

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/280805918>

Hybrid Association Rule Learning and Process Mining for Fraud Detection

Article in *IAENG International Journal of Computer Science* · April 2015

CITATIONS

55

READS

1,194

5 authors, including:



Riyanarto Sarno

Institut Teknologi Sepuluh Nopember

310 PUBLICATIONS 2,209 CITATIONS

[SEE PROFILE](#)



Tohari Ahmad

Institut Teknologi Sepuluh Nopember

69 PUBLICATIONS 637 CITATIONS

[SEE PROFILE](#)



Fernandes Sinaga

Institut Teknologi Sepuluh Nopember

2 PUBLICATIONS 71 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



WARNING CRITERION ONTOLOGY [View project](#)



Fraud Detection [View project](#)

Hybrid Association Rule Learning and Process Mining for Fraud Detection

Riyanarto Sarno, Rahadian Dustrial Dewandono, Tohari Ahmad, Mohammad Farid Naufal and Fernandes Sinaga

Abstract— Data mining and process mining provide solutions for fraud detection. The automated methods based on the historical data, however, still need an improvement. In this regard, we propose a hybrid method between association rule learning and process mining. The process mining, in this case, inspects the event log. Through an expert verification, the itemset of the association rule learning is used to generate positive and negative rules applied for compliance checking towards the testing dataset. The result then shows that the hybrid method has less false discovery rate and provides higher accuracy compared to that of the process-mining method in which the optimum accuracy lies in certain threshold of confidence level.

Index Terms— Association rule learning, fraud detection, hybrid method, process mining

I. INTRODUCTION

In company, fraud has been considered as a crucial issue. It is found, for example, that 1,388 cases of fraud have affected 1.4 billion US dollar in loss across 96 countries [1]. On average, 7% of organization gross revenue has annually lost. Related to this, the Law of 20:60:20, 20 states that 20 percent of people in companies never steal, 60 percent of them depends on case and opportunity and 20 percent is truly dishonest [2]. These surprising facts then have urged some companies to have a robust security system for fraud detection and prevention.

Most companies suffer big lose since some fraudulent schemes are not captured as soon as they are committed. Considering that fraud detection is not a simple task and as an attempt to detect such threats, the companies have obliged several security controls to staffs [2, 3] to militate against frauds. Here, Information Security Management System (ISMS) is used as an automated solution [4, 5, 6, 7] in which it can capture suspiciously unlawful schemes prohibited by company's SOP (Standard Operating Procedure).

In computer science, both data mining and process mining introduce certain promising solutions to cope with the

aforementioned problems. Research [8, 5] proposed a solution using Bayesian Network and Neural Network to determine fraudulent schemes. Here, pattern recognition, a well-known solution to evaluate both fraudulent and legitimated patterns, and Machine Learning and Genetic Algorithm, capable of inspecting abnormal activities in companies, have been applied.

Due to the increasing use of Process Aware Information System (PAIS), process mining has been introduced to bridge a gap between data mining and process modeling [9]. It is prominently known as a method to obtain essential knowledge from event logs [6] and contributes many benefits for various aspects, one of which is fraud detection [6, 11, 4]. A process-mining research has concerned to militate against any internal frauds in business process [12] using several process-mining tools in ProM (e.g., conformance checker, dotted-chart analysis, social network miner, or originator by task matrix) for investigation against fraud in the given event log. However, it has only resulted in a suspicious fraud without any method introduced to acquire the confidence level of fraud. Fraud, for some reasons, is a complex problem in which a lot of unpredictable variables (e.g. system crash or special permission) can affect the fraudulent status. Moreover, most procedures are still manual in use (e.g. by comparing fuzzy-miner model to SOP model or by searching for originators that are not in line with the concept of Segregation of Duty (SoD)).

In fraud detection, Association Rule Learning (ARL) is also applied [13] purposely to obtain the correlation of fraudulent behaviors in the transaction database in retail companies. It is functioned to correlate the information of costumers' characteristic to the suspicious frauds. Nonetheless, the research only focuses on the correlating data rather than process. Thus, it is relatively not robust for a process-based fraud.

In this paper, we have proposed a hybrid method between Association Rule Learning (ARL) and process mining by applying a process-mining investigation to obtain a number of fraud variables to generate some association rules in fraud detection. Process mining here is significantly to inspect skipped task, resource, throughput time, and decision point based on basic rules in SOP. To ensure that certain condition is precisely represents a fraud, an expert verification should be involved. The result of process mining and the expert verification are then extracted as the item set of association rule. A priori algorithm is then used to obtain all of possible Antecedences and consequence. This method automates the

Manuscript received June, 2014; revised January, 2015.

Riyanarto Sarno, Rahadian Dustrial Dewandono, Tohari Ahmad, Mohammad Farid Naufal and Fernandes Sinaga are with Department of Informatics, Institut Teknologi Sepuluh Nopember (ITS) Surabaya, Kampus ITS Keputih, Sukolilo, Surabaya, 60111, Indonesia. Telp: +6231 – 5939214 Fax: +6231 – 5913804. Email of the authors are, respectively: riyanarto@if.its.ac.id, gusdewa@gmail.com, tohari@if.its.ac.id, mohammadfaridnaufal1313@gmail.com, ndandes.02@gmail.com

detection towards new-coming fraudulent cases, which have a similar behavior with previous ones.

Our hypothesis is an investigation using process mining tendentiously to yield more False-Positive (FP) cases, i.e. legal cases captured as fraud. If a case contains a fraudulent behavior, it is considered as a fraud regardless the level of confidence towards historical data. The application of the association rules for the investigation is aimed to significantly increase the accuracy and enables the system to catch the fraud automatically. Such combined method is capable of handling a process-based fraud and creating a number of rules for investigation.

As a case study, a credit card application is inspected. To put a well-defined constraint, we clearly define SOP with respect to the application. The used dataset comprises an artificial event log with various fraudulent cases for training data and a large event log for testing data. Upon implementation, the proposed method has been evaluated for accuracy and the confidence level of the rules is variously set to obtain the effect on the robustness for fraud detection. Two evaluation scenarios have also been conducted in this research: (1) association rules directly applied to the whole batch and (2) testing data periodically processed to make the association rules recomposed in every period.

The remainder of this paper is structured as follows. Section 1 explains the essential need of this research – followed by Section 2 providing the summary of several related research. Section 3, furthermore, provides an exclusive explanation of the case study. In this section, we define the SOP and elaborate possibly fraudulent issues. Section 4 presents the proposed method, each step of which is elaborated further. Section 5 presents an evaluation procedure through an explanation about experimental design and result. At last, the conclusion of this paper is presented in Section 6.

II. RELATED LITERATURE

A. Process Mining for Fraud Detection

Process mining is an emerging field specifically to acquire knowledge from actual data recorded in an event log [10]. The event log stores important information regarding process such as what kind of task is conducted, by whom certain task is conducted, and in what time the task is started and ended [12]. The analysis of this information, in turn, can allow companies to track back the actual data and occurrences recorded in their systems.

Process mining becomes an impactful connection between business process analysis and data mining [10]. As illustrated in Fig. 1, process mining focuses on control flow analysis, while data mining concerns with large data processing and copes with data flow analysis. In this case, control flow analysis in a method, different from the data flow emphasizing on an inspection towards moving data, is to inspect the structure of process, does not move data from task to task and hardly concerns with input and output when tasks are conducted.

In studying process mining, three main activities, namely process discovery, conformance checking, and performance

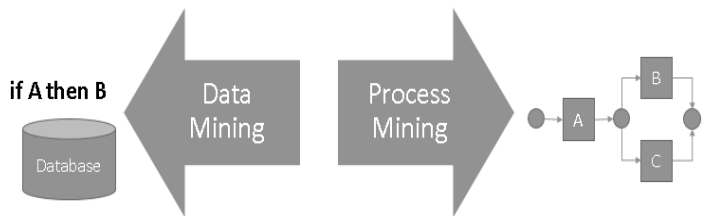


Fig. 1 Process Mining and Data Mining

analysis, are involved. Process discovery refers to a method to acquire an observed model from event log and discovery algorithms (e.g. heuristic miner, or alpha) are used to build the observed model of actual data, recorded by the event log [15, 17] and can be represented in various diagrams (e.g. Petri-Net, Fuzzy Model) [16, 18]. Conformance checking, on the other hand, essentially is to measure a deviation between real data stored in event log and standard model. This activity can be conducted by using algorithms (e.g. token-based conformance checking, or cost-based conformance checking) and can tackle the problem regarding skipped or inserted activity, noise, or wrong sequence.

Process mining enables a performance analysis to the process. To illustrate, by analyzing the location of the process bottleneck, we then can solve the bottleneck by adding more resources or by creating several alternative paths of the process. The objective of such performance analysis is to improve the quality of business processes in companies.

In cases of fraud detection, process mining contributes some advantages. Conformance checking comes to be beneficial in comparing actual data to standard model for being capable of detecting outlier. The fitness value of process model is gauged for measuring commonality, purposely to measure how close the actual data compared to the ideal model is. Considering deviation on two aspects is a necessity in performance checking. The aspects are the way to improve both model and control. It can gain a better conformance [12].

Another benefit from a process-mining based investigation is control flow analysis. It can detect skipped activity, inserted activity, and wrong sequence. It also measures how given event log is complied with a standard process model. Control flow analysis obtains the deviation of actual data in comparison with the ideal condition. In fraud detection, such deviated parts are considered as suspicious deceptions.

Furthermore, process mining encompasses various perspectives. It can acquire knowledge from event log by stressing on organizational perspective. In managerial controls, several basic principles for role management are applied, one of which is Segregation of Duty (SoD). This procedure is to obligate certain legitimated person to conduct every organizational task. To do so, the companies should ensure that different employees handle different tasks. Based on the actual data in event log, information regarding a responsible person to conduct specific tasks or unauthorized tasks can be obtained.

B. Decision Point Analysis

An event log only stores a number of attributes, fundamentally required for a process analysis. Those

attributes commonly embrace event ID, case ID, resource name, and activity name as well as start time and complete time. For some conditions, the inquiries towards external information (i.e., related data stored in tables rather than event log) are required to conduct in which the process mining, in this case, cannot handle. Hence, it has a limitation regarding fraudulent activities caused by information deception, e.g., assigning completed status when checking completeness though the application actually is not completely submitted.

In response, data-aware process mining has been introduced to handle the inability of data flow analysis in process mining by integrating process mining and data mining. It presents information regarding each activity embedded in a control flow model. Additionally, it is beneficial in decision mining analysis, which inspects if some particular tasks perform right responses based on corresponding parameters [20]. Before doing so, the event log should be aligned with the process model. Fan et al [21] introduces an alignment method for checking a precision value between event log and process model. Debreceeny and Gray [20], meanwhile, extends the method by adding a decision-tree learning to deal with a deviating behavior in the log.

Further, decision point analysis is an extended method from data-aware process mining. It is to check the correctness of related attributes. Though process mining focuses on gaining information about how events are executed, decision point analysis enables the system to associate a decision activity with the corresponding data. For instance, in “*check for loan*” task, loan amount is an important variable. The consequence of each amount category varies. Mapping attributes to activities can catch illegal cases. This procedure validates data related to fraudulent tasks.

C. Data Mining for Fraud Detection

Since anti-fraud security control requires for an automatic and more robust investigation, the implementation of computer-based methods is required. Many researches are devoted to propose such methods as an automatic solution for fraud detection [22]. Either data mining or process mining here is used to compose a robust safeguard against fraudulent cases.

Data mining is a classic computerized method in a large-data analysis that is by extracting the abstraction and by processing the data pattern. It embraces decision tree, machine learning, neural network, or association rule learning. In fraud detection, two data mining approaches (supervised and unsupervised) are involved [23]. The supervised approach estimates the models based on the sample of fraudulent and legal transactions to categorize whether new transactions are legal. In the unsupervised one, the outliers are recognized as a suspicious fraud. Such approaches predict the fraud probability in transactions.

Various data mining methods have been proposed to detect fraudulent schemes. Decision tree here is applied to predict some minor instances considered as fraud as well as cross method [24]. Neural and Bayesian networks, for example, are implemented to remove a number related

attributes. Support vector machine, in this case, is beneficial to achieve high accuracy with very little transaction data but not being able to deal with new questionnaires. Table I presents the summarized information about advantages and disadvantages of previous data mining methods concerning with fraud detection. In addition to those methods, Lorrentz [25] observes that it is natural to have majority (many samples) and minority (few samples) classes in fraud detection. This imbalance distribution, however, may influence the capability of the classifier. So, it is advised to not ignoring the minority class.

Despite the ability to inspect large dataset, data mining methods, in fact, cannot deal with process-oriented analysis. Most of them are to investigate the abnormal patterns found in the dataset. Whereas, fraudsters are likely to deceive the process by conducting tasks not complied with SOP. Process mining, in response, is introduced as a bridge between data mining and process modeling in which it can obtain a deeper inspection concerning with the process. Overall, there must be a trade-off between performance and privacy factors in the mining [26] which should be considered.

D. Association Rule Learning Applied to Fraud Detection

Association Rule Learning (ARL) is one of the unsupervised data mining methods in which an item set is defined as a collection of one or more items. Here, support refers to the ratio of the number of transactions containing the defined item set. Confidence, meanwhile, means the probability that an item set will exist and given another item set also exists in the same transaction. ARL observes a relationship among variables in a dataset.

Based on the behavior frequently found in a training dataset, the association rules are used for detection in a testing dataset. The dataset itself can imply the rules to change. In addition to its existing implementation [28, 29], association rules is also suitable for generating filters against fraud. In fraud detection, the rules based on frequent data should be taken into account in which they enhance the system to detect a similar fraud in the following batch.

Aalst et al [13] have introduced the implementation of ARL to capture frauds in credit card application. Table II lists an association between the characteristics and fraudulent status. The characteristics become antecedences, while the fraudulent status is considered as consequence.

Given some of characteristics and fraudulent status, association rules are obtained by applying a priori algorithm [30, 27]. The algorithm generates a number of item set candidates in which some strong candidates supporting more than the threshold will be chosen. Such candidates become several new association rules. C_1 - C_n here are defined as fraudulent characteristics. X represents the item set of

$$Conf(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)} \quad (1)$$

fraudulent characteristics and Y is defined as fraudulent status. $X \Rightarrow Y$ refers to a rule saying that if item set X occurs, Y is considered then. $Supp(X)$ denotes the proportion of item set X in the dataset. $Supp(X \cup Y)$, meanwhile, denotes the proportion of item set X AND Y in the dataset. $Conf(X \Rightarrow Y)$

TABLE I
DATA MINING METHODS FOR FRAUD DETECTION [23]

Reference	Method	Advantages	Disadvantages
Ehramikar (2000)	Decision tree	Predictive performance improved by increasing the number of minority instances	Only the decision tree algorithm experimented upon
Wheeler and Aitken (2000)	Case-based reasoning	Easiness for updating and maintaining model and robust to missing or irrelevant data	Requires two separate experiments;
Bolton and Hand (2001)	Outlier detection (unsupervised)	Successful in detecting local anomalies and fraudulent behavior in a continuous manner	Treating all accounts equally; not differencing between different accounts
Kim (2002)	Neural network with weighted fraud scores (unsupervised)	Increased number of detected frauds compared to a neural network only classifier	Back propagation used to train the neural networks; able to find local minima in the error function, optimal model may not always be reached
Maes (2002)	Neural & Bayesian belief networks	Improvement of fraud detection by removing highly correlated attributes	Better performance of Bayesian algorithm compared to neural networks in fraud detection
Chen (2004)	Support vector machine applied to questionnaire responded transaction data	Capable of achieving high accuracy in fraud detection with very little transaction data	Need to conduct new questionnaires whenever user behavior changes
Abdelhalim and Traore (2009)	Decision tree	Able to correctly classify 92% of the identity application fraud cases	The used data as a mix of real data collected online and synthetic data

TABLE II
ASSOCIATION BETWEEN CHARACTERISTICS AND STATUS

Case ID	X					Y	Support	Conf.
	C ₁	C ₂	C ₃	C ₄	C ₅	Fraud		
1	1	1	1	0	0	1	0.05	0.66
2	0	0	1	0	0	0	0.05	0.20
3	0	1	0	1	0	1	0.05	0.92
4	1	1	1	1	1	1	0.05	0.70

is a confidence value of rule $X \Rightarrow Y$.

Equation (1) explains that confidence value of rule $X \Rightarrow Y$ is obtained from the frequency, when X and Y appear, compared to the frequency, when only Y appears, in the event log. A threshold is set to determine whether the Antecedences affect fraud. If a case contains Antecedences and the confidence level is higher than the threshold, the case is considered as a fraud.

Research [30] introduces a number of association rules applied in both positive and negative rules. In addition to rule $X \Rightarrow Y$, it is possible to mine rules, e.g., $\neg X \Rightarrow Y$, $X \Rightarrow \neg Y$ or $\neg X \Rightarrow \neg Y$. Rule $\neg X \Rightarrow Y$, in this case, implies that if the item set X does not exist in the transaction database, item set Y occurs. Rule $X \Rightarrow \neg Y$ on the other hand denotes that if the item set X occurs, Y does not occur. Rule $\neg X \Rightarrow \neg Y$ implies that if item set X does not occur, Y does not either. Such combination of rules can be used as the negative association rules.

III. CASE STUDY

In this research, a credit card application has been investigated as a case study for being frequently found in organizations and committed variously. Several artificial event logs containing frauds with various issues have been

created as well. Fig. 2 presents a control flow model depicting the Standard Operating Procedure (SOP) of credit card application. The model also explains the information about resources and rules.

The application starts when an applicant submits a set of applications. After receiving application, a clerk will check the completion of all perquisites. The clerk furthermore asks for other information and waits until it is completely retrieved. Once the process is completed, the clerk checks to validate the income and history. Two types of checks are performed based on the loan amount of the applicant. A large amount check is performed if the loan amount is above \$500 and vice versa. This application in further process will be delivered to a manager that is in charge of making a decision about the acceptance. Once the application is accepted, the applicant is informed and a credit card will be delivered. If rejected, the applicant is informed about the rejection and the process is finally completed.

SOP embraces Physical Data Model (PDM) of activities as well as maximum and minimum time stamps for every task in company. As presented in Fig. 2, the model roles are contained in every task, for instance a role explaining about only staffs legitimated to deliver credit card. At the bottom of that figure, the roles along with their corresponding staff are presented. The rules are located in branching activities,

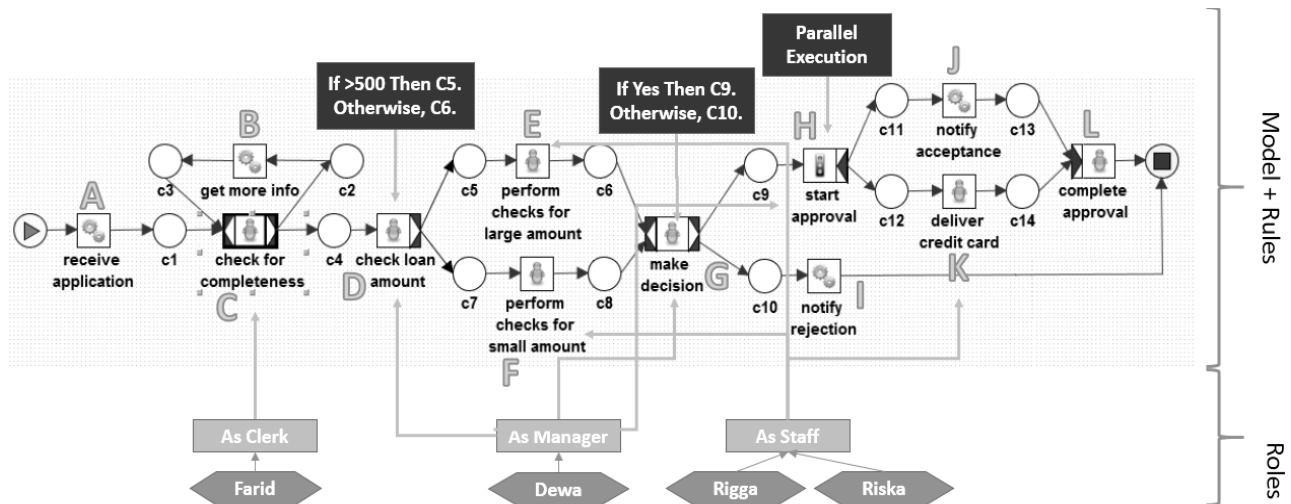


Fig. 2 Business process model and rule of Credit Card Application [12] and their respective roles

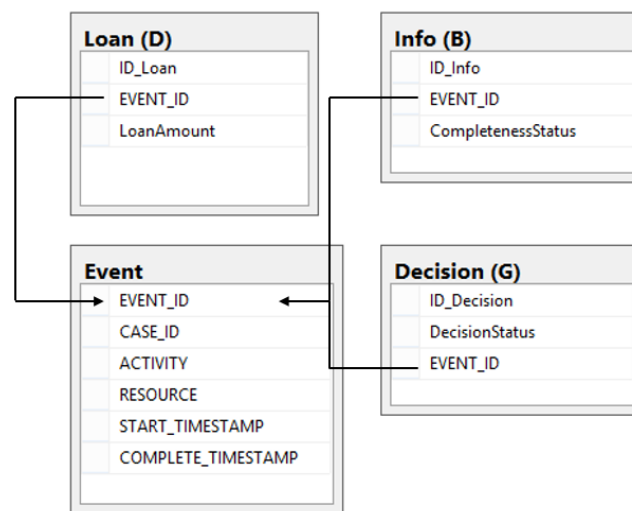


Fig. 3 PDM of Event Log Attributes

e.g. loan amount becoming a rule to perform further loan check purposely to determine, which activity the execution continues.

Throughput time should also be taken into account. Of all cases in the given event log, the average time required by an originator to proceed the task is for 10 minute. Here, we have set 5 minutes as a standard deviation in which each activity is not allowed to be conducted less than 5 minutes or more than 15 minutes. Otherwise, the case will be considered as a fraud.

Fig. 3 presents an explanatory regarding the PDM of credit card application that is about the correlation between the event log and external information stored in related tables. Event and Loan, Info and Decision are tables representing the inspected event log and its case. EVENT_ID is to be the key for other tables.

Check Activity refers to CompletenessStatus containing all prerequisites supposed to be completed and having information stored in Info. Every branching activity requires information for further investigation. Hence, table Decision is provided as well.

A number of fraudulent issues comprising skipped activity, fraudulent sequence, wrong roles, wrong attributes, or wrong time stamps are possibly found in application.

Skipped task deals with the obligated tasks not being conducted based on event log. Such tasks are located in a branching activity for being more crucial to do. An inserted activity issue is considered as a suspicious fraud - not always as fraud depending on the rules. Meanwhile, wrong roles occur due to a misrepresentation of roles. Issue regarding wrong time stamps might be possible if any wrong durations of time execution occurs. In wrong attribute issue, a further investigation towards corresponding attributes is required then.

Skipped task is considered as a suspicious fraud for any forbidden attempts committed by resource. In this case, we assume that skipped task will be considered as fraud only if committed in a branching activity (i.e. Check Loan Amount, Make Decision, and Start Approval). For instance, a procedure regarding Start Approval can be performed only if the manager has already approved in activity of Make Decision. If Make Decision is skipped or not conducted based on the event log, the system should consider it as a fraud.

In addition, the inserted activity issue is likely considered as a fraud only if not acceptable based on SOP. If a fraudster, for example, tries to commit an illegal application to steal secret information through a background process,

the system should consider it as a fraud. It is different from the error handling case. The system might record an error handling as a legal activity for not being defined in the SOP model now that it is difficult to model all errors in the system.

Based on the principle in managerial security control (e.g. Segregation of Duty), two tasks, which are potentially committed as a deception, should be conducted by different staffs with different roles. In credit card application, a complete approval activity, for example, is an important task that requires a number of prerequisites to perform. The procedure of this task involves notifying acceptance and delivering credit card application that are obligated to conduct before a manager performs a complete approval. Notifying acceptance is performed by system, while delivering credit card is by staff. In this circumstance, a dishonest manager might conduct both delivering the credit card and completing certain approval tasks. If such case is recorded in the log, system should consider it as an attempt to deceive a standard procedure.

Storing information regarding start time and complete time in the event log has a number of important purposes, one of which is to investigate the time execution of each task. A fraud might be committed by shortening certain procedures. Standard minimum and maximum duration for each task, therefore, should be defined.

Furthermore, the input-output data should be handled and analyzed properly when particular activities are executed. It is highly possible for fraudsters to commit deception towards data, which are not stored in the event log. For example, a manager will check a large amount though the loan amount is only 400. Therefore, performing further inquiry through data-aware process mining can assist to detect fraud, especially in cases of deception with manipulating data.

It is difficult to determine whether the aforementioned issues belong to fraud. There are some unpredictable variables likely to affect fraudulent status. Not every Skipped task, for instance, is considered as a fraud. This condition occurs due to a special privilege given by the manager and not defined in SOP yet. Another example is that inserted activity is not always recognized as a fraud. Interruption from malicious program can be considered as a fraud, while error handling is still legitimated.

Such inconsistencies require some verification from the expert. Process mining, in this case, helps to obtain fraudulent behavior only representing suspicions fraud in a case. Finally, an expert judgment is used to determine whether the case is a real fraud.

IV. PROPOSED METHOD

The underlying idea of the hybrid method proposed this paper is to present an automated solution for fraud detection based on learning among historical data. We, here, have figured the process-mining method in determining whether a suspicious fraud is a real fraud. The association rules generated from the training dataset can present a confidence level for recognizing whether a case is a really fraudulent case. In addition, several legal rules have been used as exceptional filters purposely to capture legal cases, which are

potential to be considered as fraud.

The proposed method consists of two methods as depicted in Fig. 4. The first method emphasizes on an investigation using a process-mining method to obtain the variables of fraud from the training dataset. Such investigation copes with four fraudulent issues, i.e. skipped task, resource misrepresentation, suspicious throughput time, and suspicious decision point. In the supervised phase, an expert is required to conduct a deeper analysis based on a number of basic rules in SOP. There are some combinations of fraud variables, which lead to fraud. Finally, the expert checks whether a case that contains such combination of fraud variables is a real fraud.

Further the second method relying on the process-mining investigation and the expert verification towards the training dataset can be conducted by applying Association Rule Learning (ARL). It generates two types of rule, i.e. suspiciously fraudulent cases that are actually fraud and legal cases suspected as fraud. The association rules are generated by implementing A priori algorithm [5]. The item sets of ARL are yielded from both process mining investigation and expert verification. Only the association rules with precise consequence (i.e. expert judgment regarding fraudulent status) are chosen as the detection rules. From such process, we have obtained the combination of antecedences leading to fraud. To increase the number of True-Negative cases, the negative association rules for capturing suspicious frauds, which are actually legal, are also applied.

The strength of the proposed method lies in the rule improvement based on the multi-batch processing and the automated process for capturing fraudulent cases. The rule improvement has been carried out using the testing dataset from the same business process (i.e. credit card application). The rules of one batch are added to the rules for the following batch. The proposed method, compared to the process-mining one, provides a significant contribution for fraud detection. It can automate the investigation, as still manually performed in the process-mining method. Even, it can deal with large-scale event log better.

A. Filtering Event Log

Event log used in a big company is a complex trail and very difficult to investigate. In fraud detection, an inspection towards task or activity that is done by human plays an important role. In addition, decision point should also be taken into account. Therefore, all of activities in the event log should be filtered to figure out whether they belong to task or decision activity.

Filtering event log differs based on the type of analysis. In skipped task analysis, event log is filtered by human. There are a number of fraudsters committing fraud by not conducting mandatory task by design. Since resource analysis performs an inspection towards human activity, the log is also filtered by human activity. In addition, investigation against fraud should take throughput time into account. It only focuses on the activities done by human since an automatic activity can perform in a definite time.

From the process model in the case study provided in Fig. 2, activities by human embrace C, D, E, F, G, K, and L. In addition, decision points are located on C, D, and G. Human

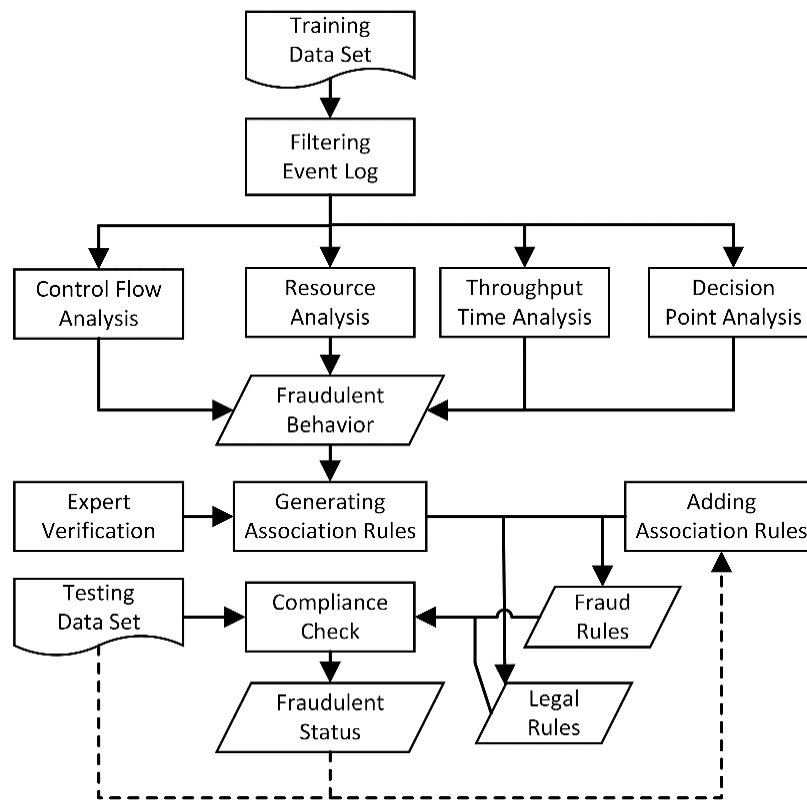


Fig. 4 The Proposed Method for Fraud Detection

activities should be taken into account in skipped task analysis, resource analysis, and throughput time analysis. In other word, if **C** or **F** is skipped in a case, it is considered as a suspicious fraud for having a fraudulent issue. **C**, **D**, or **G** is also mandatory in decision point analysis. For instance, we are able to inspect why **E** is conducted instead of **F** by verifying data regarding the loan amount in **D**.

B. Skipped Task Analysis

Control flow analysis playing an important role to handle fraudulent issues regarding skipped activities can be conducted either by manual analysis or by some plug-in automation. The analysis is performed in each manually conducted activity (usually called as tasks).

In a manual analysis, the log should firstly be discovered using a process discovery algorithm. Fuzzy-miner here can be used for being simple to compare to the SOP model [15]. After acquiring the deviation, a further analysis is conducted towards the particular parts of the process model. The lack of manual analysis lies in the reliability - particularly in determining the threshold of fuzzy-miner. If the threshold is set tight, some low-frequency tasks are not observed. On the other hand, if it is set loose, all kinds of noise are discovered and the analysis becomes more advanced.

Additionally, we have added a conformance checker module to our application purposely to assists the deviation affected by skipped tasks. It is then found that it can yield fitness, structure, and precision value. Fitness represents the behavior similarity of the event log compared to the process model. Structure, meanwhile, evaluates if the process model describes the observed process in a structurally suitable manner and precision evaluates how precisely the model describes the observed process. Such values, in this case, can

measure in advance towards the similarity between actual process and SOP.

The main objective of this analysis is to gauge discrepancy between event log and the standard process model. The different parts of the process model are likely to represent the abnormality of actual process. Fraudsters might deliberately skip certain activity to deceive the system. In addition, frauds are frequently found by inserting a malicious program to seize the confidential information. Investigating event log through this analysis can obtain such forbidden activities.

C. Resource Analysis

One of the perspectives in process mining is organizational perspective. It inspects event logs based on the resources which conduct its corresponding activities. The objective of mining through this perspective is to ensure that each task is performed by authorized resources. Mining organizational perspective requires event log containing an information regarding resources, each of which is correlated with some roles in standard process model. To keep the algorithm efficient, this analysis is only performed toward activities by human.

Table III presents the event log of credit card application mined in organizational perspective. The events are presented with the corresponding originator in charge. For instance in Case 1, activity **A** has a “system” as the information regarding originator since for being automated, while activity **C** has information whether it is conducted by Farid whose role is a clerk.

Segregation of Duty (SoD) is a basic principle in managerial security control. Each person should be different in conducting activities, which are explicitly dependent. For

TABLE III
MINING ORGANIZATIONAL PERSPECTIVE

Case ID	Trace								
1	A ^{system}	C ^{farid}	B ^{system}	C ^{farid}	D ^{dewa}	E ^{riska}	G ^{dewa}	I ^{system}	
2	A ^{system}	C ^{farid}	D ^{dewa}	F ^{riska}	G ^{dewa}	H ^{system}	J ^{system}	K ^{dewa}	L ^{dewa}
3	A ^{system}	C ^{farid}	D ^{dewa}	E ^{riska}	G ^{dewa}	H ^{system}	J ^{system}	L ^{dewa}	

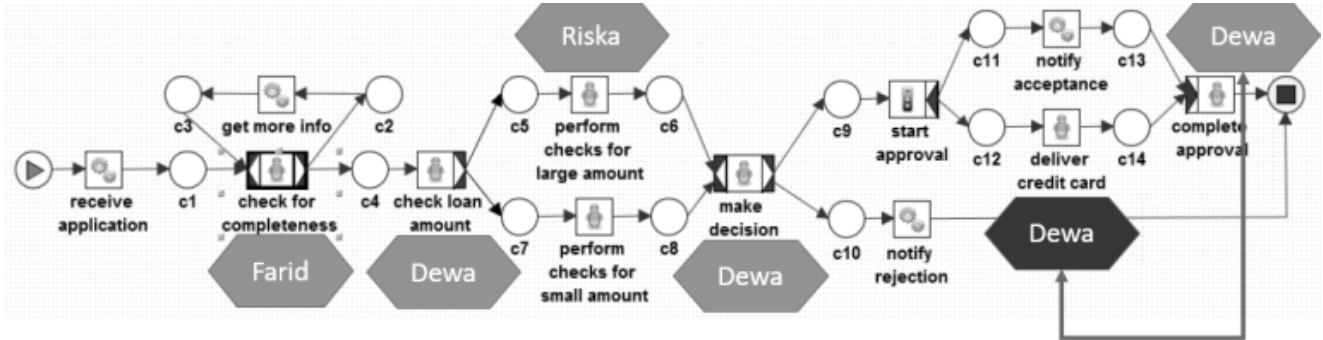


Fig. 5 Misrepresentation of Resource

instance, Case 2 in Table III as depicted in Fig. 5 shows how the originators (i.e., Dewa) conducting both “Delivery Credit Card” and “Complete Approval” activities. It cannot be tolerated based on SoD. Such trace is categorized as suspiciously fraudulent cases.

For the automation, we also have added a module to analyze the roles and the originators in process model. It provides the matrix of information regarding by whom certain activities are conducted. Compared to the SOP, we can analyze whether there is a fraudulent case caused by a resources misrepresentation.

D. Throughput Time Analysis

Throughput time is time interval between activities that can be obtained by analyzing start time and complete time recorded in the event log. The analysis is applied to tasks since fraud is hardly found in automatic activities.

As mentioned on SOP that the average of throughput time is 10 minutes, we here define that time interval of each activity is set not more than 15 minutes or not less than 5 minutes. If any tasks are not complying with the SOP, such tasks are considered as fraud. Furthermore, throughput time analysis can also be used for suspicious investigation, such as, “In which part of process a bottle neck is found?”, “Why does the manager make decision too fast?”, or other time-oriented questions.

E. Decision Point Analysis

If a case is still not captured by applying the aforementioned techniques, data aware analysis is conducted to perform further data verification towards specific cases through complex queries in relational database and to check whether the consequence is true based on the corresponding data. This analysis is able to answer the questions such as “Is the inserted loan amount in the system valid?”, “Did the clerk check all of prerequisites based on the rules?”, or “Why is applicant in certain case approved and why is another applicant rejected by the manager?”

To do so, the event log should be linked to other tables in database. Each activity is mapped to the corresponding table.

For instance, in “Check for Completeness” (see Fig. 5), three prerequisites, those are Personal Data, Organizational Data, and Historical Data have to be fulfilled. Once all prerequisites are completed, the clerk is allowed to set a “completed” status. Otherwise, the system should get more info to acquire the remaining information. A fraud might happen if the clerk commits a “completed” status, but a prerequisite remains incomplete. Even though a prerequisite actually is not completed, “Check Loan Amount” is conducted still, instead of getting more info.

To enable complex queries, we have integrated the event log with a relational database server. A custom application here is developed to perform further data investigation. The queries are performed based on specific cases. For instance, in verifying why activity “Perform Check for Large Amount” is conducted, a query to obtain loan amount can be performed.

F. Generating Association Rules

The main idea underlying the proposed hybrid method is how to generate robust filters based on knowledge regarding historical data in the event log. When mining the rules, 26 variables have been obtained from a process mining investigation meaning that the number of possible combination from such variables is 2^{26} (67,108,864). These 26 variables consist of 24 process mining and 2 expert verification variables. The former is constructed by C, D, E, F, G, K, and L (see Section 4.1) which are applied to control flow/skipped analysis, resource analysis and throughput analysis; and C, D and G which are implemented to decision point analysis. The later is performed according to the experts (i.e. system error and special permission given by a board).

ARL assists to find a number of strong association rules based on historical data as safeguards against frauds. The ARL can be performed either in single-batch or in multi-batch processing. To do so, we have generated both positive association rules and negative ones. The positive association rules here represents if certain combination of fraud variables

occur, it is positively fraud. The negative ones, on the other hand, mean that if certain combination of fraud variables exists, it is totally legal.

In fraud detection, the transactions are analogous to cases in the event log. Item sets or fraud variables are extracted from process mining investigation in four aspects of analysis as well as the expert verification towards training dataset. Process mining investigation toward a complete event log results in fraudulent behavior, a case condition based on the combination of fraud variables. Fraud variable is a typically single issue regarding fraud contained in the case (e.g. skipped task in activity **C**, or misrepresentation of resource in activity **D**). It has a binominal value: true or false. Expert verification is required to obtain information whether there are any outliers and to directly judge whether a case is fraud. The value of each case depends on the existence of fraud in certain issues. Fig. 6 displays this procedure of obtaining item sets for ARL and the respective variable values.

After the item sets are obtained, the association rules are generated by applying A priori algorithm [11] - a method to generate association rules and frequently used to operate on transaction databases. The algorithm creates candidates of item sets in which only strong item sets are chosen. Subsequently, it is pruned to new form of rules. From the investigation using the process-mining method, 24 fraudulent variables are obtained. Expert verification contributes by analyzing other 2 variables of fraud. The expert considers the fraudulent status based on some predefined rules as listed in

Table IV. In addition, some exceptional rules representing suspicious frauds, which are actually legal, are also defined. Table V presents such exceptional rules.

Upon the implementation of A priori algorithm towards such 26 variables, only the generated association rules that have consequence fraudulent status and confidence level equal or higher than minimum confidence are chosen as the strong association rules. These strong rules are used to quickly detect the fraudulent behavior of every case in the event log. Table IV presents the antecedences and the consequence from the association rules regarding fraud detection. It comprises both positive-fraud and negative-fraud rules.

In Table VI, the antecedences are obtained from process mining investigation towards the training dataset. For instance, in the first row the Antecedences include ThroughputTime[**K**], SkippedTask[**F**], and SkippedTask[**K**], Resource[**C**] representing a suspicious fraud found in particular issues (i.e., skipped task, misrepresentation of resource, wrong throughput time, and/or wrong decision point). ThroughputTime[**K**] means that the throughput time of activity **K** is not correct based on SOP. In addition, SkippedTask[**F**] as well as SkippedTask[**K**] represent that both **F** and **K** are skipped. Outlier is a variable obtained from expert verification in which the expert checks if there is a special conduction. Such antecedences, in turn, are used as filters against fraud in the testing dataset.

Association rules generate a large number of rules. It is an

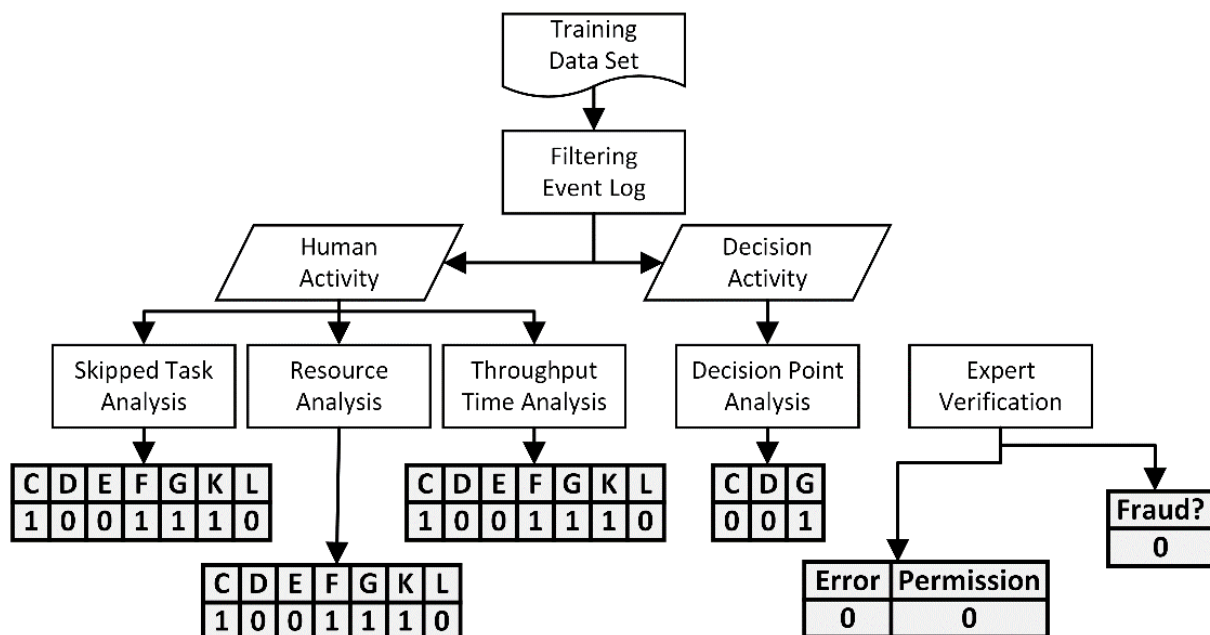


Fig. 6 Fraud Variables from Process Mining Investigation and Expert Verification

TABLE IV
BASIC RULES OF FRAUD BASED ON THE SOP

Antecedence	Consequence
Skipped task at decision point	Fraud
Misrepresentation of resource AND false throughput time at the same task	Fraud
Misrepresentation of resource AND false consequence of decision point at the same task	Fraud
Misrepresentation of resource AND skipping the following tasks	Fraud
Throughput time exceeding AND false consequence of decision point at the same task	Fraud
Shortening throughput time AND skipping the following tasks	Fraud
Misrepresentation of resource AND skipping the following tasks	Fraud

essential procedure to choose the interesting rules, which represent the combination of the antecedences leading to fraud. However, capturing fraud only with the rule of fraud tends to capture legal cases, which are wrongfully considered as fraud. To filter such False-Positive (FP) cases, we then have generated a number of exceptional rules, composed from the association regarding legal cases on the training dataset. Here, we define such rules as negative rules, meaning that they are negative for fraudulent status. In this case, both the positive and negative rules have been checked towards the testing dataset.

G. Compliance Checking using Association Rules

Compliance checking is a rapid checking employing both positive association rules and negative ones. These rules are essential for fraud detection towards new coming cases. In this research, we define the fraud variables as follows: S[X] denotes that activity X is skipped, R[X] denotes that activity X is conducted by wrong person, T[X] implies a so fast and long conduct of activity X, and D[X] implies a wrong decision taken in activity X. If a case has some exact variables in comparison to at least one of the rules, the case is directly considered as a fraud. For instance, as shown in Fig. 7, rule 1 has 3 antecedences (i.e. S[C], S [D], and R [G]), while case 1 has 4 fraudulent variables (i.e. S[C], S [D], S [G], and S [K]). The comparison between case 1 and

rule 1 is not complied, since case 1 does not contain R [G]. Nonetheless, if case 1 is checked towards rule 8, it is considered as fraud for the existence of both S[C] and S [K] on the case 1.

As shown in Fig. 7, we have created a matrix representing some generated association rules. The antecedences of the rules obtained from fraudulent behaviors are also drawn in a matrix. We firstly sort the antecedences based on their confidence levels, mainly to detect if each case in the case matrix has the exact combination of fraudulent rules presented in the matrix of rules.

Compliance checking consists of two steps. The first step is by filtering the testing dataset using positive association rules purposely to capture fraudulent cases. However, some of legal cases in this step are likely to be captured as well. To identify such legal cases, we then have performed filtering using negative association rules as the second step of compliance checking. Here, both positive association rules and negative ones are conducted through one-pass processing.

V. EVALUATION

A. Experimental Design

The evaluation in this research emphasizes on the following points: (1) comparing the advantages of the


TABLE V
BASIC RULES OF LEGAL CASES BASED ON THE SOP

Antecedence	Consequence
Changing resource AND given a permission	Legal
Throughput time exceeding AND computer error	Legal
Skipping tasks due to computer error	Legal
Manager changes staff position to conduct staff's tasks	Legal
Manager can skip certain tasks	Legal

TABLE VI
ASSOCIATION RULES WITH RESPECT TO FRAUD DETECTION

Antecedences	Consequence	Support	Confidence
ThroughputTime[K], SkippedTask [F], SkippedTask[K], Resource[C]	Fraud	0.04	1.000
SkippedTask[G], Resource[D]	Fraud	0.06	1.000
SkippedTask[L], SkippedTask[E], Resource[D]	Fraud	0.01	1.000
Resource[C], Resource[K], DecisionPoint[C]	Legal	0.03	0.600
Resource[K], DecisionPoint[C]	Legal	0.03	0.600

Case ID	S[C]	S[D]	S[E]	S[F]	S[G]	S[K]	S[L]	R[C]	R[D]	...
1	1	1	0	0	1	1	0	0	0	...
2	1	0	0	0	0	1	0	0	0	...
3	0	0	0	1	0	0	1	0	0	...
4	0	0	0	0	0	0	0	0	0	...
5	0	0	0	0	0	0	0	0	0	...



Rule ID	Antecedence
1	S[C] S[D] R[G]
2	S[E] R[C]
3	R[E] D[G]
4	S[E]
5	S[F] R[C] R[G]
6	R[E] D[G] R[G]
7	S[D] R[C]
8	S[C] S[K]

Fig. 7 Compliance Checking between Cases and Rules

proposed hybrid method to the process-mining method in the context fraud detection; and (2) measuring the accuracy of the proposed hybrid method both in single batch and in multi batches. The scenarios used in this evaluation are equal. The dataset consists of training dataset and testing dataset whose generation process is as follows.

It is analyzed that the deviation against the attributes follows Poisson distribution whose average parameter is 3. This means that, on average, there are 3 fraudulent cases each month. The Poisson distribution is used because the characteristic is analogous to the behavior of business process fraudulence. The number of fraudulent cases for each attribute is generated randomly according to the Poisson distribution. Therefore, every attribute has a different number of fraudulent cases for each month. In addition, there are 10 credit applications are processed per month. Among these, the frauds are spread over that application according to uniform (discrete) distribution. It aims to distribute the incidence of frauds on the 10 transactions in a month randomly according to the number of

occurrences of fraud in each attribute. The examples of those generated data are provided in Tables VII and VIII, respectively. For the training purpose, we generate 1000 application data comprising 700 fraud and 300 non-fraud (legal) cases by using that specified data distribution scenario randomly. These cases, in turn, become the reference on which the fraud rules and the exception rules relied on.

The behavior of process in various organizations is unique. Different companies might have a different process behavior. Hence, in the training dataset, we notice the combination of fraud variables, considered as fraud. Furthermore, the weight of each attribute may be different depending on that process behavior. In order to obtain the weight values, we use the method in [27] by classifying each attribute to lower, middle 1, middle 2 and upper according to the experts. From this process, we have final rating of each case. The rating itself is represented in a range between 0 and 1, where that in upper half means fraud.

Here, only attributes whose *supp* is greater than the specified threshold is further processed. Those selected

TABLE VII
NUMBER OF DEVIATION OF EACH ATTRIBUTE

Skip		Throughput time		Wrong resource	Wrong duty		Wrong		
sequence	decision	Min	Max		Sequence	Decision	Combine	Pattern	Decision
1	1	7	5	6	5	1	5	3	3

TABLE VIII
DISTRIBUTION OF DEVIATION OVER CASES

Cases	Skip		Throughput time		Wrong resource	Wrong duty		Wrong		
	sequence	decision	Min	Max		Sequence	Decision	Combine	Pattern	Decision
1	0	0	1	1	0	1	0	1	0	0
2	0	0	0	0	1	0	0	0	1	0
3	0	0	1	0	0	1	0	1	0	0
4	1	0	0	1	1	0	1	0	0	1
5	0	0	1	1	1	0	0	0	1	0
6	0	0	1	1	1	1	0	1	0	1
7	0	0	1	0	1	1	0	0	0	0
8	0	0	1	1	0	0	0	1	0	0
9	0	1	1	0	1	0	0	0	1	0
10	0	0	0	0	0	1	0	1	0	1

TABLE IX
ITEM SETS AND THEIR CORRESPONDING RULES FOR ARL TRAINING

Combination	Attribute combination	Support	Confidence
2-item set	Throughput min-Fraud	576	0.805594406
	Throughput max-Fraud	421	0.850505051
	Wrong resource-Fraud	449	0.757166948
	Wrong duty sequence-Fraud	398	0.752362949
	Wrong duty combine-Fraud	370	0.755102041
	Wrong pattern-Fraud	248	0.802588997
	Wrong decision-Fraud	302	0.977346278
3-item set	Throughput min-Throughput max-Fraud	348	0.96398892
	Throughput min-Wrong resource-Fraud	363	0.870503597
	Throughput min-Wrong Duty sequence-Fraud	328	0.874666667
	Throughput min-Wrong Duty combine-Fraud	298	0.853868195
	Throughput max-Wrong resource-Fraud	267	0.87254902
	Wrong resource-Wrong Duty sequence-Fraud	256	0.805031447
4-item set	Throughput min-Throughput max-Wrong resource-Fraud	214	1
	Throughput min-Wrong resource-Wrong Duty sequence-Fraud	204	0.935779817

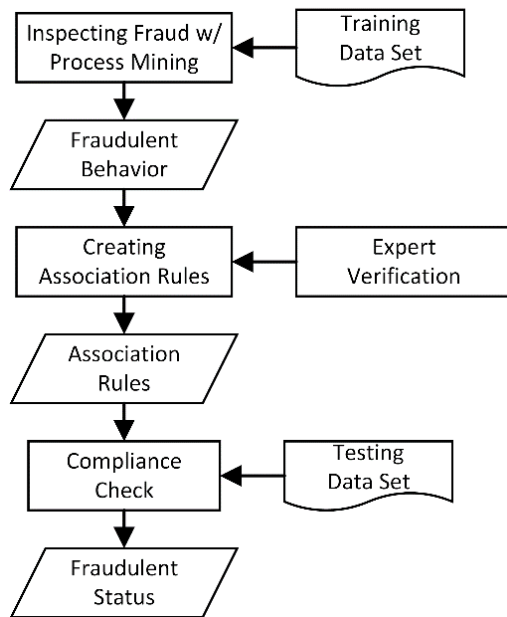


Fig. 8 Evaluation process

attributes are combined each other to have 2-item set attributes. Similarly, that with *supp* greater than the threshold is used. This step is repeated until no item set meet the threshold. In this experiment, we finish in 4-item set, as depicted in Table IX. Until this step, the training process has been finished and the rules have been obtained.

The testing dataset, meanwhile, is generated with similar behavior compared to the training dataset. By using same parameter values, new random numbers are generated. We put some constraints in randomizing the testing dataset. The event log of the testing dataset is presented based on the process behavior found in the training dataset. The testing dataset here should be generated based on the behavior of the training dataset and randomized using error tolerance. We put 100 cases for frauds and 100 others for legal cases considering that the randomly generated testing dataset might not obtain the optimum result of compliance checking. The proposed method is to catch these 100 frauds using the antecedences of rules yielded from the process-mining investigation towards the training dataset.

To highlight the significant differences between the proposed hybrid method and the process-mining one in fraud detection, we deliberately generate both training dataset and testing dataset by having fraudulent characteristics, suspiciously captured by the process-mining method. It is by considering that a process-mining method can indicate the suspicious behavior of fraud, not captured as a fraud by the proposed hybrid method since it has considered fraudulent status based on the combination of fraud variables. Hence, the significant difference lies in the value of False Discovery Rate (FDR). In this evaluation, we have proved the hypothesis that process mining tends to detect suspicious frauds, which are not actually frauds more frequently than the proposed method did. To do so, we inspect the testing dataset using both proposed method and process-mining method.

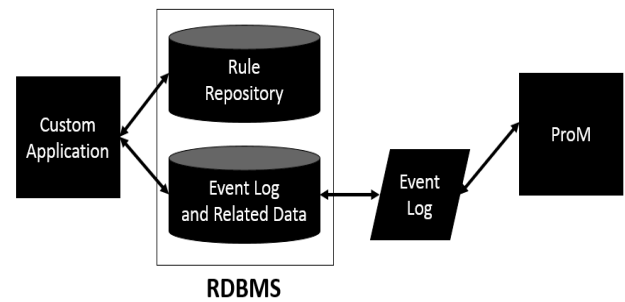


Fig. 9 Integrated Application for Evaluation

Fig.8 depicts the scenario of evaluation. After the training dataset has been inspected by process mining and fraudulent behaviors are yielded, some rules subsequently are generated through a verification of an expert. These rules are then directly tested towards the testing dataset. Finally, fraudulent status of the testing dataset depending on the rules is obtained. In this scenario, the rules are only composed by supervised learning through a process mining investigation.

To implement the evaluation scenario, we have used both ProM plug-in and a custom application to acquire fraudulent behaviors. Here, ProM processes the single file, while the custom application is integrated with Relational Database Management Server (RDBMS). Furthermore, it is required to perform further data aware analysis. Compliance check is implemented in the application by performing several queries adapted from the antecedences of the rules. Fig. 9 illustrates the implementation phase in this research.

ARL has a role to extract the association rules regarding the correlation between fraudulent behavior and fraudulent status and to store them into rule repository in RDBMS. The antecedences of the rules become filters to capture frauds. The application retrieves the rules from the rule repository and then performs specific inspection towards fraudulent behaviors as listed in the antecedences of the rules. We set various confidence levels in both scenarios of evaluation to acquire the correlation between the confidence and the accuracy of detection system, the number of the rules generated by A priori algorithm.

A. Experimental Result

In this research, Receiver Operating Characteristic (ROC) analysis is used as a framework to measure the accuracy of the proposed method. The framework here calculates four variables: True-Positive (TP), False-Positive (FP), True-Negative (TN), and False-Negative (FN). TP is obtained from the number of cases which are actually fraud and detected by the system. TN represented the number of cases which are not fraud and not detected by the system. Furthermore, FP comes from the number of legal cases detected by the system and FN is the number of fraudulent cases not detected by the system. Here, the accuracy is obtained from the proportion of TP and TN in comparison with the total number of cases.

Table X shows the result of evaluation using the proposed method. We set various values of minimum confidence to consider the value of accuracy. Based on the experiment, the optimum accuracy can be obtained with confidence greater

TABLE X
RESULT OF THE HYBRID METHOD

Min Conf.	TP	FP	TN	FN	Accuracy
0.2	86	13	87	14	0.865
0.4	86	13	87	14	0.865
0.6	86	13	87	14	0.865
0.8	73	5	87	14	0.8
1	2	0	87	14	0.445

than or equal to 0.6. The minimum confidence 0.2 results in no difference compared to the result of 0.4 and 0.6.

Both positive association rules and negative ones have been used as one-pass filtering. After acquiring suspicious frauds using positive association rules, the result are filtered using negative association rules. However, a set of cases become became an interception between the positive and negative rules. This condition comes to be the tread off in this research. The number of positive association rules and negative ones should be taken into account to capture fraudulent cases more accurately.

Being implemented in companies, two types of association rules can be performed in two-pass processing. Firstly, the positive rules have an objective to capture True-Positive cases, regardless the number of False-Positive cases. The expert can verify from those captured cases. The cases, which are really fraudulent cases, are excluded from the process. The next step is filtering the False-Positive cases by using the negative association rules purposely to filter which cases are totally legal.

VI. CONCLUSION

The process-mining method, i.e. skipped task analysis, resource analysis, throughput time analysis, and decision point analysis; inspect a number of types of fraudulent variables. Due to the inconsistency, an expert takes role to verify the existence of outlier and the fraudulent status. The fraudulent variables, followed by the outlier, become the antecedences of the ARL. The antecedences are then correlated with the fraudulent status to create the association rules with regard to fraud detection. The association rules are used to automatically filter fraudulent cases in the testing dataset. To improve the accuracy, the negative association rules are used to filter legal cases captured as fraud by the positive association rules.

The experiment has been conducted to measure the accuracy in given values of minimum confidence. The testing dataset is generated by having the similar behavior with the behavior of the training dataset. The evaluation shows that the proposed method can achieve the maximum accuracy on certain value of minimum confidence. This result is relatively better than that of process-mining method since it has less falsely detected frauds.

The rules capture the fraud with given minimum confidence. If the minimum confidence is set high, the filters become too tight to catch the fraud. If set low, many legal cases are considered as fraud. Therefore, fuzzy association rules are likely to be an alternative solution for future work.

The fraudulent status is categorized into low risk, medium risk, and high risk. It can not only detect the fraudulent case but also determine the level of fraudulent risk of the case.

REFERENCES

- [1] ACFE, Report to the Nations on Occupational Fraud and Abuse, 2012, p. 20.
- [2] P. D. Goldmann, Anti-Fraud Risk and Control Workbook, Wiley, New York, 2009, pp. 11-22.
- [3] R. Sarno and Heryanti, Developing Information Technology Policies for Enterprise Resource Planning to Improve Customer Orientation and Service, *International Journal of Computer Science and Network Security*, 10 (5) (2010), pp. 82–94.
- [4] F. Folino, G. Greco, A. Guzzo and L. Pontieri, Mining usage scenarios in business processes: Outlier-aware discovery and run-time prediction, *Data & knowl. eng.* 70 (12) (2010), pp. 1005–1029.
- [5] K. Fanning and K. Cogger, Neural network detection of management fraud using published financial data, *International J. of Intell. Syst. in Account, Finance & Management*, 7 (1) (1998), pp. 21–41.
- [6] M. Jans, J. M. v. d. Werf, N. Lybaert and K. Vanhoof, A business process mining application for internal transaction fraud mitigation, *Expert Syst. with Applications*, 38 (10) (2011), pp. 13351–13359.
- [7] R. Wheeler and S. Aitken, Multiple Algorithms for Fraud Detection, *Knowledge-Based Syst.*, 20 (8) (2000), pp. 93–99.
- [8] B. Patrick and J. H. Choi, Assessing the Risk of Management Fraud Through Neural Network Technology, *Spring*, 29 (1997), pp. 14–29.
- [9] The IEEE Task Force on Process Mining, Process Mining Manifesto, *Bus. Process Management Workshops Lecture Notes in Bus. Inf. Processing*. 99 (2012). pp. 169–194.
- [10] W.-S. Yang and S.-Y. Hwang, A process-mining framework for the detection of healthcare fraud and abuse, *Expert Syst. with Applications*, 31 (1) (2006), pp. 56–68.
- [11] W. M. v. d. Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, Springer, 2011, pp. 74-77.

- [12] D. Sánchez, M. Vila, L. Cerda and J. Serrano, Association rules applied to credit card fraud detection, *Expert Syst. with Applications*. 36 (2) (2009), pp. 3630–3640.
- [13] W. v. d. Aalst, H. Reijersa, A. Weijtersa, B. v. Dongena, A. A. d. Medeirosa, M. Songa and H. Verbeeka, Business process mining: An industrial application, *Inf. Syst.*, 32 (5) (2007), pp. 713–732.
- [14] A. d. Medeiros, Genetic process mining, *Data Min. and Knowl. Discovery*. 14 (2) (2007), pp. 245–304.
- [15] C. W. Günther and W. M. v. d. Aalst, Fuzzy Mining – Adaptive Process Simplification Based on Multi-Perspective Metrics, *BPM'07 Proceedings of the 5th International Conference on Bus. Process Management*. 5 (2007), pp. 328–343.
- [16] R. Sarno, B. Sanjoyo, A. Mukhlash and H.M. Astuti, Petri Net model of ERP business process variation for Small and Medium Enterprises, *Journal of Theoretical and Applied Information Technology* 54 (1) (2013), pp. 31–38.
- [17] M. d. Leoni and W. M. P. v. d. Aalst, Data-aware process mining: discovering decisions in processes using alignments, in *SAC '13 Proceedings of the 28th Annual ACM Symposium on Applied Computing*. 28 (2013), pp. 1454–1461.
- [18] A. Adriansyah, J. Munoz-Gama, J. Carmona, B.F. v. Dongen and W.M.P. v. d. Aalst, Alignment based precision checking, *Bus. Process Managemen Workshops*, 132 (2012), pp. 137–149.
- [19] P.-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Addison-Wesley, Minnesota, 2006, pp. 327–396.
- [20] R. S. Debreceeny and G. L. Gray, Data mining journal entries for fraud detection:, *International J. of Accounting Inf. Sys.*, 11 (2010). pp. 157–181.
- [21] W. Fan, Y. Huang and P. S. Yu, Decision tree evolution using limited number of labeled data items from drifting data streams, in *Proceedings of the Fourth IEEE International Conference on Data Min.* (2004), pp. 379–382.
- [22] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen and I. Verkamo, Finding interesting rules from large sets of discovered association rules, in *CIKM '94 Proceedings of the 3rd International Conference on Inf. and Knowl. Management*. (1994), pp. 401–407.
- [23] R. Agrawal and R. Srikant, Fast algorithms for mining association rules in large databases, in *Proceedings of the 20th International Conference on Very Large Data Bases*. (1994), pp. 487–499.
- [24] X. Wu, C. Zhang and S. Zhang, Efficient Mining of Both Positive and Negative Association Rules, *ACM Transactions on Inf. Syst.* 22 (3) (2004), pp. 381–405.
- [25] P. Lorrentz, Classification of Incomplete Data by Observation, *Engineering Letters* 18(4) (2010), pp. 316–325.
- [26] P. Lin, N. Thapa, I. St. Omer, L. Liu and Jun Zhang, Feature Selection: A Preprocess for Data Perturbation, *IAENG International Journal of Computer Science* 38 (2) (2011), pp. 168–175.
- [27] M. P. Barreiros, A. Grilo, V. Cruz-Machado, M. R. Cabrita, Applying Fuzzy Sets for ERP Systems Selection Within The Construction Industry, *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, (2010), pp. 320 - 324.
- [28] C.-F. Lee and C.-H. Tsai, Efficient Associating Mining Approaches for Compressing Incrementally Updatable Native XML Databases, *IAENG International Journal of Computer Science* 33 (1) (2007), pp. 78–85.
- [29] S. C. Abou and T.-M. Dao, Association Rules Mining Approach to Mineral Processing Control, *Engineering Letters*, 18 (2) (2010), pp. 156–164.
- [30] X. Wu, C. Zhang and S. Zhang, Efficient Mining of Both Positive and Negative Association Rules, *ACM Transactions on Information Systems*, 22 (3) 2004, pp.381-405.