

Two Stages Outlier Removal as Pre-processing Digitizer Data on Fine Motor Skills (FMS) Classification Using Covariance Estimator and Isolation Forest

by Nurul Fanani

Submission date: 26-Aug-2021 01:28AM (UTC+0700)

Submission ID: 1635876626

File name: tlier_Removal_as_Pre-processing_Digitizer_Data_on_Fine_Motor.pdf (1,020.48K)

Word count: 7225

Character count: 38039



Two Stages Outlier Removal as Pre-processing Digitizer Data on Fine Motor Skills (FMS) Classification Using Covariance Estimator and Isolation Forest

Nurul Zainal Fanani^{1,2,3*} Adri Gabriel Sooi⁴ Khamid⁵ Festa Yumpi Rahmanawati⁶
 Alex Tormasi⁷ Laszlo T. Koczy^{7,8} Surya Sumpeno^{1,2} Mauridhi Hery Purnomo^{1,2,9*}

¹Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Indonesia

²Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Indonesia

³Department of Engineering, Politeknik Negeri Jember, Indonesia

⁴ Department of Computer Science, Universitas Katolik Widya Mandira, Kupang, Indonesia

⁵Department of Informatics Engineering, Universitas Wahidiyah, Indonesia

⁶Department of Psychology, Universitas Muhammadiyah Jember, Indonesia

⁷Department of Informatics Széchenyi István University, Győr, Hungary

⁸Department of Telecommunications and Media Informatics,

University of Technology and Economics, Budapest, Hungary

⁹University Center of Excellence on Artificial Intelligence for Healthcare and Society (UCE AIHeS), Indonesia

* Corresponding author's Email: hery@ee.its.ac.id

Abstract: The increase of the classification accuracy level has become an important problem in machine learning especially in diverse data-set that contain the outlier data. In the data stream or the data from sensor readings that produce large data, it allows a lot of noise to occur. It makes the performance of the machine learning model is disrupted or even decreased. Therefore, clean data from noise is needed to obtain good accuracy and to improve the performance of the machine learning model. This research proposes a two-stages for detecting and removing outlier data by using the covariance estimator and isolation forest methods as pre-processing in the classification process to determine fine motor skill (FMS). The dataset was generated from the process of recording data directly during cursive writing by using a digitizer. The data included the relative position of the stylus on the digitizer board. x position, y position, z position, and pressure values are then used as features in the classification process. In the process of observation and recording, the generated data was very huge so some of them produce the outlier data. From the experimental results that have been implemented, the level of accuracy in the FMS classification process increases between 0.5-1% by using the Random Forest classifier after the detection and outlier removal by using covariance estimator and isolation forest. The highest accuracy rate achieves 98.05% compared to the accuracy without outlier removal, which is only about 97.3%

Keywords: Covariance estimator, Isolation forest, Outlier detection, Random forest, Fine motor skill.

1. Introduction

FMS plays an important role in the development of preschool and primary school children [1]. FMS is associated with the student's academic skills such as reading, mathematical reasoning [2], and writing [3]. Writing is a basic skill that is taught in the early cycle of education [4]. Writing requires FMS performance as in stabilizing paper positions, use of hands for

holding a pencil, and coordination between the eyes and fingers [5].

It is important to use FMS as an indicator of children's school readiness [6]. School readiness is a condition in which a child is ready to be involved in the learning experience at school. Several studies have stated that the level of FMS can be determined by children's writing abilities [7]. Through writing exercises, especially cursive handwriting, children can practice

Table 1. Related work in Outlier/Anomaly Detection using Isolation Forest

No	Author	Outlier / Anomaly Detection	Model	Data Type
1	Y.Qin [13]	Anomaly Pattern Detection	Isolation Forest	Hydrological time series dataset (Chuhe River Basin)
2	G. A. Susto [14]	Anomaly Detection	Isolation Forest	Etching real industrial dataset
3	Z.Chen g [15]	Outlier Detection	Isolation Forest and Local Outlier Factor	6 case synthetic and 6 cases real-world dataset
4	D.Xu [16]	Anomaly Detection	Isolation Forest and Local Outlier Factor	7 case public dataset
5	V. Verardi [17]	Outlier Detection	MCD	Stata dataset
6	W.S.L. Wah [18]	Anomaly Detection	Isolation Forest	Civil engineering dataset
7	F.T.Liu [19]	Anomaly Detection	Isolation Forest	Synthetic dataset
8	T.D. Nguyen [24]	Outlier Detection	Covarian Estimator	Synthetic dataset from FICM and FQCM
9	L. Puggini [25]	Anomaly Detection	Isolation Forest with Dimensionality reduction	Optical Emission Spectros copy (OES) dataset
10	Hyder Jhon [26].	Anomaluy Detection	Isolation Forest	Credit card transactions in Europe
11	N.Z. Fanani	Outlier Removal	Covariance Estimator and Isolation Forest	Real-world – Digitizer Sensor dataset

their FMS with maximum intensity [8]. These exercises force the students to think about all the details of writing. Cursive handwriting is a form of writing characterized by the relationship between letters and the relationship between syllables and word forms [9]. Besides, the results of research conducted by Comajuncocas revealed that cursive handwriting can increase the speed of the student's performance in writing and drawing [10]. Thus, schoolteachers conduct FMS assessments from children's writing using a digitizer. The digitizer records the movement of the nib and records the position of the nib [11].

This study intends to assist school teachers in conducting FMS assessments of children's writing by using a digitizer to record information on children's hand movements during the one-sentence cursive writing process. However, this paper emphasizes the explanation of the initial processing of the FMS classification process. This initial processing is important to do because the results of the recording of the writing process produce a data stream that contains a lot of outlier data.

The research to determine the level of FMS in children has been carried out previously with the process of recording data based on hanacaraka letters by using a digitizer [12]. The result of recording the data stream from this digitizer does not produce a lot of noise (outlier data) because it only records one hanacaraka letter. In this study, the data recording process using a digitizer produced many data outliers because the recorded writing process was the process of recording one sentence with continuous cursive handwriting. So, to increase the level of classification accuracy, pre-processing is needed to remove the outlier data.

The research on the handling and detection of outlier data has been carried out by using various methods including using the covariance estimator and isolation forest. The previous research using the Isolation Forest algorithm in detecting outliers in hydrological time series data conducted by Qin [13]. Qin study shows that the isolation forest algorithm has a good performance in detecting outliers in high-dimensional data. Another study compared the isolation forest algorithm with a diagrammatic approach to detect the presence of outliers in industrial data. Based on the precision and recall, the isolation forest algorithm has better performance than the diagram approach [14]. Several other studies are summarized in Table 1.

Research [15] and [16] prove that the isolation forest algorithm can significantly increase the outlier detection speed and computation time on both synthetic and real world datasets. Covariance estimator is sensitive to the detected outliers so that it

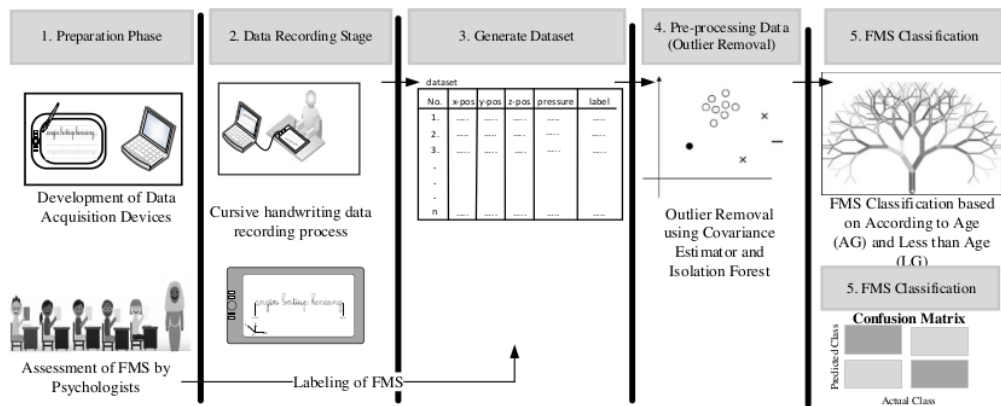


Figure. 1 The FMS Classification System with Outlier Removal

can cause a masking effect [17]. The masking effect occurs when the outliers are not detected due to the presence of other nearby outliers in the same data set [18]. To reduce the masking effect on the covariance estimator algorithm, it can be done by adding the isolation forest algorithm. The isolation forest algorithm detects outliers by partitioning the data into small groups so that each set can detect different outliers [19]. The purpose of this research is to improve the classification accuracy level of the digitizer data stream. The data stream overlaps with two stages of detection and disposal of outliers by using a covariance estimator and isolation forest. After the pre-processing stage is carried out by using these two methods, the classification process is carried out by using the Random Forest method. This method is suitable for this study because the data parameters generated from this digitizer sensor have a high correlation and are interrelated. Therefore, we need a method that uses all these parameters in the classification process. This is following the random forest method. The mechanism of the method is to create many trees from random parameters and produce the best model from the combination of these parameters.

2. Two stages outlier removal using covariance estimator and isolation forest

The proposed and applied process in this research can be seen in Fig. 1. Overall, the experimental stages consisted of six parts, namely the preparation phase, data recording stage, generate the dataset, outlier removal, FMS classification, and evaluation system.

Experiment Scenario

The initial stage in the preparation phase is the development of a data recording system. This section

consists of computers and software used to read data from digital boards and store the data in a database. The software is developed using C# language. The developed application has several features; recording information on student data, school data, age, gender. This app is also capable of reading data position, also stylus pen pressure data against the digital board in real-time. The data acquisition process is carried out during the writing process by using a digitizer, called WACOM Cintiq 13HD. The writing process is recorded by the application in real-time with a sampling frequency of 220 Hz. The data that can be obtained from this digital board is the value of $x\text{-pos}$ which has a value between 0-1365, $y\text{-pos}$ has a value between 0-767, $z\text{-pos}$ has a value between 0-1023, and pressure (p) has a value between 0-1023. This specification is needed to obtain observations in the form of x , y , z , and depth or pressure. These four features play an important role in obtaining an observation as detail as possible in anticipating the completeness of the acquisition results data later. The second required device is a PC or laptop to receive all recorded data of acquisition result by using Wacom stylus. The basis for selecting these two devices is based on the minimum specifications used by [20]. In addition to direct tools for data acquisition, the participant's ergonomics in recording data was also prepared, which aims to provide comfort and reduce student's stress levels in doing cursive handwriting. This ergonomic determination is following the model developed by [21].

In this preparation stage, a psychologist tests participants to determine the condition of their fine motor skills which will be used as a comparison in this study. A label of FMS class from a psychologist shows the condition of a student's FMS based on age in two categories, namely FMS according to age (AG) and FMS less than age (LG). From this activity, the

Table 2. Visualization sample data recorded from the digitizer

Participant	Data Visualization	class
1		AG
2		AG
3		LG
4		LG

psychologist will evaluate the condition of the student's FMS through hand strokes

Second, after all, hardware equipment has been prepared and calibrated, the next step is to determine the flow of data recording scenario to be observed. The stages are as follows: compiling a sentence that can represent all aspects of cursive handwriting. The aspects that must be involved are the representation of up, down, horizontally, connecting letters, and so on. This is important to prepare so that the dataset contains observations that can represent various events in cursive handwriting. The basis for composing sentences that can represent various aspects of cursive handwriting is based on research [22].

Third, the next process is data recording. In this process, there are three stages to be done. a) Introducing the use of data acquisition tools. This must be done to equalize basic skills in using tools. Participants or students are guided on how to use the device and try short words. After the participants are proficient in using it according to the prepared threshold, then the actual recording process is applied. b) Each student who records cursive handwriting using WACOM writes the start and end times. The note-taking process occurs automatically when the stylus rubs against the surface of the electronic whiteboard. The data to be recorded and become observations consists of no, timestamp, x, y, z, p. The x, y, z are the coordinates of the stylus to the electronic keyboard, while p is the pressure generated by the stylus. Each student's data is separated and will be labeled according to the arrangements that have been prepared, namely the various classes. The recording process was implemented for a total of 99 participants. c) Still in the process of data recording, after all those 99 participants did cursive handwriting, the data was cleaned and grouped according to labels. This is known as the Ground Truth process or FMS labeling. This process is based on previous research that has also been carried out by [12].

The fourth stage is the initial processing of the dataset obtained from digitizer sensor reading. The label obtained from psychology is combined with the data obtained from the digitizer, so that the data visualization is in Table 2. The color on the graph shows the pressure at position (x, y). The green color indicating the pressure at position (x, y) is high pressure. The blue color indicates medium pressure and the red color indicates low pressure. Meanwhile, most of the values in the z-pos variable are 0. The visualization in Table 2, also shows that the participants made several mistakes when did writing cursive. These mistakes are plotted as random patterns outside of the letters that form the sentence "angin bertuip kencang". This random pattern will be

detected and removed using our proposed method. To produce clean data to obtain a high-level of accuracy, we propose two preprocessing stages by using the covariance estimator and isolation forest as methods for detecting and removing data outliers.

The covariance estimator (CE) plays an important role in data processing [23]. Classically, the covariance matrix is estimated using the maximum likelihood estimator by assuming the data following the multivariate normal distribution. The Maximum Likelihood Estimation (MLE) looks for the parameters (μ, Σ) which maximizes the probability of obtaining sample data R as shown in Eq. (1) [24].

$$MLE = \max_{\mu, \Sigma} \sum_{i=1}^n \log[f(R_i | \mu, \Sigma)] \quad (1)$$

Under the assumption of normality, solutions form closed (μ_e, Σ_e) is denoted as Eq. (2) and Eq. (3):

$$\mu_e = \sum_{i=1}^n \frac{1}{n} R_i \quad (2)$$

$$\Sigma_e = \sum_{i=1}^n \frac{1}{n} R_i^t R_i - \mu_e^t \mu_e \quad (3)$$

Isolation Forest is a tree-based model development used to detect outliers [25]. The basic idea of this method is to consider outliers as few and different data [26]. The isolation forest isolates the data that are considered as outliers by studying the characteristics of the data and measuring the susceptibility of the data to isolation [14]. The outliers were isolated by partitioning the spaces in the parameter into sections that were scored according to their susceptibility to isolation [27]. The isolation forest steps are as follows for given data set $X = \{x_1, x_2, \dots, x_w\}$ with X is the attributes [28]:

1. Choosing a random dimension in the data
2. Randomly selecting values in the dimensions
3. Drawing a straight line through the values in the selected data and separate the data.
4. Repeating steps 1 - 3 until the number of trees is specified
5. Calculating the anomaly value using Eq. (4)

$$c(w) = \begin{cases} 2H(w-1) - \frac{2(w-1)}{n} & \text{for } w > 2 \\ 1 & \text{for } w = 2 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

4

where $H(i)$ is the harmonic number in the range $\ln(i) + 0.5772156649$ (Euler's constant), $c(w)$ is

<p>Algorithm 1 Pseudocode for Two Stages Outlier Removal</p> <p>Input : X_{train} : Train Dataset ; X_{test} : Test Dataset; Y_{train} : Label for training dataset; Y_{test} : Label for testing dataset; ; n_{tree}: number of trees generated in random forest. Ouput : M : Model, Accuracy, Recall, Precision, Specificity</p> <p>Pseudocode for Building the Model</p> <ol style="list-style-type: none"> 1. $X_{train} = CovarianceMatrix(X_{train})$ 2. $X_{train} = IsolationForest(X_{train})$ 3. $M = RandomForestTraining(X_{train}, Y_{train}, n_{tree})$ 4. RETURN Model, Accuracy, Precision, Specificity <p>Pseudocode for FMS Classification with Two Stages Outlier Removal</p> <ol style="list-style-type: none"> 1. $X_{test} = CovarianceMatrix(X_{test})$ 2. $X_{test} = IsolationForest(X_{test})$ 3. $Y_{test} = FMSClassification(M, X_{test})$ 4. Accuracy = calculateAccuracy(Y_{test}^*, Y_{test}) 5. Recall= calculateRecall(Y_{test}^*, Y_{test}) 6. Precision = calculatePrecision(Y_{test}^*, Y_{test}) 7. Specificity = calculateSpecificity(Y_{test}^*, Y_{test})
--

the average of $h(x)$ which is used to normalize $h(x)$. An anomaly score is defined in Eq. (5):

$$s(x, w) = 2^{-\frac{E(h(x))}{c(w)}} \quad (5)$$

$E(h(x))$ is the average $h(x)$ of a collection of trees.

- $E(h(x)) = 0, s = 1$
- $E(h(x)) = w - 1, s = 0$
- $E(h(x)) = c(w), s = 0$

The following conditions give the specific value of the anomaly score.

the range of values is $0 < s \leq 1$ and $0 < h(x) \leq w - 1$, if s approaches 1 it is considered an anomaly, if less than 0.5 is considered normal, and if $s \approx 0,5$, then the whole sample does not really have a different anomaly.

The characteristics of the data stream from the cursive-writing reading using this digitizer produce data with a high enough density and a fairly closely related feature which is the reason for using this covariance estimator and isolation forest method. In this study, the combination of these two methods was carried out at the initial processing to remove outlier data to increase the level of accuracy.

In this experiment, covariance estimator and isolation forest will be used. These two anomaly detectors will be arranged sequentially before being

Table 3. Confusion matrix

Actual	Classification	
	+	-
+	True Positive (TP)	False Negatives (FN)
-	False Positive (FP)	True Negative (TN)

forwarded to the training or modeling process. The anomaly detection process by the two algorithms arranged sequentially will use certain parameters, such as what percentage of anomalies will be discarded. In the following process, there are two main experiments carried out, namely the classification without detecting anomalies and using two stages outlier removal with pseudocode can be seen in Algorithm 1.

In the experiment without anomaly detection, several classifiers will be used, namely, Random Forest, k-NN, Naïve Bayes, Neural Network, and SVM. Each classifier will use the same parameters in both experiments to maintain the uniformity of the classifier parameters. The Random Forest classifier will use the parameter number of trees 10, 20, and 30. Random Forest is a machine learning method for classifying tree-based data [29]. Random forest is also known as decision tree cumulative so it is called "forest" [30]. Each tree unit in the forest gives a vote to determine the most likely class label [31]. In the computation process, the random forest method does not require a long time and can recognize non-linear data patterns [32]. The random forest also uses regression methods and the task of constructing multiple decision trees [33]. With the following formula as shown in Eq. (6):

$$\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(x') \quad (6)$$

The k-NN classifier will use the parameter number of neighbors (k) 2, 3, 4, and 5 with the Euclidean metric, weight distance. The Neural Network classifier uses the parameter number of neurons in hidden layers: 100, ReLu for the activation function, Adam for the Solver function; and the maximum iterations: 100. The SVM classifier was tested on all types of existing kernels, namely Linear, Polynomial with the value g: auto, c: 0, d: 3, RBF with a value of g: auto, and Sigmoid with a value of g: auto and c: 0. With the cost and regression loss epsilon respectively 1 and 0.1. All classifiers will use the same k-fold cross-validation, namely 10-fold cross-validation. After the experimental process without anomaly detection, the experiment with

Table 4. Testing with the Random Forest method

Random Forest Trial			
Trees	Accurate	Recall	Specificity
10	97.85%	98.19%	97.52%
20	97.95%	98.19%	97.71%
30	98.05%	98.29%	97.81%

Table 5. Testing with the k-NN method

K-NN trial			
Total k	Accuracy	Recall	Specificity
k=2	96.30%	97.15%	95.48%
k=3	96.30%	96.86%	95.75%
k=4	96.25%	96.86%	95.66%
k=5	96.15%	96.66%	95.65%

Table 6. Testing with the SVM method

SVM Trial			
Kernel	Accurate	Recall	Specificity
linear	51.05%	51.16%	50.96%
polynomial	49.60%	49.55%	49.64%
rbf	48.30%	48.15%	48.43%
sigmoid	53.70%	54.13%	53.35%

anomaly detection was done. The training set is denoted as X . The response is denoted by Y , bagging repetition is denoted by (B iteration). The amount of training data is denoted as n . Samples with replacement contents are denoted by X_b, Y_b . The regression tree is denoted as f_b in X_b, Y_b . After the training process, the prediction is denoted as x' .

Fifth, the next stage is to carry out the data training process according to the scenario prepared in the fourth step. Here, the dataset will be trained with labels that have been determined in the ground truth or FMS labeling process. The whole process is carried out sequentially and recorded in the form of a confusion matrix tab. In addition to recording and tabulation of confusion matrix results from the classification process, it is used to calculate parameters such as Accuracy, Recall, and Sensitivity. This confusion matrix is useful for measuring how well the classification model has been made [34]. The confusion matrix is $n \times n$ in size, where n is the number of different classes [35]. The confirmation matrix can determine the accuracy, recall, and specificity obtained from the values of several parameters, such as True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) [36, 37]. The confusion matrix table is shown in Table 3.

After knowing the values of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), the results of accuracy, Eq. (7) to Eq. (9).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$Specificity = \frac{TN}{TN + FP} \quad (9)$$

Sixth, the final process of the whole research is to compare all the results planned in the fourth stage and the implementation in the fifth stage. This process will provide information about the hypothesis that has been stated before the research is done, namely: Anomaly detection is expected to significantly improve the classification performance, which results in the availability of fine motor skill classification model from cursive handwriting.

The data sample in Table 2 has several outlier data that can affect the classification results so that the outlier data needs to be removed. TSOR could be implemented to handle those outliers. After handling the outlier data, testing the classification method was also implemented. The classification method used is random forest, k-NN, Naïve Bayes, NN, and SVM. The classification system would be evaluated to determine the accuracy, recall, and specificity values.

3. Experiment result and analysis

FMS classification uses trials of several classification methods and the handling of outliers in the data. The data were obtained from elementary school students cursive writing in grade 2 based on the value of x-pos, y-pos z-pos, and its pressure, where x-pos, y-pos z-pos are the positions of the writing with coordinates x, y, z on the digitizer and p axes is the pen pressure on the digitizer in writing. The data are categorized into two classes, namely FMS according to age (AG) and FMS less than age (LG).

The x-pos and y-pos data samples form the words "angin bertiup kencang" in AG class. The data in Fig. 2 is employed to detect outlier, while the data in Fig. 3 is a graph of the data after removing the outliers by using the covariance estimator method.

Covariance Estimator is a method that looks at the outlier data based on the covariance value. Covariance assesses the outlier data based on the distribution of the data or the standard deviation of the overall data. Based on the Y-axis coordinates, the data shown in Fig. 2 have a center of distribution between 260 to 320. The data that exceed this interval include the data exceeding the distribution of data in general. The data have the potential to become data outliers. Outliers in Fig. 2 shows the data that exceed

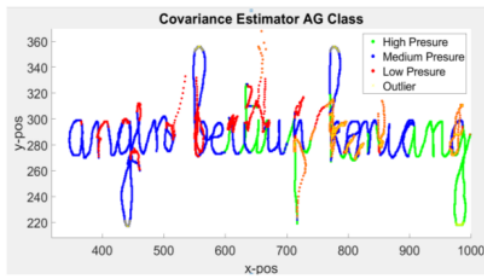


Figure. 2 Outlier detection using covariance estimator

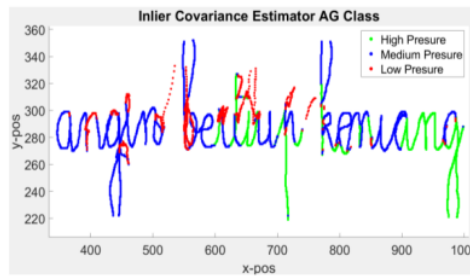


Figure. 3 Outlier remover by covariance estimator method

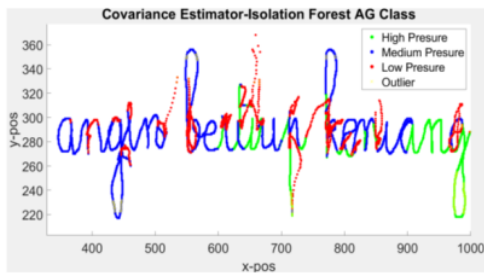


Figure. 4 Outlier detection using covariance estimator-isolation forest

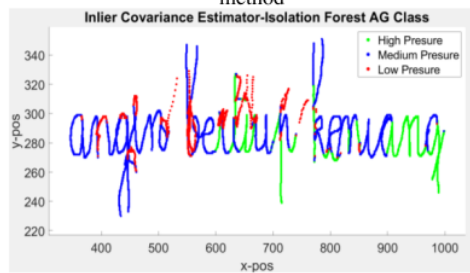


Figure. 5 Outlier remover by covariance estimator-isolation forest

the estimate or tolerance of the covariance that has been obtained. In addition to looking at the Y-axis, the outlier data assessment depends on the X-axis coordinates. The data assessment based on the X-axis is different from the assessment seen from the Y-axis. Each letter group has a different covariance and covarian estimate. For example, a collection of the letter g in the word “kembang” has a distribution of points whose coordinates are far from a normal set of points. Based on the outliers that have been detected in Fig. 2, new data is obtained in Fig. 3

Based on Fig. 4, the isolation forest detects outliers based on its partition tree so that it can detect outliers in each partition. The isolation forest method can detect more outliers than the covariance estimator method. Based on the distribution of the outlier graph, most of the detected outliers are in the range y 420 to 480. The addition of the isolation forest method to detect outliers in inlier data by the covariance estimator method in Fig. 5 shows that the isolation forest method can detect outliers that are not detected by the covariance estimator method so that the resulting inlier data can represent the characteristics of the data on each label well and produce a high accuracy value in the classification process.

The initial data were about 392923, which were then applied pre-processing to remove data that were considered as outliers. The outlier handling was done by using Covariance estimator and isolation forest sequentially.

After handling the outlier data, the data were classified using several methods, namely random forest, k-NN, Naïve Bayes, NN, and SVM. The results of the accuracy of each experiment at those methods are shown in Tables 4 to 6.

Based on [table 4](#) with the Random Forest classification, it can be seen that the best accuracy results are obtained with the number of trees equal to 30. The number of accuracies, recall, and specificity reaches 98.05%, 98.25%, 97.81%.

The k-NN classifier as seen in [Table 5](#), produces the best accuracy at k = 3 with an accuracy of 96.3%, a recall result of 96.86%, and a specificity of 95.75%. While testing by using the Neural Network classifier, the best results were obtained on the number of hidden layers 100 with an accuracy of 77.32%, 73.72% recall, and 80.04% specificity as seen in [Table 7](#). While in [Table 6](#) with the SVM classifier, the best accuracy results were obtained in the sigmoid kernel with the results of 53.7% accuracy, 54.13% recall, and 53.35% specificity.

The accuracy results of the outlier handling testing and classification using several methods are shown in [Table 7](#). From [Table 7](#), it can be seen that the best accuracy results by using the combination of handling outliers between Covariance Estimator and Isolation Forest obtained the best results in the Random Forest method with an accuracy value of 98.05% with the difference between the best accuracy results compared to the accuracy results without

Table 7. The classification accuracy results from handling outlier data by several methods

Handling outliers Test	Number of raw data	Number of preprocessed data	Random Forest	k-NN (k=3)	Naïve Bayes (NB)	Neural Network (NN)	SVM
Covariance Estimator [24]	392952	373304	97.80%	95.80%	61.40%	74.30%	45.80%
Isolation Forest [19]	392952	353661	97.80%	95.70%	60.60%	73.30%	46.20%
Covariance Estimator+Isolation Forest	392952	335975	98.05%	96.30%	61.90%	77.32%	53.70%
Without Handling outlier [12]	392952	392952	97.30%	95.30%	61.50%	74.30%	55.70%
Average			97.83%	95.78%	61.35%	74.81%	50.35%

handling outliers of 1.2%. The highest accuracy value in the k-NN method with k equal to 3 is 96.3%. The difference between the highest accuracy and the accuracy without handling outliers is 1%. Whereas in the naïve Bayes and neural network methods, the highest accuracy obtained was 61.5% and 77.32%. The difference in accuracy results without the handling of outliers and the highest accuracy results in the SVM method achieved 2%. However, in the SVM method, the accuracy results without handling outliers are higher than the accuracy results with Covariance Estimator and Isolation Forest outlier handling methods. Whereas in the SVM method, the accuracy results without outlier handling were higher than the accuracy results with Isolation Forest outlier handling methods.

The test results by using one type of Handling outlier in the form of Covariance Estimator or Isolation Forest contributed to the increased accuracy, especially in the Random Forest and k-NN methods with an increase in accuracy ranging from 0.4% to 0.5%. In testing by using one type of Handling outlier, the Naïve Bayes, Neural Network, and SVM classifiers did not increase in accuracy, even decreased the level of accuracy.

Based on Table 7, the outlier handling by using hybrid covariance estimator-isolation forest can remove more outliers than other methods as many as 56977 outliers. With a smaller amount of data compared to other methods, the hybrid covariance estimator-isolation forest (CEIF) can produce the highest accuracy value on the random forest classification method, k-NN, Naïve Bayes, and Neural Network.

Based on the accuracy values in Table 7, the CEIF outlier handling method has a good performance compared to other methods. Therefore, a comparison of the CEIF method as an outlier handling with other sequential data is carried out. It is based on the accuracy results in the classification process by using the RF, KNN, NB, NN, SVM methods [39]. The

comparison of the accuracy result of several methods based on music and handwriting data is shown in Fig. 6.

Based on Fig. 6, the classification method works better after the outlier handling was done by using the CEIF method. In data [38] which is sequential EEG data, the classification process with the addition of CEIF method as outlier handling is higher than without applying outlier handling. The use of the CEIF method is more suitable for cursive handwriting data. This is because data retrieval at the recording stage using a digitizer still has many outliers which results in the classification model being unable to recognize data patterns in each class.

The implementation of the isolation forest algorithm has several weaknesses in certain datasets. The isolation forest uses only a subset of dimensions to form an isolation tree. Each subset usually cannot detect the presence of outliers in low dimensions. The weakness of isolation forest can be overcome with the iNNE algorithm. The iNNE algorithm uses ensemble size as the main parameter. A large ensemble size will produce a variety of isolation models so that it can detect outliers in high and low dimensions. It is expected that further research can implement the iNNE algorithm to obtain better accuracy results [27]. For further research, other methods such as SOTA [39], One-Class SVM [40] and LSTM [41] which have been tested in data with sequence characteristics will be used to detect outliers in the FMS data stream from the digitizer sensor.

4. Conclusion

This study reveals that the handling of outlier data by using the Covariance Estimator and Isolation Forest that we propose is successful in increasing the accuracy of Fine Motor Skill classification by 1.2% compared to the use without handling outlier data. The increase in the highest level of accuracy is obtained from the classification using the Random

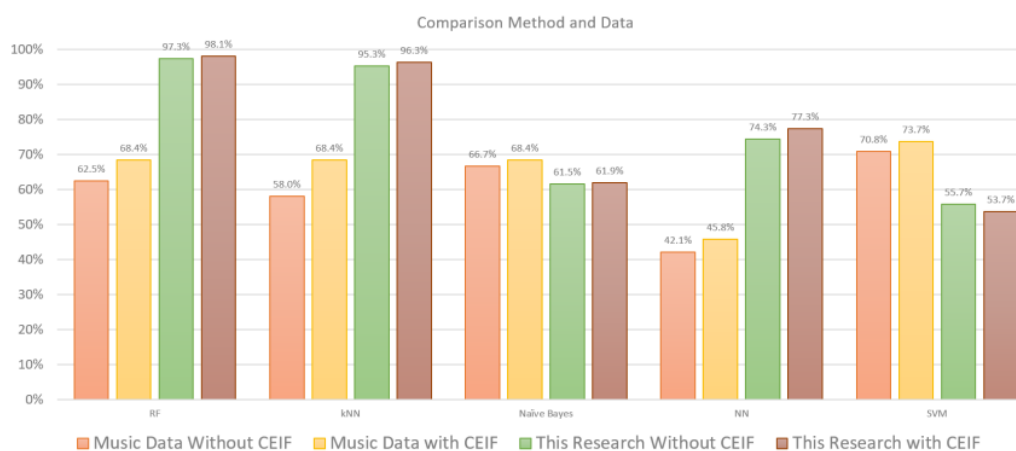


Figure. 6 Graph of comparison data

Forest method. Handling outlier data using one method, namely the covariance Estimator method only or Isolation Forest alone, cannot significantly improve the accuracy of the Neural Network and Naïve Bayes classification methods. However, there is still an increase in the classification process when using a combination of two methods of handling outlier data, namely the covariance Estimator and Isolation Forest.

2

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, Nurul Zainal Fanani; methodology, Nurul Zainal Fanani, Adri Gabriel Sooi and Alex Tormasi; software, Nurul Zainal Fanani and Khamid; validation, Nurul Zainal Fanani, Festa Yumpi Rahmanawati, Surya Sumpeno, and Mauridhi Hery Purnomo; formal analysis, Nurul Zainal Fanani, Laszlo T. Koczy, Surya Sumpeno, and Mauridhi Hery Purnomo; investigation, Nurul Zainal Fanani; resources, Nurul Zainal Fanani, Khamid and Adri Gabriel Sooi; data curation, Nurul Zainal Fanani, and Alex Tormasi; writing—original draft preparation, Nurul Zainal Fanani; writing—review and editing, Nurul Zainal Fanani; visualization, Nurul Zainal Fanani; supervision, Surya Sumpeno, and Mauridhi Hery Purnomo; project administration, Nurul Zainal Fanani. All authors read and approved the final manuscript.

Acknowledgments

The research for this paper was supported by the Indonesian Ministry of Education and Culture

through Doctoral Dissertation Research Grant and Enhancing International Publication Program. The lead and the corresponding author acknowledges the financial support funded by the Science and Technology Center of Artificial Intelligence for Healthcare and Society (PUI AIHeS), Indonesia.

References

- [1] S. Suggate, E. Pufke, and H. Stoeger, “Do fine motor skills contribute to early reading development?”, *J. Res. Read.*, Vol. 41, No. 1, pp. 1–19, 2018.
- [2] U. Fischer, S. P. Suggate, J. Schmir, and H. Stoeger, “Counting on fine motor skills: links between preschool finger dexterity and numerical skills”, *Dev. Sci.*, Vol. 21, No. 4, p. e12623, 2018.
- [3] P. Martzog, H. Stoeger, and S. Suggate, “Relations between Preschool Children’s Fine Motor Skills and General Cognitive Abilities”, *J. Cogn. Dev.*, Vol. 20, No. 4, pp. 443–465, 2019.
- [4] L. Taverna, M. Tremolada, B. Tosetto, L. Dozza, and Z. S. Renata, “Impact of Psycho-Educational Activities on Visual-Motor Integration, Fine Motor Skills and Name Writing among First Graders: A Kinematic Pilot Study”, *Children*, vol. 7, no. 4, p. 27, 2020.
- [5] C. Axford, A. V. Joosten, and C. Harris, “iPad applications that required a range of motor skills promoted motor coordination in children commencing primary school”, *Aust. Occup. Ther. J.*, Vol. 65, No. 2, pp. 146–155, 2018.
- [6] D. Grissmer, K. J. Grimm, S. M. Aiyer, W. M. Murrah, and J. S. Steele, “Fine motor skills and early comprehension of the world: Two new school readiness indicators”, *Dev. Psychol.*, Vol.

- 46, No. 5, pp. 1008–1017, Aug. 2010.
- [7] M. Kadar, F. Wan Yunus, E. Tan, S. C. Chai, N. A. Razaob Razab, and D. H. Mohamat Kasim, “A systematic review of occupational therapy intervention for handwriting skills in 4–6 year old children”, *Aust. Occup. Ther. J.*, vol. 67, no. 1, pp. 3–12, 2020.
- [8] M. S. Julius, R. Meir, Z. S. Nissim, and E. A. Japha, “Children’s ability to learn a motor skill is related to handwriting and reading proficiency”, *Learn. Individ. Differ.*, Vol. 51, pp. 265–272, Jun. 2016.
- [9] C. Semeraro, G. Coppola, R. Cassibba, and D. Lucangeli, “Teaching of cursive writing in the first year of primary school: Effect on reading and writing skills”, *PLoS One*, Vol. 14, No. 2, p. e0209978, 2019.
- [10] A. Comajuncosas, M. F. Zanuy, J. S. Casals, and M. P. Tresserra, “Preliminary Study on Implications of Cursive Handwriting Learning in Schools”, *Smart Innovation, Systems and Technologies*, pp. 339–344, 2018.
- [11] R. Senatore and A. Marcelli, “Do handwriting difficulties of Parkinson’s patients depend on their impaired ability to retain the motor plan? A pilot study”, In: *Proc. of 18th Biennial Conf. of the Int. Graphonomics Society (IGS’17)*, pp. 139–142, 2017.
- [12] N. Z. Fanani, A. G. Sooai, S. Sumpeno, and M. H. Purnomo, “Penentuan Kemampuan Motorik Halus Anak dari Proses Menulis Hanacaraka Menggunakan Random Forest”, *J. Nas. Tek. Elektro dan Teknol. Inf.*, Vol. 9, No. 2, pp. 148–154, 2020.
- [13] Y. Qin and Y. Lou, “Hydrological time series anomaly pattern detection based on isolation forest”, In: *Proc. of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pp. 1706–1710, 2019.
- [14] G. A. Susto, A. Beghi, and S. McLoone, “Anomaly detection through on-line isolation forest: An application to plasma etching”, In: *Proc. of 2017 28th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, pp. 89–94, 2017.
- [15] Z. Cheng, C. Zou, and J. Dong, “Outlier detection using isolation forest and local outlier factor”, In: *Proc. of the conference on research in adaptive and convergent systems*, pp. 161–168, 2019.
- [16] D. Xu, Y. Wang, Y. Meng, and Z. Zhang, “An improved data anomaly detection method based on isolation forest”, In: *Proc. of 2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, Vol. 2, pp. 287–291, 2017.
- [17] V. Verardi and C. Dehon, “Multivariate outlier detection in Stata”, *Stata J.*, Vol. 10, No. 2, pp. 259–266, 2010.
- [18] W. S. L. Wah, J. S. Owen, Y. T. Chen, A. Elamin, and G. W. Roberts, “Removal of masking effect for damage detection of structures”, *Eng. Struct.*, Vol. 183, pp. 646–661, 2019.
- [19] F. T. Liu, K. M. Ting, and Z. H. Zhou, “Isolation-based anomaly detection”, *ACM Trans. Knowl. Discov. from Data*, Vol. 6, No. 1, pp. 1–39, 2012.
- [20] A. G. Sooai, K. Yoshimoto, H. Takahashi, S. Sumpeno, and M. H. Purnomo, “Dynamic Hand Gesture Recognition on 3D Virtual Cultural Heritage Ancient Collection Objects Using k-Nearest Neighbor”, *Eng. Lett.*, Vol. 26, No. 3, pp. 356–363, 2018.
- [21] A. G. Sooai, P. Batarius, Y. C. H. Siki, P. A. Nani, N. M. R. Mamulak, E. Ngaga, U. D. Rosiani, S. Sumpeno, M. H. Purnomo, and S. D. B. Mau, “Comparison of Recognition Accuracy on Dynamic Hand Gesture Using Feature Selection”, In: *Proc. of 2018 International Conference on Computer Engineering, Network and Intelligent Multimedia*, pp. 270–274, 2018.
- [22] A. G. Sooai, A. N. Rumaksari, K. Khamid, N. Z. Fanani, S. Sumpeno, and M. H. Purnomo, “Deteksi Gestur Lengan Dinamis pada Lingkungan Virtual Tiga Dimensi Koleksi Warisan Budaya”, *J. Nas. Tek. Elektro dan Teknol. Inf.*, Vol. 7, No. 4, 2018,
- [23] X. Hua, Y. Cheng, H. Wang, and Y. Qin, “Robust covariance estimators based on information divergences and riemannian manifold”, *Entropy*, Vol. 20, No. 4, p. 219, 2018.
- [24] T. D. Nguyen and R. E. Welsch, “Outlier detection and robust covariance estimation using mathematical programming”, *Adv. Data Anal. Classif.*, Vol. 4, No. 4, pp. 301–334, 2010.
- [25] L. Puggini and S. McLoone, “An enhanced variable selection and Isolation Forest based methodology for anomaly detection with OES data”, *Eng. Appl. Artif. Intell.*, Vol. 67, pp. 126–135, 2018.
- [26] H. John and S. Naaz, “Credit card fraud detection using local outlier factor and isolation forest”, *Int. J. Comput. Sci. Eng.*, Vol. 7, pp. 1060–1064, 2019.
- [27] T. R. Bandaragoda, K. M. Ting, D. Albrecht, F. T. Liu, Y. Zhu, and J. R. Wells, “Isolation - based anomaly detection using nearest - neighbor ensembles”, *Comput. Intell.*, Vol. 34, No. 4, pp. 968–998, 2018.
- [28] Y. Zeng, H. Chen, C. Xu, Y. Cheng, and Q. Gong, “A hybrid deep forest approach for outlier

- detection and fault diagnosis of variable refrigerant flow system”, *Int. J. Refrig.*, Vol. 120, pp. 104–118, 2020,
- [29] P. A. A. Resende and A. C. Drummond, “A survey of random forest based methods for intrusion detection systems”, *ACM Comput. Surv.*, Vol. 51, No. 3, pp. 1–36, 2018.
- [30] R. Geetha, S. Sivasubramanian, M. Kaliappan, S. Vimal, and S. Annamalai, “Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis using Random Forest Classifier”, *J. Med. Syst.*, Vol. 43, No. 9, 2019.
- [31] X. Xu and W. Chen, “Implementation and performance optimization of dynamic random forest”, In: *Proc. of 2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pp. 283–289, 2017.
- [32] A. Chaudhary, S. Kolhe, and R. Kamal, “An improved random forest classifier for multi-class classification”, *Inf. Process. Agric.*, Vol. 3, No. 4, pp. 215–222, 2016.
- [33] A. Sankaran, A. Jain, T. Vashisth, M. Vatsa, and R. Singh, “Adaptive latent fingerprint segmentation using feature selection and random decision forest classification”, *Inf. Fusion*, Vol. 34, pp. 1–15, 2017.
- [34] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, “An improved method to construct basic probability assignment based on the confusion matrix for classification problem”, *Inf. Sci. (Ny)*, Vol. 340–341, pp. 250–261, 2016,
- [35] S. Visa, B. Ramsay, A. L. Ralescu, and E. V. D. Knaap, “Confusion Matrix-based Feature Selection”, *MAICS*, Vol. 710, pp. 120–127, 2011.
- [36] A. Z. Foady, D. C. R. Novitasari, A. H. Asyhar, and M. Firmansjah, “Automated Diagnosis System of Diabetic Retinopathy Using GLCM Method and SVM Classifier”, In: *Proc. of 2018 5th Int. Conf. Electr. Eng. Comput. Sci. Informatics*, pp. 154–160, 2019.
- [37] D. C. R. Novitasari, R. Hendradi, R. E. Caraka, Y. Rachmawati, N. Z. Fanani, A. Syarifudin, T. Toharudin, and R. C. Chen, “Detection of COVID-19 chest x-ray using support vector machine and convolutional neural network”, *Commun. Math. Biol. Neurosci.*, Vol. 2020, 2020.
- [38] J. Sobierajewicz, R. Naskręcki, W. Jaśkowski, and R. H. J. V. D. Lubbe, “Do musicians learn a fine sequential hand motor skill differently than non-musicians?”, *PLoS One*, Vol. 13, No. 11, pp. 1–22, 2018.
- [39] G. Vandewiele, F. Ongenaes, and F. D. Turck, “GENDIS: Genetic Discovery of Shapelets”, *Sensors*, Vol. 21, No. 4, p. 1059, 2021.
- [40] S. Dreiseitl, M. Osl, C. Scheibböck, and M. Binder, “Outlier detection with one-class SVMs: an application to melanoma prognosis”, In: *AMIA Annual Symposium Proceedings*, Vol. 2010, p. 172, 2010.
- [41] F. Karim, S. Majumdar, H. Darabi, and S. Chen, “LSTM fully convolutional networks for time series classification”, *IEEE Access*, Vol. 6, pp. 1662–1669, 2017.

Two Stages Outlier Removal as Pre-processing Digitizer Data on Fine Motor Skills (FMS) Classification Using Covariance Estimator and Isolation Forest

ORIGINALITY REPORT

5%

SIMILARITY INDEX

3%

INTERNET SOURCES

4%

PUBLICATIONS

3%

STUDENT PAPERS

PRIMARY SOURCES

- 1** Submitted to Suan Sunandha Rajabhat University
Student Paper 2%
- 2** www.inass.org
Internet Source 1%
- 3** Dian Candra Rini Novitasari, Ahmad Zoebad Foeady, Muhammad Thohir, Ahmad Zaenal Arifin, Khoirun Niam, Ahmad Hanif Asyhar. "Automatic Approach for Cervical Cancer Detection Based on Deep Belief Network (DBN) Using Colposcopy Data", 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC), 2020
Publication 1%
- 4** Fei Tony Liu, Kai Ming Ting, Zhi-Hua Zhou. "Isolation-Based Anomaly Detection", ACM Transactions on Knowledge Discovery from Data, 2012
Publication 1%

Exclude quotes On

Exclude matches < 1%

Exclude bibliography On