# Comparing the Performance of Multivariate Hotelling's $T^2$ Control Chart and Naive Bayes Classifier for Credit Card Fraud Detection

**Ichwanul Kahfi Prasetya[1], Devi Putri Isnawarty[1], Abdullah Fahmi[1], Salman Alfarizi Pradana Andikaputra[1], Wibawati[1]**

[1]Department of Statistics, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia

* Corresponding author: ichwankahfi@gmail.com

**Received:** 2 September 2023
**Revised:** 1 March 2024
**Accepted:** 7 March 2024

**ABSTRACT – Credit card is a transaction tool using a card which is a substitute for legitimate cash in transactions. The use of computer technology is needed for various kinds of electronic transactions. In the world of technology, the term machine learning is not new and technological developments are increasingly rapid in recent years. Statistical process control method (SPC) is one of the measuring instruments used to improve the performance of public services. Hotelling $T^2$ control chart is a method in SPC that can be used to control the process. Methods that are widely used in the detection and classification of documents one of them is Naive Bayes Classifier (NBC) which has several advantages, among others, simple, fast and high accuracy. Those two methods will be used to detecting o2utlier of this dataset. The study used the credit card fraud registry with some PCA as independent variables. The size of fraud transaction is very small which represented only 0.172% of the 284,807 transactions. This research will use Area Under Curve (AUC) as the performance goodness test. A comparison of the accuracy of NBC and Hotelling's $T^2$ predictions shows that the performance of the $T^2$ Hotelling method is better in detecting outliers than the NBC method.**

**Keywords– Credit Fraud Detection, Confusion Matrix, Hotelling's $T^2$ Control Chart, Naive Bayes Classifier, Stratified K-Fold Cross Validation**

## I. INTRODUCTION

Credit card is a transaction tool using a card which is a substitute for legitimate cash in transactions. Credit cards as a means of payment are growing rapidly in Europe. The main factor that supports the use of credit cards is none other than the conditions that are leading to a reduction in the use of cash. With the existence of credit cards, it makes consumers easier for practical to transact and meet their needs at various ages. With this development, the user's lifestyle will also change to adjust the turnover of daily transactions. This change in lifestyle also depends on how users view the existence of modern payment, one of which is a credit card [1].

The use of computer technology is needed for various kinds of electronic transactions. In the world of technology, the term machine learning is not new and technological developments are increasingly rapid in recent years. Statistical process control (SPC) method is one of the measuring instruments used to improve the performance of public services. Hotelling's $T^2$ control chart is a method in SPC that can be used to control the process.

Methods that are widely used in the detection and classification of documents one of them is Naive Bayes Classifier (NBC) which has several advantages, among others, simple, fast and high accuracy. The Naive Bayes Classifier (NBC) method for classification or categorization of text uses word attributes that appear in a single document as the basis for its classification. The advantages of the use of Naive Bayes Classifier in the classification of documents can be seen from the process that takes action based on the data that has been there before. Therefore, the classification of documents by this method can be adjusted according to the nature and needs [2].

The study used the credit card fraud registry, which consisted of 284,807 transactions made by credit card holders in Europe over a two-day period, obtained from the Kaggle dataset. The dataset information contained a very unbalanced data set, containing 492 fraudulent transactions, which represented only 0.172% of the 284,807 transactions. For some information about the characteristics of V1, V2, ..., V28 is the main component obtained with PCA. The "time" attribute contains the seconds elapsed between each transaction in the data log. Attribute "amount" is the number of transactions, this attribute can be used as a paid learning. The 'class' attribute is a response variable and takes the value 1 if fraud occurs and 0 is not fraud. The presentation of the problems faced in this study is a collection of data with unbalanced categories, which compares 99.80% of the major categories and 0.2% of the minor categories of the overall transactions that occurred. So that researchers will apply the method of Hotelling's $T^2$ and Naive Bayes Classifier (NBC) for classification and sought which method is best.

## II. MATERIAL AND METHODS

### A. Fraud Detection

Fraud is fraud committed in the presentation of a company's financial statements. In general, there are three things that encourage fraud, namely encouragement, opportunity, and justification for the actions taken (rationalization) [3]. According to Bank Indonesia a credit card is a card-based payment instrument that can be used to make payments for obligations arising from an economic activity, including shopping transactions and cash withdrawals, where the cardholder's payment obligations are met first by the acquirer or issuers and cardholders are obligated to make payments at the agreed time, either by payment in one lump sum or by payment in installments. Credit card fraud has two types, namely offline fraud and online fraud [4].

Forms of credit card fraud and crime can be in the form of phishing, skimming, carding, cracking, credit card theft, extrapolation, and telephone fraud. Credit card theft is an offline crime where criminals steal someone else's credit card, then the credit card is used to transact anywhere. The number of incidents of fraud and credit card crimes shows the public. This fraud affects and affects all parties, both banks and credit card holders.

### B. Stratified K-fold Cross Validation

K fold cross validation is used to estimate prediction error in evaluating model performance. The data is divided into nearly equal k subsets. Models in the classification were trained and tested as many as k. In each iteration, one of the subsets will be used as training data and testing data [5].

Stratified cross validation is a technique of separating or dividing data by ensuring that in training data and testing data there must be representatives from all classes with the same percentage. Stratified is done to ensure that each fold is a good representation of the data. The data sharing method using K-fold cross validation which is commonly used is not suitable when applied to classification problems with unbalanced data. This is because the distribution of data into K-folds has a uniform probability distribution so that one or more folds will have few or no examples from the minority class. [6]

### C. Hotelling's T2 Control Chart

Hotelling's $T^2$ diagram is a control chart that is used if in a control process together the average value of the sample in each observation with the characteristics that are examined more than one [7]. The Hotelling's $T^2$ control chart is used when two or more characteristics are technically dependent or suspected to be related. Hotelling's $T^2$ statistics for individual observations are obtained by the following formula.

$$T^2 = (X - \overline{X})' S^{-1} (X - \overline{X}) \quad (1)$$

Where :

$T^2$ : Hotelling's $T^2$ statistic value, X : the average value of the sample in each observation, and S is covariance matrix

In this study, the control chart used is the individual Hotelling's $T^2$ control chart. The control limit used is UCL (Upper Control Limit) and the value of LCL (Lower Control Limit) for normal data.

$$UCL = \frac{p(m+1)(m-1)}{m^2 - mp} F_{\alpha,p,m-p}$$

$$LCL = 0$$

But when the dataset that used is large m > 100. We need to use chi-squared distribution for UCL [2].

$$UCL = \chi^2_{\alpha,p}$$

$$LCL = 0 \quad (2)$$

### D. Naïve Bayes Classifier

Naïve Bayes Classifier is a classification method rooted in Bayes' theorem. The classification method using probability methods and naive Bayes classifier statistics predicts future opportunities based on previous experience. The Naïve Bayes Classifier uses a very strong assumption of independence from each condition or event. Where each of the instructions is independent of each other. With these assumptions, the following equation applies [8]:

$$P(H|X) = P(H)\prod_{i=1}^{n} P(X_i|H) \quad (3)$$

The Naïve Bayes Classifier algorithm is an algorithm used to find the highest probability value which is then classified into the most appropriate category [9]. The following are the stages of the algorithm of the nave Bayes classification method. The t variable is a set of keyword weight documents represented by the tm attribute for m=1,2,...,M where M is the total number of keywords. while yk is the set of categories.

1. Calculating the value of $P(y_k)$ on the training data using equation (4) below.

$$P(y_k) = \frac{|train_k|}{|train|} \quad (1)$$

Explanation

$P(y_k)$ = probability of k category transactions

$train_k$ = the number of training data category k

$train$ = the number of training

2. Each word probability of each category is calculated during training in equation (5).

$$P(t_m \mid y_k) = \frac{t_{mk} + 1}{\left| \sum_{m=1}^{M} t_{mk} + \sum_{m=1}^{M} \sum_{k=-1,1} t_{mk} \right|} \quad (2)$$

Explanation:
$t_{mk}$ = TF-IDF weights on the m variable categorized k; $m$ = 1,2,...,$M$; $k$ = -1, 1
$\Sigma t_{mk}$ = total weight of data category k
$\Sigma\Sigma t_{mk}$ = total weight of data

3. Classify into category groups by calculating the highest probability of the word tm in the yk category using equation (6).

$$Y_{MAP} = \underset{y_k = Y}{\arg\max}\, P(y_k)\prod_{m=1}^{M} P(t_m \mid y_k) \tag{3}$$

### E. Confussion Matrix

The test is carried out with a test tool, namely the Confusion Matrix to determine the correct distribution of the predicted data against the actual data [6]. The Confusion Matrix table is shown in the Table 1.

**Table 1** Confusion Matrix

| Label | Variables Name |
|---|---|
| *True Positive (TP)* | The number of positive that are considered positive |
| *True Negative (TN)* | The number of negative that are considered negative |
| *False Positive (FP)* | The number of positive that are considered negative |
| *False Negative (FN)* | The number of negative that are considered positive |

There are some performance goodness test of detection and classification data.

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

$$F1 = \frac{2\times Precision \times Recall}{Precision+Recall} \tag{9}$$

$$Accuracy = \frac{TP+FP}{TP+TN+FP+FN} \times 100\% \tag{10}$$

Meanwhile, for imbalanced data, the measurement of classification accuracy used is Area Under Curve (AUC). AUC is an indicator of ROC (Receiver Operating Characteristic) curve performance which can be summarized into a classifier into one value [10]. Here is the formula for calculating AUC :

$$AUC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{11}$$

### F. Experimental Work

The data source used is secondary data. The secondary data was taken from the Kaggle.com website dataset entitled Credit Card Fraud Detection. The dataset contains transactions made with credit cards in September 2013 by cardholders in Europe. The variables used in this study consisted of the response variable (Y) and predictor variable (X) which are presented in Table 2.

The data structure used in this study after text preprocessing is carried out which is presented in Table 3. The data structure used is xij, where i=1,…,n is the number of transactions, j=1, 2, …, k is the number of main components obtained by PCA.

The analytical steps used to achieve the objectives are as follows.

1. Pre-processing data
2. Dividing data on training and testing using stratified 10-fold cross validation.
3. Classify the data using $T^2$-Hotelling
4. Classify the data using Naive Bayes Classifier
5. Comparing classification goodness
6. Make conclusions and suggestions

**Table 2** Data Variables

| Variabel | Variables Name | Data Scale |
|---|---|---|
| *Y* | Category<br>0 = Normal Transaction<br>1 = Fraud Transaction | Nominal |
| *X1* | PCA 1 | Ratio |
| *X2* | PCA 2 | Ratio |
| *X3* | PCA 3 | Ratio |
| *X4* | PCA 4 | Ratio |
| *X5* | PCA 5 | Ratio |
| *X6* | PCA 6 | Ratio |
| *X7* | PCA 7 | Ratio |
| *X8* | PCA 8 | Ratio |
| *X9* | PCA 9 | Ratio |
| *X10* | PCA 10 | Ratio |
| *X11* | PCA 11 | Ratio |
| *X12* | PCA 12 | Ratio |
| *X13* | PCA 13 | Ratio |
| *X14* | PCA 14 | Ratio |
| *X15* | PCA 15 | Ratio |
| *X16* | PCA 16 | Ratio |
| *X17* | PCA 17 | Ratio |
| *X18* | PCA 18 | Ratio |
| *X19* | PCA 19 | Ratio |
| *X20* | PCA 20 | Ratio |
| *X21* | PCA 21 | Ratio |
| *X22* | PCA 22 | Ratio |
| *X23* | PCA 23 | Ratio |
| *X24* | PCA 24 | Ratio |
| *X25* | PCA 25 | Ratio |
| *X26* | PCA 26 | Ratio |
| *X27* | PCA 27 | Ratio |
| *X28* | PCA 28 | Ratio |
| *X29* | Amount | Ratio |
| *X30* | Time | Ratio |

**Table 3** Data Structure

| Transactions | Kelas (y) | PCA 1 ($x_1$) | PCA 2 ($x_2$) | … | Time ($x_{30}$) |
|---|---|---|---|---|---|
| 1 | $y_1$ | $x_{1,1}$ | $x_{1,2}$ | … | $x_{1,30}$ |
| 2 | $y_2$ | $x_{1,1}$ | $x_{2,2}$ | … | $x_{2,30}$ |
| ... | ... | ... | ... | ... | ... |
| $n$ | $y_n$ | $x_{n,1}$ | $x_{n,2}$ | … | $x_{n,30}$ |

## III. RESULTS AND DISCUSSION

### A. Data Characteristics

The response of "Credit Card Fraud Detection" dataset is consist of normal and fraud transactions with the total 284,807 transactions. The distribution of the response data can be shown at Figure 1.

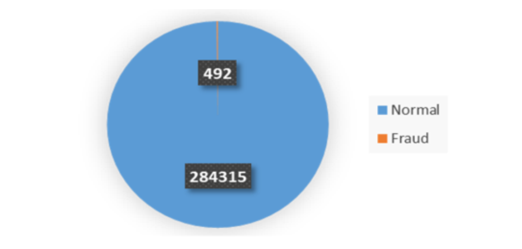


**Figure 1** Credit Card Transaction Status Data Distribution

Based on the Figure 1, this dataset have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the negative class (frauds) account for 0.172% of all transactions. Given the class imbalance ratio, confusion matrix accuracy is not meaningful. However measuring the accuracy using the Area Under Curve (AUC) is more recommended for the unbalanced classification.

### B. *Hotelling's $T^2$ Classification*

Hotelling's $T^2$ classification method is used to be able to observe quality and process control. The response variable used is a class classification variable in the form of normal or fraudulent transactions. While the predictor variables are the main components of the PCA results that have been categorized and have been described in Table 1. In the $T^2$ method, an alpha significance value is needed which will affect the control limits of this diagram. The following are the results of the $T^2$ Hotelling experiment using several alphas ranging from 0.1% to 50%.

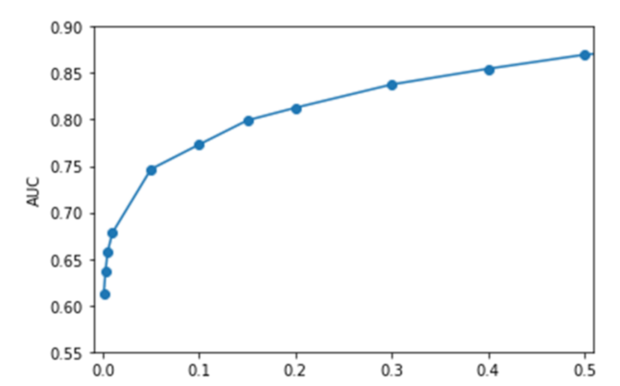


**Figure 1** Optimum Alpha Determination on $T^2$ Hotelling Method

Based on the graph, it is known that the greater the level of significance, the greater the value of the accuracy of the AUC so that the best alpha value is the value with the highest AUC, namely alpha = 50%. So the Hotelling's $T^2$ method this time will use a 50% alpha significance level. After confirming the alpha value, then the training and testing data sharing method is carried out using the stratified 10-fold cross validation method. The following is the accuracy of the classification results for each 10-fold using the Hotelling's $T^2$ method with alpha = 0.5.

**Table 4** AUC of Hotelling's $T^2$ Method

| Fold | AUC |
|---|---|
| 1 | 82.91 |
| 2 | 86.78 |
| 3 | 87.81 |
| 4 | 88.21 |
| 5 | 84.89 |
| 6 | 89.24 |
| 7 | 88.02 |
| 8 | 86.11 |
| 9 | 85.10 |
| 10 | 87.26 |
| Average | 86.633 |

Based on Table 4, it shows that the AUC of Hotelling's $T^2$ in the credit card fraud class is 86.633% in the testing data. The highest AUC testing value is in the 4th fold with an AUC value of 89.24%.

The determination of the classification in the Hotelling's $T^2$ method is based on the $T^2$ score and the Upper Control

Limit (UCL). If the $T^2$ value does not exceed the UCL value, then the transaction data is labeled positive or in control. Meanwhile, if the transaction data is above the UCL, then the transaction data is given a negative label or out of control. The following Table 5 is an example of some calculations of $T^2$ values on some testing data.

**Table 5** Hotelling's $T^2$ Statistics and Label

| Number of Testing data | $T^2$ Score | UCL | Label |
|---|---|---|---|
| 1 | 25 | 29.34 | Positive |
| 2 | 13 | 29.34 | Positive |
| 3 | 18 | 29.34 | Positive |
| 3 | 16 | 29.34 | Positive |
| 4 | 11 | 29.34 | Positive |
| ... | ... | ... | ... |
| 28478 | 43 | 29.34 | Negative |
| 28479 | 11 | 29.34 | Positive |
| 28480 | 21 | 29.34 | Positive |
| 28481 | 28 | 29.34 | Positive |

Here are the the result of Hotelling's $T^2$ control chart on testing data in Figure 2.

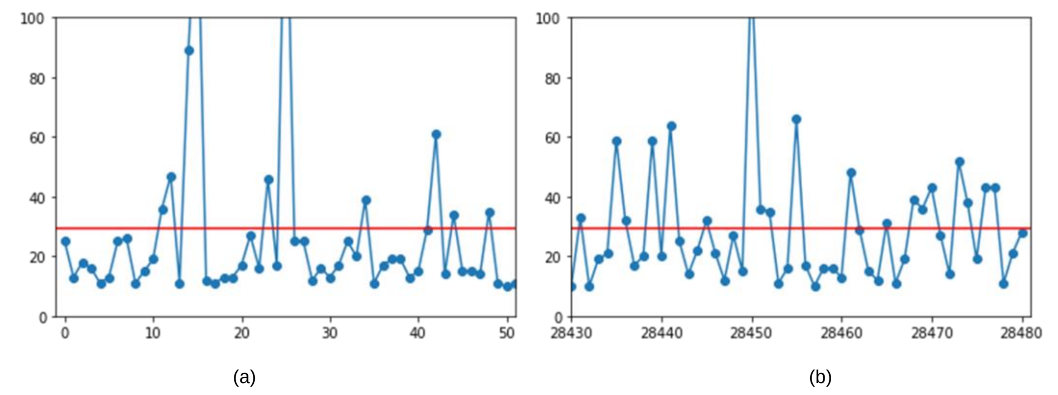


**Figure 2** Hotelling's $T^2$ Control Charts on First 50 (a) and Last 50 (b) of the Testing Data

Furthermore, measurement of classification goodness is carried out using a confusion matrix as shown in Table 6.

**Table 6** Confussion Matrix of Hotelling's $T^2$

| Actual | Prediction | |
|---|---|---|
| | **Positive** | **Negative** |
| Positive | 22.822 | 5549 |
| Negative | 1 | 39 |

Table 6 shows that the accuracy of $T^2$ classification for positive data that is classified as positive is 22822 data, while 5549 positive data are incorrectly predicted to be negative. Meanwhile, as many as 49 negative data were correctly classified into negative data and only 1 was incorrectly classified as positive. Based on the confusion matrix, the classification accuracy is obtained as follows.

**Table 7** Accuracy and AUC of Hotelling's $T^2$ Method

| Data | Accuracy | AUC |
|---|---|---|
| Testing | 80.51% | 89.24% |

Based on Table 7, it is known that the measure of classification accuracy on unbalanced data (balance) is AUC. In the Hotelling's $T^2$ method, the AUC classification performance value is consistent at 89.24% where almost all negative data can be predicted well.

### C. *Stratified K-fold Cross Validation*

Naïve Bayes Classifier method is used to find out the goodness of the classification results. The response variable or Y variable used is a class classification variable in the main components of the PCA results that have been categorized and described in Table 1. Based on the training and testing data sharing method, the 10-fold cross validation method is used. Here is the goodness of the classification results for each 10-fold.

Based on Table 8, it shows that the AUC of NBC in the credit card fraud class is consistent at 50% in both training data and testing data. Here are the example of NBC calculation by using the 1st fold.

**Table 8** AUC of NBC

| Fold | AUC |
|---|---|
| 1 | 50.00 |
| 2 | 50.00 |
| 3 | 50.00 |
| 4 | 50.00 |
| 5 | 50.00 |
| 6 | 50.00 |
| 7 | 50.00 |
| 8 | 50.00 |
| 9 | 50.00 |
| 10 | 50.00 |
| Average | 50.00 |

By using NBC, the YMAP value is obtained for each transaction to determine the label or prediction of that transaction. Here are the YMAP values and transaction categories from some testing data on Table 9.

**Table 9** NBC Calculation

| Data | Positive Probability | Negative Probability | Label |
|---|---|---|---|
| 1 | 0.9986 | 0.0014 | Positive |
| 2 | 0.9987 | 0.0013 | Positive |
| 3 | 0.9988 | 0.0012 | Positive |
| ... | ... | ... | ... |
| 284806 | 0.9988 | 0.0012 | Positive |
| 284807 | 0.9988 | 0.0012 | Positive |
| 284808 | 0.9989 | 0.0011 | Positive |

Based on Table 9, all the data was labelled by NBC into positive data. Furthermore, measurement of classification accuracy is carried out using a confusion matrix as shown in Table IX.

**Table 10** NBC Calculation

| Actual | Prediction | |
|---|---|---|
| | Positive | Negative |
| Positive | 28.432 | 0 |
| Negative | 49 | 0 |

Table 10 shows that all data are predicted to be in the positive class where 28432 are correctly predicted to be positive while 49 actual data are negative and incorrectly predicted to be positive. It can be said that this method is not suitable for detecting negative classes (outliers). Based on the confusion matrix, the classification accuracy is obtained as follows:

**Table 11** Accuracy and AUC of NBC Method

| Data | Accuracy | AUC |
|---|---|---|
| Testing | 99.83% | 50.00% |

Based on Table 11, it is known that the measure of classification accuracy on unbalanced data (balance) is AUC. In the NBC method, the AUC classification performance value is consistent at 50% because all data is predicted to be positive class so that no negative data is predicted correctly. So it can be concluded that the NBC method is not suitable for detecting outliers.

### D. *Classification Goodness Comparison*

After knowing the results of each classification goodness in the Hotelling's $T^2$ and Naïve Bayes Classifier (NBC) methods, the next step is to compare the results of the two methods. The following are the results of the comparison of those two methods based on the accuracy value in Table 7 and Table 11.

**Table 12** Goodness Comparison of Hotelling's $T^2$ and NBC Method

| Data | Average of AUC | |
|---|---|---|
| | NBC | $T^2$ Hotelling |
| Testing | 50.00% | 86.63% |

Table 12 shows the prediction goodness value of Hotelling's $T^2$ method is better in detecting outliers than the NBC method. Although the accuracy of Hotelling's $T^2$ prediction is not yet perfect, it is still slightly better than the NBC method which failed to detect a single outlier.

## IV. RESULTS AND DISCUSSION

Based on the results of the analysis and discussion, the following conclusions are obtained. The dataset is highly unbalanced with the negative class (frauds) account for 0.172% of all transactions. However, measuring the accuracy using the Area Under Curve (AUC) is more recommended for the unbalanced classification. The results of the accuracy of predictions using the Naïve Bayes Classifier (NBC) on the testing data obtained an average AUC value of 50% and failed to detect outliers. The results of the accuracy of predictions using Hotelling's on testing data obtained an average AUC value of 86.63% and good at detecting outliers even though with a fairly large level of significance. A comparison of the accuracy of NBC and Hotelling's $T^2$ predictions shows that the performance of the Hotelling's $T^2$ method is better in detecting outliers than the NBC method. Although the accuracy of Hotelling's $T^2$ prediction is not yet perfect, it is still slightly better than the NBC method which failed to detect a single outlier. Suggestions for the next research to use a robust estimator on Hotelling's $T^2$ so that it is expected to get a better classification model with a low level of significance. In addition, the Naive Bayes Classifier method is not suitable for detecting outliers for this dataset, so it is recommended to use another machine learning methods such as Support Vector Machine. Also, the MEWMA type of chart as in [11]-[13] can be considered.

## REFERENCES

[1] M. Fauzan, "Gaya Hidup Nasabah dan Keputusan Penggunaan Kartu Kredit," *Esensi: jurnal Bisnis dan Manajemen*, vol. 7, no. 2, pp. 181-192, 2017.

[2] B. Indonesia, "Edukasi Kartu kredit," 2020. [Online]. Available: https://www.bi.go.id/id/edukasi-perlindungan-konsumen/edukasi/produkdan-jasa-sp/kartu kredit/Contents/Default.aspx. [Accessed 10 October 2022].

[3] K. Chaudhary, J. Yadav and B. Mallick, "A review of Fraud Detection Techniques: Credit Card," *Internal Journal of Computer Applications*, vol. 45, no. 1, pp. 39-44, 2012.

[4] Nurhayati, I. Soekarno and M. Cahyono, "A Study of Hold-Out and K-Fold Cross Validation for Accuracy of Groundwater Modeling in Tidal Lowland Reclamation Using Extreme Learning Machine," in *2nd International Conference on Technologu, Informatics, Management, Engineering & Environment*, 2014.

[5] R. Kohavi, G. John, R. Long, D. Manley and K. Peger, "MLC++: A machine learning library in C++, in \Tools with Articial Intelligence," *IEEE Society Press*, 1994740-743.

[6] J. Han, M. Kamber and J. Pei, Data Mining Concepts, USA: Morgan Kaufmann, 2012.

[7] M. Erpianti, "ANALISIS PENGENDALIAN KUALITAS DATA DENGAN METODE T2 HOTELLING INDIVIDUAL," *Buletin Ilmiah Math, Stat, dan Terapannya*, vol. 09, no. 03, pp. 431-436, 2020.

[8] D. Spiegelhalter and K. Rice, "Bayesian Statistics," *Scholarpedia*, vol. 4, no. 8, p. 5230, 2006.

[9] R. Feldman and J. Sanger, The Text Mining Handbook:Advanced Approaches in Analyzing Unstructured Data, New York: Cambridge University Press, 2007.

[10] M. Bekkar, H. K. Djemma and T. Alitouche, "Evaluation Measure for Models Assesment over Imbalanced Data Sets," *Journal of Information Engineering and Applications*, vol. 3, pp. 27-38, 2013.

[11] N. Sulistiawanti, M. Ahsan, and H. Khusna, "Multivariate Exponentially Weighted Moving Average (MEWMA) and Multivariate Exponentially Weighted Moving Variance (MEWMV) Chart Based on Residual XGBoost Regression for Monitoring Water Quality.," Eng. Lett., vol. 31, no. 3, 2023.

[12] H. Khusna, M. Mashuri, S. Suhartono, D. D. Prastyo, and M. Ahsan, "Multioutput least square SVR based multivariate EWMA control chart: The performance evaluation and application," Cogent Eng., Oct. 2018, doi: 10.1080/23311916.2018.1531456.

[13] H. Khusna, M. Mashuri, Suhartono, D. D. Prastyo, and M. Ahsan, "Multioutput Least Square SVR Based Multivariate EWMA Control Chart," J. Phys. Conf. Ser., vol. 1028, no. 1, p. 12221, 2018.