluded which classifiers are the most optimal for the classification of tourist attractions in this research.

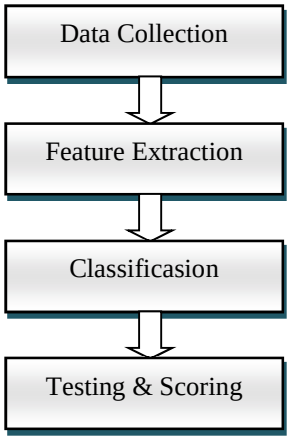

Fig. 1 Steps of the research

A. Data Collection

We used data for training samples collected based on manual interpretations of tourist attractions image data from various internet sources. This study's training sample data is limited to 5 famous tourist attractions in Indonesia: Danau Toba, Nusa Penida, Raja Ampat, Tanah Lot, and Wakatobi. Each of these locations is represented by seven images whose partial samples can be seen in Figure 2.

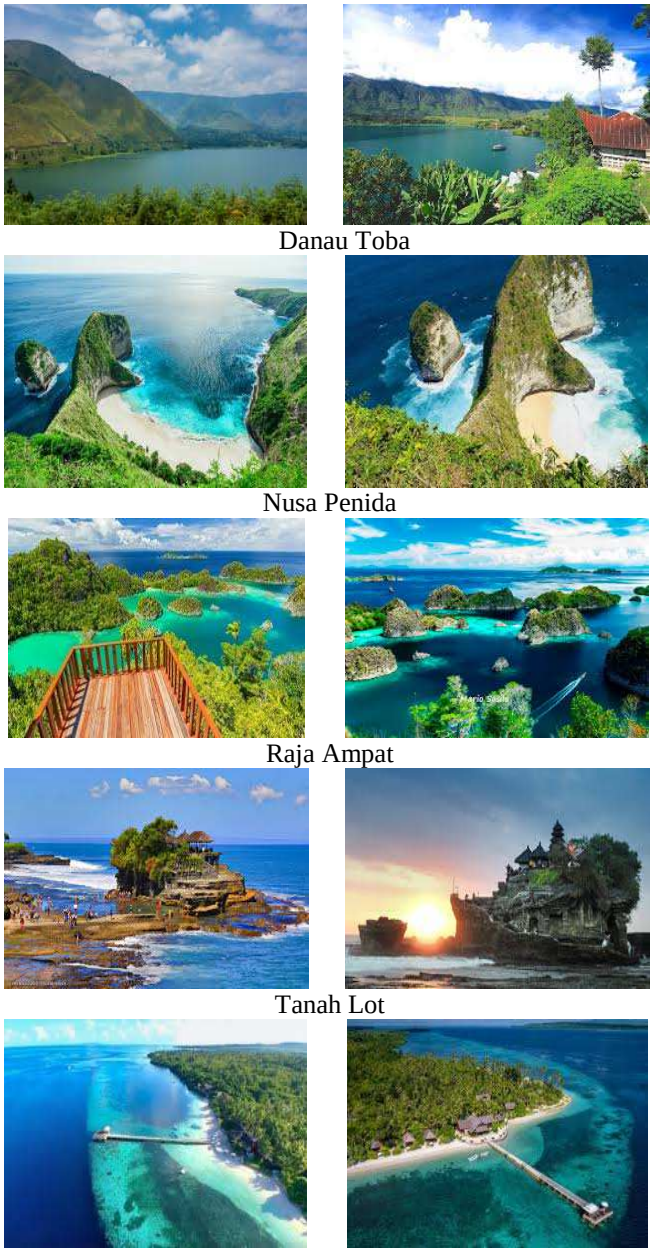

Fig. 2 Samples of training data images

As an illustration of the data used in this research, the average size of the image file used is 10.9 Kbyte with the smallest file 6.7 Kbyte, and the largest is 18.7 Kbyte while the pixel size of the average image file used as data in this research was 50,367 pixels with the smallest size of 50,232 pixels and the largest 50,615 pixels.

B. Feature Extraction

Each prepared image is calculated with its feature vector with the Squeezenet deep learning model [11]. This process is incorporated into the image embedding process, which in addition to producing features (Table 1), the output of this process produces data categories, image names, image sizes, and file sizes of each image (Table 2).

TABLE I
EXAMPLE OF FEATURES FROM EMBEDDING PROCESS

Image Name	n0	n1	n2	n3
danautoba01	812.355	970.285	245.972	699.557
danautoba02	805.179	896.337	694.684	618.124
danautoba03	116.382	134.646	759.583	8.186
danautoba04	940.785	120.711	817.567	741.033
danautoba05	105.181	129.733	770.038	641.499
danautoba06	90.274	129.363	702.906	750.937
danautoba07	907.657	146.487	531.898	897.612
nusapenida01	161.247	134.799	27.434	309.427
nusapenida02	859.301	995.068	416.425	367.967
nusapenida03	122.785	145.041	603.355	528.384
nusapenida04	137.407	116.591	519.359	343.107
nusapenida05	121.742	115.029	520.254	456.517
nusapenida6	170.357	111.992	763.277	42.749
nusapenida7	114.031	109.668	511.231	701.407
rajaampat1	425.578	127.765	295.035	599.793
rajaampat2	35.662	535.801	-136.011	-0.313587
rajaampat3	827.373	15.702	271.101	312.206
rajaampat4	397.578	115.063	319.399	-0.118268

TABLE II
EXAMPLE OF DATA IMAGE FROM EMBEDDING PROCESS

Category	Image Name	Size	Width	Height
danau toba	danautoba01	6669	300	168
danau toba	danautoba02	12051	275	183
danau toba	danautoba03	9758	275	183
danau toba	danautoba04	8272	275	183
danau toba	danautoba05	7959	275	183
danau toba	danautoba06	11402	300	168
danau toba	danautoba07	7005	294	171
nusa penida	nusapenida01	11937	275	183
nusa penida	nusapenida02	10430	252	200
nusa penida	nusapenida03	13141	301	167
nusa penida	nusapenida04	10930	275	183
nusa penida	nusapenida05	12559	275	183
nusa penida	nusapenida6	10460	301	167
nusa penida	nusapenida7	8074	259	194
raja ampat	rajaampat1	18726	348	145
raja ampat	rajaampat2	8601	275	183
raja ampat	rajaampat3	17541	275	183
raja ampat	rajaampat4	13096	275	183
raja ampat	rajaampat5	14479	299	168
raja ampat	rajaampat6	11596	275	183
raja ampat	rajaampat7	11501	302	167
tanah lot	tanahlot1	8652	275	183
tanah lot	tanahlot2	16600	300	168

C. Classification

The classification methods used in this study are kNN, LR, RF, and SVM. The classification method is a method of grouping data with a label or target class so that it is categorized into guided learning. The goal of supervised learning is that the target label or data acts as a 'supervisor' who oversees the learning process in achieving a certain level of accuracy or precision.

The nearest K-neighbor (kNN) is an algorithm that performs to classify data based on data learning, taken from the nearest neighbor (nearest neighbor). With k is the number of closest neighbors. The nearest neighbor is collecting data with data when there are many dimensions, and this space is divided into sections that represent the criteria for learning data. Each learning data is represented as points in many spaces. The new data, which is further classified, is projected on many-dimensional spaces that have contained c points of learning data. The classification process is done by looking for the nearest neighbor c point.

We need to determine the number of k nearest neighbors used to classify new data to use the nearest neighbor algorithm. The number of k should be an odd number, for example, k = 1, 2, 3, and so on. Determining the value of k is considered based on the amount of data available and the size of the data's dimensions. The more data available, the k number chosen should be lower. However, the larger the dimensions of the data, the k number chosen should be higher.

Logistic regression is an approach to making predictive models as well as linear regression. The difference is in logistic regression, and researchers predict the dependent variable that has a dichotomy scale. The logistic regression coefficient can estimate the odds ratio for each independent variable in the model. The odds ratio is a measure of the probability increase for one category compared to another. We could find out how the increase in the dependent variable score is reviewed by individual predictors when the other predictors are constant through odd ratios.

In machine learning, we often hear about the Random Forest method used to solve problems. The Random Forest method is one method in the Decision Tree. A decision tree is a classification method that uses a tree structure, where each node represents an attribute, and its branches represent the value of the quality. In contrast, the leaves are used to describe the class.

This method is a prevalent method to use because the model results are easy to understand. This method is called a decision tree because the rules formed are like the shape of the tree. Trees are formed from the binary recursive sorting process in data groups so that the value of the response variables in each data group makes the sorting results more homogeneous.

Making predictions is one of the expertise and advantages of the Vector Support Machine. Being able to classify and regression in a case is the ability possessed by Support Vector Machine (SVM). Although SVM has a basic linear classification principle, SVM has now been developed by researchers so that SVM can also work on non-linear problems by adding kernel concepts to high-dimensional workspaces. Please note that something called a hyperplane will be searched in a high-dimensional space that can maximize the distance between several data classes.

Linear SVM is the separation of data according to its linear. The best separator can not only separate data but also margins. Furthermore, so SVM can separate data that is not only linearly fed, SVM will be modified. SVM is supposed to work better than the Neural Network. Both have been successfully used in pattern recognition. From ordinary people to scientists have applied this method in solving problems in daily life. It has been proven that SVM provides excellent work results in many implementations. SVM is very easy to explain simply for businesses looking for the best hyperplane that is useful for classifying two classes in

the input space. The classification problem can be solved by finding a line or hyperplane that separates the two groups

D. Testing and Scoring

The classification results of each method used in this study were tested with 5-fold cross-validation and 0.75 training data using random sampling techniques. The result is the value of AUC, CA, F1, precision, and recall for each method used, namely kNN, LR, RF, and SVM.

III. Results and Discussion

This research used "Orange" tools [12]. Orange is open-source software for processing Data Analytics / Data Mining. Workflow classification of tourist attractions by comparing kNN, LR, RF, and SVM classifiers can be seen in Figure 3.

In experiments using kNN, the number of neighbors = five was used, the metric used is Euclidean and Weight = Uniform. In the experiment using LR, we used regularization = Ridge(L2) and strength C=1. In the experiment using RV, the number of trees = 10 and growth control were used using settings that did not include subsets less than 5. While in experiments using SVM, RBF kernel, regression loss epsilon = 0.1, cost = 1, tolerance was used. numeric = 0.001 and iteration limit = 100.

Cross-Validation is one technique for evaluating/ validating the accuracy of a model built on a particular dataset. K-fold is one of the popular Cross-Validation methods by folding the data as much as K and repeating the experiment as much as K as well. Then, experimenting with using data already on the partitions will be repeated five times (K = 5). However, the Test partition's data position is different in each iteration. For example, the first iteration of the Test in the initial partition position, continue repeating the second partition in the second position test, and so on.

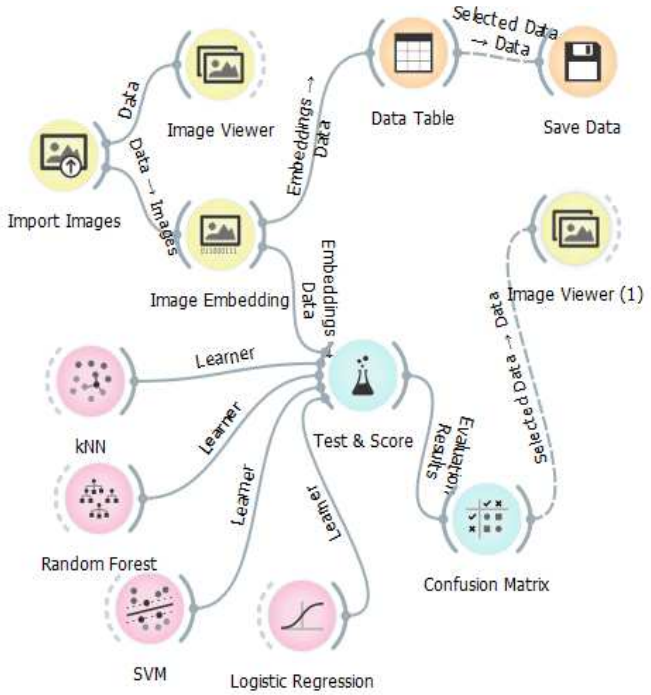

Fig. 3 The workflow of tourist attractions classification

Tables 3 and 4 present the results of each sampling technique.

TABLE III
Results of 5-Fold Cross Validation Experimental

Method	AUC	CA	F1	Precision	Recall
kNN	0.999	0.914	0.913	0.931	0.914
LR	1.000	1.000	1.000	1.000	1.000
RF	0.956	0.800	0.794	0.801	0.800
SVM	1.000	0.971	0.971	0.975	0.971

TABLE IV
Results of Random Sampling Experimental

Method	AUC	CA	F1	Precision	Recall
kNN	0.998	0.900	0.902	0.919	0.900
LR	1.000	1.000	1.000	1.000	1.000
RF	0.959	0.800	0.798	0.819	0.800
SVM	1.000	0.922	0.924	0.919	0.900

From Tables 3 and 4, the LR method has the best precision with an average precision of 100%, followed by the kNN and SVM method with the precision of 91.9% and RF with the precision of 81.9%. If analyzed from the image data, this is more due to the considerable amount of noise in samples taken from images on the internet. In contrast, RF is not sensitive to noise, which shows its ability to deal with unbalanced data.

Images that are not well predicted from the results of the proportion of predicted each method can be seen from Table 5 to Table 8. From the experiment results, it can be concluded that the LR method can predict all data correctly. The SVM method has been able to predict 24 images or 96% accurately. The kNN method has been able to predict 22 images or 88% correctly. Besides, the lowest in the RF method only able to predict 24 images or 68% accurately.

TABLE V
The Proportion of The Predicted with the KNN Method

		Predicted				
		danau toba	nusa penida	raja ampat	tanah lot	wakatobi
Actual	danau toba	77.8 %	0.0 %	0.0 %	0.0 %	0.0 %
	nusa penida	0.0 %	87.5 %	0.0 %	0.0 %	0.0 %
	raja ampat	11.1 %	12.5 %	100.0 %	0.0 %	0.0 %
	tanah lot	11.1 %	0.0 %	0.0 %	100.0 %	0.0 %
	wakatobi	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %

TABLE VI
The Proportion of The Predicted with The LR Method

		Predicted				
		danau toba	nusa penida	raja ampat	tanah lot	wakatobi
Actual	danau toba	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %
	nusa penida	0.0 %	100.0 %	0.0 %	0.0 %	0.0 %
	raja ampat	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %
	tanah lot	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %
	wakatobi	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %

TABLE VII
THE PROPORTION OF THE PREDICTED WITH THE RF METHOD

Actual	Predicted: danau toba	nusa penida	raja ampat	tanah lot	wakatobi
danau toba	57.1 %	0.0 %	28.6 %	11.1 %	0.0 %
nusa penida	0.0 %	85.7 %	14.3 %	0.0 %	0.0 %
raja ampat	28.6 %	14.3 %	42.9 %	11.1 %	0.0 %
tanah lot	0.0 %	0.0 %	0.0 %	77.8 %	0.0 %
wakatobi	14.3 %	0.0 %	14.3 %	0.0 %	100.0 %

From Table 5, there are three images that the kNN cannot predict correctly. One picture of Raja Ampat predicted as Danau Toba and one image predicted as Nusa Penida and one image of Tanah Lot predicted as Danau Toba.

TABLE VIII
THE PROPORTION OF THE PREDICTED WITH THE SVM METHOD

Actual	Predicted: danau toba	nusa penida	raja ampat	tanah lot	wakatobi
danau toba	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %
nusa penida	0.0 %	87.5 %	0.0 %	0.0 %	0.0 %
raja ampat	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %
tanah lot	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %
wakatobi	0.0 %	12.5 %	0.0 %	0.0 %	100.0 %

Figure 4 shows images that the kNN cannot predict correctly. From Table 7, eight images cannot be correctly predicted by the RF method, as follows:

- two Danau Toba images predicted as Raja Ampat and Tanah Lot;
- one Nusa Penida image predicted as Raja Ampat;
- three Raja Ampat images predicted as Danau Toba, Nusa Penida, and Tanah Lot; and
- two Wakatobi images that are predicted as Danau Toba and Raja Ampat.

Fig. 4 Pictures that are not correctly predicted by the kNN

Images that RF cannot predict correctly can be seen in Figure 5. From Table 8, there is one image that cannot be predicted correctly by the SVM method, which is a Wakatobi image that is predicted as Nusa Penida. Figure 6 gives images that the kNN cannot predict accurately. From the results of the four methods with the baseline feature that can be seen in Tables 3 and 4, we have tried to optimize the results of each method using different variables.

Fig. 5 The picture that is not correctly predicted by the RF

Fig. 6 The picture that is not correctly predicted by the SVM

To maximize the precision value of the kNN method, we experimented with comparing three other metrics (Manhattan, Chebyshev, and Mahalanobis) with Euclidean Metric as the baseline. The experimental results using the 5-fold cross-validation technique for these four metrics can be seen in Table 9. It turns out that using Chebyshev Metric, the kNN method can increase its accuracy to 0.956 compared to Euclidean as the baseline. Manhattan Metric has the same value in this case, while Mahalanobis has shallow precision values, so it is not recommended to classify natural appearance images.

TABLE IX
THE PRECISION OF KNN WITH METRIC VARIATIONS

Metric	Precision
Euclidean	0.931
Manhattan	0.931
Chebyshev	0.956
Mahalanobis	0.380

Although the LR method has obtained an absolute value in the previous experiment, we have tried to compare Ridge Regularization, which is used at baseline, with other Regularization, namely Lasso Regularization. The results show that Lasso Regularization produces lower precision values (0.950) compared to Ridge Regularization (Table X).

TABLE X
THE PRECISION OF LR WITH REGULARIZATION VARIATIONS

Regularization	Precision
Ridge	1.000
Lasso	0.950

In the baseline experiment using the RF method, we have used the number of trees = 10, which produces precision = 0.801 with the 5-fold cross-validation technique. We try to test the RF method by replacing the variable value with a value of 7 to 13. The results can be seen in Table XI. It is understood that the highest value for the RF method in the data used in this study is 0.897 utilizing the number of trees = 11.

TABLE XI
THE PRECISION OF KNN WITH NUMBER OF TREES VARIATIONS

Number of Trees	Precision
7	0.770
8	0.738
9	0.795
10	0.801
11	0.897
12	0.796
13	0.776

For the SVM method, we have experimented with different kernels, namely Linear, Polynomial, and Sigmoid, to be compared with the RBF Kernel as the baseline. The experimental results using the 5-fold cross-validation technique can be seen in Table XII. It turns out that using the Linear Kernel, the precision of the SVM method can achieve absolute values, while the Sigmoid Kernel results are the same as the RBF and the Polynomial Kernel produces precision values lower than RBF, which is 0.656.

TABLE XII
THE PRECISION OF SVM WITH KERNEL VARIATIONS

Kernel	Precision
RBF	0.975
Linear	1.000
Polynomial	0.656
Sigmoid	0.975

IV. CONCLUSION

From our previous explanation and analysis, it can be concluded that the Logistic Regression method's performance is very reliable to classify natural images as was done in this study. The LR method can classify image data that cannot be correctly classified by other methods such as kNN, SVM, and RF. However, SVM also shows good performance by only making one mistake on the results of its classification. It is generally taught that LR method has the best precision with an average precision of 100%, followed by the kNN and SVM method with a precision of 91.9% and RF with a precision of 81.9%.

Variations of the variables used in the experiment also determine each method's precision. Chebyshev Metric has the highest precision value in the kNN method, and Ridge Regularization has the highest precision value in the LR method. The number of best on the RF method is 11, and Linear Kernel is the Kernel that gets the best precision value on the SVM method.

REFERENCES

[1] Simpson, T. Mcintire, and M. Sienko, "An improved hybrid clustering algorithm for natural scenes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 2, pp. 1016–1032, 2000.

[2] L. Breiman, Random forests. Machine Learning, 45, 5–32. 2001.

[3] K. Anwar, A. Harjoko, and S. Suharto, "A New Method for Measuring Texture Regularity based on the Intensity of the Pixels in Grayscale Images," *International Journal of Computer Applications*, vol. 137, no. 7, pp. 1–5, 2016.

[4] Y. Wicaksono, R.S. Wahono, and V. Suhartono, "Color and Texture Feature Extraction Using Gabor Filter – Local Binary Patterns for Image Segmentation with Fuzzy C-Means", *Journal of Intelligent Systems*, vol. 1, no. 1, pp 15-21, 2015.

[5] O. R. Indriani, E. J. Kusuma, C. A. Sari, E. H. Rachmawanto, and D. R. I. M. Setiadi, "Tomatoes classification using K-NN based on GLCM and HSV color space," *2017 International Conference on Innovative and Creative Information Technology (ICITech)*, 2017.

[6] E. H. Abdelfattah, "Using the Logistic Regression to Predict Saudi's Kidney Transplant Rejection Patients," Biometrics & Biostatistics International Journal, vol. 5, no. 2, 2017.

[7] L. A. Ahmed, "Using logistic regression in determining the effective variables in traffic accidents," Applied Mathematical Sciences, vol. 11, pp. 2047–2058, 2017.

[8] J. Panyavaraporn and P. Horkaew, "Classification of Alzheimer's Disease in PET Scans using MFCC and SVM," International Journal on Advanced Science, Engineering and Information Technology, vol. 8, no. 5, p. 1829, 2018.

[9] R. V. K. Reddy and U. R. Babu, "Efficient Handwritten Digit Classification using User-defined Classification Algorithm," International Journal on Advanced Science, Engineering and Information Technology, vol. 8, no. 3, p. 970, 2018.

[10] H. W. Nugroho, T. B. Adji, and N. A. Setiawan, "Random Forest Weighting based Feature Selection for C4.5 Algorithm on Wart Treatment Selection Method," International Journal on Advanced Science, Engineering and Information Technology, vol. 8, no. 5, p. 1858, 2018.

[11] F.N. Iandola, M.W. Moskewicz, K.. Ashraf, S. Han, W.J. Dally, and K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size. CoRR, abs/1602.07360., 2016.

[12] J. Demsar, T. Curk, A. Erjavec, C. Gorup, T. Hocevar, M. Milutinovic, M. Mozina, M. Polajnar. M. Toplak, A. Staric, M. Stajdohar, L. Umek, L. Zagar, J. Zbontar, M., Zitnik, and B. Zupan. Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research 14(Aug), pp. 2349−2353. 2013.