khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Enhanced Image Classification by Eliminating Outliers with the Combination of Feature Selection and K-means Techniques

Nina Sevani[1*], Lukas Cuvianto[1], Jessica Octaviany[1], Albert Salomo[1]
[1]Informatics Department
Krida Wacana Christian University
Jakarta, Indonesia
*nina.sevani@ukrida.ac.id

Abstract-Accurate image classification will yield valuable information to support decision-making. Support Vector Machine (SVM) is a widely used technique to achieve high classification accuracy. However, data outliers can reduce the SVM's accuracy. To resolve this problem, the K-Means clustering method is used to eliminate the outliers by checking the proximity between data and clustering the data. Nevertheless, one of the challenges of using K-Means is the sensitivity of the initial centroid selection which is done randomly. Therefore, this study combines the use of K-Means, feature extraction with VGG-16 deep learning architecture, and feature selection using the Chi^2 technique to get better classification accuracy. The combination of these methods is empirically proven to increase the accuracy of three image dataset about 20%. The results demonstrate that using these methods in conjunction can also reduce the amount of time needed for image classification. Nevertheless, label information is not taken into consideration in this study. Therefore, in the future, this research can still be developed by applying other standards and adding information labels in the feature selection process.

Keywords: K-Means Clustering, Feature Selection Chi^2, Feature Extraction VGG-16

Article info: *submitted April 25, 2023, revised November 30, 2023, accepted December 3, 2023*

1. Introduction

Data can take many forms in the modern digital world, including text, statistics, photos, and videos. Image processing, which is one of the many methods and techniques used to process different kinds of data, involves handling image data. Image processing is used to produce a better image or to obtain the information that is contained in the image [1]. Therefore, image processing such as identifying or classifying images is very important in various fields such as agriculture, health, education, and various other fields. Given that important information is generated from an image, identification can help decision-making, planning, and interpretation [2], [3].

One of the image identification methods that give better results because it can predict high-dimensional datasets is the Support Vector Machine (SVM) [4] which provides informations such as the composition of texture on the surface structure, changes of the intensity, or brightness. Gray level co-occurence matrix (GLCM). A previous study showed that the SVM method gives better results in the case of skin cancer detection than the KNN and Random Forest methods [5]. Color, texture, and complicated sizes are just a few examples of the factors that might slow down and complicate the SVM recognition process when applied to image data. For this reason, superfluous features are eliminated from image identification during using the feature extraction and feature selection technique. Feature extraction is used to change the initial feature into a new feature form, while saving only informative features. Feature selection works by evaluating each feature and then removing superfluous features for determining label data. Removing superfluous features can shorten training time and reduce model complexity while still producing good accuracy [6]which will promote the development and application of precision medicine. Considering the natural order of genes, a new classification method that combines fused lasso and elastic net as regularization for linear support vector machine (SVM) [7].

Nevertheless, in the process of data classification, removing superfluous features does not equate to removing outlier data [8]-[12]. This causes the accuracy still not optimal. Outliers are the detection of anomalies/ abnormalities in the dataset. Removing the outliers is a very important step in the pre-processing phase to get

clean, noise-free, and consistent data. The same problem was also found in the study of patient classification to detect outliers from SVM and optimization of Density-Based penalty with SVM [8], [9].

One way to handle outliers can be done by removing them by utilizing the proximity of the data in one dataset before carrying out the classification process. Assuming that adjacent data are similar and data that are "separated" from other data are outliers. Therefore, clustering methods can be used to see the proximity of the data. Clustering will form data groups, where the data separated from any group can be labeled as outliers and the data will then be removed from the dataset. Of the various clustering methods, K-Means clustering is a fairly reliable clustering method. Previous research related to outlier analysis on K-Means using the Local Outlier Factor (LOF) method, showed that K-Means clustering has a lower minimum error value combined with outlier analysis. This can happen, due to the similarity of entities that are included in the number of clusters using the K-Means method with better outlier analysis [13]. So, it can be concluded that K-Means with outlier analysis will give more optimal results and also can resolve the weakness of SVM. K-Means Clustering method is used to identify images that have the same characteristics. K-Means clustering works by determining the center point of each cluster, which is called the centroid. Behind the sophistication of K-Means clustering, there are several weaknesses, one of them is the sensitivity in determining the initial centroid because the selection is done randomly. If the initial selection of centroid is wrong, it can affect the process of method performance and the results from partition data [14]. Therefore, K-Means clustering needs to be optimized, so that the accuracy performance can be better. Research on optimization of K-Means clustering has already been done, where K-Means was carried out to optimize PSO. The results of this study indicate that K-Means techniques combined with PSO give an accuracy of 47.33% for the identification of plant leaf images [15]. Although this result is still considered not good enough, considering the determination of centroid position on K-Means is a problem that still needs optimization.

The development of deep learning architecture, start to be widely used for image processing. It can be used to overcome the problem of sensitivity to centroid selection and also outliers detection. Among the various deep learning, VGG-16 is one of the CNN (Convolutional Neural Network) architectures which has 16 layers and is widely used for image processing [16]. Several studies have revealed that this method allows getting high accuracy even by using a small number of samples. Several examples of the use of VGG-16 in several research showed an accuracy that resulted in more than 90% and even reached 100%, such as the detection of eggplant disease with an accuracy of 99.4% [17], CT Scan images of the human brain with 100% accuracy [18], and classification of salak fruit quality with an accuracy of 95.83%. This happens because this architecture is the first transfer learning architecture that has succeeded in classifying ImageNet images with high accuracy [19].

The description makes it clear that the main problem that usually occurs during the image classification process is the presence of data outliers that could potentially obstruct the classification process and reduce its accuracy. The existence of these outliers can be overcome by grouping or clustering, which is using the K-Means method. However, this also often creates new problems in the sensitivity centroid determination process. Feature extraction using deep learning architecture provide the possibility to overcome this sensitivity problem. To improve accuracy in image classification, feature selection techniques can also be applied to filter only relevant features that will be used by machine learning models. This study looks at those issues and attempts to classify image by combining strategies for feature extraction utilizing deep learning, feature selection, and clustering. The deep learning architecture used is VGG-16 as a feature extractor to improve the performance of the K-Means clustering method.

2. Methods

This section will describe the procedures in implementing the recommended strategy into practice in order to address the issues that have been discussed. The procedures consist of four primary stages: feature extraction, feature grouping, feature selection, and model development.

a. Workflow of the Procedures

Figure 1 is a flowchart of the four stages of the procedures used in this study. The method begins with data collection and proceeds to feature extraction using VGG16. The result of the feature extraction will be used for feature grouping using K-Means to eliminate the outliers. The next stage is feature selection using Chi2 technique to select only the significant features to be used for training the Support Vector Machine (SVM) as the classification model. The accuracy of the model then will be measured using confusion matrix.

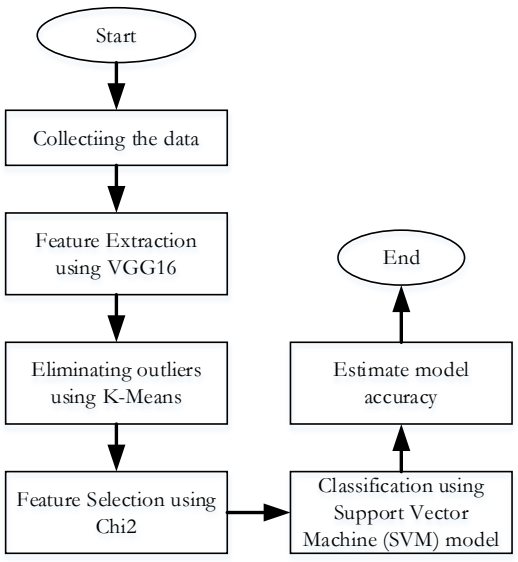


Figure 1. Flowchart

b. Data Collection

This study uses three datasets, namely: amazon, dslr, and webcam [20]. The dimensions of each of these datasets will be 150 x 150 x 3. The total data used are 958 amazon data, 157 dslr data, and 295 webcam data. We split the dataset to 80% for training and 20% for testing.

c. Feature Extraction using VGG16

Feature extraction is a data preprocessing stage with the goal is to reduce the dimensions of the dataset. In this study, the VGG16 as one of the CNN-based architectural models will be used for feature extraction. Feature extractors work by extracting several features without losing important and relevant information. Usually, image datasets are preprocessed using this feature extraction technique to produce feature map. Feature extraction has the benefit of allowing for feature reduction without sacrificing crucial data.

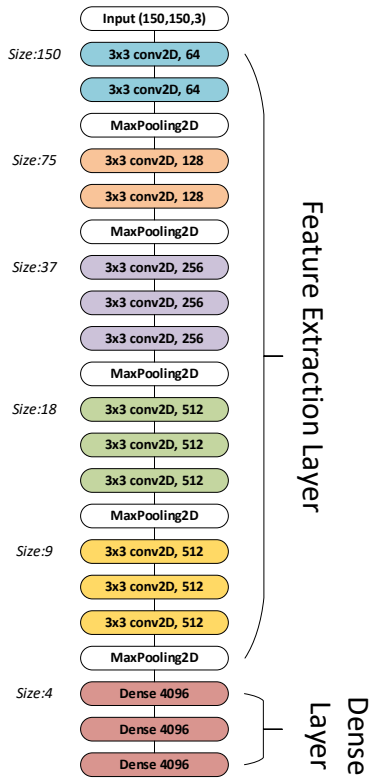


Figure 2. VGG-16 Architecture

VGG16 architecture used in this research has 16 convolution layers as shown in Figure 2 and has been trained with a 1000 class image dataset (ImageNet). Layers in VGG16 consist of an input layer, Convolutional Neural Network (CNN), max-pooling layer, dense layer (fully connected layer), and SoftMax output layer[19].

VGG-16 is merely a feature extractor at this point. Since a classification layer is not required for this research, the parameter include_top = False and input shape = (150, 150, 3) are employed.

d. K-Means Clustering

Clustering is a method for dividing data into a group or cluster according to the maximum similarity of characteristics [21]. K-Means Clustering is a method for grouping unsupervised learning data by dividing the data into several group [21]. Figure 3 represent the stages along with the flowchart of the algorithm of K-Means Clustering [22], [23]:

1) Determine the number of k clusters to be formed randomly.
2) Determine the initial centroid (cluster center point) randomly from the available objects for the number of k clusters. The formula used to calculate the centroid in iterations can be seen in Equation 1.

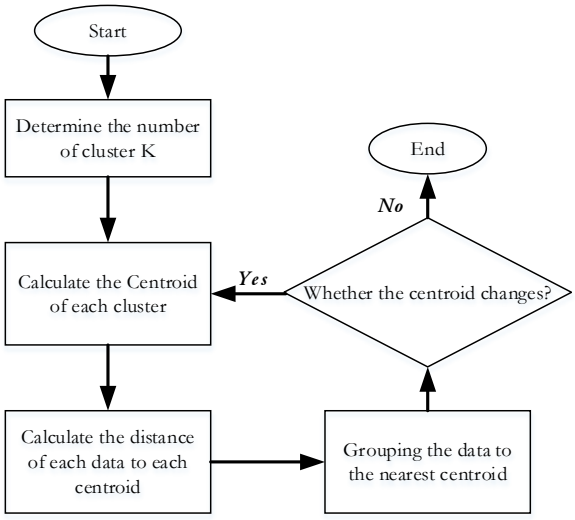
$$v = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

where:
v = centroid in the cluster
n = the number of objects that are members of the cluster
x_i = object
3) Calculate the distance of each data to each centroid using the Euclidean Distance formula as shown in Equation 2.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

where:
d = Euclidean Distance
x = object value
y = object value
n = total object
x_i = object to-
y_i = object to-
4) Clustering the data into the nearest centroid by considering the proximity of the distance between the data and the centroid. Then, iterate continuously until you find a new centroid with the calculations from the equation above.
5) If the position of the new centroid is changed, repeat point c until it reaches a converged condition.

K-Means is being used in this study to categorize the features and exclude outliers, or features that do not belong in any groups. There is no relationship between the feature and the other features, as indicated by these rejected features, often known as outliers.



Figures 3. K-Means Flowchart

e. Feature Selection

Feature Selection is a technique to reduce dimension (dimensionality reduction) by reducing the number of features from a dataset. The main purpose of feature selection is to maximize accuracy results by only use the significant features for the data classification. This technique is very useful especially for datasets with large dimensions such as image datasets [7].

This study uses the Chi^2 methodology, often known as chi-squared as a non-parametric test in a study, as the feature selection method. Equation 3 show how Chi^2 works by comparing between features.

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

where:
x^2 = chi-square value
O = observed value
E_i = expected value

By using the Chi^2, not only the accuracy level can be rise up, but it is also can reduce the time complexity for model training.

f. Support Vector Machine Classification

Support Vector Machine (SVM) is a supervised learning technique for non-linear problems such as classification and regression. It works by predicting data based on patterns from the training data process [24], [25]. SVM works by creates a dividing line to separate between classes which are called a hyperplane.

Two stages make up the classification: the training and testing phase. The results of the training phase is a model for the classification that build from the training dataset [24]. Meanwhile, the testing phase will be used to test the accuracy of the model using testing dataset. Table 1 presents the size from each dataset that used for the training and testing phase.

Table 1. Dataset Split

Dataset	Total	Train (80%)	Test (20%)
Amazon	791	632	159
Dslr	119	95	24
Webcam	242	193	49

g. Confusion Matrix

Confusion Matrix is one of the techniques for measuring the performance of both accuracy, precision-recall, and F-1 score in classification case [26]. The Confusion Matrix table consists of four matrices containing True Positive, True Negative, False Positive, and False Negative. These four matrices represent the results of the classification with the following interpretation [26].

1) True Positive (TP), when case A is predicted to be Positive and the value is True
2) True Negative (TN), when case A is predicted to be Negative and the value is True
3) False Positive (FP), when case A is predicted to be Positive and the value is False
4) False Negative (FN), when case A is predicted to be Negative and the value is False.

Even though the confusion matrix can calculate the level of accuracy, precision, recall, and F-1 score, but in this study only compare the accuracy between using the proposed method and using only SVM. The formula of accuracy can be seen in Equation 4 [26]:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (4)$$

where:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

3. Results

From the experiment using three datasets, it can be seen that the proposed method can increase the accuracy results in the classification. There is an increase of around 20% in each dataset from the value range of 65% to 80% to the range of 85% to 100%. Where on the Amazon dataset, before applying the proposed method is only achieved an accuracy of 70%, but after applying the proposed method, it could achieve an accuracy of 91%. Likewise, the DSLR and Webcam datasets also experienced a significant increase in accuracy after applying the proposed method. To attain the highest level of accuracy in the image classification, this study suggests combining multiple methodology. The combination of the methodology will focus on eliminating uninformed features that can interfere with the classification process. Discussion of the results of the proposed method can be seen from several aspects:

a. Comparison of total data (instances)

The use of K-means clustering in this study aims to eliminate outliers. Outliers are data that are not included in the existing cluster, so they are considered not to have the same pattern as other data in the dataset group. Outliers can reduce the classification results due to errors in determining the label during the classification process. K-means clustering method's implementation will decrease the amount of data required for the next step. In this step, the eliminating process is implemented in the data or instances and does not change or reduce the number of features. The results of this step can be seen in Figure 4.

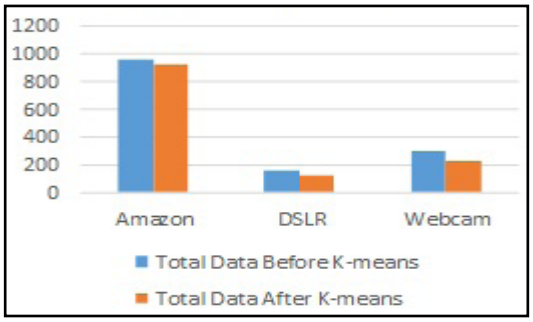


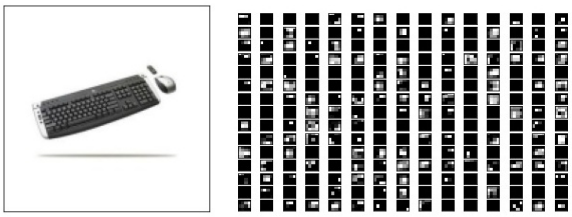
Figure 4. Comparison of Total Data Before and After Application of K-means Clustering on 3 datasets

The total data before K-means clustering are 958, 157, and 295 data for Amazon, DSLR, and Webcam respectively. The total data after applying the K-means clustering method are 921, 124, and 223 for Amazon, DSLR, and Webcam. These results show that the Amazon dataset only gets a 4% reduction from the original data. This value is relatively small compared with other datasets that can achieve about 20% data reduction. Because the quality of the images in the datasets varies greatly, there is a high value of data reduction in DSLR and Webcam. Although DSLR has high resolution and low noise, it uses natural lighting to take the images. Natural lighting also can affect the image quality. Webcam also has many outliers because the image has low resolution and contains much noise. Since the Amazon dataset was obtained from a photo studio, the final image was nearly identical in quality.

Even though the amount of data has decreased, it still has a substantial impact on accuracy level in the end. Since the reduction in data also results in decrease in the overall number of features.

b. Comparison of total feature

As explained in the methodology section above, this study using VGG16 as the feature extraction. The result of feature extraction with VGG16 for Amazon dataset as shown in Figure 5(a) for the original image and 5(b) after the feature extraction.

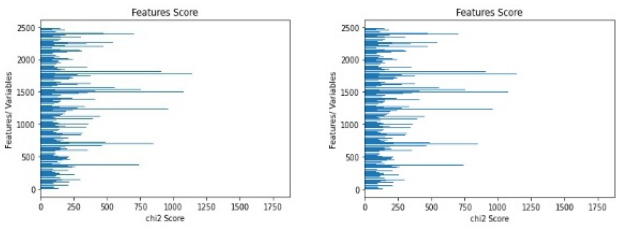


(a) Original Image **(b) After Feature Extraction**

Figure 5. Amazon Feature Data Extraction

The output of the convolutional layer extraction process is a feature map, also known as an activation map. A feature map is a snippet of data that stores important information for the next classification process. Feature extraction will transform the original data and keeps only the most significant features, the whole feature set is reduced as a result of the process. Each dataset's initial total feature count was 67,500. This procedure results in an overall feature of 8,000 for each dataset. Because there are uninformative features in the image, like the background, the feature reduction is quite important. This will affect how the clustering procedure chooses the appropriate centroid to increase accuracy results.

In addition to the use of feature extraction, the reduction of uninformative features is also carried out by feature selection using the Chi^2 technique. The feature selection technique used will select features according to the specified standard or criteria. The average Chi^2 value of every feature in the pertinent dataset served as the study's standard. Features that either meet the requirement or have a Chi^2 value greater than the dataset's average feature set will be the only ones chosen. On the other hand, features that do not reach the standard will not be used in the next process because these features are considered not to contain important information and can worsen the level of accuracy. The feature selection results for the three sample datasets can be seen in Figure 6. In each image, it can be seen that Figure 6(a) shows the initial features which amounted to 8000 features and Figure 6(b) shows the number of features that have been selected, which is in the range of 2000 features.



(a) Initial Features **(b) Selected Features**

Figure 6. Amazon Feature Data Selection

In general, the reduction in features after feature selection occurs is very significant, ranging from 8,000 to around 2,000 features. The total features after using feature selection are 2,481 for the Amazon dataset, 2,733 for the DSLR dataset, and 2,707 for the Webcam dataset. These results show that using the proposed method

can reduce the number of features for the classification process. Reduction occurred on the entire dataset, from the initial 67,500 to a decrease at each stage, until finally there were only 2,481 to 2,733 features. The feature reduction for each step is shown in Figure 7.

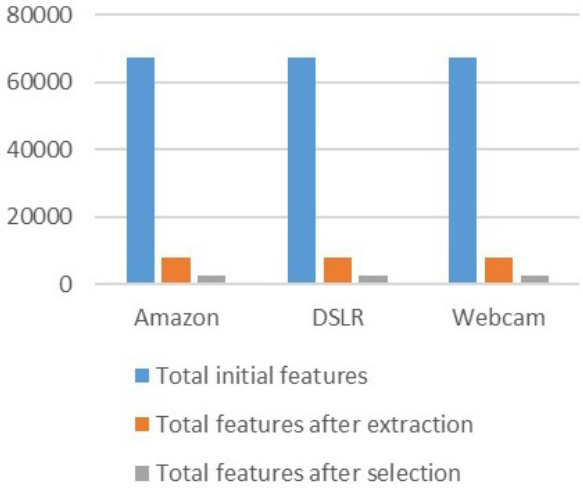


Figure 7. Comparison of Total Features Before and After Feature Selection in 3 Datasets

c. Performance measurement with Confusion Matrix

Based on Figure 8, there was a significant increase in accuracy results after applying the method proposed in this study. For example, the accuracy before applying the method to the Amazon dataset was only 70%, the DSLR dataset was 69%, and the Webcam dataset was 78%. While the accuracy after using the proposed method, the Amazon dataset reaches 91%, the DSLR dataset reaches 85%, and Webcam dataset reaches 97%. Without method means that the datasets directly used for the classification process, without reducing features and removing outliers. Using method means that the datasets carry out the process of feature extraction, feature selection, and clustering method to reduce the features and remove the outliers

Figure 8. Comparison of Accuracy Before and After Application of the Proposed Method on 3 Datasets

The range of 15% to 20% growth is indicated by these datasets. The smallest accuracy increase is 16% on DSLR and the largest is 21% on Amazon. When compared to other datasets, Amazon's accuracy rises more because of the dataset's more consistent image quality, which prevents excessive data loss like the K-means findings seen in Figure 4.

4. Conclusion

Based on the results of the research conducted, it can be concluded that the use of a combination of VGG16 (Feature Extractor), Chi^2 (Feature Selection), and K-Means Clustering methods can produce better accuracy than using only the Support Vector Machine (SVM) method in image classification. The use of a combination of these methods can solve the problem of early centroid detection on K-Means and data outliers on SVM. K-Means and VGG16 as the feature extractor were also proven to reduce the total number of features in each dataset. This has an impact on increasing accuracy as seen by the accuracy result of the three approaches combined, which is between 85% and 97%, compared to only the SVM method's 69% to 78%.

The limitation of this proposed method is in the process of figuring out which standard to apply when choosing features. The proposed method only uses the average as a standard in feature selection, while other standards can still be used. In addition, by only using the average standard, this method also does not pay attention to label information during feature selection.

In the future, this method can be further developed by trying to apply other standards to determine significant features. This method can also be developed by adding label information in selecting features to be used for the classification process.

References

[1] S. V Khedaskar, M. A. Rokade, B. R. Patil, and T. P. N, "A Survey of Image Processing and Identification Techniques," *Viva-Tech International Journal for Research and* Innovation, vol. 1, no. 1, pp. 1–10, 2018.

[2] Y. F. Kao and R. Venkatachalam, "Human and Machine Learning," *Computational Economics.*, vol. 57, no. 3, pp. 889–909, 2021, doi: 10.1007/s10614-018-9803-z.

[3] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets.*, vol. 31, no. 3, 2021, doi: 10.1007/s12525-021-00475-2.

[4] N. Neneng, K. Adi, and R. Isnanto, "Support Vector Machine Untuk Klasifikasi Citra Jenis Daging Berdasarkan Tekstur Menggunakan Ekstraksi Ciri Gray Level Co-Occurrence Matrices (GLCM)," *Jurnal Sistem Informasi Bisnis*, vol. 6, no. 1, p. 1, 2016, doi: 10.21456/vol6iss1pp1-10.

[5] A. Murugan, S. A. H. Nair, and K. P. S. Kumar, "Detection of Skin Cancer Using SVM, Random Forest and kNN Classifiers," *Journal of Medical Systems.*, vol. 43, no. 8, 2019, doi: 10.1007/s10916-019-1400-8.

[6] Y. Shen, C. Wu, C. Liu, Y. Wu, and N. Xiong, "Oriented Feature Selection SVM Applied to Cancer Prediction in Precision Medicine," *IEEE Access*, vol. 6, pp. 48510–48521, 2018, doi: 10.1109/ACCESS.2018.2868098.

[7] S. Jain and A. O. Salau, "An image feature selection approach for dimensionality reduction based on kNN and SVM for AkT proteins," *Cogent Engineering.*, vol. 6, no. 1, 2019, doi: 10.1080/23311916.2019.1599537.

[8] J. Mourão-Miranda, D. R. Hardoon, T. Hahn, A. F. Marquand, S. C. R. Williams, J. S-Taylor, *et al.*, "Patient classification as an outlier detection problem: An application of the One-Class Support Vector Machine," *Neuroimage*, vol. 58, no. 3, pp. 793–804, 2011, doi: 10.1016/j.neuroimage.2011.06.042.

[9] Y. Liu, J. Lian, M. R. Bartolacci, and Q. A. Zeng, "Density-based penalty parameter optimization on C-SVM," *The Scientific World Journal*, vol. 2014, 2014, doi: 10.1155/2014/851814.

[10] X. Y. Zhang, P. Yang, Y. M. Zhang, K. Huang, and C. L. Liu, "Combination of classification and clustering results with label propagation," *IEEE Signal Processing Letters.*, vol. 21, no. 5, pp. 610–614, 2014, doi: 10.1109/LSP.2014.2312005.

[11] T. Chakraborty, "EC3: Combining clustering and classification for ensemble learning," *Proceeding - IEEE Industrial Conference on Data Mining, ICDM*, vol. 2017-Novem, no. 9, pp. 781–786, 2017, doi: 10.1109/ICDM.2017.92.

[12] Y. Alapati and K. Sindhu, "Combining Clustering with Classification: A Technique to Improve Classification Accuracy," *International Journal of Computer Science Engineering.*, vol. 5, no. 06, pp. 336–338, 2016, [Online]. Available: https://en.wikipedia.org/wiki/Feature_selection.

[13] P. A. Ariawan, "Optimasi Pengelompokan Data Pada Metode K-means dengan Analisis Outlier," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 5, no. 2, pp. 88–95, 2019, doi: 10.25077/teknosi.v5i2.2019.88-95.

[14] M. S. Nasution and N. Fadillah, "Deteksi Kematangan Buah Tomat Berdasarkan Warna Buah dengan Menggunakan Metode YCbCr," *Jurnal Nasional Informatika dan Teknologi Jaringan*, vol. 3, no. 2, pp. 147–150, 2019, doi: 10.30743/infotekjar.v3i2.1059.

[15] F. Y. Bisilisin, Y. Herdiyeni, and B. P. Silalahi, "Optimasi K-Means Clustering Menggunakan Particle Swarm Optimization pada Sistem Identifikasi Tumbuhan Obat Berbasis Citra," *Jurnal Ilmu Komputer dan Agri-Informatika*, vol. 3, no. 1, p. 37, 2017, doi: 10.29244/jika.3.1.37-46.

[16] H. Shen, Y. Deng, W. Xu, and C. Zhao, "Rate maximization for downlink multiuser visible light communications," *IEEE Access*, vol. 4, pp. 6567–6573, 2016, doi: 10.1109/ACCESS.2016.2614598.

[17] A. Krishnaswamy Rangarajan and R. Purushothaman, "Disease Classification in Eggplant Using Pre-trained VGG16 and MSVM," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-020-59108-x.

[18] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour, "Unifying Representations and Large-Scale Whole-Body Motion Databases for Studying Human Motion," *IEEE Transaction on Robotics.*, vol. 32, no. 4, pp. 796–809, 2016, doi: 10.1109/TRO.2016.2572685.

[19] R. Rismiyati and A. Luthfiarta, "VGG16 Transfer Learning Architecture for Salak Fruit Quality Classification," *Telematika*, vol. 18, no. 1, p. 37, 2021, doi: 10.31315/telematika.v18i1.4025.

[20] H. Li, S. J. Pan, R. Wan, and A. C. Kot, "Heterogeneous transfer learning via deep matrix completion with adversarial kernel embedding," *33rd AAAI Conference on Artificial Intelligence (AAAI-19)*, pp. 8602–8609, 2019, doi: 10.1609/aaai.v33i01.33018602.

[21] S. Oktarian, S. Defit, and Sumijan, "Clustering Students' Interest Determination in School Selection Using the K-Means Clustering Algorithm Method," *Jurnal Informasi dan Teknologi.*, vol. 2, pp. 68–75, 2020, doi: 10.37034/jidt.v2i3.65.

[22] F. Nasari and C. J. M. Sianturi, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokkan Penyebaran Diare Di Kabupaten Langkat," *Cogito Smart Journal*, vol. 2, no. 2, pp. 108–119, 2016, doi: 10.31154/cogito.v2i2.19.108-119.

[23] G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 5, no. 1, pp. 17–24, 2019, doi: 10.25077/teknosi.v5i1.2019.17-24.

[24] A. Liani, "Analisis Perbandingan Kernel Algoritma Support Vector Machine dalam Mengklasifikasikan Skripsi Teknik Informatika berdasarkan Abstrak," *Journal of Information System*, vol. 5, no. 2, pp. 240–249, 2020, doi: 10.33633/joins.v5i2.3715.

[25] C. Chazar and B. Erawan, "Machine Learning Diagnosis Kanker Payudara Menggunakan Algoritma Support Vector Machine," *Jurnal Informatika dan Sistem Informasi*, vol. 12, no. 1, pp. 67–80, 2020, doi: 10.37424/informasi.v12i1.48.

[26] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Information Sciences.*, vol. 507, pp. 772–794, 2020, doi: 10.1016/j.ins.2019.06.064.