

# IDENTIFICATION OF FRAUD ATTRIBUTES FOR DETECTING FRAUD BASED ONLINE SALES TRANSACTION

Solichul Huda

Lecturer, Department of Computer Science, University of Dian Nuswantoro, Jl. Imam Bonjol No.207,  
Semarang, 50131, Indonesia  
solichul.huda@dsn.dinus.ac.id

Aripin

Lecturer, Department of Computer Science, University of Dian Nuswantoro, Jl. Imam Bonjol No.207,  
Semarang, 50131, Indonesia  
arifin@dsn.dinus.ac.id

Mohammad Farid Naufal

Lecturer, Department of Computer Science, Universitas Surabaya, Jl. Raya Rungkut Kali Rungkut,  
Surabaya, 60293, Indonesia  
faridnaufal@staff.ubaya.ac.id

Vanny Martianova Yudianingtias

Lecturer, Department Humanities, University of Dian Nuswantoro, Jl. Imam Bonjol No.207,  
Semarang, 50131, Indonesia  
vannyningtias7@gmail.com

## Abstract

Fraud can occur in transactions or in business processes. Process-mining approach can be used to detect fraud in business processes. Methods of fraud detection based on process mining have been proposed in several previous studies. The methods used included throughput time analysis, wrong pattern analysis and decision analysis. However, errors still occur in identifying fraud in online sales transactions. Thus, we propose the distant events method and distant events method for analysis of the distance between activities and analysis of the number of activities to identify any fraud issue in online sales transaction process. We analyze the activities carried out by consumers starting from the selection of the good, ordering the good, and the payment of the good activities. Then, the methods analyzes the time gap between activities and the total number of activities for each sales transaction. Furthermore, we did conformance on the results of the analysis with the Standard Operating Procedure (SOP). The violation of the SOP in a sales transaction processes are called the distant events and added event attributes. Prior to the delivery of the purchased good, this fraud detection system will determine these identified attributes of fraud as fraud or not. If the purchase process is identified as fraud, the delivery process of the good bought is canceled. This study proves that by using analysis on the activities distance and analysis on the added activities in online sales transactions, it can reduce false negatives and also increase the accuracy by 0.8 than that by previous methods.

**Keywords:** Fraud Detection, Process Model, Fraud Attribute, Process Mining, Online Sales

## 1. Introduction

Fraud is a criminal act where a person's or a company's money, property or profit is intentionally and illegally taken by theft. This study focused on fraud since it is one of the main causes of organizational and corporate losses, including any companies which perform online sales transaction. This [1] is an extensively developed paper which has been accepted in the 3rd international conference (MECnIT) 2020 and also published in the conference proceeding. Approximately 5% of annual organizational and corporate income losses is caused by fraud [2]. Fraud has contributed to more than 70 trillion dollars in losses. The damage is very significant for large as well as small companies, including those using online transactions [3].

Fraud can be detected when the detection system works properly. For example, if an SOP (standard operating procedure) deviation performed by employees is detected as early as possible; the company can adjust the employees' work patterns to reduce the possibility of such fraud. In some cases, a solution can be proposed by testing the suitability of the business processes with the SOP [4]. Companies potentially encounter financial losses when its anti-fraud protection cannot detect all fraud occurrences. In process mining, such SOP is known as a process model.

Data mining and fraud detection have been studied for decades and various methods have been used in these studies, such as a neural network algorithm in [5], a self-organizing maps algorithm in [6], the Dempster-Shafer theory in [7], Bayesian learning algorithms, classification models in [8], and empirical analysis and web service collaboration in [9]. In addition, control flow analysis, role and performance analysis, association rule learning, hybrid ARL and process mining, Fuzzy Multiple Attribute Decision Making (MADM) [10] and behavior models [11] have been applied in process mining.

Fraud in business processes can be identified by applying process mining, including performance, event sequence, control-flow and role analysis. In addition, fraud detection can also be done by data mining combined with association rule learning (ARL), and data mining come up with process mining (hybrid methods). Then, the results are analyzed based on business processes to identify process model violations [12].

One study operated an ARL algorithm to analyze the correlation between fraud and role behavior in transaction data of credit cards [12]. The result of this study showed that the behavior of the originator (the user who runs an event) can be consistent with the character of fraud behavior and/or fraud perpetrators. This study spotted fraud by estimating process model deviations applying a non-fuzzy method, where the conditions (such as not fraud, between fraud and not fraud, fraud, definitely fraud and very definitely fraud) are not decided.

A hybrid method combining an ARL algorithm and process mining was proposed in [12]. However, the method in [12] only achieved 85% accuracy in detecting online transaction fraud. This low accuracy may be due to the fact that Fraud attributes are unable to identify all process model violations. In this study, we investigated SOP violations in online sales transactions. The research hypothesis was that having more complete Fraud attributes in the business process of online sales transactions enables more accurate online sales fraud detection.

This paper is organized as follows. Section 2 presents previous research of process mining for fraud detection. Section 3 illustrates the identification of fraud attributes in online transactions. Section 4 contains the implementation of fraud detection. Section 5 contains an evaluation of the proposed method and a discussion of its framework and performance. Finally, the conclusion based on the results of the proposed method is presented in Section 6.

## 2. Process Mining for Fraud Detection Method

### 2.1. Process Mining

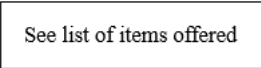
Process mining consists of: discovery, enhancement, and conformance checking [13]. Conformance checking is employed to analyze the prevalence of a case in a process model [14]. Comparison between process instances in event logs and a process model has been proposed in [4]. This work concerned the development of a conformance method for fraud detection. It employed statistical tools to analyze the business process.

Event logs can be defined as a collection of recorded activities of an information system. An event log consists of a set of  $L \subseteq C$  cases, where  $c_1$  and  $c_2 \in L$ . By applying event logs, information such as case ID, activity name, resource, start time stamp and complete time stamp can be obtained [13]. This information is the minimum information that must be available in an event.

#### 2.1.1. Event

Process information in event logs is stored in an event. For instance, if  $E$  stands for all events, then every event  $e \in E$ . Every event has at least information about the name of the event, the name of the originator who ran it, and the time when he administered the event.

The concept, lifecycle, timestamp, activity and resource are defined in the event. The concept is used to define the name of the event; the lifecycle indicates the transition; the timestamp shows the date and time of execution; the activity is the name of the event; and the resource or the originator is the name of the executor who performs the event. Meanwhile, the start and complete timestamps indicate the start and end times of implementing the event. Fig 1 illustrates an example of an event named "See list of items offered". This event is run at 10:11 and finished at 10:15. The originator running the event is named "Welly". This means the event "See list of items offered" has been executed by Welly for four minutes.



See list of items offered

Figure 1. An example of event in online sales

### 2.1.2. Case

A case is a set of execution sequences of  $\sigma \in E$  events where there is only one event in the case for  $1 \leq i \leq |\sigma|$ :  $\sigma(i) \neq \sigma(j)$ . Cases and events are written in event logs. The case code is usually defined in the form of a concept and a value in the case. Fig 2 shows an example of a case.

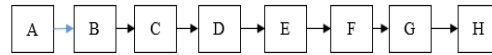


Figure 2. An example of case in online sales

where A refers to “See list of items offered”, B “select Handphone”, C “Cast on delivery”, D “Enter the buyer’s date”, E “payment by M-banking method”, F “Report of payment process”, G “Evident of receiving goods” and H “Information of seller about recipient’s identity”.

### 2.2. Fraud Detection Model

There are three types of process mining: discovery, conformance, and enhancement (refinement) [15] [16]. It is necessary to use the most suitable method to analyze SOP violations in a process model [10] [12]. In this study, the authors used conformance by comparing the cases in the event log with the process model. Moreover, statistical tools can be applied to analyze business processes. This research focused on developing a conformance method for fraud detection.

Process-based fraud is a form of fraud that may be identified through a process that finds deviations from the business process [9]. Detecting Fraud in business processes can be conducted from three different perspectives. Firstly, from the perspective of the business process, Fraud can be detected by comparing various business processes with models. Secondly, from the perspective of the business rules, Fraud can be detected by analyzing every process that deviates from the business rules. Thirdly, from an organizational perspective, Fraud can be detected by analyzing each originator who deviates from the segregation of duties (SOD) or job separation [17].

There are several advantages in employing process mining to detect Fraud. For instance, conformance checking can be used to compare business processes with process model. In addition, this method is able to detect skipping events, which are identified as suspected fraud [10]. In addition, it is able to control and analyze the flow of the business processes. Accordingly, the sequence of the business processes can be analyzed applying this method. As in the previous method, if there is a process that deviates from the process sequence, it is identified as suspicious.

### 2.3. Fraud Attributes

Sarno *et al.* [12] proposed four Fraud attributes, i.e. skipped event, wrong throughput time, wrong resource, and wrong decision. Huda *et al.* in [10] proposed eleven Fraud attributes; they are skip event, skip decision, throughput time min, throughput time max, wrong resource, wrong duty sequence, wrong duty decision, wrong duty combination, wrong decision, wrong pattern and parallel event. However, these attributes cannot identify all types of process model deviations in online sales transactions. In this study, we proposed five fraud attributes to detect fraud in online sales transaction. These attributes were obtained from the companies’ event logs serving online sales transaction.

### 2.4. Case Study

In the case study, the business processes for online sales transactions were investigated to detect online fraud behavior. The result was utilized to identify violation against process model. The process instance and the business rules were investigated to get a number of Fraud attributes. First, the buyer will see a list of items offered. Then, the buyer determines the purchasing method, transfer or cash on delivery (COD). After that, the buyer enters the buyer’s data, including name, address, city, contact number and e-mail address. When the buyer chooses the transfer payment model, the buyer decides the payment method, mobile banking or ATM. The buyer will make a payment according to the choice of payment method. Then, the ATM or mobile banking application will provide a report of the payment process that was carried out and information about the shipping. After the process of sending the goods, there is a menu showing evidence of receiving goods or returning goods. Receiving goods provides information to the seller about the recipient’s identity, while returning goods provides information about the goods and the reason for the return. Table 1 shows the event names of online sales.

We developed this model of online transaction business process based on the procedures of online transaction sales applied in the company which was being investigated in this case study. The formation of this business process model used heuristic algorithm [1]. In this study, we developed this process model of online transaction.

Table 1: Event names of the online sales

No	Events name
1	See list of items offered
2	Select items
3	Determines the purchasing method
4	Enter the buyer's date
5	Chooses the transfer payment model
6	Provide a report of the payment process and shipping
7	Evident of receiving goods
8	Information of seller about recipient

### 3. Identifying Attributes of Fraud in Online Transaction

This research proposed methods for detecting online sales transaction fraud. The methods were established based on fraud attributes identified from the companies' big data in the form of event logs. Furthermore, the methods for detecting fraud were made according to the fraud attributes identified in the training data. We identified violations which occurred in case violation against process model. Process model violations which occurred were further analyzed to be determined as attribute of fraud.

#### 3.1. Standard Time for Time Distance (interval) execution between events

Analysis of time distance (interval) between events was performed by analyzing the time lag between events. This analysis requires a standard time between 2 events and the allowable tolerance time. In this study, the standard time referred to standard time, lower tolerance time or upper tolerance time. In this study, the standard time distance between events was calculated by the average time distance between events obtained from the training data. Meanwhile, the standard time to execute an event was determined used Eq. (1) as in [18].

$$T_i = \overline{X}_i \pm (Tol_i + Ci_i) \quad (1)$$

where  $T_i$  is the standard time to execute event  $i$ ,  $Tol$  is tolerance time to execute event  $i$ ,  $Ci$  is the confidence interval value of event  $i$ , and  $\overline{X}_i$  is the average time to execute event  $i$ . However, the value of  $\overline{X}_i$  in this study was replaced by the standard time to execute event  $i$  as specified in the process model. Table 2 illustrates the standard distance time between events.

Table 2: Standard Distance Time between Events

Events name	Standard Distance Time between Events
See list of items offered - Select items	120 minute
Select items - determines the purchasing method	10 minute
determines the purchasing method - Enter the buyer's date	65 minute
Enter the buyer's date - chooses the transfer payment model	4 minute
chooses the transfer payment model - provide a report of the payment process and shipping	2 minute
provide a report of the payment process and shipping - Evident of receiving goods	3 days
Evident of receiving goods - Information of seller about recipient	2 days

#### 3.2. Analysis of Process Model Violations

We used 10,000 event logs from companies conducting online sales transactions. Moreover, we did analyze the event logs and all corrupted or incomplete data were discarded or not used. This step was executed so that the event logs were in normal conditions. We divided the event logs into 6,000 and 4,000 of training data and testing data respectively.

Conformance was performed to compare the business instance to the business process. In this conformance, we used 6,000 training data of event logs. In addition, a case that violated the business process was identified as attributes of fraud. Each attribute in the case was filled with the accumulated value of violations that occurred, for example, added event was filled with the number of new event that were added in one case; different pattern was filled with the number of events whose order was different from the business process, and so on.

Table 3: Type of Process Model Violation

Attribute	Description
Different pattern	Cases whose order was different from the event sequence in the process model
Distant Event	the time between events exceeded the standard time between events
Added event	the number of events increased compared to the number events in the business process
Parallel events	events that ran concurrently with another event
Throughput time short	an event took less time than the standard event time
Throughput time long	an event took more time than the standard event time

This study identified six types of process model violations in the event log of online sales transactions. The identification was done by analyzing training data compared to the process model. The type of violations consisted of throughput time short, throughput time long, added event, distant event, different pattern, and parallel event. Throughput time short was indicated when an event took less time than the standard event time. Throughput time long was located when an event took more time than the standard event time. Cases where the number of events increased compared to the number events in the business process were verified as added event. When the time between events exceeded the standard time between events, distant event was identified. Cases whose order was different from the event sequence in the process model were confirmed as different pattern. Finally, events that ran concurrently with another event violating the process model were indicated as parallel event. Table 3 describes six types of process model violation which were identified from training data.

### 3.3. Determining Attributes of Fraud in Online Sales Transaction

Having confirmed the attributes of Fraud, the correlation between each type of violation with the weight of fraud was calculated. Types of violations that had a significant correlation to fraud were considered as an attribute of Fraud. These attributes of fraud were the basis for determining the fraud weight. In Eq. (2) was used to calculate the correlations:

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2][n(\sum Y^2) - (\sum Y)^2]}} \quad (2)$$

where  $r$  is the value of the correlation coefficient,  $X$  is an observation of types of violations,  $Y$  is an observation of variable Fraud,  $n$  is the number of paired observations  $Y$  and  $X$ .

The attribute correlation test used training data from online transactions consisting of 3400 fraud cases and 2600 non-fraud cases. The weight assessment of fraud on the training data was carried out by experts. The fraud attributes identified in the conformance test stage were then tested to obtain the fraud attributes. Meanwhile, the experts also assessed the weight of the fraud of cases that violated the process model. This correlation test was used to obtain fraud attributes that have a significant influence on fraud. Fraud attributes that possess significant influence on fraud were decided as fraud attributes.

Table 4: Result of Correlation Test on Attributes

Attribute	Correlations	
Different pattern	0.81	Significant
Distant Event	0.901	Significant
Added event	0.861	Significant
Parallel events	0.431	Not significant
Throughput time short	0.791	Significant
throughput time long	0.706	Significant

From the correlation test of each of these violations it was found that five types of violations had a significant relation to fraud, while the other one type of violations did not have a significant relation to fraud. The five attributes that had a significant relation to fraud were: throughput time short, throughput time long, different

pattern, distant event and added event. Consequently, these five attributes were defined as Fraud attributes. The other one attribute, parallel event, was not identified as Fraud attributes. Table 4 describes the result of correlation test.

The study tested the validity of five attributes using 4-fold cross-validation method. The test resulted validity of the five attributes for detecting process model violation. Of these five Fraud attributes, the first three attributes have been identified in previous studies [11], [19] while the latter two attributes, added event and distant event, were confirmed by this study. The result of validation test and the description of each attribute are elaborated in Table 5 and Table 6.

Table 5: Result of Validation Test

Attribute	Validation test result
Different pattern	Valid
Distant Event	Valid
Added event	Valid
Throughput time short	Valid
Throughput time long	Valid

Table 6: Description of Fraud Attributes

Attribute	Description	Example				
Distant Event	The time distance between the execution of the event and the execution of the previous event is longer than the standard time distance	The time distance between the time of execution of ‘See item’ and that of ‘Select item’ is longer than the standard time distance				
Added events	There is a new event that does not conform with the business process	The ‘Confirm account number’ event is not in the SOP				
Throughput time short	The event execution time is shorter than the minimum standard event time	<div>The<table><tr><td>Event name</td><td>Standard time</td></tr><tr><td>Input order</td><td>15 minutes</td></tr></table><p>execution of the ‘input order’ event took only 8 minutes instead of 15 minutes</p></div>	Event name	Standard time	Input order	15 minutes
Event name	Standard time					
Input order	15 minutes					
Throughput time long	The event execution time is longer than the maximum standard event time	The execution of the ‘input order’ event took 35minutes				
Different pattern	The case pattern is different from the business process pattern	<div><div>D→F→G</div><p>business process pattern</p><div>D→G→F</div><p>Case pattern</p></div>				

### 3.4. Create Algorithm of Fraud Methods

The study was expected to contribute to establishing five methods of fraud detection, i.e. added event method, distant event method, different pattern method, throughput time shorts method and throughput time long method. These methods were used to identify process model violation in any existing case. The details of methods algorithms are shown in Table 7.

Table 7: Algorithm of Fraud Detection Method

<p><b>1. Added Event Method Algorithm</b></p> <p>Develop SOP of business process in the form of wf-net  <math>\alpha(\text{SOP}) = (\text{PL}, \text{TL}, \text{FL})</math>  Calculating the transition from SOP and value of <math>\text{PL}_s</math> and <math>\text{TR}_s</math> from the SOP of business process  <math>\text{PLs} = (\text{as}, \text{Pc}(\text{As}, \text{Bt}) \mid \text{as} \in \text{As} \cup (\text{Pc}(\text{As}, \text{Bt}), \text{bt} \mid \text{bt} \in \text{Bt})</math>  <math>\text{TRs} = \text{TR}_{i \text{ from SOP} \mid i = \sum \text{as}}</math>  Determining the <math>i^{\text{th}}</math> of event <math>e</math> from case <math>c</math> of case <math>C</math> set  <math>e_{i^{\text{th}}} \text{ from case } c = e_{i^{\text{th}}} \text{ from SOP} \mid e \in E</math>  Quantifying the event sequence of SB  <math>\text{SBs} \in \alpha(\text{SOP}) = \sum \text{Bt} = 1, \text{Bt} \neq \text{FL}</math>  Evaluating the event decision of DB  <math>\text{DBs} \in c = \sum \text{Bt} &gt; 1, \text{Bt} \in \text{FL}</math>  Analyzing the added event by comparing event <math>i^{\text{th}}</math> from case <math>C</math> to event <math>i^{\text{th}}</math> from the SOP  <math>e_{j^{\text{th}}} \text{ from case } i^{\text{th}} \neq e_{j^{\text{th}}} \text{ from SOP} \rightarrow \sum \text{FINDe}_{j^{\text{th}}} \text{ from SOP}</math>  Attributes of added event will increase if there is an increasing event in the case  <math>(e_{j^{\text{th}}} \text{ from SOP} \neq e_{j^{\text{th}}} \text{ from case } i^{\text{th}}) \rightarrow \text{SBs} \in c \rightarrow \text{added} + 1</math></p>
<p><b>2. Distant Event Algorithm</b></p> <ul style="list-style-type: none"> <li>Determining throughput time standard of Tst event <math>e</math> from case <math>C</math> of the SOP  <math>\text{Tst}_i = T_{\text{from event } i^{\text{th}} \text{ in SOP}}</math></li> <li>Calculating the values of tolerance from event <math>e</math> of case <math>C</math> set  <math>\text{Tol} = \sigma_{\text{event } i^{\text{th}}} + \text{CI}_{\text{event } i^{\text{th}}}</math></li> <li>Counting the execution time of event <math>T</math> toward event <math>i^{\text{th}}</math>  <math>T_i^{\text{th}} = T_{\text{end from event } i^{\text{th}}} - T_{\text{start from event } i^{\text{th}}}</math></li> <li>Analyze the throughput time of event <math>e</math>. If the throughput time of event <math>e</math> is less than the minimum standard of event limit, the minimum throughput time attribute will increase by 1. If the throughput time of the event is higher than the maximum standard of event limit, the maximum throughput time attribute is increased by 1  <math>T_i &gt; \text{Tst}_i + \text{Tol} \rightarrow \text{Distant} + 1</math></li> </ul>
<p><b>3. Throughput Time Algorithm</b></p> <ul style="list-style-type: none"> <li>Establishing throughput time standard Tst event <math>e</math> from case <math>c</math> of SOP  <math>\text{Tst}_i = T_{\text{from event to } i \text{ di SOP}}</math></li> <li>Calculating the tolerance value of event <math>e</math> from set case <math>C</math>  <math>\text{Tol} = \sigma_{\text{event to } i} + \text{CI}_{\text{event to } i}</math></li> <li>Counting execution time of event <math>T</math> to event <math>i</math>  <math>T_i = T_{\text{end from event to } i} - T_{\text{start from event to } i}</math></li> <li>Analyzing throughput time of <i>event</i> <math>e</math>. If throughput time of event <math>e</math> is smaller than the minimal standard event value, the attribute of throughput time is added by 1 at minimum. Moreover, if <i>throughput time of event</i> is more than maximal limit of standard event, then, the attribute of <i>throughput time</i> is maximum added by 1.  <math>T_i &gt; \text{Tst}_i + \text{Tol} \rightarrow \text{Tlong} + 1</math>  <math>T_i &lt; \text{Tst}_i - \text{Tol} \rightarrow \text{Tshorts} + 1</math></li> </ul>
<p><b>4. Different pattern</b></p> <ul style="list-style-type: none"> <li>Creating the SOP of business process in wf-net  <math>\alpha(\text{SOP})_i = (\text{PL}, \text{TL}, \text{FL})</math></li> </ul>

- Calculating the value of FI from business process SOP

$$FLs = (as, Pc(As, Bt) \mid a \in As \cup (P(As, Bt), bt \mid bt \in Bt)$$

- Counting event  $e$  to  $i$  at case  $c$  from set case  $C$

$$e_i = e_i \text{ from case } c \mid e \in E;$$

- Analyzing wrong pattern by comparing FLs to PLc, if it is not similar, it means there is a different pattern occurrence.

$$FL_{S_{i+1}} \neq e_i \rightarrow \text{Different\_pattern} + 1$$

### 3.5. Create Pseudo-code for Fraud Detection Methods

As a guide for creating applications in the implementation of this method, we arranged the pseudo-code of fraud detection method that we proposed. The details of Pseudo-code methods are described in Table 8.

Table 8. Pseudo-code for Fraud Detection Methods

1. Pseudo-code of added event method
Pseudo-code of added event Read(log) Read(case:cd_case) Read(event:name;time;transition;resource) Read(PNML) Read(transition:name) If(transition=2) SB=event.Sequence else DB=event.Decision for(j.tableLog[j]<sum.transition;j++) { if(tableLog[j] equal transition[j]) added=false else for (a; tableLog <sum.transition;a++) if(tableLog[a] equal transition[j]) { add=false break } else add=true break } if(event=SB) added=added+1 }
2. Pseudo-code of distant event method
Pseudocode of distant events Read(log) Read(case:cd_case) Read(event:name;time;transition;resource) Read(PNML) Read(transition:name;time) Tst=time Read(time_tolerance) Tol=time_tolerance For(i;table.log<sum.transition;i++) { T = time. Complete.tableLog[i] - time.start.tableLog[i]



<pre> If (T[i]&gt;(Tst[i]+Tol[i]))     Tdistant=Tdistant+1 } </pre>
<b>3. Pseudo-code of throughput time method</b>
<pre> Pseudocode of Throughput time Read(log)     Read(case:cd_case)     Read(event:name;time;transition;resource) Read(PNML)     Read(transition:name;time) Tst=time Read(time_tolerance) Tol=time_tolerance For(i;tablelog&lt;sum.transition;i++){     T = time. Complete.tableLog[i] - time.start.tableLog[i]     If (T[i] &lt; (Tst[i]-Tol[i]))         Tshorts=Tshorts+1     elseif (T[i]&gt;(Tst[i]+Tol[i]))         Tlong=Tlong+1} </pre>
<b>4. Pseudo-code of Different Pattern Method</b>
<pre> Pseudocode of different pattern Read(log)     Read(case:cd_case)     Read(event:name;time;transition;resource) Read(PNML)     Read(transition:name;time) For(i;tableLog&lt;sum.transition;i++){     e=tablelog[i]     FLs=transition[i+1]     If(e[i] ≠ FLs[i+1])         Different_pattern=Different_pattern+1} </pre>

### 3.6. Determine the Attribute Importance

Attributes to detect fraud in business processes can be given different weights. One of the solutions to the weighting problem is modified digital logic (MDL). MDL is a method to determine importance weights. Experts discuss the importance of each attribute compared to other attributes to determine the attribute's importance weight. In addition, three experts give an assessment of the importance of each attribute compared to other attributes. Experts value each attribute with '1', '2' or '3'. '3' is used to show that an attribute is more important than other attributes, '2' is used to state that the attributes are equally important, whereas '1' is occupied to assess that the attribute is less important. Attribute importance weighting is similar to the research conducted by [9]. Table 9 shows the results of the experts' assessment.

The attribute importance weights were calculated using Eq. (3), taken from [17], where  $p$  is a positive decision and  $j$  is the number of attributes.

$$W_j = \frac{P_j}{\sum_{j=1}^n P_j} \quad (3)$$

Based on Table 9, it can be stated that Throughput time short has Important, Throughput time long has Important, Different pattern has Important, Added event has Very Important, and Distant Event has Very Important. where I refers to Important and VI means Very Important.

### 3.7. Fuzzy Logic of Violation Rate and Attribute Importance Weight

The five attribute values were converted into attribute importance weights. In this case, method in [10] was employed to calculate the Fraud rating. In [10] an example of an attribute importance weight indicating a violation is if a case has a Fraud rating of 0.2.

As in [10] each Fraud attribute was initialized by employing the following linguistic variables: low, middle and high. Attribute importance weights were then specified using the following linguistic variables: very weak, weak, fairly important, important and very important. The weight of each Fraud attribute was adopted from [11].

Table 9: Experts Opinion in MDL

Attributes	Throughput time short (A1)	Throughput time long (A2)	Different pattern (A3)	Added event (A4)	Distant Event (A5)	Weights	Linguistic
Throughput time short (A1)	2	2	2	1	1	0.13	I
Throughput time long (A2)	2	2	2	1	1	0.13	I
Different pattern (A3)	2	2	2	1	1	0.13	I
Added event (A4)	3	3	3	2	2	0.21	VI
Distant Event (A5)	3	3	3	2	2	0.21	VI
						1.00	

#### 4. Implementation of Fraud Detection in Online Sales Transaction

This study contributed to the new attributes for detecting fraud. Moreover, to test the performance of the methods, this study carried out experimental test following the stages conducted in [10]. Fraud detection in this study consisted of nine stages. Fig 3 shows an illustration of the Fraud detection process.

The stage Fraud detection in this study can be described as follows:

##### Step 1. Added Event Analysis

Each case in the process model contains an event, starting from the beginning to the end of the event. If there was a new event in a case, it was labeled as added event.

##### Step 2. Distant Event Analysis

There is a standard value for the time distance between two events in a transaction. If the time distance between two events was larger than the standard time distance, then it was labeled as distant event.

##### Step 3. Throughput Time Analysis

There is also a standard time for performing an event in a transaction. If the execution of an event took longer than the standard time, then it was labeled as throughput time long. In contrast, if the time to execute an event was shorter than the standard time, then it was labeled as throughput time short.

##### Step 4. Different Pattern Analysis

Implementing the transaction process sequence in a business process should match the process order in the process model. If the execution of the processes sequence did not match the process model, it was labeled as different pattern.

##### Step 5. Conversion of Fraud Attribute Values to Fuzzy Values

The previous steps determined the respective Fraud attribute values, which were then converted to fuzzy values using method in [10]. The results were further grouped into low, middle and high deviation.

##### Step 6. Calculation of Fuzzy Values of Attribute Importance Weights

The attribute's importance weights were needed to calculate the attribute rating. These attribute importance weights were obtained by converting the attribute importance values to fuzzy values. The weight of each attribute importance was adopted from [10].

##### Step 7. Calculation of Attribute's Rating

The fuzzy numbers of the attributes' rating were received by multiplying the fuzzy numbers of the attribute values with the fuzzy numbers of the attribute importance values. This formula was used to calculate the attribute rating as in [10].

#### Step 8. Calculation of Attribute's Rating Script Value and Fraud Rating.

The attribute script number was applied to calculate the Fraud rating. Since  $S$  is considered as the script number of the attribute rating. To get the number of the script from the attribute rating and Fraud rating are using method in [10].

#### Step 9. Fraud Determination

The Fraud rating level was used to decide if a transaction was suspected as Fraud. The expert opinions were implemented to figure out the Fraud rating levels according to the method proposed in [22]. In the study, a case with a Fraud rating of greater than or equal of 0.4 is classified as fraud, while a case with a Fraud rating smaller than of 0.4 is not categorized as fraud.

Hence, Fraud can be detected by establishing Fraud rating-based fraud categories. Cases with a Fraud rating higher than 0.4 are identified as fraud, while cases with a Fraud rating lower than 0.4 are assessed as not fraud. Considering the actual fraud occurrence, experts may decide that fraud occurs at a Fraud rating of 0.5. Consequently, the not fraud category can be changed to between 0.01 and 0.5 Fraud ratings for Fraud detection. Therefore, changing the Fraud threshold can be utilized for Fraud detection.

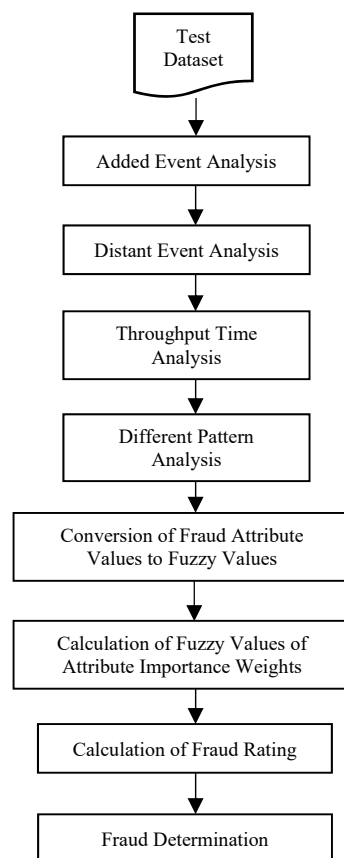


Figure 3: Illustration of Fraud Detection Process

## 5. Methods Evaluation and Discussion

### 5.1. Experimental Design

The experiment in this study employed data testing consisting of 4,000 event logs from companies serving online sales transactions. We proposed five methods in the experiments, i.e. Added Event analysis, Distant Event analysis, throughput time shorts analysis, throughput time long analysis and different pattern analysis. These methods were used to identify any case that violates the process model. The evaluation of the proposed method is described in Figure 4.

The result of the test data analysis showed that there were 822 cases violating the Process model. Case ID 1821 had three fraud attributes: different pattern, throughput time short and throughput time long. Meanwhile, case ID 2115 had two fraud attributes: added event and distant event. An example of the test dataset result is presented in Table 10.

Fraud attributes were selected by referring to the maximum value of each attribute in the training data (for example, distant event has a maximum value 3, and a minimum value of 1. Thus, distant event 3 belongs to the high violation category, distant event 2 belongs to the middle violation category, and distant event 1 belongs to the low violation category). In Table 10, several attributes have 1 value. This indicates that there was one violation. Any cases that violates an attribute in the process model (such as added event and distant event) is established as high violation [4].

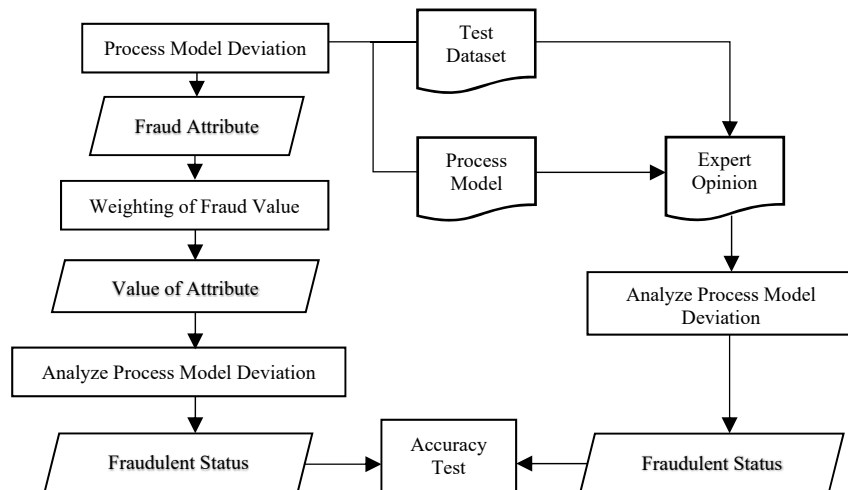


Figure 4: Evaluation Process

Fraud attributes were selected by referring to the maximum value of each attribute in the training data (for example, distant event has a maximum value 3, and a minimum value of 1. Thus, distant event 3 belongs to the high violation category, distant event 2 belongs to the middle violation category, and distant event 1 belongs to the low violation category). In Table 10, several attributes have 1 value. This indicates that there was one violation. Any cases that violates an attribute in the process model (such as added event and distant event) is established as high violation [4].

Table 10: Example of Test Dataset Result

Case Id	Added Events	Distant event	Different Pattern	Throughput time short	Throughput time Long
1821			1	1	1
2115	1	2			
2117	1	1			
2119	1	1			
2561					3
2810					3
2812					1
2817				3	1
2831				1	1
2890				3	1
3125			1	1	2
3224				3	1
3521				2	

Table 11: Example of Attribute value

Case Id	Added Events	Distant event	Different Pattern	Throughput time short	Throughput time Long
1821			Low	Low	Low
2115	High	Middle			
2117	High	Low			
2119	High	Low			
2561					High
2810					High
2812					Low
2817				High	Low
2831				Low	Low
2890				High	Low
3125			Low	Low	Middle
3224				High	Low
3521				Middle	

## 5.2. Results of the Experiment and Discussion

The online transaction fraud analysis method used in [12] and the method proposed in this study were compared. The evaluation consisted of two scenarios: (1) analyzing the testing data using the method in [12], and (2) analyzing the testing data using the method proposed in this study[12]. Meanwhile, experts also analyzed the testing data using their own methods. The accuracy and sensitivity evaluation from the two methods were used to analyze the effectiveness of both methods. In Eq. (4) was used to calculate accuracy, while Eq. (5) was used to calculate sensitivity.

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

$$sensitivity = \frac{TP}{TP+FN} \quad (5)$$

Sarno *et al.* [12] proposed ten attributes for fraud detection, however, in detecting fraud in online sales transactions these attributes only had 85% accuracy. Hence, this study proposes a new method for detecting fraud in online sales transactions. The method uses five attributes: added event, distant event, throughput time short, throughput time long and different pattern. Furthermore, added event analysis, and distant event analysis, throughput time analysis, and different pattern analysis were used to analyze the testing data. A Fraud rating of 0.01-0.4 was categorized as not fraud, and a Fraud rating higher than 0.4 was categorized as fraud. Fraud rating was conducted as done in [10] to find out the performance of the two methods proposed in this study.

The experiment in this study used data collected from the event logs of online transactions from 2018 to 2020. The data were then grouped into training data and testing data, i.e. 6000 cases and 4000 cases, respectively. The business processes in the testing data were analyzed to get the cases that violated the process model. The identified violations were determined as fraud attributes.

Receiver operating characteristic (ROC) was applied to measure the accuracy of the fraud detection methods. Measuring the accuracy of the methods was conducted by counting true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP means experts and the proposed method produced the same determination when a case was categorized as fraud. TP also indicates that experts and the proposed method produced the same categorization when a case was not fraud. If the experts decided a case was fraud while the method determined that that case was not fraud, then it was a FN. FP can also mean that the experts confirmed a case as not fraud while the proposed method determined a case as fraud.

To test the performance of the fraud detection methods we proposed, we compared the performance of our methods to fraud detection methods in [12] in analyzing fraud in testing data. At first, we analyzed fraud in testing data using methods in [12]. After that, we analyzed the same testing data using our methods. The accuracy comparison of both methods will show the performance of our methods.

The result of the testing data evaluation showed that there were 822 cases that violated the process model. The results of the expert discussions showed that with the method in [12] there were 229 cases identified as true positive, 593 cases as false negative, and 3178 cases as true negative. Meanwhile, by using the proposed method there were 548 cases identified as true positive, 274 cases as false negative, and 3178 cases as true negative. By using Eq. (3) and Eq. (4), the method in [12] had 0.85 accuracy and 0.28 sensitivity, while the proposed method had 0.93 accuracy and 0.67 sensitivity. A summary of the testing data evaluation is given in Table 12.

Table 12: Result of (Proposed) Methods Evaluation

Method	ROC variables				Accuracy	Sensitivity
	TP	FP	FN	TN		
Previous Method	229	0	593	3178	0.85	0.28
The proposed method	548	0	274	3178	0.93	0.67

In the evaluation of the methods, the result of accuracy value depends on the Fraud threshold value which has been set. If the threshold value is lowered, the fraud detection accuracy value will decrease because the false negative will increase. If the threshold value is increased, the false positive value will increase. Thus, in order to make the evaluation accurate, the fraud rating threshold value is determined the same as in [12]. In addition, comparative data were obtained from the experts whose similar same competence as those in the study of [12].

When evaluating fraud detection using the proposed methods, there were two fraud attributes that could not be detected by the previous methods, i.e. the interval between events and the addition of new events in a case. As an example in case '2117', there was an event "*confirm by telephone*" asking the seller's mobile banking ownership. From the data of fraud occurrence, it was found that most fraud was preceded by the event even though referring to Table 1, there was not this event in the process business of online sales. This kind of business model violation has never been indicated in previous studies.

In the testing data, fraud events were dominated by the addition of events. The addition of these events is mostly in the form of confirmation styles by prospective buyers via telephone call or Short Message Service (SMS). Based on analysis in this study, the online sales application already provides facilities to check the activities carried out by the buyers and the steps carried out by the sellers, for example, the application provides features for payment and delivery of goods. This application also presents online facilities to serve confirmation by buyers by eliminating direct contact between buyers and sales. Such feature capable to detect business process fraud is the first contribution of our proposed methods.

In addition, this method is able to detect fraud styles in the form of distance time between events exceeding the standard time distance between events. Table 2 shows the standard running time between events.

In the data testing, there were several frauds, one of the indicators of which was the interval/time distance in running between two events that was longer than the standard time. For example in Case '2119', the distance between the "*select list goods*" event and the "*determines the purchasing method*" event is two hours. Referring to Table 2, the standard time distance of the two events is only 10 minutes. Therefore, case '2119' was suspected as fraud. Previous studies never identified such fraud attribute yet; thus, the distant event method is the other contribution of this study in detecting fraud in online sales transactions.

Table 13: Advantages and Disadvantages of Previous Method and the Proposed Method

Method	Advantage	Disadvantage
Previous method	Ready to use	1. Low accuracy, unable to identify violations in online transactions 2. The method consists of eleven attributes, so it needs nine methods
The proposed methods	1. Able to detect violations in business processes of online transactions 2. The method consists of five fraud attributes, so it needs four models only	Need to re-identify new forms of violation attributes

In accordance to the evaluation on these two methods, i.e. added event and distant event, it can be stated that the use of these two methods to detect fraud in online sales transactions can identify violations of the process model in the form of time spent to run the transaction or business activities and time intervals to carry out the activities. Consequently, false identification of fraud in online sales transactions can be reduced.

The comparison of the previous method from [12] with the method proposed in this study shows that the latter was able to decrease the number of false negatives. The decrease in false negatives was due to the fact that the proposed method detected new deviations from the process model, i.e. added event and distant event were added as attributes. The proposed method also achieved better accuracy (0.08). Based on the provided data, it can be inferred that there are both advantages and disadvantages when using either the method from [12] or the proposed method, as shown in Table 13.

## 6. Conclusion

In this study, we analyze business process violations in online sales transactions. We use a process mining approach to reanalyze business processes which violate the process model. We developed a conformance checking technique to identify the violation of the process model of online sales transaction. In the event logs analysis, we identified 5 (five) types of business process violations or fraud attributes, i.e. added events, distant events, throughput time short, throughput time long, and different patterns. Then, we developed an algorithm to identify these 5 (five) fraud attributes. Furthermore, we publish the algorithms in the ProM application. In this paper, two new methods were proposed to investigate the occurrence of process model violations which were not identified in previous studies, i.e. added events and distant events. This is one of the advantages of using the process mining approach, i.e. fraud is detected without any loss. Based on this study, the methods succeeded to improve the accuracy of fraud detection by reducing the number of false negatives. Moreover, the results of the experiment show that this method of fraud attribute can be implemented to detect fraud in online sales transaction applications.

## References

- [1] S. Huda, Aripin, M.F. Naufal, V. Martianova Yudianingias, Anisti, "Fraud Patterns Classification: A study of Fraud in business Process of Indonesian Online Sales Transaction," in MECnIT 2020 - International Conference on Mechanical, Electronics, Computer, and Industrial Technology, 212–217, 2020, doi:10.1109/MECNIT48290.2020.9166644.
- [2] E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, **50**(3), 559–569, 2011, doi:10.1016/j.dss.2010.08.006.
- [3] I. AMARA, A. BEN AMAR, A. JARBOUI, "Detection of Fraud in Financial Statements: French Companies as a Case Study," *International Journal of Academic Research in Accounting, Finance and Management Sciences*, **3**(3), 2013, doi:10.6007/ijarafms/v3-i3/34.
- [4] M. Jans, J.M. Van Der Werf, N. Lybaert, K. Vanhoof, "A business process mining application for internal transaction fraud mitigation," *Expert Systems with Applications*, **38**(10), 13351–13359, 2011, doi:10.1016/j.eswa.2011.04.159.
- [5] S. Maes, K. Tuyls, B. Vanschoenwinkel, "Credit Card Fraud Detection Using Bayesian and Neural Networks," Maciunas RJ, Editor. *Interactive Image-Guided Neurosurgery*. American Association Neurological Surgeons, (March), 261–270, 1993.
- [6] M. Bansal, S.C.S.E. Dept, "Credit Card Fraud Detection Using Self Organised Map," *International Journal of Information & Computation Technology*, **4**(13), 1343–1348, 2014.
- [7] S. Panigrahi, A. Kundu, S. Sural, A.K. Majumdar, "Credit card fraud detection: A fusion approach using Dempster-Shafer theory and Bayesian learning," *Information Fusion*, **10**(4), 354–363, 2009, doi:10.1016/j.inffus.2008.04.001.
- [8] S. Aihua, T. Rencheng, D. Yaochen, "Application of classification models on credit card fraud detection," in *Proceedings - ICSSSM'07: 2007 International Conference on Service Systems and Service Management*, (1997), 2–5, 2007, doi:10.1109/ICSSSM.2007.4280163.
- [9] C.C. Chiu, C.Y. Tsai, "A web services-based collaborative scheme for credit card fraud detection," *Proceedings - 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service, EEE 2004*, 177–181, 2004, doi:10.1109/eee.2004.1287306.
- [10] S. Huda, R. Sarno, T. Ahmad, "Fuzzy MADM approach for rating of process-based fraud," *Journal of ICT Research and Applications*, **9**(2), 111–128, 2015, doi:10.5614/itbj.ict.res.appl.2015.9.2.1.
- [11] S. Huda, R. Sarno, T. Ahmad, "Increasing accuracy of process-based fraud detection using a behavior model," *International Journal of Software Engineering and Its Applications*, **10**(5), 175–188, 2016, doi:10.14257/ijseia.2016.10.5.16.
- [12] R. Samo, R.D. Dewandono, T. Ahmad, M.F. Naufal, F. Sinaga, "Hybrid association rule learning and process mining for fraud detection," *IAENG International Journal of Computer Science*, **42**(2), 2015.
- [13] W.M.P. Van Der Aalst, A.K.A. De Medeiros, "Process mining and security: Detecting anomalous process executions and checking process conformance," *Electronic Notes in Theoretical Computer Science*, **121**(SPEC. ISS.), 3–21, 2005, doi:10.1016/j.entcs.2004.10.013.
- [14] R. Sarno, K.R. Sungkono, "Hidden markov model for process mining of parallel business processes," *International Review on Computers and Software*, **11**(4), 290–300, 2016, doi:10.15866/irecos.v11i4.8700.
- [15] W.M.P. Van Der Aalst, H.A. Reijers, M. Song, "Discovering social networks from event logs," *Computer Supported Cooperative Work*, **14**(6), 549–593, 2005, doi:10.1007/s10606-005-9005-9.
- [16] V. Huser, "Book Review," *Journal of Biomedical Informatics*, **45**(5), 1018–1019, 2012, doi:10.1016/j.jbi.2012.06.007.
- [17] J. Stoop, "A case study on the theoretical and practical value of using process mining for the detection of fraudulent behavior in the procurement process," *Process Mining and Fraud Detection*, Netherlands, Twente University / BASE / Google Scholar, (December), 22–63, 2012.
- [18] F. Sinaga, R. Sarno, "Business Process Anomali Detection using Multi-Level Class Association Rule Learning," *IPTEK Journal of Proceedings Series*, **2**(1), 121–122, 2016, doi:10.12962/j23546026.y2015i1.1135.
- [19] A. Shemshadi, H. Shirazi, M. Toreihi, M.J. Tarokh, "A fuzzy VIKOR method for supplier selection based on entropy measure for objective weighting," *Expert Systems with Applications*, **38**(10), 12160–12167, 2011, doi:10.1016/j.eswa.2011.03.027.
- [20] L.A. Zadeh, "Fuzzy sets," *Information and Control*, **8**(3), 338–353, 1965, doi:10.1016/S0019-9958(65)90241-X.
- [21] S. Vats, G. Vats, R. Vaish, V. Kumar, "Selection of optimal electronic toll collection system for India: A subjective-fuzzy decision

- making approach,” *Applied Soft Computing Journal*, **21**, 444–452, 2014, doi:10.1016/j.asoc.2014.04.006.
- [22] M.P. Barreiros, A. Grilo, V. Cruz-Machado, M.R. Cabrita, “Applying fuzzy sets for ERP systems selection within the construction industry,” in *IEEM2010 - IEEE International Conference on Industrial Engineering and Engineering Management*, 320–324, 2010, doi:10.1109/IEEM.2010.5674473.

## Authors Profile



Solichul Huda has doctoral from computer science program of Institut Teknologi Sepuluh Nopember (ITS) Surabaya in 2017. He has been a researcher and a lecturer at University of Dian Nuswantoro since 1998. He has expertise in process mining method for data security and computer security. His research interests include fraud detection, cyber crime detection methods and computer security



**Aripin** earned his bachelor degree in Information Systems from Dian Nuswantoro University, Semarang in 1997 and received an M. Kom degree in 2004 from the Informatics Department of the Computer Science Faculty of Universitas Dian Nuswantoro Semarang, Indonesia (e-mail: arifin@dsn.dinus.ac.id). He earned a doctor's degree in Electrical Engineering from Institut Sepuluh Nopember Surabaya (ITS), Surabaya in 2017. He works as a lecturer in the Informatics Engineering Department of Dian Nuswantoro University, Semarang. He is an IAENG member (member number: 170713). His research interests include natural language processing and human-computer interaction.



Mohammad farid naufal received his bachelor's degree (2014) and master's degree (2016) from Institut Teknologi Sepuluh Nopember. Currently, he is a full-time lecturer at the Universitas Surabaya, Faculty of Engineering, Department of Informatics Engineering. His research currently focuses on Artificial Intelligence, Machine Learning, and Deep Learning, and Computer Vision.



Vanny Martianova Yudianingtias received her Bachelor Degree in English Language and Literature, Master in Linguistics and Master in Administration Science in 2005, 2011 and 2016 respectively from Diponegoro University, Indonesia. She has been an English practitioner focusing in academics writing and translation as well as teaching in Dian Nuswantoro University and Diponegoro University. Her research interest includes linguistics anthropology, media studies, Islamic studies, psycholinguistics, pragmatics, critical discourse analysis, artificial intelligent and translation.