DOI: 10.15849/IJASCA.251130.21

Int. J. Advance Soft Compu. Appl, Vol. 17, No. 3, November 2025
Print ISSN: 2710-1274, Online ISSN: 2074-8523

The Performance of Data Cleaning on Machine Learning Model for Mapping Land Cover

Dyah E Herwindiati, Janson Hendryli, Albertus Sulaiman, Hongi Nagaputra

Faculty of Information Technology, Universitas Tarumanagara, Jakarta, Indonesia
e-mail: dyahh@fti.untar.ac.id
Faculty of Information Technology, Universitas Tarumanagara, Jakarta, Indonesia
e-mail: jansonh@fti.untar.ac.id
National Research and Innovation Agency, Indonesia
e-mail: albe002@brin.go.id
College of Information Engineering, Yangzhou University, Jiangsu, China
e-mail: hongi.putra11@gmail.com

Abstract

This study investigates the effect of robust data cleaning on the accuracy of land cover classification using machine learning. We apply the minimum vector variance (MVV) method to detect and remove outliers from multispectral Landsat 8 imagery of the Bogor region, Indonesia, and train a quadratic discriminant analysis (QDA) model for land cover mapping. MVV identifies and removes 37,163 anomalous pixels from a dataset of 595,586 pixels, improving the quality of training data without requiring a positive-definite covariance matrix. After cleaning, the QDA model achieves a higher F1-score compared to the model trained on the original dataset. Visual inspection of the resulting maps for four Bogor sub-regions confirms clearer class boundaries and reduced confusion between green and partially green areas. These results demonstrate that robust preprocessing, particularly outlier detection with MVV, significantly enhances the reliability of land cover mapping and has the potential to be broadly applied to remote sensing-based classification tasks.

Keywords: *data cleaning, land cover, minimum vector variance, outlier detection, quadratic discriminant analysis.*

1 Background

Data cleaning is a crucial phase in data processing after data collection. It involves identifying and correcting errors in the dataset, such as dealing with missing data, removing redundancy, and handling outliers, as highlighted by [1]. Data cleaning aims to ensure that the results of analytical data processing are achieved accurately and realistically. They are the critical aspects that all researchers and practitioners in machine learning, data science, and data mining should be well-informed about.

Data anomalies, often referred to as outliers, are found in almost every data set. They are usually assumed as errors or noises of various kinds. One of the effects of outliers is that the results of analytical data processing will be biased due to biased parameter estimates. Furthermore, outliers can cause overfitting in the machine learning algorithms. The study

Received 26 September 2025; Accepted 10 November 2025

of outliers continues to develop today and has become a hot topic in data mining, offering a wealth of opportunities for further exploration and learning.

The study of outliers has long attracted many researchers. As early as a century ago, [2] proposed a criterion for rejecting outlying observations. However, defining an outlier for general situations takes work to formulate. Some definitions that are often used in research studies are those from [3], [4], [5], [6], [7], [8], and [9]. This paper uses the definition given by [6] that an outlier is one or more data that is 'inconsistent' from other data groups.

There are several approaches to identify outliers in a multivariate case. The first approach, non-robust distance-based identification proposed by [10], transforms random vectors into random variables. The most popular transformation is the Mahalanobis distance. The second approaches use distances that are built based on robust distances, such as the minimum volume ellipsoid and the minimum covariance determinant introduced by [11], the feasible solution algorithm which was introduced by [12], the fast minimum covariance determinant by [13], and the blocked adaptive computationally efficient outlier nominators by [14]. Another approach is through a projection pursuit as proposed by [15] and [16].

This paper discusses data cleaning related to outlier detection. Our research examines the benefits of the outlier detection process in machine learning models for land cover mapping. This research uses the depth function for robust minimum vector variance estimation of a multivariate location-scale parameter method proposed by [17] for cleaning the data from outliers and by [18] who proposed a minimum vector variance method to determine robust estimator with a high breakdown point.

The machine learning model used in this research is the quadratic discriminant analysis. The model tries to find a nonlinear decision boundary between different classes and has been shown to have a significant advantage in producing informative data visualizations, especially for territorial maps [19]. The practical application of this model in the land mapping process, which is based on the Landsat 8 satellite data, is a crucial point of interest for researchers and practitioners in machine learning, data science, and data mining. This study also aims to show the performance of land cover mapping results after performing a preprocessing step, that is data cleaning for outliers.

2 Data Preprocessing

Data preprocessing means preparing the raw data to ensure that the data is clean and structured. This is an essential step which has been shown to significantly improve the model accuracy [20][21][22]. The preprocessing steps commonly consists of data cleaning, data transformation, and feature selection. Data cleaning involves identifying and removing missing data, irrelevant data, or outliers. The original data can also be transformed to another form of data to make it more suitable for the prediction model. The transformation can be simple transformations, such as data scaling, normalization, encoding, or other complex transformations. Furthermore, a feature selection step can be performed on the original or transformed data by selecting only the relevant variables or features.

The preprocessing steps depend on the characteristics of the data and the needs of the machine learning model. For example, preprocessing techniques for text dataset in a sentiment classification task involve steps such as stop words removal to discard unimportant words and tokenization to transform the text data into numbers or vectors that are more suitable as inputs for any machine learning models [20]. Some models also require categorical variables to be encoded, features to be scaled, or benefits from feature selection to address multicollinearity problems.

Outliers in the data can occur because of faulty instruments, human error, or natural deviations in the population. For example, failure in the previous Landsat 7's scan line corrector instruments left gaps in the satellite data. A machine learning model should learn better when the original data have been preprocessed, i.e., the faulty values need to be discarded or filled with secondary data. This step is usually called the data cleaning step. The inaccurate data points may skew statistical measures and lead to inaccurate predictions [23], so the preprocessing step here may aim to remove the faulty data. Generally, there is no single outlier detection approach that covers all kinds of scenario and data set, so an algorithm or method that is suitable for the specific data set has to be considered. The preprocessing explored in this research is the data cleaning step using the robust minimum vector variance method.

3 Robust Minimum Vector Variance for Data Cleaning

3.1 The Robust Minimum Vector Variance Algorithm

Minimum vector variance (MVV) is a robust measure for identifying outliers, which was proposed by [18]. The excellent characteristics of MVV, as mentioned by [18], is that the MVV computation is simple and efficient. Apart from that, the requirement that the covariance matrix must be positive and definite is not required for MVV.

The MVV concept is derived from the multivariate dispersion measure vector variance (VV) [24][25]. Geometrically, VV is a square of the diagonal length of a parallelotope generated by all variance of the p variable. Computational MVV adopted the good properties of the C-step algorithm [13].

The MVV was proposed by modifying the C-step using the criterion of minimizing the square trace of the covariance matrix known as MVV. The concentration step or C-step was proposed by [13] as the fast minimum covariance determinant (FMCD). The FMCD algorithm, which was built on the multivariate generalized variance (GV) dispersion or covariance determinant, is a robust and efficient method. The high breakdown point, and the use of an effective and efficient C-step underscore its effectiveness. As explained by [18], the difference between FMCD and MVV is in the use of multivariate dispersion. While FMCD uses the covariance determinant (also often referred to as generalized variance), MVV uses VV as the measure of dispersion.

Suppose $\vec{X}_1, \vec{X}_2, \vec{X}_3, \ldots, \vec{X}_n$ denotes a random sample of size n picked from a p-variate distribution having location parameter $\vec{\mu}$ and positive definite covariance matrix Σ. To compare the structure of GV with the structure of VV, we assume $S_{p\times p}$ as in Eq. 1.

$$S_{p\times p} = \begin{bmatrix} S_{11} & \cdots & S_{1p} \\ \vdots & \ddots & \vdots \\ S_{p1} & \cdots & S_{pp} \end{bmatrix} \tag{1}$$

The GV is defined as the determinant of S and the VV is defined as in Eq. 2. When it comes to the VV formula, its unique features stand out. It can effectively measure multivariate dispersions even when the covariance matrix is singular. Moreover, its computation process is remarkably efficient, as noted by [18].

$$\mathrm{Tr}(S^2) = \sum_{i=1}^{p} S_{ii}^2 + 2\sum_{i<j}\sum_{j=1}^{p} S_{ij}^2 \tag{2}$$

Consider a data set $X = \{\vec{X}_1, \vec{X}_2, \vec{X}_3, \dots, \vec{X}_n\}$ of p-variate observations and $H \subseteq X$. We define T_{MVV} and C_{MVV} as the MVV estimators for the location parameter and covariance matrix, respectively. These two estimators are determined based on the set H that consists of $h = \left[\frac{n+p+1}{2}\right]$ data which gives covariance matrix C_{MVV} of minimum $\mathrm{Tr}(C_{\text{MVV}}^2)$ among all possible sets of h data. Therefore, we can define T_{MVV} and C_{MVV} as in Eq. 3 and Eq. 4.

$$T_{\text{MVV}} = \frac{1}{h}\sum_{i \in H} \vec{X}_i \tag{3}$$

$$C_{\text{MVV}} = \frac{1}{h}\sum_{i \in H}\left(\vec{X}_i - T_{\text{MVV}}\right)\left(\vec{X}_i - T_{\text{MVV}}\right)^t \tag{4}$$

The MVV algorithm is as follows [18]:

1. Form an arbitrary set H_{old} that consists of $h = \left[\frac{n+p+1}{2}\right]$ data.
2. Compute the mean vector $\bar{\vec{X}}_{H_{\text{old}}}$ and covariance matrix $S_{H_{\text{old}}}$ of all data in H_{old}. Then, for $i = 1, 2, \dots, n$, compute Eq. 5.

$$d_{H_{\text{old}}}^2(i) = d_{H_{\text{old}}}^2\left(\vec{X}_i, \bar{\vec{X}}_{H_{\text{old}}}\right) = \left(\vec{X}_i - \bar{\vec{X}}_{H_{\text{old}}}\right)^t S_{H_{\text{old}}}^{-1}\left(\vec{X}_i - \bar{\vec{X}}_{H_{\text{old}}}\right) \tag{5}$$

3. Sort the computations from the smallest to the largest. The order gives a permutation π on the index of observations. Let the result of sorting as $d_{H_{\text{old}}}^2(\pi_1) \leq d_{H_{\text{old}}}^2(\pi_2) \leq \cdots \leq d_{H_{\text{old}}}^2(\pi_n)$.
4. Form a set H_{new} that consists of h observations of index $\pi_1, \pi_2, \dots, \pi_h$.
5. Compute $\bar{\vec{X}}_{H_{\text{new}}}$, $S_{H_{\text{new}}}$, and $d_{H_{\text{new}}}^2\left(\vec{X}_i, \bar{\vec{X}}_{H_{\text{new}}}\right)$ as in Eq. 5.
6. If $\mathrm{Tr}\left(S_{H_{\text{new}}}^2\right) = \mathrm{Tr}(S_{H_{\text{old}}}^2)$, the process is finished. Otherwise, if $\mathrm{Tr}\left(S_{H_{\text{new}}}^2\right) < \mathrm{Tr}(S_{H_{\text{old}}}^2)$, the process is continued until the k-th iteration when $\mathrm{Tr}\left(S_{H_{\text{new}}}^2\right) = \mathrm{Tr}(S_{H_{\text{old}}}^2)$.
7. Suppose that S_{H_k} is the covariance matrix obtained from the k-th iteration. At the end of the k-th iteration, we obtain $\mathrm{Tr}\left(S_{H_1}^2\right) \geq \mathrm{Tr}\left(S_{H_2}^2\right) \geq \cdots \geq \mathrm{Tr}\left(S_{H_{k-1}}^2\right) = \mathrm{Tr}\left(S_{H_k}^2\right)$.

The MVV estimators for location parameters and covariance matrices are $T_{\text{MVV}} = \bar{\vec{X}}_{H_{\text{new}}}$ and $C_{\text{MVV}} = S_{H_{\text{new}}}$, respectively, on the k-th iteration. The robust Mahalanobis distance between $\vec{X}_i$ and T_{MVV} based on MVV is written as $dR_{\text{MVV}}\left(\vec{X}_i, T_{\text{MVV}}\right)$ and is defined on the quadratic form as Eq. 6 for $i = 1, 2, \dots, n$. The data that give large $dR_{\text{MVV}}\left(\vec{X}_i, T_{\text{MVV}}\right)$ value will be labeled as outlier or assumed as candidates of outliers.

$$dR_{\text{MVV}}\left(\vec{X}_i, T_{\text{MVV}}\right) = \left(\vec{X}_i - T_{\text{MVV}}\right)^t C_{\text{MVV}}^{-1}(\vec{X}_i - T_{\text{MVV}}) \tag{6}$$

3.2 The Breakdown Point of Minimum Vector Variance

The breakdown point is a quantitative measure to describe the concept of robustness. It measures how much data can be changed to infinity before being meaningless and crushed to bits. Several researchers, such as [26], [27], [28], and [29], gave interpretations of breakdown point both from the context of a population and from the context of a sample. This paper uses the interpretation given by [29] which is from the context of a sample. The breakdown point is defined in more detail as the smallest fraction of data, which causes the value of the estimator to be infinity when the value of all data in the fraction is changed to be infinity.

Consider a data set $X = \{\vec{X}_1, \vec{X}_2, \vec{X}_3, \dots, \vec{X}_n\}$ of p-variate observations and $H \subseteq X$. Then, $T_n(X)$ and $C_n(X)$ are the robust estimators for the location parameter and covariance matrix, respectively. Suppose the estimator $T_n(X)$ becomes $T_n(X^*)$ if the value of m data are changed. The breakdown point is defined as in Eq. 7 and Eq. 8 [29] which measures the greatest difference between $T_n(X^*)$ and $T_n(X)$.

$$\text{bias}\left(m, T, \vec{X}\right) = \sup_{X^*} ||T_n(X^*) - T_n(X)|| \tag{7}$$

$$\varepsilon_n^*\left(T, \vec{X}\right) = \min\left\{\frac{m}{n}\ \text{bias}\left(m, T, \vec{X}\right)\ \text{infinite}\right\}$$

Assume the m data for which the values are changed to be infinity imply that $\text{bias}\left(m, T, \vec{X}\right)$ is infinite. If the value of $(m - 1)$ data among them are changed to be infinity do not imply that the $\text{bias}\left(m, T, \vec{X}\right)$ is infinite, then the breakdown point is $\frac{m}{n}$.

A brief description of the formula is summarized and simulated as follows: a random data size $n = 100$ is generated from a mixture of p-variate multivariate normal distribution $(1 - \varepsilon)\ \mathrm{N}_p(\mu_1, I_p) + \varepsilon\ \mathrm{N}_p(\mu_2, I_p)$ where $p = 5$ and ε is the contaminant level. In this experiment, $\mu_1 = 0$, $\mu_2 = 10e_{1\times p}$, and $e_{1\times p} = (1, 1, 1, \dots)^t$.

The breakdown point in robust statistics is a pivotal concept defined as the highest proportion of anomaly observations or outliers that an estimator can tolerate before its results become arbitrarily inaccurate. A higher breakdown point signifies a more robust estimator, meaning it is more resistant to outliers and can handle data with significant errors or inconsistencies.

From Fig. 1, it is evident that the breakdown of MVV estimator exceeds 0.5, a point at which it becomes damaged and inconsistent due to the presence of outliers. The figure illustrates that the MVV estimator, a key tool in robust statistics, is at risk if outliers are present in more than half of the data.

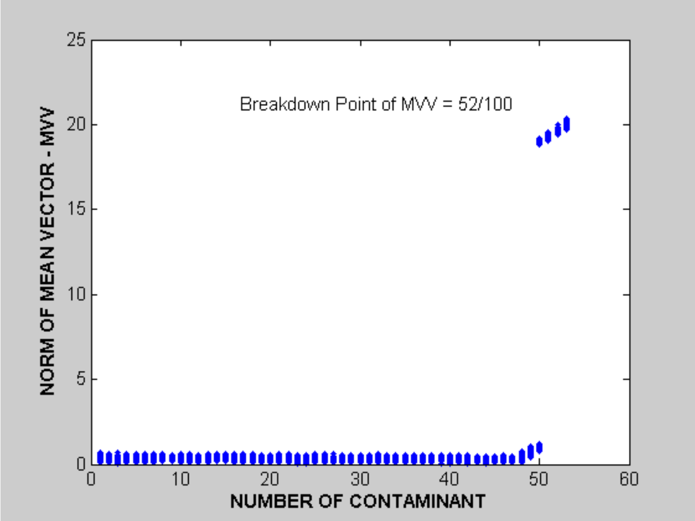


Fig. 1: The simulation of breakdown points of MVV.

3.2 The Sensitivity of the Classic and Robust MVV Estimator

The occurrence of one or more outliers shifts the mean vector toward the outliers, and the covariance matrix becomes inflated. An observation is called influential if its deletion would cause significant changes in the estimation. Outliers can be considered as an influential observation that could significantly change the estimator. The estimator is said to be insensitive if no significant change is due to removing the outliers. This paper presents a simple discussion on the computation and the theoretical distribution. The sensitivity of robust estimator is also explained thoroughly in [30] and [31].

Theorem 1 *Suppose $\vec{X}_1, \vec{X}_2, \vec{X}_3, \dots, \vec{X}_n$ are random sample of size n of a probability distribution having mean $\vec{\theta} \in \mathbb{R}^p$ where $p \geq 2$ is an integer and the covariance matrix $\boldsymbol{\Sigma}$ is of positive definite. Then the random vector is as in Eq. 8 where $\vec{\bar{X}}_n = \sum_{j=1}^{n} \vec{X}_j$ and $\boldsymbol{C}\boldsymbol{C}^t = \boldsymbol{\Sigma}$.*

$$\vec{Y}_n = \sqrt{n}\,\boldsymbol{C}^{-1}\left(\vec{\bar{X}}_n - \vec{\mu}\right) \rightarrow \sim \mathrm{N_p}\left(\vec{0}, \boldsymbol{I}_p\right) \tag{8}$$

Consider the data set $X = \left\{\vec{X}_1, \vec{X}_2, \vec{X}_3, \dots, \vec{X}_n\right\}$ of p-variate. The scatter matrix of sample A is as in Eq. 9 where the sample's mean vector $\vec{\bar{X}} = \frac{1}{n}\sum_{j=1}^{n} \vec{X}_j$. From Eq. 9, the scatter matrix A is of Wishart distribution with parameter $\boldsymbol{\Sigma}$ and the degree of freedom $n - 1$, written as $A \sim W_p(\boldsymbol{\Sigma}, n-1)$. Furthermore, A is independent of $\vec{\bar{X}}$.

$$A = \sum_{j=1}^{n}\left(\vec{X}_j - \vec{\bar{X}}\right)\left(\vec{X}_j - \vec{\bar{X}}\right)^t \tag{9}$$

Define A_{-i} as the scatter matrix removing the i^{th} observation (outlying observation) and is formulated as in Eq. 10 where $\vec{\bar{X}}_{-i} = \frac{1}{n-1}\sum_{i \neq j=1}^{n} \vec{X}_j$.

$$A_{-i} = \sum_{i \neq j=1}^{n}\left(\vec{X}_j - \vec{\bar{X}}_{-i}\right)\left(\vec{X}_j - \vec{\bar{X}}_{-i}\right)^t \tag{10}$$

The scatter matrix A_{-i} is also of Wishart distribution with parameter Σ and the degree of freedom $n - 2$, and $A_{-i} \sim W_p(\Sigma, n-2)$. The ratio of scatter matrix as the consequence of the removal of i^{th} observation is given by Eq. 11.

$$R_i = \frac{|A_{-i}|}{|A|} \tag{11}$$

The estimator is said to be insensitive to an outlier when $R_i >$ beta$\left(\frac{n-p-1}{2}, \frac{p}{2}\right)$. R_i is a constant that is used as a measure of the sensitivity of the i^{th} observation. The R_i value ranges from 0 to 1. The R_i close to 1 means that removing the i^{th} observation does not affect the estimator. This condition is called an insensitive estimator. On the other hand, if R_i is close to 0, the estimator is very sensitive to the deletion of the i^{th} observation. A good estimator is an estimator that is not sensitive to an outlier, namely an estimator with R_i close to 1.

In data analysis, the problem of $k > 1$ outliers are commonly found where the masking effect problem is unavoidable. Suppose the group I consists of k outliers. The scatter matrix A_{-I} as a consequence of the removal of the I^{th} group is of distribution $A_{-I} \sim W_p(\mathbf{\Sigma}, m)$. Similar with the case of single outlier, the ratio of scatter matrix as a consequence of the removal of the I^{th} group can be formulated as Eq. 12.

$$R_I = \frac{|A_{-I}|}{|A|} \tag{12}$$

The estimator is said to be insensitive to k outliers $(k > 1)$ when $R_I > \prod_{i=1}^{h} u_i$ where $u_i \sim$ beta$\left(\frac{m-p+i}{2}, \frac{p}{2}\right)$, $i = 1, 2, \dots, k$. Moreover, if the value of R_I is close to 1, then it means that there are no significant changes due to the removal of k observations on the group I.

The distribution of the classical approach is well known. It is different from that of the robust approach which is not easy to compose. Usually, we have to use a simulation approach to get it. This section also discusses the sensitivity and approximated distribution of the robust approach.

Let the data set $X_n = \{\vec{X}_1, \vec{X}_2, \vec{X}_3, \dots, \vec{X}_n\}$ of p-variate observations. If the observations taken from a subset $H \subseteq X$ that consists of h data points, then $\vec{X}_1, \vec{X}_2, \vec{X}_3, \dots, \vec{X}_h$ are random sample of size h and of distribution $\mathrm{N}_p(\vec{\mu}, \Sigma)$ where h is assumed as $h = \left[\frac{n+p+1}{2}\right]$. The location and scale estimator can then be computed as in Eq. 13 and Eq. 14. Based on the central limit theorem, if $\vec{X}_1, \vec{X}_2, \vec{X}_3, \dots, \vec{X}_n \sim \mathrm{N}_p(\vec{\mu}, \Sigma)$, then the distribution S^R can be approximated by Eq. 15.

$$\vec{\bar{X}}^R = \frac{1}{h} \sum_{i \in h} \vec{X}_i \tag{13}$$

$$S^R = \sum_{i \in h} \left(\vec{X}_i - \vec{\bar{X}}\right)\left(\vec{X}_i - \vec{\bar{X}}\right)^t \tag{14}$$

$$m\, c^{-1}\, S^R \sim W(m, \Sigma) \tag{15}$$

Then, it can be concluded that the scatter matrix A^R is as in Eq. 16 where $A^R \sim \frac{1}{m} W(m, \Sigma)$. The value of c can be approximated as 1. Whereby, [30] predicted the values of m by simulation approach and are listed in Table 1. Finally, the formula of $R_i^R = \frac{|A_{-i}^R|}{|A^R|}$ is approximated by $R_i^R \sim \frac{mp}{m(m-p+1)} F_{m,m-p+1}$.

$$A^R = \sum_{j \in h} \left(\vec{X}_j - \vec{\bar{X}}^R\right)\left(\vec{X}_j - \vec{\bar{X}}^R\right)^t = c^{-1} \frac{S^R}{m} \tag{16}$$

Table 1: The prediction of m from [30]

Dimension (p) and Size (n)	m_{pred}
$p = 5, n = 50$	12.89
$p = 10, n = 100$	33.13
$p = 10, n = 500$	126.71
$p = 20, n = 1000$	298.35

The illustration in Table 2 shows the considerable difference in sensitivity between the robust MVV estimator and the classical estimator from a simulation that is generated from multivariate data having sizes $p = 5$ and $n = 50$ with $k = 3$ outliers. It is shown that the MVV estimator is insensitive to outliers.

Table 2: Comparison of the sensitivity of MVV and classical robust estimators

	Method	
	Classical	***Robust MVV***
A	*27.1628*	*0.040128*
A_{-I}	*0.807614*	*0.039319*
R_{-I}	*0.029732*	*0.979824*
Cutoff	*0.999722*	*0.096965*
Sensitivity to outliers	***Very sensitive***	***Insensitive***

4 Quadratic Discriminant Analysis

Quadratic discriminant analysis or QDA is a statistical tool that is used for classification and is an extension of linear discriminant analysis (LDA) [19][32]. Both have been used successfully for various classification tasks, such as diabetes disease classification [33], radar signal classification [34], and industrial process fault diagnosis [35]. The discriminant analysis can also be used as dimensionality reduction methods [36] and commonly used due to its simplicity and effectiveness [37].

Both LDA and QDA strives to find a decision boundary that best separates the classes in the data. The decision boundary in the LDA is calculated in a linear function where it assumes that the variance between all classes is equal. Meanwhile, the decision boundaries between classes in the QDA can be non-linear, or in this case, quadratic.

Suppose we have training samples represented by vectors of m features and class labels. The objective of the training phase of a machine learning model is to calculate the discriminant functions for each class, $\{f_1, f_2, \ldots, f_c\}$, that are being used to determine the decision boundaries [38]. In order to assign a class label to the unknown samples, the discriminant functions are compared to get the maximum score, i.e., $f_i(x) > f_j(x)$ with $i, j = 1, 2, \ldots, c$ and $i \neq j$. To simplify, assume we have two classes $\{\omega_1, \omega_2\}$ and, consequently, two discriminant functions f_1 and f_2, then the decision boundary can be calculated as in Eq. 17.

$$\operatorname{sign}(\phi(x)) = \operatorname{sign}(f_1(x) - f_2(x)) = \begin{cases} \omega_1 & \text{if } \phi(x) \geq 0 \\ \omega_2 & \text{if } \phi(x) < 0 \end{cases} \tag{17}$$

Let X be the input data which consists of N samples of which has m features. The algorithm to build the QDA classifier is as follows [38]:

1. Compute the mean of each class $\mu_i = \frac{1}{n_i}\sum_{i=1}^{n_i} x_i$ where μ_i denotes the mean of the i^{th} class and $x_i \in \omega_i$.
2. Calculate the priori probability of each class: $P(\omega_i) = \frac{n_i}{N}$.
3. Calculate the covariance matrix for each class: $\Sigma_i = \frac{1}{n_i}\sum_{x\in\omega_i}(x - \mu_i)(x - \mu_i)^T$.
4. Calculate the discriminant functions f_i for all classes ω_i as in Eq. 18.

$$f_i(x) = -\frac{\Sigma_i^{-1}}{2}(x^T x + \mu_i^T \mu_i - 2\mu_i^T x) - \frac{m}{2}\ln(2\pi) - \frac{\ln|\Sigma_i|}{2} + \ln\big(P(\omega_i)\big) \tag{18}$$

The decision boundary between two classes, ω_1 and ω_2, can then be represented as in Eq. 19 where $W = -\frac{1}{2}(\Sigma_1^{-1} - \Sigma_2^{-1})$ represents the coefficient of the quadratic term $x^T x$, $= \mu_1^T \Sigma_1^{-1} - \mu_2^T \Sigma_2^{-1}$ denotes the slope, and $W_0 = -\frac{1}{2}(\mu_1^T \Sigma_1^{-1}\mu_1 - \mu_2^T \Sigma_2^{-1}\mu_2 + \ln|\Sigma_1| - \ln|\Sigma_2|) + \ln\frac{P(\omega_1)}{P(\omega_2)}$ is the bias [38].

$$\phi(x) = x^T W x + w^T x + W_0 \tag{19}$$

5 Remote Sensing Data for the Experiments

The first Landsat satellite was launched on July 23, 1972, as a collaboration between NASA and the United States Geological Survey (USGS) [39]. The early Landsat-1 satellite imagery has been used in geological mapping and hydrological studies [40], mineral exploration [39], and agricultural resources monitoring [41]. Currently, the satellite is in its ninth series (Landsat-9) which, together with Landsat-8, provides valuable Earth observation data every 8 days. Since 2008, the satellite image data have been freely available to the public and can be accessed or downloaded through the USGS website [42].

The Landsat-8 satellite is equipped with two main instruments: the operational land imager (OLI) and thermal infrared sensor (TIRS). The OLI is a sensor that captures multispectral images with a total of 9 spectral and panchromatic bands. The first band is a coastal or aerosol band, bands 2 to 7 are the visible, near-wave infrared, and short-wave infrared bands, and band 9 is specifically designed to detect cirrus clouds. Moreover, the panchromatic band 8 has a lower spatial resolution of 15 meters, instead of 30 meters in the other spectral bands [43]. Meanwhile, the TIRS captures thermal infrared images and consists of two bands (band 10 and 11) with 100 meters spatial resolution. The Landsat-9 was launched on September 27, 2021, more than 8 years after its predecessor, with nearly identical, but improved, instruments and provides redundancy and continuation of the Landsat-8 [44]. Fig. 2 shows the example of Landsat-9 satellite imagery from the OLI sensor.

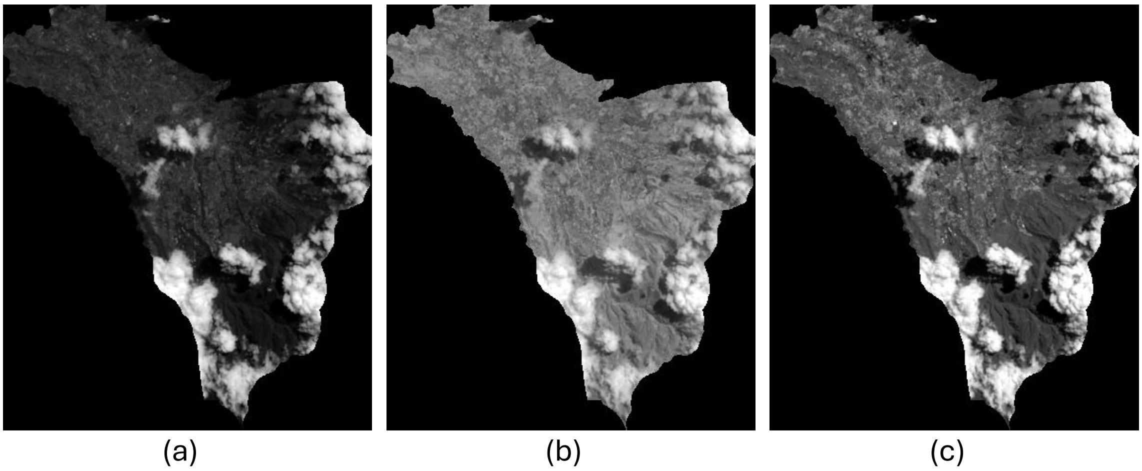


Fig. 2: Example of Landsat-9 OLI data showing the Cisarua regency in Indonesia; (a) band 3 visible green; (b) band 5 near-infrared; (c) band 7 short-wave infrared 2.

The study area for the experiment is the Bogor area in West Java, Indonesia. Previous studies from [45][46] have already shown the importance of urbanization in the area for the surrounding cities. The Cigudeg, Cisarua, Nanggung, and Sukajaya regency are excluded from the training set and used for the land mapping as the testing area. The total pixels for training the QDA model are 595,586 pixels with a composition of 239,049 pixels (40.14%) for the green area, 214,586 pixels (36.03%) for the partial green area, and 141,951 pixels (23.83%) for the impervious area. Moreover, each pixel in the data set is represented by 6 spectral bands (3 visible bands, 1 near infrared band, and 2 short-wave infrared bands).

6 Experiments and Results

Many factors influence the success of a machine learning model prediction. Data processing which is a collection of techniques for enhancing the quality of raw data, is not just a key step but a crucial one in the machine learning process [47]. In this step, raw data is converted into a format that can be understood and evaluated by computers and machine learning algorithms. These algorithms play a crucial role in automatically extracting knowledge from machine-readable data. The quality of the data determines the performance of the model. Therefore, the potential impact of improving the dataset quality on the model's accuracy is significant. To overcome issues such as noisy, redundant, or missing data, preprocessing is not just beneficial but necessary. In the experiments, data preprocessing involved data cleaning using the minimum vector variance method. Data cleaning is performed before the machine learning process. The flowchart in Fig. 3 illustrates the experimental process.

The data set that contains 595,586 pixels are cleaned using the MVV method. After that, the performance of the QDA model trained using the cleaned data is compared to the model that is trained using the original data set. The MVV detects outliers from the data set and removes 37,163 pixels. The final data set after data cleaning consists of 558,423 pixels. The data set is split 70% as the training data and 30% for the model validation ensuring that the proportions of each classes are preserved. The classification models are compared using the precision, recall, and F1-score metrics on the validation data. Meanwhile, the training data are used to train the QDA model. The QDA is implemented using Python programming language, and the Scikit-Learn library using regularization parameter 0.5.

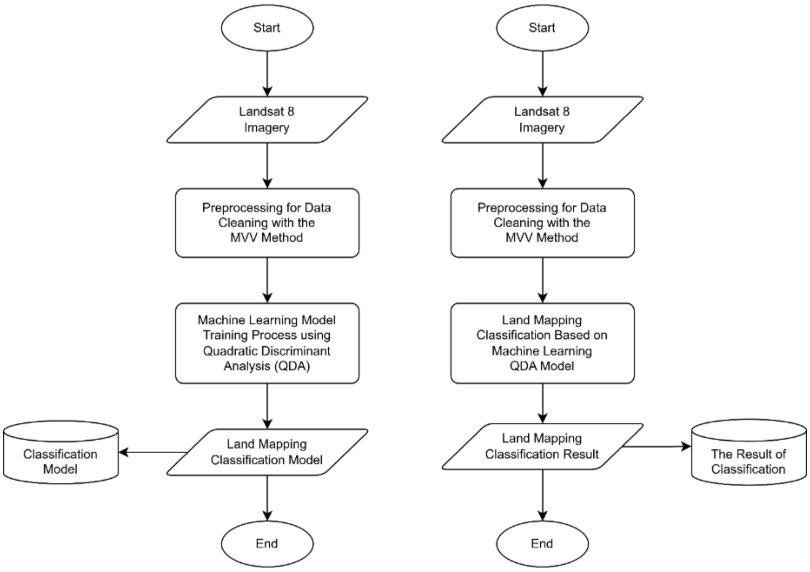

Start
Landsat 8 Imagery
Preprocessing for Data Cleaning with the MVV Method
Machine Learning Model Training Process using Quadratic Discriminant Analysis (QDA)
Land Mapping Classification Model
Classification Model
End
Start
Landsat 8 Imagery
Preprocessing for Data Cleaning with the MVV Method
Land Mapping Classification Based on Machine Learning QDA Model
Land Mapping Classification Result
The Result of Classification
End

Fig. 3: The flowchart of the experiments

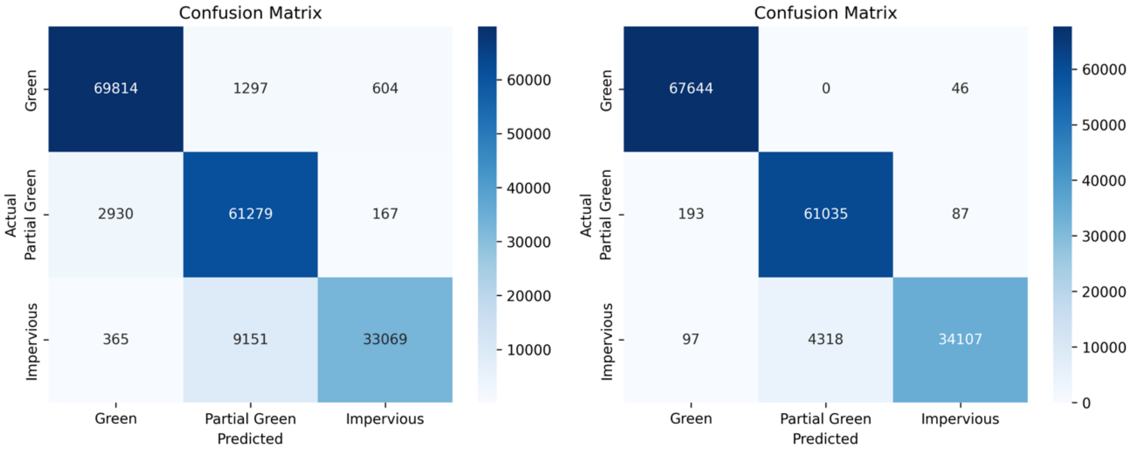
Table 3 presents the F1-scores obtained from both the original and cleaned datasets across ten independent experiments. The results consistently show that the cleaned dataset yields higher F1-scores compared to the original dataset, indicating a notable improvement in model performance after data preprocessing. It also suggests that cleaning the data substantially enhances the model's ability to balance precision and recall. Furthermore, the small standard deviation of 0.0011 demonstrates that the improvement is stable and reproducible across multiple trials, highlighting the robustness of the data cleaning process.

Table 3: Comparison of the F1-scores obtained from the original and cleaned datasets across ten experiments.

# Experiments	*F1-score from Original Data Set*	*F1-score from Cleaned Data*	*Difference*
1	*0.9100*	*0.9662*	*0.0562*
2	*0.9117*	*0.9668*	*0.0551*
3	*0.9115*	*0.9672*	*0.0557*
4	***0.9105***	***0.9674***	***0.0569***
5	*0.9122*	*0.9660*	*0.0538*
6	***0.9107***	***0.9674***	***0.0567***
7	*0.9101*	*0.9666*	*0.0565*
8	*0.9131*	*0.9670*	*0.0539*
9	*0.9119*	*0.9668*	*0.0549*
10	*0.9118*	*0.9667*	*0.0549*
		Mean	*0.0555*
		Std. Deviation	*0.0011*

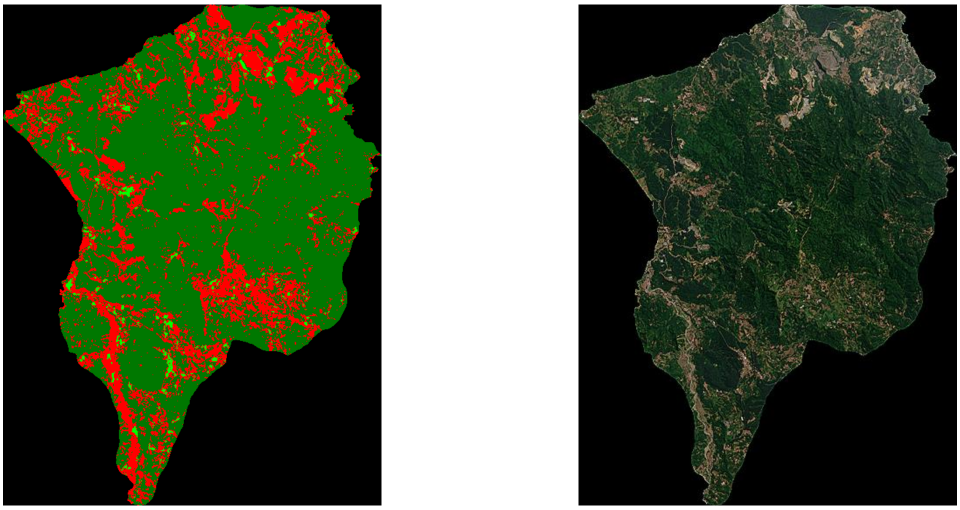
To evaluate whether the improvement in F1-scores after data cleaning is statistically significant, a paired *t*-test was performed comparing the results from the original and cleaned datasets across ten experiments. Since each pair of scores represents the same experimental condition before and after cleaning, the paired *t*-test is appropriate for this analysis. The test yielded $t(9) = 159.9423$ with a p-value of 1.6×10^{-15}, which is far below the 0.05 significance threshold. This confirms that the observed improvement in F1-score is highly significant and not due to random variation. Therefore, the data cleaning process demonstrably enhanced the model's classification performance.

The confusion matrix as in Fig. 4 shows that the second model that is trained using cleaned data appears to have better class separation with fewer misclassifications. Both models perform well on the green class with high true positive count, but some confusion clearly exists with the partial green class in the model without data cleaning. Data cleaning using the MVV completely eliminated the confusion between green and partial green class. It suggests better decision boundaries on the second model. However, a lot of impervious area are falsely detected as partial green in both models. They also perform moderately well on partial green class.



(a) (b)

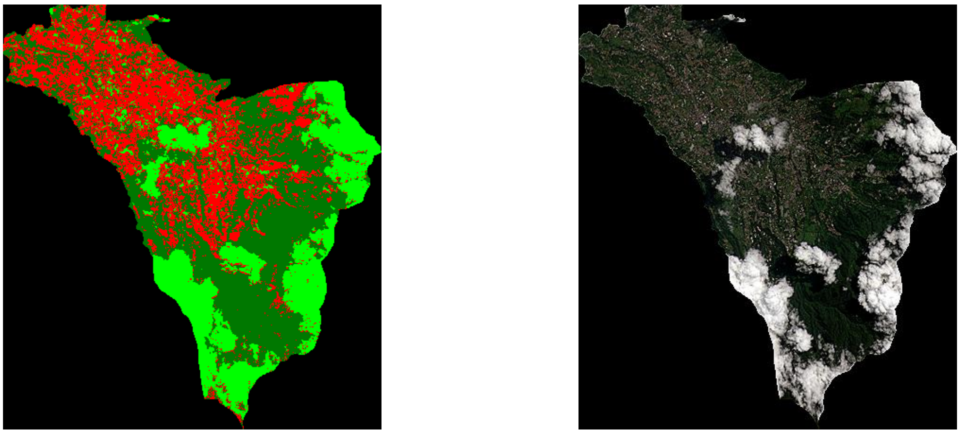
Fig 4: The confusion matrix of (a) model trained with the original data set (without cleaning); and (b) model trained using the cleaned data.



(a) (b)

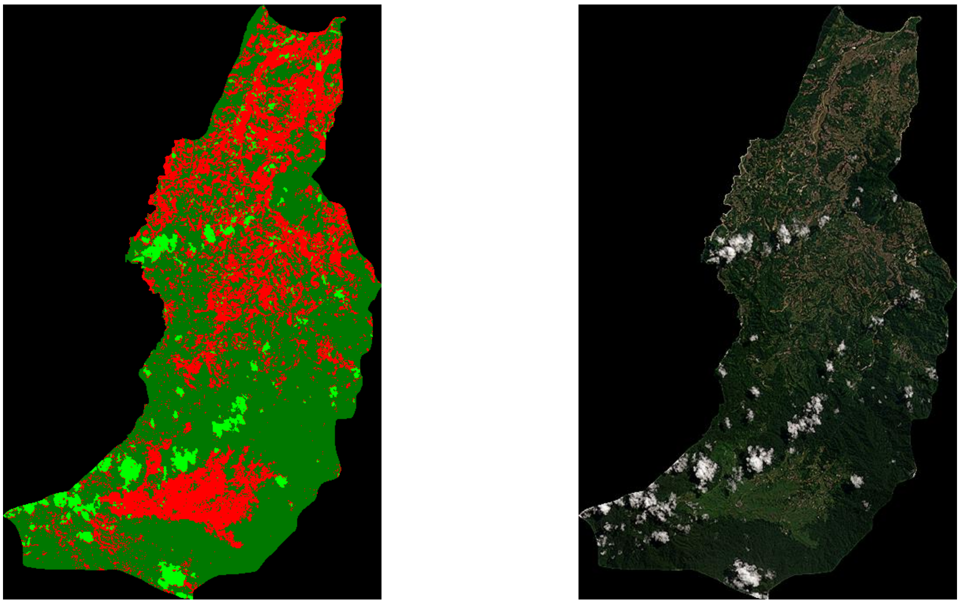
Fig 5: The mapping overlay for the Cigudeg area showing (a) the mapping results and (b) the composite images from Landsat data.

The best model, which is the QDA model trained using data that have been cleaned using MVV, is implemented for land mapping in the 4 areas in Bogor to evaluate the model performance visually. The Landsat 8 images are from April 2020. Fig. 5, 6, 7, and 8 show the mapping results for the Cigudeg, Cisarua, Nanggung, and Sukajaya regency, respectively, where red areas denote the impervious area, dark green areas denote the green area, and light green denotes the partial green area or the transitional zones. These overlays visually show the spatial distribution of land categories in the regions. Comparing each images side by side allows direct visual comparison where the QDA predictions align with the actual surface features, such as vegetation, roads, and buildings.



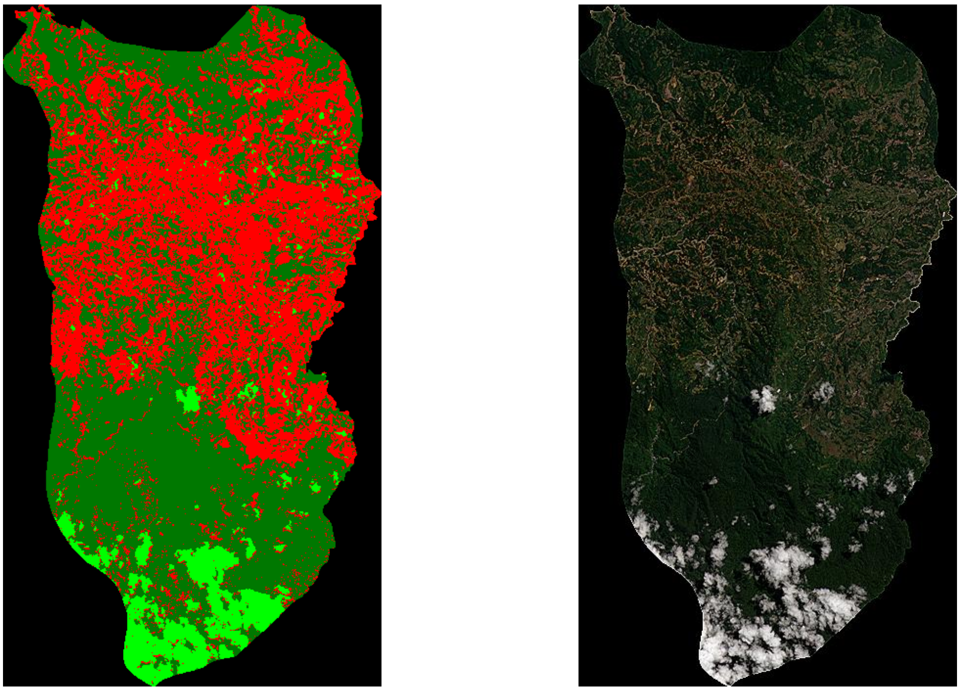
(a) (b)

Fig 6: The mapping overlay for the Cisarua area showing (a) the mapping results and (b) the composite images from Landsat data.



(a) (b)

Fig 7: The mapping overlay for the Nanggung area showing (a) the mapping results and (b) the composite images from Landsat data.



(a) (b)

Fig 8: The mapping overlay for the Sukajaya area showing (a) the mapping results and (b) the composite images from Landsat data.

From the total area of 177.47 km^2 in Cigudeg, the model identifies 131.12 km^2 (73.8%) of green area, 3.16 km^2 of partial green area, and 43.19 km^2 of impervious area. Cigudeg also has the largest percentage of green areas among the four regencies. In contrast, Cisarua has the smaller proportion of green areas at 42.21% of its total 71.59 km^2. It shows significant partial green (18.63 km^2) and impervious (22.74 km^2) coverage, suggesting a more mixed-use landscape. Nanggung presented more balanced land profile, with 63.07% of its 157.91 km^2 classified as green, along with smaller areas of partially green (8.88 km^2) and impervious (49.43 km^2) land. The mapping result of Sukajaya indicates an area with substantial human influence. With a total area of 166.16 km^2, it has 51.4% green coverage. It also contained 12.58 km^2 of partially green and 68.18 km^2 of impervious area.

Moreover, the QDA parameters from the best model that is trained using cleaned data can be found in Eq. 20, 21, and 22 where the discriminant function is defined as $x^T Q x + w^T x + c = 0$.

$$Q = \begin{bmatrix} 0.133 & 0.065 & -0.005 & 0.035 & -0.022 & -0.093 \\ 0.065 & -0.005 & -0.066 & -0.052 & 0.029 & 0.002 \\ -0.005 & -0.067 & -0.122 & 0.017 & 0.082 & 0.023 \\ 0.035 & -0.052 & 0.017 & 0.002 & 0.004 & -0.002 \\ -0.022 & 0.029 & 0.082 & 0.004 & -0.058 & -0.009 \\ -0.094 & 0.002 & 0.023 & -0.002 & -0.009 & 0.049 \end{bmatrix} \tag{20}$$

$$w = (-0.006 \quad 0.020 \quad 0.060 \quad 0.148 \quad -0.014 \quad -0.051) \tag{21}$$

$$c = 6.94959725 \tag{22}$$

7 Conclusion

This research highlights a critical phase of data preprocessing, specifically robust outlier detection for data cleaning and enhancing machine learning performance in remote sensing applications. By employing the minimum vector variance method (MVV), we successfully removed anomalous observations from Landsat 8 imagery, resulting in a cleaner dataset that enabled the quadratic discriminant analysis model to generate more accurate and visually consistent land cover maps. The F1-score improvement from 0.91 to 0.9662 illustrates how eliminating outliers sharpens class boundaries and reduces misclassification, particularly between green and partially green areas.

The findings emphasize that preprocessing can be a decisive factor in model accuracy. MVV's robustness makes it a practical tool for large multivariate datasets. The MVV estimator is a high-breakdown-point estimator. The estimator can tolerate a large proportion of outliers in the data before its performance is significantly broken. Future work may explore integrating MVV with other classification methods or extending it to multi-temporal satellite data for change detection. Overall, the study confirms that robust data cleaning is a prerequisite for reliable land cover mapping and should be prioritized in remote sensing and machine learning pipelines. While this study focuses on the Bogor region, the underlying principles of the MVV-based cleaning method are not region-specific. Therefore, the approach could be adapted to other geographic or climatic contexts, provided that sufficient regional calibration and model training is performed. Future research should validate the method in regions with different topographies and weather patterns to assess its broader applicability.

References

[1] Borrohou, S., Fissoune, R., & Badir, H. (2023). Data Cleaning Survey and Challenges – Improving Outlier Detection Algorithm in Machine Learning. *Journal of Smart Cities and Society, 2,* 125-140.

[2] Irwin, J. O. (1925). On a Criterion for the Rejection of Outlying Observations. *Biometrics*, *17*(3), 238-250.

[3] Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations Samples. *Technometrics*, *11,* 1-21.

[4] Hawkins, D. M. (1980). *Identification of Outliers 2nd ed.* New York: Chapman and Hall.

[5] Beckman, R. J., & Cook, R. D. (1983). Outliers, *Technometrics*, *25,* 119-149.

[6] Barnett, V., & Lewis, T. (1984). *Outliers in Statistical Data 2nd ed.* New York: John Wiley.

[7] Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association*, *85*(411), 633-639.

[8] Iglewicz, B., & Hoaglin, D. C. (1993). *Volume 16: How to Detect and Handle with Outliers*. Milwaukee: Quality Press.

[9] Becker. C., & Gather, U. (1999). The Masking Breakdown Point of Multivariate Outlier Identification Rules. *Journal of the American Statistical Association*, *94*, 947-955.

[10] Derquenne, C. (1992). Outlier Detection Before Running Statistical Methods, *Siam*, *34*(2), 323-326.

[11] Rousseeuw, P. J. (1985). Multivariate Estimation with High Breakdown Point. *Mathematical Statistics and Applications* (pp. 283-297). D Reidel Publishing Company.

[12] Hawkins, D. M. (1994). The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data. *Computational Statistics and Data Analysis*, *17*, 197-210.

[13] Rousseeuw, P. J., & van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, *41*, 212-223.

[14] Billor, N., Hadi A. S., Velleman, P. F. (2000). BACON: Blocked Adaptive Computationally Efficient Outlier Nominations. *Computational Statistics and Data Analysis*, *34*, 279-298.

[15] Pan, J. X., Fung, W. K., & Fang, K. T. (2000). Multiple Outlier Detection in Multivariate Data Using Projection Pursuit Technique. *Journal of Statistical Planning and Inference*, *83*, 153-167.

[16] Pena, D., & Prieto, J. F. (2001). Multivariate Outlier Detection and Robust Covariance Matrix Estimation. *Technometrics*, *43*(3), 286-300.

[17] Herwindiati, D. E., Hendryli, J., & Hiryanto, L. (2017). Impervious Surface Mapping Using Robust Depth Minimum Vector Variance Regression. *European Journal of Sustainable Development*, *6*(3), 29-39.

[18] Herwindiati, D. E., Djauhari, M. A., & Mashuri, M. (2007). Robust Multivariate Outlier Labeling. *Journal of Communication in Statistics Simulation and Computation*, *36*(6), 1287-1294.

[19] Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear Discriminant Analysis: A Detailed Tutorial. *AI Communications*, *30*(2), 169-190.

[20] Alam, S., & Yao, N. (2019). The Impact of Preprocessing Steps on the Accuracy of Machine Learning Algorithms in Sentiment Analysis. *Computational and Mathematical Organization Theory*, *25*(3), 319-335.

[21] Huang, J., Li Y. F., & Xie, M. (2015). A Systematic Analysis of Data Preprocessing for Machine Learning-based Software Cost Estimation. *Information and Software Technology*, *67*, 108-127.

[22] Ibrahimi, E., Lopes, M. B., Dhamo, X., Simeon, A., Shigdel, R., Hron, K., Stres, B., D'Elia, D., Berland, M., & Marcos-Zambrano, L. J. (2023). Overview of Data Preprocessing for Machine Learning Applications in Human Microbiome Research. *Front*, *14*.

[23] Hodge, V., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, *22*, 85-126.

[24] Alt, F. B., & Smith, N. D. (1988). 17 Multivariate Process Control. *Handbook of Statistics*, *7*, 333-351.

[25] Djauhari, M. A. (2007). A Measure of Multivariate Data Concentration. *Journal of Applied Probability and Statistics*, *2*(2), 139-155.

[26] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1985). *Robust Statistics*, New York: John Wiley.

[27] Huber, P. J. (1980). *Robust Statistics*. Massachusetts: John Wiley.

[28] Kotz, S., & Johnson, N. L. (1985). *Encyclopedia of Statistical Sciences*, *6*, (pp. 110-122), New York: John Wiley.

[29] Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley.

[30] Hardin, J., & Rocke, D. M. (2004). Outlier Detection in Multiple Cluster Settings Using Minimum Covariance Determinant Estimator. *Computational Statistics and Data Analysis, 44*, 625-638.

[31] Herwindiati, D. E. (2008, July). The Advantage of Robust Measure for Mining Multivariate Outliers. In *Proceedings of the 2008 International Conference on Information Theory and Statistical Learning*, (pp. 29-35).

[32] Ghosh, A., SahaRay, R., Chakrabarty, S., & Bhadra, S. (2021). Robust Generalized Quadratic Discriminant Analysis. *Pattern Recognition*, *117*.

[33] Araveeporn, A. (2022). Comparing the Linear and Quadratic Discriminant Analysis of Diabetes Disease Classification Based on Data Multicollinearity. *International Journal of Mathematics and Mathematical Sciences*, *1*.

[34] Guo, S., & Tracey, H. (2020). Discriminant Analysis for Radar Signal Classification. *IEEE Transactions on Aerospace and Electronic Systems, 56*(4), 3134-3148.

[35] Li, H., Jia, M., Mao, Z. (2023). Dynamic Feature Extraction-based Quadratic Discriminant Analysis for Industrial Process Fault Classification and Diagnosis. *Entropy*, *25*(12).

[36] Laiadi, O., Ouamane, A., Benakcha, A., Taleb-Ahmad, A., & Hadid, A. (2020). Tensor Cross-View Quadratic Discriminant Analysis for Kinship Verification in the Wild. *Neurocomputing*, *377*, 286-300.

[37] Cai, T., Tony, & Zhang, L. (2021). A Convex Optimization Approach to High-Dimensional Sparse Quadratic Discriminant Analysis. *The Annals of Statistics*, *49*(3), 1537-1568.

[38] Tharwat, A. (2016). Linear vs. Quadratic Discriminant Analysis Classifier: A Tutorial. *International Journal of Applied Pattern Recognition*, *3*(2), 145-180.

[39] Wulder, M. A., Roy, D. P., Radeloff, V. C., Loveland, T. R., Anderson, M. C., Johnson, D. M., Healey, S., et al. (2022). Fifty Years of Landsat Science and Impacts. *Remote Sensing of Environment*, *280*.

[40] Capolupo, A., Saponaro, M., Mondino, E. B., & Tarantino, E. (2020). Combining Interior Orientation Variables to Predict the Accuracy of RPAS-SFM 3D Models. *Remote Sensing*, *12*(17).

[41] Draeger, W. C. (1973). Agricultural Applications of ERTS-1 Data. *NASA Goddard Space Flight Center Symposium on Significant Results Obtained from the ERTS-1*, *1*.

[42] Hemati, M. A., Hasanlou, M., Mahdianpari, M., & Mohammadimanesh, F. (2021). A Systematic Review of Landsat Data for Change Detection Applications: 50 Years of Monitoring the Earth. *Remote Sensing*, *13*(15).

[43] Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M. C., Kommareddy, A., Pickens, A., et al. (2021). Mapping Global Forest Canopy Height through Integration of GEDI and Landsat Data. *Remote Sensing of Environment*, *253*.

[44] Masek, J. G., Wulder, M. A., Markham, B., McCorkel, J., Crawford, C. J., Storey, J., & Jenstrom, D. T. (2020). Landsat 9: Empowering Open Science and Applications through Continuity. *Remote Sensing of Environment*, *248*.

[45] Adhikurnia, S., Herwindiati, D. E., Hendryli, J., & Susilo, V. V. (2024). PCA-SVM for Effective Land Cover Mapping in Bogor Area. In *Proceedings of the 2024 9th International Conference on Information Technology and Digital Applications*, (pp. 1-5). IEEE.

[46] Budi, D., Herwindiati, D. E., & Hendryli, J. (2021). Land Use Change Using Least Absolute Shrinkage and Selection Operator Regression in Jakarta's Buffer Cities. In *Proceedings of the 2021 11th Symposium on Computer Applications & Industrial Electronics*, (pp. 30-35). IEEE.

[47] Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data Preprocessing for Supervised Learning. *International Journal of Computer Science*, *1*(1).

Notes on contributors

Dyah E Herwindiati completed her doctoral program in Robust Statistics at the Department of Mathematics of Institut Teknologi Bandung (ITB) in 2006, one of Indonesia's leading educational institutions. Since earning her doctorate, she has been at the forefront of developing innovative, robust algorithms specifically designed for image processing applications, which have enhanced the accuracy and reliability of data analysis in various fields. In 2014, her expertise and contributions to the field were recognized when the Indonesian Minister of Education officially approved her as a professor specializing in informatics engineering. Her research has had a significant impact on several domains, particularly in data science and machine learning. Her contributions to academia have enriched the learning experience and fostered collaborations that bridge theoretical research and practical applications in the field of informatics. Currently, she has been promoted by the Ministry of Education, Research, and Technology of the Republic of Indonesia to the rank of full professor.

Janson Hendryli received his master's degree in computer science from Universitas Indonesia and is currently a faculty member at Universitas Tarumanagara, Jakarta, where he teaches courses in programming and deep learning. His research interest includes the application of machine learning models for various scenes, e.g., environmental and urban studies. He is also active in the industry, specifically in application development and project management.

Albertus Sulaiman received B.S. on Meteorology (ITB), M.S. on Elementary Particle Physics (UI), and Ph.D. on Theoretical Physics from Bandung Institute of Technology. His research interests are on environmental monitoring and modeling in tropical area, such as watershed, peatland, nonlinear atmospheric, and oceanic phenomena especially in equatorial region. He is currently Head of Research Center for Climate and Atmosphere, National Research and Innovation Agency, Indonesia.

Hongi Nagaputra received his Bachelor of Science (S.Kom.) degree in Informatics Engineering from Universitas Tarumanagara, Jakarta, Indonesia, in 2024. He is currently pursuing a master's degree in software engineering at Yangzhou University, where he primarily researches physiological signal processing and atrial fibrillation signal classification using machine learning.