Learning Fraud Detection from Big Data in Online Banking Transactions: A Systematic Literature Review

Indrajani[1], Harjanto Prabowo[2], Meyliana[3]
*[1,3]School of Information System, Bina Nusantara University, KH. Syahdan, 9,
11480 Jakarta, Indonesia.
[2]Bina Nusantara University, KH. Syahdan, 9,
11480 Jakarta, Indonesia.
indrajani@binus.ac.id*

***Abstract*— The implementation of fraud detection in online banking transactions on big data is one of the most important strategies applied by banks to protect their transactions and highly related to algorithms. In fact, it is not easy to successfully implement this strategy because it requires a huge investment and is influenced by complexity algorithms, training, and testing. The frauds bring fatal impact, such as destruction of the banking reputation, banking loss, and state financial loss. One target of the fraud perpetrators in banking is online banking transactions. Security has become a major issue in the online banking transaction. Furthermore, the research of fraud is switching to big data and turns out that online banking data are stored in the database operational and big data. This study aims to find out what kind of algorithms fraud detection for online banking transactions using a systematic literature review to the 25 relevant papers.**

***Index Terms*— Fraud; Fraud detection system; Big data; Online banking transaction; Algorithm.**

I. INTRODUCTION

The attack on online banking system will be one of the most widespread methods of stealing money from bank and user account [1] [2] . The number of crimes committed is increased rapidly around the world, including Indonesia, apart from all the technical measures taken by the bank. Online banking fraud can be done internally by fraud perpetrators or externally [1].

Issues of the specific fraud detection system (FDS) are not published by banks [2]. Nevertheless, the fact that the bank is applying the FDS, lately some vendor of software and hardware provided an amount of information about FDS banking and sold it in the banking worldwide For example, Proactive Risk Manager (PRM) is the FDS used by the top 20 banks in the world that are disseminated in more than 40 countries. Similarly, SAS Fraud Management System reportedly is used in debit and credit card answers FDS in more than 43,000 online banking sites [2].

According to Gregory Trompeter (2012) has been evolving systematic reviews in their research to describe the fraud and how the fraud occurs within the organization [3]. According to Mirjana et al. (2014) in the systematic review in their research defined one of the important opinions on the machine learning, which is that the modern science found the designs and make predictions of a large amount of data, with the goal of making a learning system from experience. The growth of data that is available every day is a good purpose to believe that machine learning, data mining, and other related fields will become more extensive as the crucial components for the innovation of technology [4]. Their research question is how to research of fraud has been established during the years and algorithms to detect fraud. The research available was elevated several methods such as supervised learning which is comprised of K-nearest neighbor (k-NN), Decision tree (DT), Neural Networks (NNS), Support vector machines (SVMs), and Bayesian statistics, and then the Unsupervised Learning which are consists of K-means clustering and principal component analysis (PCA), and Reinforcement Learning [3][4][5][6][7][8] [9][10]. By reviewing the literature, there are quite a lot who marked about the algorithm in online banking transaction and match some of the literature such as what was on papers [11][12][13][14][15][16] . However, the amount of papers in comparison to all only ranged between 1-25 papers. The highest number was five papers, which can be considered up to date because they are all published in 2015.

The research of this paper shows the use of different algorithms with the application of the rule base system (RBS). The approach of RBS used by previous summarized FDS system has been criticized on a number of users. One of the negatives was based on knowledge barrier attainment phenomenon, in which the process of transferring knowledge to the enterprise system is indirect (connecting experts and developer's knowledge), industry intensive, and usually limited to a specific context. The maintenance of this system has also been defined as laborious, costly, and challenging. This is because the typical adjustments RBS will involve expert knowledge and domain experts. Usually, if the new acquisition of knowledge required expert and knowledge developer, then keeping the system would require experts to guarantee that the latest changes do not make old knowledge unacceptable. Those factors and the fact that the maintenance in this case was done as an additional task for the acquisition of knowledge creating the maintenance of the system is time wasting and expensive project. Another major technical limitation of FDS profitable systems is that they are fragile. Instability occurred when RBS did not recognize when knowledge is not adequate for certain cases. Weakness also associated with a deficiency of common intelligence in an expert system. These limitations fake a serious threat to the prospective loss of large sums of money [2].

The motivation of this study is to create a research of fraud detection from big data in online banking transactions which has a complete literature review, a structured approach to systematic literature review (SLR), a significant and innovative value (being one of the updates). The research question of this study is what algorithm is used in implementing fraud detections from big data in online banking transactions? While the purpose of this study are to identify and study the algorithm fraud detections from big data in online banking transactions and to see if there is a change in fraud detections algorithm from big data in online banking transactions.

The contribution for academic community such as proposing a hierarchical feature of fraud from a large volume of bank transactions as input, proposing a robust statistical model to detect and recognize fraud action from big data or a large volume of bank transaction data and prove of concept of deep learning application in banking industry. Research paper to international journal/conference. Meanwhile, the contribution for banking industry is a robust statistical model to detect and recognize fraud action from a large volume of bank transaction data to reduce potential loss of banking industry and improving public trust toward banking industry.

II. THEORETICAL FOUNDATIONS

A. Fraud Detection

Fraud means to obtain services, goods, and money by the unethical way, and grow a problem in the world today. The deals of fraud with cases involving criminal purposes mostly are difficult to identify [7] [17] [18] [19] [20] [21].

There are various types of fraud, which are numerous and diverse as financial institutions and technology products, as shown in figure 1 [7]. Type of fraud divided eight categories such as technology ATM and internet, transaction products credit and debit cards and checks. In this paper, focus on ATM and internet such as transfer funds, pay bills, check saving account balance, paying instalments, and purchase [7].

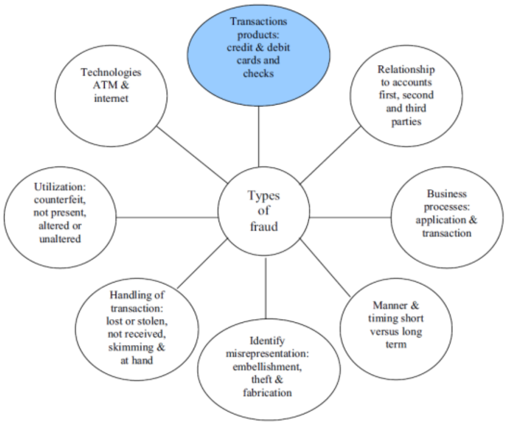
Transactions products: credit & debit cards and checks
Technologies ATM & internet
Relationship to accounts first, second and third parties
Utilization: counterfeit, not present, altered or unaltered
Types of fraud
Business processes: application & transaction
Handling of transaction: lost or stolen, not received, skimming & at hand
Identify misrepresentation: embellishment, theft & fabrication
Manner & timing short versus long term

Figure 1: Type of fraud [7].

B. Big Data

Increasing of data storage capabilities, computational processing power, and availability of increased volumes of data are causing the rise of Big Data. Most of organization do not have enough computing resources and technologies to process the data. Big data refers to data sets whose size is beyond the ability of typical database software to capture, store, manage and analyze [22].

Big Data is also associated with other specific complexities referred as the four Vs: Volume, Variety, Velocity, and Veracity. Instead of the four Vs, Big Data also covered some other factors, such as the data variability (inconsistency), velocity (speed in which new data is generated) and complexity (structures, relationships, etc.) [4][12].

Computing is more efficient and easier to layered in parallel and distributed fashion which from the point of view of computational is a clear challenge to machine learning technique to back into simple focus technique (such as K-means). In high-dimensional data, “big” means not only in the number of cases but also in a number of features that are used to describe them. The great potential to revolutionize all aspects of our society offered by Big Data is not an ordinary task because information is growing large and rapidly which is unprecedented, non-traditional data volumes requires both the development of advanced technology and interdisciplinary teams to work together closely. An important role in Big Data analysis and knowledge discovery is achieved by learning machine techniques and advances available computational electricity. They are widely employed to support the predictive power of Big Data in the field such as search engines, banking, medicine, astronomy, and basis for an economic innovation revolution [12].

C. Online Banking Transaction

For the purpose of convenience of the users, not all transactions are marked at risk. The risky financial transactions are as the following transfer of funds between bank accounts. Transfer in terms of the funds from a bank account to another domestic or international bank account, massive payments, and credit card payments. This payment refers only to credit card that is issued by other banks [23] [24] [25].

Online banking transactions include transactions in ATM and internet banking. ATM and internet banking have nearly the same features. Internet banking have some features such as financial transactional and non-financial transactional. For example, financial transaction such as an account to account transfer, paying a bill, and payment, meanwhile non-financial such as purchase, payment, e-Commerce Payment, fund transfer, account information, credit card information, consumer credit information, investment product information, other information, transaction status, transaction history, administration, and E-Mail [25]. Furthermore, ATM is a terminal deployed by bank or other financial institutions, which enables the customers to withdraw cash, to make a balance enquiry, to order a bank statement, to make a money transfer or to deposit cash. The ATMs are basically self-service banking terminals and are aimed at providing fast and convenient services to customers [25].

III. METHODOLOGY

This study uses SLR approach proposed by Mirjana et al. (2014) and Iqbal Muhammad (2015). This approach is divided into several sections, namely: defining research questions outlined in the introduction, determine the source of the study, reach the findings by using keywords, data mining, and analyzing the findings to answer the research question [4] [5].

A. Search Process

After defining the research question, settle the source (digital library OR database) that is indispensable in creating SLR. The selected sources for SLR are as follow Science Direct (www.sciencedirect.com), Google Scholar, IEEEXplore Digital Library (http:/ieeexplore.ieee.org), ACM (http://dl.acm.org/), Springer Link (link.springer.com), and Emerald Insight (www.emeraldinsight.com).

The use of keywords is applied to find paper relating to defined research questions. The use of this keyword is enabled by adding Boolean operators such as AND, OR, NO. All sources mentioned above have a keyword-based search engines. Searching for keywords, the defined search strings are:

- (‘fraud’ OR ‘detection’) AND ((‘big’ AND ‘data’) OR (‘online’ AND ‘banking’ AND ‘transactions’))
- (‘fraud’ OR ‘detection’) AND ((‘big’ AND ‘data’) OR (‘automatic’ AND ‘teller’ AND ‘machine’) OR (‘internet’ AND ‘banking’))
- (‘fraud’ OR ‘features’) AND ((‘big’ AND ‘data’) OR (‘online’ AND ‘banking’ AND ‘transactions’))
- (‘fraud’ OR ‘features’) AND ((‘big’ AND ‘data’) OR (‘automatic’ AND ‘teller’ AND ‘machine’) OR (‘internet’ AND ‘banking’))

Thus, to look for these papers, the keywords are: 'fraud detection', 'fraud features',' big data ',' online banking transaction, 'automated teller machine',' ATM ', internet banking'. The detailed results are presented in table 1.

B. Inclusion and Extraction

Right after entering a keyword into the source, papers relating to the certain keywords and total summary of all these papers are shown as 'study found'. The next step is to read the title of the paper. When the title is not enough to determine whether to include the paper as the candidate or not, then the abstract is read. If the titles and abstracts match with previously defined research questions, then this paper will be downloaded for further investigations. The amount of paper that is downloaded is called 'prospective studies'. All 'prospective studies' results papers will be read thoroughly to find answers to research questions. These papers are the ones that will be used in research as 'the chosen study'.

The paper was issued based on their publication date (before 2004) and place of publication (journals and conference except the papers mentioned) for the following reasons to acquire new knowledge and information from previous studies and to obtain a quality literature from reliable sources. Paper copies of the same study were also issued in the SLR. In addition, this SLR should concentrate on fraud detection in the banking implementation.

C. Data Extraction

The literature study was initiated on Sep 2015 and examined 205 papers. By the source page numbers that are summarized in Table 1. Among the 205 papers examined, 42 papers have the titles related to the study and basic abstract. However, after being studied further, only 25 papers that can be included on the basis of this study. Other papers appear in the source because some articles have been mentioned in the 'Study chosen' by the author. This paper, however, is not included in the previously mentioned sources.

Table 1
Number studies in selected sources

No.	Source	Studies Found	Candidate Studies	Selected Studies
1	Science Direct	105	12	7
2	Google Scholar	68	15	5
3	IEEE	19	9	8
4	ACM	10	4	4
5	Springer	2	1	1
6	Emerald	1	1	0
	Total	205	42	25

IV. RESULT AND DISCUSSION

There are demographic and trend characteristics such as source of publications, most prolific authors, most productive institutions, authors’ academic background (discipline of authors) , publication trends (frequency of publications), background of authors, university affiliation according to country, researched industries and countries, and researched institution size. First, inn source of publication, there are plenty of published papers about fraud detection from big data in online banking transactions on various sectors or industries. Some of the journals or conferences are: Expert System With Application (#4), the another journal or conferences (#1). Overall, there are a total of 25 journals or conferences. The second, as seen from the writer’s analysis perspective, there are 131 authors who have written 25 papers in total. The most consistent author in writing about fraud detection is Olivia Caelen (#2). The other 76 authors wrote 1 paper each. The third, the most productive institutions are Wordline (Brussel, Belgium). The rest just generally produced one paper each. The fourth, the 77 authors are categorized into 40 science fields. If regrouped, six big science groups can be identified. Among these fields, the fraud detection clearly falls within the computer science and management field. This is because; in order to succeed implementing fraud detection will always depend on its technology. The fifth, the most productive years are 2012 (seven papers). Next, 2015 with five papers, and 2014 three papers. The sixth, authors who contribute on fraud detection area in online banking transactions tend to write with research approaches, so that most of them own academic background (86%). The rest just collaborate with industry or public sector experts. The seventh, Out of 42 institutions which contribute on fraud detection from big data in online banking transaction, the largest comes from the USA with 11 institutions and 20 authors. The list is followed by the Belgium with 5 institutions and 8 authors. This is the case because both of these developed countries have been implementing fraud detection very well for a long time now.

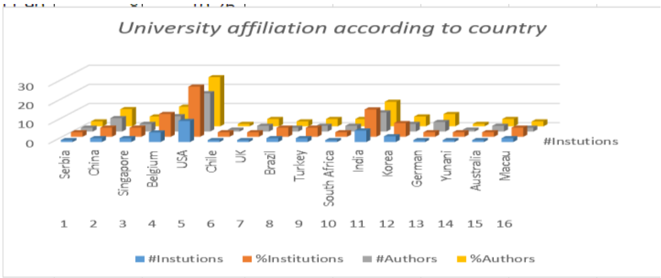

Figure 2: MDS University affiliation according to country

Moreover, the eighth, Industry/sector areas which are researched by the authors are getting more complete and more tested in various countries. India is the most country wrote papers (5 papers). The list is followed by the USA with 4 papers and China with 3 papers. Thus, it points out that fraud detection covers almost every industry all over the world. Company's awareness towards the importance of fraud detection in online transactions. The last, research companies falling in the medium up to large-scale range produced 24 papers. This is followed by general producing 1 papers.

In lists keywords which are used in searching papers related to fraud detection study from big data in online banking transaction. This keyword finds 202 papers in six sources. Among the data, it can be seen that the most frequently used is 'fraud detection big data online banking transactions automatic teller machine ATM internet banking IB' and followed by 'fraud detection big data online banking transactions internet banking IB'. Paper data which are found in other keyword are overlapping data.

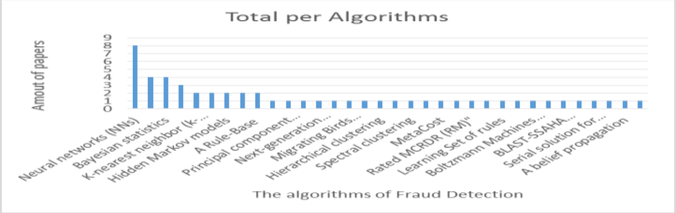

Figure 3: Algorithms of Fraud Detection from Big Data in Online Banking Transactions based on analysed literature frequency

Algorithm fraud detection from big data in online banking transaction are listed below top 2 from 25 papers and brief explanation is given:

A. Neural Network

The algorithm, which is a mathematical algorithm that is used Neural Network (NN) to make the process of learning. There are several types of algorithms used, namely: first, Backpropagation, a learning algorithm (learning algorithm) used by NN on supervised methods. One form is the delta learning rule. Second, Delta learning rule, a learning algorithm (learning algorithm) used by NN on supervised method, wherein the change of weight is obtained by multiplying the input, error and learning rate. Third, Forward propagation, an algorithm where the output neurons only propagated in one direction from input to output. Fourth, Hebb learning rule, an algorithm used by supervised learning, especially in perceptron, where changes in weight is obtained by multiplying the input, output and learning rate. Fifth, Simulated annealing, a special type of learning algorithm, specifically for the NN type of feedback.

In this technique, we create an artificial neural network that aims to simulate the workings of neurons inside a human cell. Neuron itself a very important function in the human body because it plays an important role in receiving and processing the signal. In the modern era, the technique of the neural network is learning the techniques of the most popular and effective. Neural network has many advantages such as can perform distributed calculations, can tolerate noise in the input, and the ability to learn.

B. Decision Tree

Decision Tree is a tree structure, where each node of the tree represents the attributes that have been tested, each branch is a division of the test results, and leaf node (leaf) represents a particular class group. The top-level node of a Decision Tree is the root node (root) are usually in the form of attributes that most have the greatest influence on a particular class. In general Decision Tree conduct a search strategy from the top down to the solution. In the process of classifying data is unknown, the attribute value will be tested by tracing the path from the root node (root) to the end node (leaf) and then would have predicted class owned by a certain new data

Decision Tree Users can create a tree that is over-complex and so cannot generalize the data well. There is an elusive concept with a decision tree, because the decision tree cannot explain easily, but easy to understand and implement. It only takes a little preparation data, using the concept of White Box models and can handle numerical and categorical data.

V. IMPLICATIONS

Based on the detailed discussion above, there are two algorithms found almost in every browsed literature, which are: neural network and decision tree. Big data of banking transaction that impossible to be detected manually or using conventional transaction analysis. Fraud actors might impersonate as a good bank customers. Fraud actors might use various banking transaction modes to disguise the fraud actions. The problem still important as research topic. The fraud cases tends to increase that cause many tangible and intangible damage for banking industry.

VI. LIMITATION

This paper has a limitation in that the number of databases is limited much because of restricted access. The amount of papers needs to be augmented mainly excavated from a credible database and published in the last twelve years.

VII. CONCLUSION

This study found that there are fraud detection on online banking transactions' algorithm based on investigation on a big data. A new combination of algorithm was found i.e. Bayesian and Neural Networks (in table 3). This paper managed to extract 34 algorithms. The algorithm Neural Network is the top algorithm for fraud detection (8 papers), is followed by Decision Tree and Bayesian (4 papers), and the other 18 algorithm is investigated in 1 paper. The research aims not only to help academics and researchers in fraud detection study on big data in online banking transactions in companies, but also assisting practitioners in field implementation.

ACKNOWLEDGMENT

The author would like to thank you for the in-depth reviews and suggestions from reviewers on earlier versions of this paper. Their comments have substantially helped to improve the paper.

REFERENCES

[1] K 1. R. Khande: Online Banking in India, “Attacks and Preventive Measures to Minimize Risk,” *ICICES2014*, no. 978, 2014.

[2] O. O. Maruatona, P. Vamplew, and R. Dazeley , “Prudent Fraud Detection in Internet Banking,” *Third Cybercrime Trust. Comput. Work*, pp. 60–65, 2012.

[3] T. Gregory, C. Tina, D. Naman, J. Keith, and R. Richard, “A Synthesis of Fraud Related Research. Am,” *Account. Assoc. J.* , pp. 63, 2012.

[4] M. Ivanović and M. Radovanović, “Modern machine learning techniques and their applications,” *International Conference on Electronic, Communication, and Network*, 2014.

[5] I. Muhammad and Z. Yan, , “Supervised Machine Learning Approaches : A Survey,” *ICTACT Journal on Soft Computing*, pp. 946–952, 2015.

[6] M. A. Alsheikh, S. Lin, D. Niyato, and H. Tan , “Machine Learning in Wireless Sensor Networks : Algorithms , Strategies , and Applications,” *Communications Surveys & Tutorials, IEEE*, pp.1–23, 2015.

[7] J. P. Linda Delamaire, Hussein Abdou , “Credit card fraud and detection techniques : a review,” *Banks Bank Syst*., vol. 4, no. 2, 2009.

[8] S. B. E. Raj, A. A. Portia, and A. Sg , “Analysis on Credit Card Fraud Detection Methods,” *International Conference on Computer, Communication and Electrical Technology*, pp. 152–156, 2011.

[9] M. Paliwal and U. A. Kumar , “Neural networks and statistical techniques : A review of applications,” *Expert Syst. Appl*., vol. 36, no. 1. pp. 2–17, 2009.

[10] A. Dal, O. Caelen, Y. Le Borgne, S. Waterschoot, and G. Bontempi, “Learned lessons in credit card fraud detection from a practitioner perspective,” *Expert Syst. Appl*., vol. 41, no. 10, pp. 4915–4928, 2014.

[11] M. A. B. Bella, J. H. P. Eloff, and M. S. Olivier , “A fraud management system architecture for next-generation networks,” *Forensic Science International*, vol. 185, pp. 51–58, 2009.

[12] X. Chen, S. Member, and X. Lin, “Big Data Deep Learning : Challenges and Perspectives,”. *2014 IEEE. Translations and content mining are permitted for academic research only*, vol. 2, pp. 2169-3536, 2014.

[13] E. Duman, “A Novel and Successful Credit Card Fraud Detection System,” *IEEE 13th International Conference on Data Mining Workshops*, 2015.

[14] R. Rieke, M. Zhdanova, J. Repp, R. Giot, and C. Gaber , “Fraud Detection in Mobile Payments Utilizing Process Behavior Analysis,” *Int. Conf. Availability, Reliab. Secur*, pp. 662–669, 2013.

[15] V. Van Vlasselaer, C. Bravo, O. Caelen, T. Eliassi-rad, L. Akoglu, M. Snoeck, and B. Baesens, “APATE : A novel approach for automated credit card transaction fraud detection using network-based extensions,” *Decis. Support Syst.*, vol. 75, pp. 38–48, 2015.

[16] L. Zhang, J. Yang, and B. Tseng, “Online Modeling of Proactive Moderation System for Auction Fraud Detection. International World Wide Web Conference Committee,” pp. 669–678, 2012.

[17] C. Classifier, J. Kim, K. Choi, G. Kim, and Y. Suh, “Expert Systems with Applications Classification cost : An empirical comparison among traditional classifier,” *Expert Syst. Appl*., vol. 39, no. 4, pp. 4013–4019, 2012.

[18] R. C. Newman: Cybercrime , Identity Theft , and Fraud , “Practicing Safe Internet - Network Security Threats and Vulnerabilities,” *InfoSecCD Conference* , 2006.

[19] K. Ramakalyani and D. Umadevi, “Fraud Detection of Credit Card Payment System by Genetic Algorithm,” *International Journal of Scientific & Engineering Research*, vol. 3, no. 7, pp. 1–6, 2012.

[20] T. Tian, J. Zhu, F. Xia, X. Zhuang, and T. Zhang, “Crowd Fraud Detection in Internet Advertising,” *International World Wide Web Conference Committee*, 2015.

[21] F. Louzada and A. Ara, “Expert Systems with Applications Bagging k-dependence probabilistic networks : An alternative powerful fraud detection tool,” *Expert Syst. Appl.*, vol. 39, no. 14, pp. 11583–11592, 2012.

[22] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, “Deep learning applications and challenges in big data analytics,” pp. 1–21, 2015.

[23] V. Aggelis and W. P. Sa, “Offline Internet Banking Fraud Detection. ARES,” , 2006.

[24] A. Leung, Z. Yan, and S. Fong , “On designing a flexible e-payment system with fraud detection capability,” *Proc. - IEEE Int. Conf. E-Commerce Technol. CEC*. pp. 236–241, 2004.

[25] S. S. Mhamane and L. M. R. J. Lobo, “Internet banking fraud detection using HMM,” *Third Int. Conf. Comput. Commun. Netw. Technol*. pp. 1–4 , 2012.