Research Article

ARTIFICIAL INTELLIGENCE AND BIG DATA IN FRAUD DETECTION[1*]

Mehmet Emre ÖZENGEN[1], Ali OKATAN[2]
Can BALKAYA[3]

[1]Student at Big Data Analytics and Management Master Program, Bahcesehir University, Istanbul, mehmetemre.ozengen@bahcesehir.edu.tr, ORCID: 0000-0001-6610-6463

[2]Professor at Department of Software Engineering, Istanbul Aydın University, Istanbul, aliokatan@aydin.edu.tr, ORCID: 0000-0002-8893-9711

[3]Professor at Department of Civil Engineering, Istanbul Aydın University, Istanbul, canbalkaya@aydin.edu.tr, ORCID: 0000-0002-0689-2746

ABSTRACT

Artificial intelligence is used for many purposes nowadays. With the developments in technology, the fraudsters develop their methods. On the other hand, artificial intelligence methods are used in fraud detection for increasing the efficiency of corporations. AI and big data play an important role in real time data enrichment, deep learning integration and decisions. There are ten artificial intelligence methods explained which are used for fraud detection. Each method has its unique bases and it can not be said that there is only one optimal method. In this research, the methods are briefly explained, and a comparison is done for accuracy of methods. Supervised machine learning, unsupervised machine learning or semi-supervised machine learning as well as adaptive machine learning techniques against adaptive attacks with the advantage of big data and artificial intelligence are discussed with effectiveness usage for the future applications.

Keywords: *Artificial intelligence, Big data, Fraud detection, Supervised machine learning, Unsupervised machine learning, Adaptive machine learning.*

[1*] Received: 05.06.2021 - Accepted: 01.08.2021
DOI: 10.17932/EJEAS.2021.024/ejeas_v01i2001

1. INTRODUCTION

Fraud can be defined as illegally obtaining services, goods, or money belonging to other people or organizations and it is one of the greatest challenges for business and organizations. All the systems containing financial transactions are subject to fraud and most of its different forms are determined as a kind of crime in laws. Rapidly increasing volume of e-commerce is attracting the fraudsters equipped with new technologies and the technological defense tools against fraud become more crucial every day. Preventing and detecting fraud is becoming more important more than ever.

Fraud detection is always both related to and fed by data mining and text mining even before the emergence of 'Big Data' phenomenon. However, before the Big Data research techniques developed, there were limited set of ways to develop algorithms to analyze huge amounts of data.

In this research paper, it is aimed to present main ideas of the papers which reviewed the most common and most practiced artificial intelligence techniques of fraud detection. Machine learning models for fraud detection as supervised machine learning models (SMLM), unsupervised machine learning models (UMLM) and semi-supervised machine learning models (SSMLM) as well as adaptive machine learning techniques against adaptive attacks are also discussed with the advantage of using big data and artificial intelligence for the future applications.

2. FRAUD TYPES AND DETECTION METHODS

2.1 FRAUD TYPES

Fraud is a broad term that includes many different types. Submitting fake documents while applying for a job can be defined as a fraud. On the other extent, making manipulations in the financial tables of a big multinational corporation can also be defined as a fraud. In this research, we determined the limit of fraud as financial frauds. Financial fraud can be determined with the help of two factors: The financial gain and an illegal method implementation. Limiting the framework to financial fraud gives the advantage of using fiscal terms and scales.

Financial fraud has several subsets. There are mainly three industries vulnerable to never ending fraud attempts. The first is banking. The instruments generally attached to fraud cases are listed as the credit cards, the mortgages and complex cash transactions involving money laundering. Fraud can occur in the appliance or distribution phases of credit transactions. The second sector is insurance. The most probable fraud attacks may be on the healthcare and auto insurance instruments. The last industry involving the greatest risk of fraud is the telecommunication industry. There are mainly two areas in that sector. Namely, subscription fraud in which the fraudsters obtain telecommunication accounts without paying and the other is superimposed fraud, in which the legally registered customers pay the fraudsters' expenses.

Detection of fraud is basically a classification problem and if it is not done efficiently, it may be costly for the firms. Because of that, many artificial intelligence techniques aim to increase efficiency in classification [1]. In AI terms, classification can be defined as predicting of a result with the use of inputs. To implement modelling to classify, there must be a training dataset. By benefitting the training dataset, the success of the model can be measured in a test dataset. There are types of classification namely, binary classification, multi-class classification, multi-label classification and imbalanced classification. Artificial intelligence fraud detection techniques generally use imbalanced classification in which most of the training dataset belong to normal values and the minority is labelled as abnormal. The reason of that is in real life, the fraud cases make up a very small proportion of the whole cases.

Early fraud detection works were mainly built on statistical methods such as logistic regressions and neural networks. After that, data mining techniques were implied. Finally, the hybrid methods are the main way of fraud detection. Therefore, it is very natural that the techniques are evolving and will be improved in the future.

There are mainly two drawbacks to make fraud detection research more challenging for the researchers. The first is fraud detection techniques are mostly specialized for every different companies. The second is that accessing the real-world data is not so easy because of the privacy issues. Therefore, many scholars attempt to make research on different sectors and put weight on comprehensive analysis.

2.2 ARTIFICIAL INTELLIGENCE TECHNIQUES USED FOR FRAUD DETECTION

In this section, the techniques are summarized to maintain a theoretical base.

Bayesian Belief Networks: A Bayesian belief network uses a classifier to calculate for all possible classes and inserts the value X into the class with the highest probability. In this way, the network is shown to classify each sample into a class that it is most likely to belong to [2].

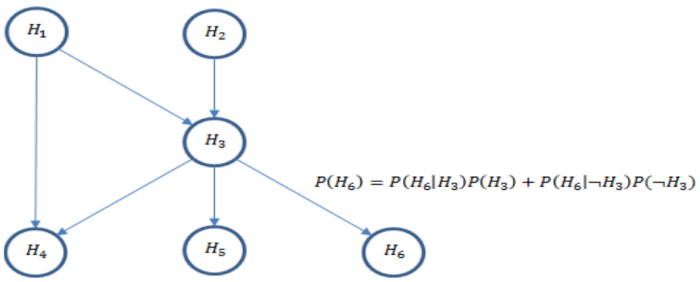

Figure 1: Bayesian Belief Networks Method [3].

Bayesian Belief Networks: A Bayesian belief network uses a classifier to calculate for all possible classes and inserts the value X into the class with the highest probability. In this way, the network is shown to classify each sample into a class that it is most likely to belong to [2].

Logistic Regression: It is statistical method of classifying binary data by using a linear model. It is generally used for predicting of the probability of a case is whether fraudulent or not [4].

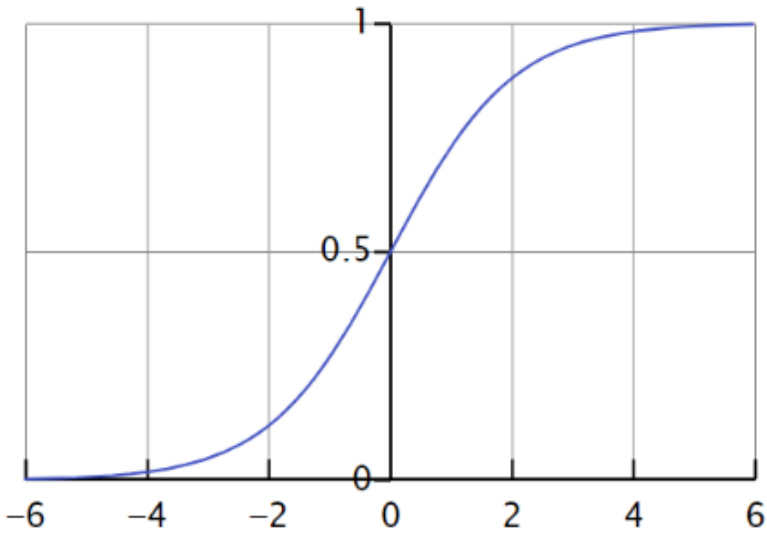

Figure 2: Logistic Regression Method [5].

Neural Network: An artificial network of neurons or nodes is used in this method. The connections of the neurons are modeled as weights. Positive values of weights represent excitatory connection, and the negative values represent inhibitory connections. After all the weights are aggregated, a function controls the amplitude of the output. This technique is known for its relevance to the predictive modelling which is also used to predict the cases are fraudulent or not.

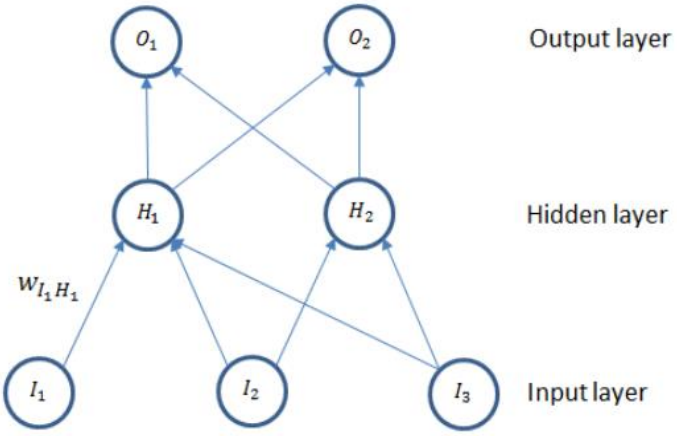

Figure 3: Neural Network Method [3].

Support Vector Machine: This method is a kind of machine learning algorithms and is used for both classification and regression analysis. This method enables complicated non-linear problems to be solved by linear classification without increasing the demand of computational complexity [3].

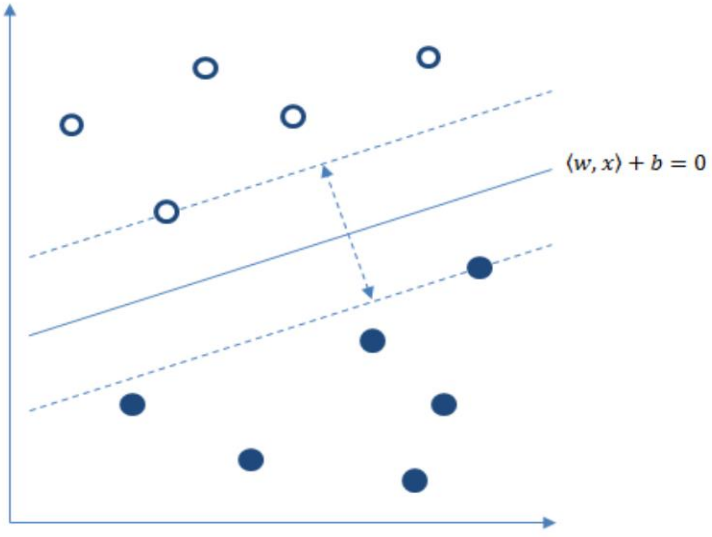

Figure 4: Support Vector Machine Method [3].

Genetic Algorithms and Programming: This method has the aim of solving problems by evolving an initially random set of possible solutions, through the application of operators inspired by natural genetics and natural selection, such that in time the best solutions would last only. In other words, it is the search of the most optimal program among other algorithms. It selects different parts of programs and produces new generation of programs while combining them. Genetic algorithms are like neural networks in that they require no prior knowledge of the problem domain and are capable of detecting underlying relationships between the samples [6].

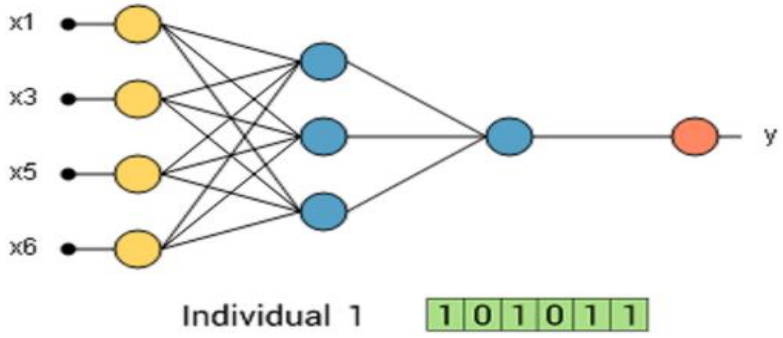

Figure 5: Genetic Algorithms and Programming Method [7].

Decision Trees, Forest: Decision trees are a technique used to make classifications or predictions on data. It uses a tree with internal nodes representing binary choices on attributes and branches representing the outcome of that choice. The nodes are created by artificial intelligence by using the dataset and it makes decision branches until it is eventually sorted into a mutually exclusive subgroup [2].

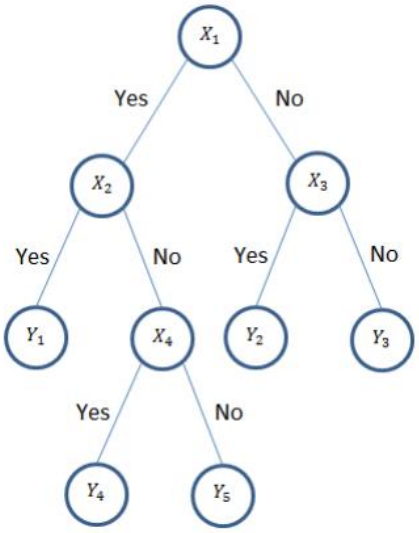

Figure 6: Decision Tree Method [3].

Group Method of Data Handling: It is a deep learning data mining algorithm that calculating optimal solutions through a series of models and increases the accuracy of the model. It has an inductive nature, unlike the other deductive methods. It aims to minimize the coders' influence on result of modelling through a set of several algorithms including parametric, clusterization, analogues complexing, rebinarization and probability algorithms. It finds interpretable relations in dataset and selects effective features.

Text Mining: It is a kind of data mining based on plain text data. It filters out the stop words like 'the', 'is', or 'a'. After that, it reduces the derived forms of words into their roots. Finally, it analyzes the data according to the frequencies of words. It basically transforms the text-based data into a quantitative dataset.

Self-Organizing Map: This method is a kind of artificial neural network method which is built on a single matrix of neurons. A high-dimensional data is reduced into a 2-dimensional matrix form. The difference between self-organizing maps and artificial neural networks is that the first applies competitive learning, whereas the latter applies error-correcting learning.

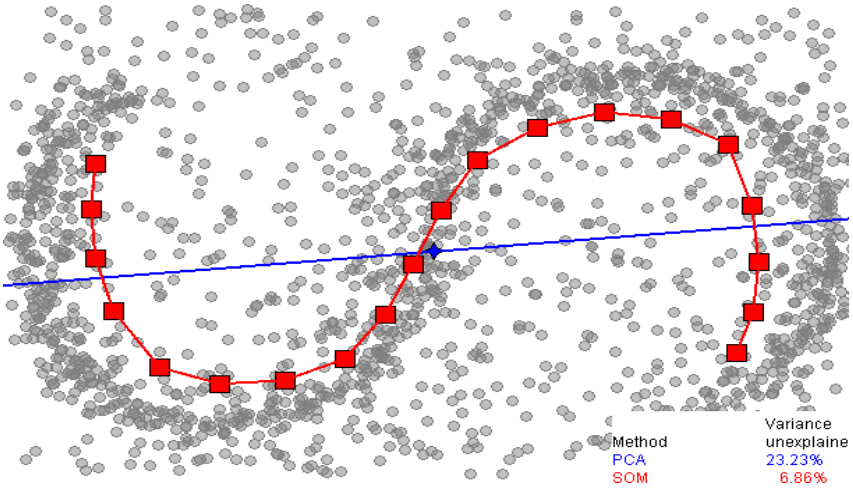

Figure 7: Self-Organizing Map Method [8].

Process Mining: The goal of process mining is to turn event data into insights and actions. Process mining is an integral part of data science, fueled by the availability of data and the desire to improve processes. Process mining techniques use event data to show what people, machines, and organizations are really doing. Process mining uses these event data to answer a variety of process-related questions. Process mining techniques such as process discovery, conformance checking, model enhancement, and operational support can be used to improve performance and compliance [9].

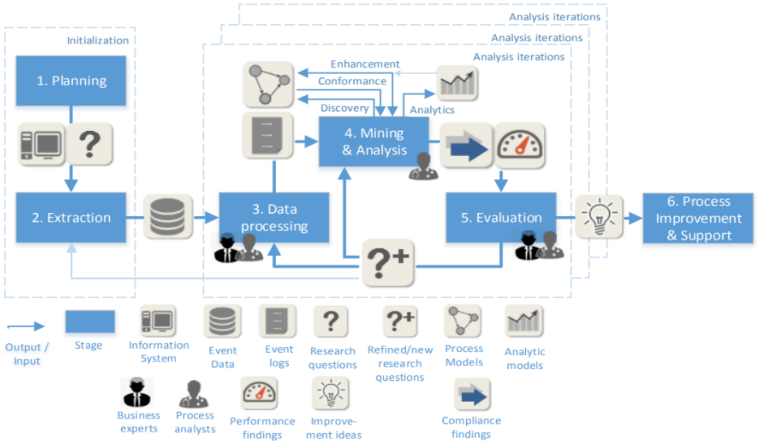

Figure 8: Process Mining Method [10].

Artificial Immune System: It is a class of artificially intelligent, rule-based machine learning systems inspired by the principles and processes of an immune system of clever creatures like humans. The algorithms of this system are modeled very similarly to an immune system and it was inspired by the concept that learning and memory for use in problem-solving. There are four techniques that can be classified as artificial immune system: Clonal selection algorithm, negative selection algorithm, immune network algorithm, dendritic cell algorithm. The first is used for optimization, the second for anomaly detection, third for clustering and visualization, the fourth for multi-scale processing.

Hybrid Methods: They are combinations of the afore-mentioned methods. They can be constructed in a number of ways. The first model's outputs may be applied to another method as an input. One method may be applied as a pre-processing method, while the other makes the essential part. They are constructed to make tailored and specifically targeted solutions.

2.3 OVERVIEWING THE ACCURACY OF METHODS

It is researched that the success of methods in different studies. However, the most comprehensive study was done by West et al. and it shows the comparison of different methods [3]. Interpreting the values in the table, the most accurate methods differ from the datasets. In a study of credit card fraud, the accuracies are very similar. However, when inspecting a dataset of financial statements of manufacturing firms, Bayesian belief networks are more accurate than decision trees and neural networks. In other datasets, the best methods differ also.

Table 1: The comparison made on the accuracies of the methods [3].

Fraud Investigated	Method Investigated	Accuracy
Credit card transaction fraud from a real-world example	Logistic model (regression) Support vector machines Random forests	96.6-99.4% 95.5-99.6% 97.8-99.6%
Financial statement fraud from a selection of Greek manufacturing firms	Decision trees Neural networks Bayesian belief networks	73.6% 80% 90.3%
Financial statement fraud with financial items from a selection of public Chinese companies	Support vector machine Genetic programming Neural network (feed forward) Group method of data handling Logistic model (regression) Neural network (probabilistic)	70.41-73.41% 89.27-94.14% 75.32-78.77% 88.14-93.00% 66.86-70.86% 95.64-98.09%
Financial statement fraud with managerial statements for US companies	Text mining	95.65%
Financial statement fraud with managerial statements for US companies	Text mining Text mining and support vector machine hybrid	45.08-75.41% 50.00-81.97%
Financial statement fraud with managerial statements for US companies	Text mining and decision tree hybrid Text mining and Bayesian belief network hybrid Text mining and support vector machine hybrid	67.3% 67.3% 65.8%
Financial statement fraud with financial items from a selection of public Chinese companies	CDA CART Neural network (exhaustive pruning)	71.37% 72.38% 77.14%

3. MACHINE LEARNING MODELS FOR FRAUD DETECTION

Machine learning can be used with more effectiveness as 1. Supervised Machine Learning Models (SMLM), 2. Unsupervised Machine Learning models (UMLM) and 3. Semi-Supervised Machine Learning Models (SSMLM) against adaptive attacks. Machine learning models need to collect big data to detect fraud. The model analyzes all the input data gathered and extracts the required features. The machine learning model that receives training sets that teach it to predict the probability of fraud. Then, it creates fraud detection machine learning models.

In case of supervised machine learning an algorithm that learns to perform a task from known examples as training data. A supervised learning model is based on predictive data analysis and the accuracy depends on the training set provided for it. SMLM needs large amount of labeled data and has difficulties in detection of unknown data. Labeled data to train the models are the important and quantity of data and quality of the data is the biggest limitation in the supervised machine learning. In a supervised learning model, all input information has to be labeled as good or bad.

On the other hand, unsupervised machine learning will be the future of machine learning for detection of unknown attacks. UMLM has an algorithm that learns to identify linkages and patterns in the data without prior knowledge of what to look for and does not require labeled training data using auto-label and auto-rules generation. Generation large set of features, performing correlation analysis, graph analysis to link fraudulent clusters, identifying attack rings and assigning confidence score and categorizing are the basic steps in UMLM. An unsupervised learning model continuously processes and analyzes new data and updates its models. It learns to notice patterns and decide whether they're parts of legitimate or fraudulent operations. Deep learning in fraud detection is usually associated with unsupervised learning algorithms.

Semi-supervised learning models are somewhere between supervised and unsupervised learning models. SSLM works for cases where labeling information is either impossible or too expensive and requires the labor of human experts.

Effectiveness is increased in SMLM and UMLM with the increasing time however with the advantage of big data, computational time and with the advantage of decision systems in artificial intelligence, these systems can be used effectively for fraud detections.

Adaptive machine learning techniques can also be effective solution for the analysis of datasets and supervised and unsupervised or even semi-supervised machine learning.

4. CONCLUSION

Fraud detection is a very challenging and important subject to explore for increasing efficiency in some industries hence the number of fraud cases increasing with the technology. In this research, the artificial intelligence methods for fraud detection are categorized and reviewed. Some of the methods have statistical approaches, however, some of them have computational approaches. These techniques can be used alone or in a combination. Even though the performances of methods may differ according to

datasets, all of them have unique implementations. Every firm can make a hybrid of these methods according to their own needs. Supervised machine learning, unsupervised machine learning or semi-supervised machine learning as well as adaptive machine learning techniques against fraud detection and adaptive attacks with the advantage of big data and artificial intelligence can be used effectively for the future applications.

REFERENCES

[1] E. Duman, and M.H. Ozcelik, "Detecting credit card fraud by genetic algorithm and scatter search.", Expert Systems with Applications 38, pp 57-63, 2011.

[2] E. Kirkos, C. Spathis, and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements.", Expert Systems with Applications 32, pp 995-1003, 2007.

[3] J. West, and M. Bhattacharya, Intelligent financial fraud detection: a comprehensive review, Computers & Security, pp 47-66, 2015.

[4] E. Ngai, Y. Hu, Y. Wong, Y.Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature", Decision Support Systems 50, pp 559-69, 2011.

[5] Logistic regression. (2021, January 03). Retrieved January 28, 2021, from https://en.wikipedia.org/wiki/Logistic_regression

[6] P. Ravisankar, V. Ravi, G. Raghava, and I. Bose, Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Systems 50, pp 491-500, 2011.

[7] Genetic algorithms for feature selection. (n.d.). Retrieved January 28, 2021, from https://www.neuraldesigner.com/blog/genetic_algorithms_for_feature_selection

[8] Self-organizing map. (2021, January 03). Retrieved January 28, 2021, from https://en.wikipedia.org/wiki/Self-organizing_map

[9] Van der Aalst, W., Process Mining: Data Science in Action, p 77, 2016.

[10] M. L. van Eck, X. Lu, S.J.J. Leemans, and Wil M.P. van der Aalst, "PM^2: A Process Mining Project Methodology.", Advanced Information Systems Engineering, pp 297-313, 2015.