CAUCHY: Jurnal Matematika Murni dan Aplikasi
Volume 9(1) (2024), Pages 119-128
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Comparison between Statistical Approaches and Data Mining Algorithms for Outlier Detection

Annisa Putri Utami[1], Anwar Fitrianto[1,2*], Khairil Anwar Notodiputro[1]

[1] Department of Statistics and Data Science, Faculty of Mathematics and Natural Sciences, Institut Pertanian Bogor, Indonesia
[2]Institute of Engineering Mathematics, Universiti Malaysia Perlis, Malaysia

Email: anwarstat@gmail.com

ABSTRACT

Outliers are observation values that are very different from most observations. The presence of outliers in data can have a negative impact on research but can contain important information for other research. So, identifying outliers before conducting data analysis is a crucial thing to do. Outlier detection methods/techniques were first pioneered by researchers in statistics. However, due to rapid technological advances which have an impact on the ease of collecting extensive data, the development of outlier detection techniques is now handled mainly by researchers in the field of computer science (data mining) using computing facilities. This research aims to examine the results of simulation studies by comparing methods for identifying several outliers using statistical approaches and data mining algorithm approaches in various predetermined data scenarios. Based on the scenario carried out, the outlier detection method using a statistical approach is generally better than the outlier detection method using a data mining-based approach. Suggestions for further research are to improve the data mining method by focusing more on statistical analysis apart from focusing on data processing computing time so that the expected results of outlier detection are faster and more precise.

Keywords: distance-based methods; masking; outlier; outlier detection method; swamping

INTRODUCTION

Outliers is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism [1]. There is a rule of thumb that says that observation points that deviate more than three times the standard deviation from the mean of a normal distribution can be identified as outliers [2]. Outliers are often caused by human error, such as errors in data collection, recording, or entry [3].

The existence of outliers in data can have a negative impact on research but can contain important information for other research [4]. Estimating parameters on data that contains outliers will produce biased estimated values, causing the interpretation of analysis results to be inaccurate. Because many test statistics are sensitive to outliers, the ability to detect outliers is an important part of the initial stage of data analysis to produce

Submitted: December 31, 2023 Reviewed: January 19, 2024 Accepted: February 01, 2024
DOI: http://dx.doi.org/10.18860/ca.v9i1.25450

better statistical analysis. Since outliers affect data analysis, some studies delete outlier data before conducting data analysis. Nevertheless, outliers should be carefully handled, as the unjust removal of extreme values can increase the Type I error rate [5].

An example where an outlier has the potential to contain important information is in cases of indications of fraud in the misuse of credit cards in financial transaction data, which results in sudden changes in usage patterns. Another case is that in the production process, a defective product will have characteristics that are different from the characteristics of a standard product, this can signal that there may be damage to the machine. Outlier detection also finds applications in environmental monitoring, structural monitoring, intrusion detection, and fake news detection [6]. By identifying outlier data, researchers can estimate the cause of the emergence of these values; if it is caused by human error, such as in the process of recording or measuring observations, researchers can reconfirm the recording process or, if possible, improve the measurement tools.

The field of statistics is a pioneer in techniques for identifying outlier data. Statistical methods in outlier detection can be classified into two categories: parametric and non-parametric techniques [7]. Parametric techniques assume knowledge of the underlying distribution and estimate parameters from given data, for example, Gaussian-based methods and regression-based methods. Meanwhile, non-parametric techniques generally do not assume anything regarding the data, so the model structure is determined from the given data [8].

This research will focus on a method based on a multivariate normal distribution so that observation points that do not match the model are thought to be outlier observations. Assuming multivariate normality, an approach for more accurate outlier detection is to use Mahalanobis squared distance [9]. Detection of outliers with the Mahalanobis distance is carried out based on the observed distance to the cut-off value; in this case, the cut-off value is a value from the chi-square distribution with p degrees of freedom, where p is the number of variables from the multivariate data [10].

Although outlier detection methods or techniques were first pioneered by researchers in the field of statistics, and most of the existing statistical techniques are univariate [11]. However, as a result of rapid technological advances which have an impact on the easy accumulation of extensive data and the need for fast information from that data, the development of outlier detection techniques is now handled mainly by researchers from the field of computer science (data mining) using computing facilities [12].

One of the earliest studies on the development of algorithms for distance-based outlier detection in data mining was DB-outlier, which was introduced by Knorr [13]. Knorr defines a point o as an outlier if less than p points are within a distance d of the point o; this means that an outlier needs to have p or fewer points within its distance d. In this approach, researchers can set the percentage of outliers in the data (p) and distance (d); determining these values is done by trial and error, so it requires several iterations. Overcoming this weakness, Ramaswamy [14] proposed a new formulation for distance-based outliers, which is based on the distance of a point from its kth nearest neighbor, so it does not require researchers to determine the distance value (d). Ramaswamy's definition of outliers does not consider the information contained in the k neighborhoods of a point and thus cannot correctly differentiate between dense or sparse neighborhoods. Angiulli [15] made a slight modification to the definition of outliers, outliers are points that have the most significant weight value, and the weight is calculated from the sum of the distances between a point and its k nearest neighbors.

However, when it comes to real data, determining parameters is not an easy thing. Zhang [16] proposed a new definition of an outlier, called Local Distance-based Outlier Factor (LDOF). LDOF uses the relative distance of an object to its neighbors to measure the extent to which the object deviates from its environment. Based on the the comparison results, top-n LDOF works relatively stably at various parameter values because the method is less sensitive to parameter values.

Even though it starts with the basis of statistical reasoning, most research in the field of data mining focuses more on the computational efficiency (processing time) of algorithms compared to the quality of the statistics produced [17]. In 2018, Zimek and Filzmoser conducted a review of outlier detection methods from these two approaches to determine the combined point of view or connection between computer science and statistics in the development of outlier detection techniques carried out by researchers from the field of data mining. However, this research is only based on an intuitive perspective. Therefore, this research will focus on examining in more depth the outlier detection technique using a statistical approach and the outlier detection technique using a data mining approach by testing the accuracy in various data scenarios. Outliers in the regression model can be located in the response variable (Y), which is called a vertical outlier, or located in the response variable (X), which is called the leverage point [18].

The statistics-based outlier detection method that will be used is the Mahalanobis squared distance with the MCD (Minimum Covariance Determinant) estimator, while the data mining algorithm approach uses the LDOF method. This research aims to examine the results of a simulation study from a comparison of methods in identifying multiple outliers with statistical approaches and data mining algorithm approaches in various predetermined data scenarios.

METHODS

The simulation data consists of two independent variables (x_1 and x_2) and one dependent variable (y). Independent variable data is generated from a multivariate normal with a mean vector μ, variance matrix $\sigma^2 I$, and correlation ρ. In this study, the mean and variance are 0 and 1, with N=10000. The dependent variable data is generated using the following model:

$$y_i = \beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \varepsilon_i \quad (1)$$

Where:

y_i : The variable value i-th observation
x_{1i}, x_{2i} : Values of the 1st and 2nd independent variables in the i-th observation
$\beta_{0i}, \beta_{1i}, \beta_{2i}$: Model coefficients
ε_i : Error of the i-th observation generated from the distribution $N(0,1)$

Comparison of the performance of the two methods is measured through various simulation scenarios. Details of the simulation data generation scenario are presented in the following table:

Table 1. Scenarios for generating simulation data

Scenario	Based Method	
	Statistics	**Data Mining Algorithms**
Types of Outliers	*vertical outlier, laverage point*	*vertical outlier, laverage point*
Proportion of Outliers (α)	5%, 10%	5%, 10%
Correlation between variables x (ρ)	0.1, 0.8	0.1, 0.8
Outlier distance (δ)	3σ, 5σ	3σ, 5σ

From Table 1, it is known that the total data generated for each method is 16 data clusters, where each combination will be repeated 1000 times.

Analysis Stages

The stages carried out in this research are as follows:

Simulation Method

Data generation and analysis were carried out using R Software. The simulation procedures carried out in this research were:

[Simulation data generation stage]

1. Generate population data for cluster x of size $N \times p$ with $p = 2$ and $N = 10000$. The cluster population data is generated with a multivariate normal distribution $N(\mu, \Sigma)$, with $\mu = (10 \quad 20)$ and two varying matrix conditions as follows:
$$\Sigma_1 = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$
2. Generate data for the variable y with the following equation (1), the parameters for the population $(\beta_0, \beta_1, \beta_2)$ in this study are $\beta_0 = 0$ and $\beta_1 = \beta_2 = 1$ with $\varepsilon_i \sim N(0,1)$ where $i = 1,2, \dots, n$.
3. Take a sample of 100 data.

[Data contamination stage]

4. Create a contaminated data cluster, with the following scenario:
   Where α is the proportion of outliers (5% and 10%) from the large number of data samples, and δ is the distance of outliers (3 and 5).
   - Outliers on the y-axis (Vertical Outliers)
     (outlier on right side)
     a) Sort the y response data from largest to smallest
     b) Take the top data as much as $(\frac{1}{2}\alpha \times n)$ data
     c) Calculate the value of $\hat{\sigma}_y$
     d) Replace the data with data that has been contaminated using the equation $y_i^* = y_i + \delta\hat{\sigma}_y$

     (outlier on left side)
     e) Sort the y response data from smallest to largest
     f) Take the top data as much as $(\frac{1}{2}\alpha \times n)$ data
     g) Calculate the value of $\hat{\sigma}_y$
     h) Replace the data with data that has been contaminated using the equation $y_i^* = y_i + \delta\hat{\sigma}_y$
   - Outliers on the x-axis (Laverage point)
     (outlier on right side)
     a) Sort the values of the variable x_1 from largest to smallest
     b) Take the top data as much as $(\frac{1}{2}\alpha \times n)$ data
     c) Calculate the value of $\hat{\sigma}_{x_1}$
     d) Replace the data with data that has been contaminated using the equation $x_i^* = x_i + \delta\hat{\sigma}_{x_1}$

     (outlier on left side)
     e) Sort the values of the variable x_1 from smallest to largest
     f) Take the top data as much as $(\frac{1}{2}\alpha \times n)$ data
     g) Calculate the value of $\hat{\sigma}_{x_1}$

h) Replace the data with data that has been contaminated using the equation $x_i^* = x_i - \delta\hat{\sigma}_{x_1}$

[Outlier detection stage]

5. Detect outliers with 2 approaches, as follows:

i. First approach (Statistical Approach)

The MCD estimator is generated from the Fast-MCD algorithm with the following steps [19]:

a) Take a number h of different observations at random. From n observations, a new set $\binom{n}{h}$ will be generated. The optimal h value satisfies $(n + p + 1)/2$.

b) Define the first set as H_1. Based on the set H_1, calculate the mean vector and covariance matrix$(\bar{x}_1, S_1)$ using the following formula:

$$\bar{x} = \frac{1}{h}\sum_{i=1}^{h} x_i = \frac{x_1 + x_2 + \cdots + x_h}{h} \tag{2}$$

$$S = \frac{1}{h-1}\sum_{i=1}^{h} (x_i - \bar{x})(x_i - \bar{x})' \tag{3}$$

c) Calculate the mahalanobis distance using the formula:

$$d_1(i) = \sqrt{(x_1 - \bar{x}_1)' S_1^{-1} (x_i - \bar{x}_1)} \tag{4}$$

d) Sort $d_1(i)$ from smallest value to largest value.

e) Define a new subset with H_2, such that $\{d_1(i); i \in H_2\} \coloneqq \{(d_1)_{1:n}, (d_1)_{2:n}, \ldots, (d_1)_{h:n}\}$, where $(d_1)_{1:n} \leq (d_1)_{2:n} \leq \cdots \leq (d_1)_{h:n}$.

f) Calculate the mean vector and covariance matrix $(\bar{x}_2, S_2)$ from H_2.

g) Compare $\det(S_2)$ with $\det(S_1)$. If $\det(S_2) > \det(S_1)$ repeat the steps in points (a) to (f) until it is found that $\det(S_{m+1}) \leq \det(S_m)$. S_{MCD} covariance matrix with the smallest determinant.

h) Calculate the mahalanobis distance for each observation using the formula:

$$d_{MCD}(i) = \sqrt{(x_i - \bar{x}_{MCD})' {S_{MCD}}^{-1} (x_i - \bar{x}_{MCD})} \tag{5}$$

i) An *i*-th observation (x_i) is said to be an outlier if

$$d_{FastMCD}(i) > \chi^2_{p,(0.975)} \tag{6}$$

ii. Second approach (Data Mining Approach)

Top-n LDOF (Local Distance-based Outlier Factor) algorithm. LDOF works relatively stable at various parameter values k, this research uses the optimum value k.

Input: For the given dataset D, enter the value of k.

a) For each object p in D, take its k-nearest neighbors.

b) Calculate the LDOF for each object p.
Objects with LDOF > 1 are outliers.

c) Sorts N objects according to their LDOF values.

Output: first n objects with the highest LDOF value.

[Method evaluation stage]

6. Calculate the accuracy value of identifying outlier data.
   Outlier detection performance can be evaluated using the confusion matrix method, which is presented in the following table:

Table 3. *Confusion Matrix*

Actual	Prediction	
	Normal data	**Outlier data**
Normal data	*True Positives* (TP)	*False Negatives* (FN)
Outlier data	*False Positives* (FP)	*True Negatives* (TN)

With the formula:

$$\text{Ac}curacy\ (\%) = \frac{\text{TP} + \text{TN}}{\text{total observations}} \times 100\% \tag{7}$$

7. Repeat 1000 times from step 3 to step 6.
8. Calculate the average of the accuracy values of the two outlier detection methods.

$$\text{Average accuracy} = \frac{1}{r}\sum_{k=1}^{r}[\text{accuracy}_k] \tag{8}$$

where: r : Number of simulation repetitions

9. Calculate the masking level and swamping level, using the following formulas [20]. Masking and swamping are common problems related to outliers. Masking means some outliers are not identified, while swamping means some non-outliers are identified as outliers, and ae calculated as follows.

$$masking = \frac{\text{Number of outliers identified as inliers}}{\alpha \times n \times r} \tag{9}$$

$$swamping = \frac{\text{Number of inliers identified as outliers}}{[n - (\alpha \times n)] \times r} \tag{10}$$

10. Repeat step 1 to step 9 for all existing data combinations.

RESULTS AND DISCUSSION

In the case of outliers on the *x*-axis (Leverage point)

A. Against the proportion of outliers

The results of the first simulation process, which focused on the proportion of outliers in the data, obtained the following values:

Table 4. Simulation results focus on the proportion of outliers on the *x*-axis

	Proportion of Outliers			
	5%		10%	
	Statistics	Data Mining	Statistics	Data Mining
Accuracy	0.93831	0.84049	0.89842	0.81488
Masking	0	0	0	0.00004
Swamping	0.00006	0.00016	0.00011	0.00020

Based on simulations, it was found that, in general, the level of accuracy of statistics-based methods is better than data mining-based methods. The higher the proportion of outliers, the lower the method's accuracy in detecting outliers. The performance of statistics-based outlier detection methods produces smaller masking values than data mining-based methods, so data statistics-based methods

successfully detect outlier observations even though there are adjacent outlier observations. Similar to the swamping value, the value produced from statistics-based methods shows a slight possibility that normal data is suspected to be an outlier. In contrast, for data mining-based methods the possibility tends to be greater for data detection errors to occur.

B. Against outlier distance

The simulation process with a focus on the distance of the generated outliers produces the following values:

Table 5. Simulation results focus on the distance of outliers on the *x*-axis

	Outlier Distance			
	3σ		5σ	
	Statistics	Data Mining	Statistics	Data Mining
Accuracy	0.918343	0.829288	0.918335	0.826088
Masking	0	0.000035	0	0.000009
Swamping	0.000089	0.000181	0.000088	0.000187

Based on the table above, in general, the level of accuracy of statistics-based methods is better than that of data mining-based methods. For outlier distances of 3σ and 5σ, the level of accuracy obtained is not that different. Still, it can be said that the farther the outlier distance, the lower the method's accuracy in detecting outliers. The performance of statistics-based outlier detection methods produces smaller masking values than data mining-based methods. Therefore, it can be said that data statistics-based methods successfully detect outlier observations even though there are adjacent outlier observations. The swamping value resulting from statistics-based methods tends to be lower than data mining-based methods, indicating a slight possibility that there is average data that is suspected to be an outlier. In contrast, for data mining-based methods, the occurrence of data detection errors is still more remarkable.

C. Regarding the correlation with variable X

The simulation results focusing on the correlation with variable X are as follows:

Table 6. Simulation results focus on the correlation of X variables

	Correlation of Variable X			
	0.1		0.8	
	Statistics	Data Mining	Statistics	Data Mining
Accuracy	0.918318	0.834630	0.918420	0.820745
Masking	0	0.000025	0	0.000019
Swamping	0.000088	0.000176	0.000088	0.000192

Table 6 shows that the level of accuracy of statistics-based methods is much better than data mining-based methods. The more significant the correlation between the X variables, the lower the accuracy of the two methods of detecting outliers. The statistics-based outlier detection method produces a masking value smaller than the data mining-based method, so the data statistics-based method successfully detects outlier observations even though there are nearby outlier observations. The swamping value resulting from the data mining-based method is high, indicating that normal data is suspected to be an outlier. Hence, errors still occur in detecting outliers in the data.

In the case of an outlier on the y -axis (Vertical outlier)

A. Against the proportion of outlier

The results of the simulation process where the outliers are on the y-axis and the simulation focuses on the proportion of outliers in the data, the following values are obtained:

Table 7. Simulation results focus on the proportion of outliers on the *y*-axis

	Proportion of Outliers			
	5%		10%	
	Statistics	Data Mining	Statistics	Data Mining
Accuracy	0.995573	0.842785	0.995395	0.814545
Masking	0	0	0	0.000006
Swamping	0.0000046	0.000165	0.0000051	0.000205

Table 7 shows that the accuracy values of statistical methods are generally better than those of data mining methods. In this case, the masking value in the statistics-based method produces a value of zero, whereas in the data mining-based method, the more significant the proportion of outliers in the sample data, the higher the error value in detecting outliers. The swamping value from the statistics-based approach method is much smaller than the swamping value from the data mining-based method.

B. Against outlier distance

The simulation process with a focus on the distance of the generated outliers produces the following values:

Table 8. Simulation results focus on the outlier distance on the *y*-axis

	Outlier Distance			
	3σ		5σ	
	Statistics	Data Mining	Statistics	Data Mining
Accuracy	0.995573	0.842785	0.995395	0.814545
Masking	0	0	0	0.000006
Swamping	0.0000046	0.000165	0.0000051	0.000205

Based on Table 8, statistical methods' accuracy is better than data mining methods. In this simulation scenario, the masking value in the statistics-based method produces a value of zero. In contrast, the data mining-based method shows that the closer the distance of the outlier to normal data, the higher the error value in detecting the outlier. The swamping value in the table above shows that the swamping value from the statistics-based approach method is relatively much smaller than the swamping value from the data mining-based method.

C. Regarding the correlation with variable X

The simulation results focusing on the correlation with variable X are as follows

Table 9. Simulation results focus on the correlation of X variables

	Correlation of Variable X			
	0.1		0.8	
	Statistics	Data Mining	Statistics	Data Mining
Accuracy	0.99554	0.836165	0.99542	0.821165
Masking	0	0.000005	0	0.000001
Swamping	0.0000048	0.000177	0.0000049	0.000194

Based on the table above, the accuracy level of statistical methods is better than that of data mining methods. This can be seen from the distribution of accuracy values, which is almost close to one. The masking value in the statistics-based method produces a value of zero. In contrast, the data mining-based method shows that the further the correlation between the X variables, the higher the error value in detecting outliers. The swamping value from the statistics-based approach method is relatively much smaller than the swamping value from the data mining-based method.

CONCLUSIONS

Based on all the scenarios that have been tried, the outlier detection method using a statistical approach is better than the outlier detection method using a data mining-based approach. Therefore, the suggestion that can be given for further research is to make improvements to the data mining method by focusing more on statistical analysis apart from focusing on data processing computing time so that the expected results of outlier detection are not only fast but also more precise.

REFERENCES

[1] D. M. Hawkins, *Identification of Outliers*. London: Chapman and Hall, 1980.
[2] V. Kotu and B. Deshpande, *Data Science*, Second. Morgan Kaufmann, 2019.
[3] J. W. Osborne and A. Overbay, "The power of outliers (and why researchers should ALWAYS check for them)," *Practical Assessment, Research, and Evaluation*, vol. 9, 2004, doi: 10.7275/QF69-7K43.
[4] K. Wada, "Outliers in official statistics," *Jpn J Stat Data Sci*, vol. 3, no. 2, pp. 669–691, Dec. 2020, doi: 10.1007/s42081-020-00091-y.
[5] M. Bakker and J. M. Wicherts, "Outlier Removal and the Relation with Reporting Errors and Quality of Psychological Research," *PLoS ONE*, vol. 9, no. 7, p. e103360, Jul. 2014, doi: 10.1371/journal.pone.0103360.
[6] E. Panjei, L. Gruenwald, E. Leal, C. Nguyen, and S. Silvia, "A survey on outlier explanations," *The VLDB Journal*, vol. 31, no. 5, pp. 977–1008, Sep. 2022, doi: 10.1007/s00778-021-00721-1.
[7] Ch. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, "An outliers detection and elimination framework in classification task of data mining," *Decision Analytics Journal*, vol. 6, p. 100164, Mar. 2023, doi: 10.1016/j.dajour.2023.100164.
[8] Z. Niu, S. Shi, J. Sun, and X. He, "A Survey of Outlier Detection Methodologies and Their Applications," in *Artificial Intelligence and Computational Intelligence*, vol. 7002, H. Deng, D. Miao, J. Lei, and F. L. Wang, Eds., in Lecture Notes in Computer Science, vol. 7002. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 380–387. doi: 10.1007/978-3-642-23881-9_50.
[9] A. Smiti, "A critical overview of outlier detection methods," *Computer Science Review*, vol. 38, p. 100306, Nov. 2020, doi: 10.1016/j.cosrev.2020.100306.
[10] J. Majewska, "Identification of Multivariate Outliers – Problems and Challenges Of Visualization Methods," *Informatyka i Ekonometria*, vol. 4, pp. 69–83, 2015.
[11] S. A. Shaikh and H. Kitagawa, "Efficient distance-based outlier detection on uncertain datasets of Gaussian distribution," *World Wide Web*, vol. 17, no. 4, pp. 511–538, Jul. 2014, doi: 10.1007/s11280-013-0211-y.

[12] X. Xu, H. Liu, L. Li, and M. Yao, "A Comparison of Outlier Detection Techniques for High-Dimensional Data:," *IJCIS*, vol. 11, no. 1, p. 652, 2018, doi: 10.2991/ijcis.11.1.50.
[13] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications," *The VLDB Journal The International Journal on Very Large Data Bases*, vol. 8, no. 3–4, pp. 237–253, Feb. 2000, doi: 10.1007/s007780050006.
[14] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient Algorithms for Mining Outliers from Large Data Sets," 2000.
[15] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-based detection and prediction of outliers," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 2, pp. 145–160, Feb. 2006, doi: 10.1109/TKDE.2006.29.
[16] K. Zhang, M. Hutter, and H. Jin, "A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data." arXiv, Mar. 18, 2009. [Online]. Available: http://arxiv.org/abs/0903.3257
[17] A. Zimek and P. Filzmoser, "There and back again: Outlier detection between statistical reasoning and data mining algorithms," *WIREs Data Min & Knowl*, vol. 8, no. 6, p. e1280, Nov. 2018, doi: 10.1002/widm.1280.
[18] A. M. Baba, H. Midi, M. B. Adam, and N. H. A. Abd Rahman, "Detection of Influential Observations in Spatial Regression Model Based on Outliers and Bad Leverage Classification," *Symmetry*, vol. 13, no. 11, p. 2030, Oct. 2021, doi: 10.3390/sym13112030.
[19] P. J. Rousseeuw and K. V. Driessen, "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, Aug. 1999, doi: 10.1080/00401706.1999.10485670.
[20] E. A. Mahmood, H. Midi, S. Rana, and A. G. Hussin, "Robust Circular Distance and its Application in the Identification of Outliers in the Simple Circular Regression Model," *Asian J. of Applied Sciences*, vol. 10, no. 3, pp. 126–133, Jun. 2017, doi: 10.3923/ajaps.2017.126.133.