AI-based model for fraud detection in bank systems

Ahmed Al-Fatlawi*, Ahmed A. Talib Al-Khazaali, Sajjad H. Hasan

Department of Computer Techniques Engineering University of AlKafeel Al-Najaf, Iraq

Emails: ahmed.fatlawi@alkafeel.edu.iq; ahmed.ali@alkafeel.edu.iq; sajad.hadi@alkafeel.edu.iq

Abstract

Due to the very high direct or indirect costs of fraud, banks and financial institutions seek to accelerate the recognition of the activities of fraudsters. The reason for this is its direct effect on serving the customers of these institutions, reducing operating costs and remaining as a reliable and valid financial service provider. On the other hand, in recent years, with the development of information and communication technology, electronic banking has become very popular. In the meantime, it is inevitable to use fraud detection techniques to prevent fraudulent actions in banking systems, especially electronic banking systems. In this paper, a method has been developed that leads to the improvement of fraud detection in information security and cyber defense systems. The main purpose of fraud detection systems is to predict and detect false financial transactions and improve the intrusion detection system using information classification. In this regard, the genetic algorithm, which is known as one of the stochastic optimization methods, is used. At the end, the results of the genetic algorithm have been compared with the results of the decision tree classification and the regression tree. The simulation results show the effectiveness and superiority of the proposed method.

Keywords: Artificial intelligence; Intrusion detection system; Genetic algorithm; Banking system; Information security; Cyber defense; Fraud detection

1. Introduction

Fraud, a practice as old as human civilization itself, has now evolved into a global multi-million-dollar enterprise, with its financial scale continuously on the rise. Recent years have witnessed the proliferation of new technologies, which have furnished fraudsters and criminals with a multitude of avenues for perpetrating fraudulent activities. The introduction of novel information systems, while offering numerous advantages, also presents an increased potential for criminals to engage in fraudulent acts. In response to this challenge, fraud detection techniques not only aim to identify and analyze fraudulent activities within organizations but also strive to predict future fraudulent behavior by examining user or customer conduct, thereby mitigating the risk of fraud [1-3]. The advent of modern technologies has spurred significant transformations in the realm of banking. Consequently, traditional banks are compelled to adapt by either modernizing their operations or focusing on the integration of new technologies. In the last decade, the ease of online payments has created great opportunities in e-commerce. Although e-commerce has boomed in recent years, it is still a playground for fraudsters who try to exploit the transparency of online purchases and transfers through credit cards [4-6].

Any electronic banking system must consider issues such as authentication, confidentiality, integrity, non-repudiation, and other security factors and ensure that only authorized persons can access authorized, confidential information and customer accounts, and that transaction records are untraceable. The issue of trust in the online banking environment is more important than in offline banking because creation and cultivation of trust is important when uncertainty and risk are pervasive [7]. The extraordinary growth in the number of internet transactions, especially for online purchases, which recently led to an inherent increase in fraudulent activities, makes it necessary to design and use a fraud detection system for all internet services of financial institutions, especially banks, to reduce risk and losses. Therefore, fraud detection has become one of the most important research topics in these years. Also, due to the value of information and the increase of huge amount of data as daily operations, they use technologies such as data mining to transfer knowledge from data. For this reason, data mining is becoming an important pole for many business organizations, including the banking sector [8-10].

Doi: https://doi.org/10.54216/FPA.140102
Received: September 29, 2023 Revised: November 15, 2023 Accepted: December 29, 2023

According to the said content, with the increasing dependence of banks and payment systems on internet platforms and the like, the ways of cheating in these systems will also increase. The word "fraud" here refers to the misuse of an organization's profit system without necessarily leading to direct legal consequences. These frauds are increasing in an incredible way, and annually include losses of billions of dollars for the capital owners as well as the banks that implement these modern systems. Therefore, it should be possible to identify frauds through methods, methodologies and even algorithms in the fields of statistics, mathematics, data mining, etc., or prevent them by recognizing patterns. It is impossible to comment definitively about the legitimacy of a transaction or request. The least expensive option is to extract possible evidence of fraud from the available data, which is done by mathematical algorithms [11].
With much research that have been done in different societies, especially advanced societies, the analysis engines in these solutions and software are based on artificial security systems, artificial intelligence, audit, database, parallel and distributed computing, economics, professional systems, fuzzy logic, genetic algorithms, machine learning, neural networks, pattern recognition, statistics, visualization, and other fields are obtained. "Fraud detection" involves monitoring the behavior of millions of users to predict, detect, or prevent undesirable behavior. There are different solutions to detect fraud with the given definition, some of them are innovative optimization process in the field of data [12-15, 17-18].
The purpose of using the genetic algorithm to detect fraud in this paper is to achieve better solutions over time. This algorithm has been used in data mining, especially for variable selection. Genetic algorithm is a member of the family of computational models inspired by the evolution process. This algorithm encodes the potential solutions of a problem in the form of simple chromosomes and then apply combinatorial operators on these structures. Genetic algorithm is often known as a method for optimizing functions, although the scope of using these methods is much wider than this.

2. Problem Modelling

In this method, more studies are conducted on fraud detection methods in the information security system and cyber defense of private banks. Studies are also done to get familiar with the MATLAB programming language, as well as the implementation of the genetic algorithm in the MATLAB software.
The main source of information for the implementation of the records registered in the server of a private bank in the country. In addition, a dataset in Germany has been used. Also, in this paper, two methods of deep thinking and survey study, which include publications, Persian and Latin papers related to the subject, have been widely used.

2-1- The investigated parameters

The investigated features in the mechanism for detecting fraud in banking systems include precision, recall, accuracy, and score criterion, which are given in equations 1 to 4, respectively.

$$\text{Precision} = \frac{tp}{tp + FP} \tag{1}$$

$$\text{Recall} = \frac{tp}{tp + fn} \tag{2}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \tag{3}$$

$$\text{F}_{score} = \frac{2tp}{2tp + fp + fn} \tag{4}$$

2-2- Genetic algorithm

Genetic algorithm is an iterative process that creates the best solution value for the given features by generating a predetermined number of generations. The implementation mechanism of the genetic algorithm to detect fraud behaves as an intrusion detection system [16]. In this way, the current values of the parameters are determined and compared with the critical values in the dataset parameters. This process is shown in Figure 1.
Considering the need for a classification method in addition to the genetic algorithm for features classification, the decision tree method has been used as an innovative method.

2-3- Database

The required data is obtained from the following sites and databases:

Doi: https://doi.org/10.54216/FPA.140102
Received: September 29, 2023 Revised: November 15, 2023 Accepted: December 29, 2023

Statlog (German Credit Data) Data Set
This data set describes the classification of different people by a set of features such as good or bad credit card risks. This dataset classifies people with features such as good or bad credit risks. It specifies the classification in two formats along with a cost matrix.

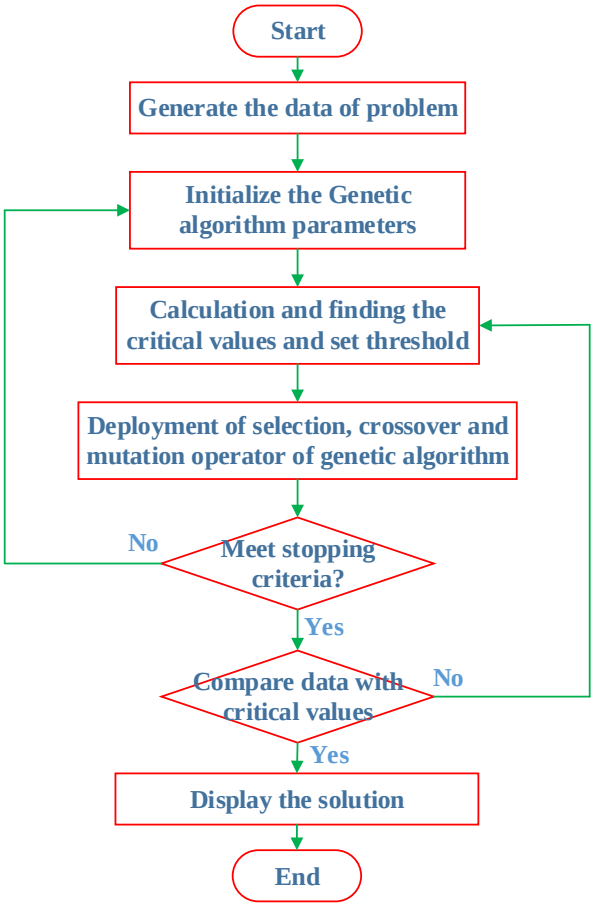

Figure 1: The implementation of genetic algorithm mechanism to detect the fraud in the banking system

Table 1: Features of the dataset

Dataset type	multivariable
Characteristic features	A group of integers
Related tasks	Classification
Number of samples	1000
Number of features	20
Missing amounts	N/A
Area	Financial
Creation Date (Donation)	1994-11-17
Number of web views	253439

Dataset information is as follows:
Two datasets are provided. The original dataset in a form by Professor Hoffman contains a series of symbolic/classification features that are present in the German data file.
For an algorithm that requires numerical features, the University of Strathclyde has produced a numerical file of German data. This file has been edited and several indicator variables have been added to it, which are suitable for algorithms that cannot be categorized. For example, one of the characteristics is coded as an integer. This format is used by statlog. This dataset requires the use of a cost matrix.
The rows represent the real classification, and the columns represent the predicted classification. It is worse when a customer class is considered good even though they are bad than when a customer class is considered bad even though they are good.
The variables are divided into two qualitative and quantitative categories: the amount of bank credit in the said bank, the deposit rate multiplied by the percentage of disposable income, the current employment of users from

Doi: https://doi.org/10.54216/FPA.140102
Received: September 29, 2023 Revised: November 15, 2023 Accepted: December 29, 2023

the current date, the age of users in the current year, the number of credits available in the said bank, the number of people who are responsible for maintaining (repairs), existing checking account status, stored credit card records, savings/equity account, marital status and gender, guarantor/debtor status, housing status, occupation, and users' phone numbers.

3. Simulation results

In this research paper, we employ the retentive approach of cross-validation to partition the dataset. This cross-validation technique entails the extraction of a portion of the training data, which is subsequently employed as a test dataset. Initially, the model undergoes training using the training set, and then it is tasked with predicting outcomes based on the test set. This basic form of cross-validation proves particularly valuable when dealing with extensive datasets or necessitating a swift and straightforward validation process. The procedure for data partitioning within this method is as follows:

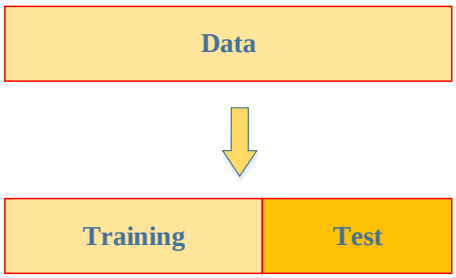

Figure 2: Data dividing method in cross validation method.

The common approach for data retention often includes partitioning a dataset, with around 20-30% allocated for testing, while the remaining portion is designated for training. These proportions may be adjusted as needed. Increasing the proportion of test data can increase the model's susceptibility to errors since it receives less training exposure, whereas a lower percentage of test data might introduce undesirable bias towards the training data. This deficiency in training or introduced bias can result in inadequate model training.

Also, in this paper, since the purpose of feature selection was with the help of genetic algorithm, the training data is divided into training and validation categories. Accordingly, the dataset used in this paper is divided into the following three categories:

- Training 50%
- 20% validation
- 30% test

Considering that the main goal of this paper is to detect and classify fraud in information security and cyber defense, criteria related to classification should be used in this paper to evaluate the proposed method. Considering that, the purpose of this research was data classification, the disturbance matrix, and the relationships in it were used to evaluate the proposed method. The disturbance matrix is shown in Table 2.

Table 2: Features of the dataset

		Real label	
		Positive class	Negative class
Predicted	Positive class	True Positive (TP)	False Positive (FP)
	Negative class	False Negative (FN)	True Negative (TN)

In this paper, three classification algorithms have been investigated, these three algorithms are:

- Decision tree and classification (DT)
- Support vector machine with kernel (RBF)
- Aggregated (or combined) classification tree with the number of base learners 15 (AdDT)

In the following, the results obtained on training, validation and test data are presented and compared.

3-1- Results obtained on training data

In this section, the results obtained on the training data will be presented based on the classifications used. Accordingly, the disturbance matrix will also be given. In Tables 3 to 5, the first means the results obtained before selecting the feature, and the second means the results obtained after selecting the feature.

Table 3: Results obtained on training data

Doi: https://doi.org/10.54216/FPA.140102
Received: September 29, 2023 Revised: November 15, 2023 Accepted: December 29, 2023

	Classified tree		SVM		AdDT	
	before	after	before	after	before	after
precision	89.13	90.40	77.20	76.60	77.00	77.00
recall	95.31	95.71	90.29	90.29	93.43	90.57
accuracy	90.73	91.03	79.80	79.20	78.04	79.45
score criterion	92.73	93.31	84.72	84.38	85.05	84.65

Based on the above table, it is clear on the training data:

- The classification tree has performed better. Also, in most cases, feature selection has led to better results.
- AdDT blended learning has led to better results.

Following are the disturbance matrices for the three classifiers used after feature selection on the training data.

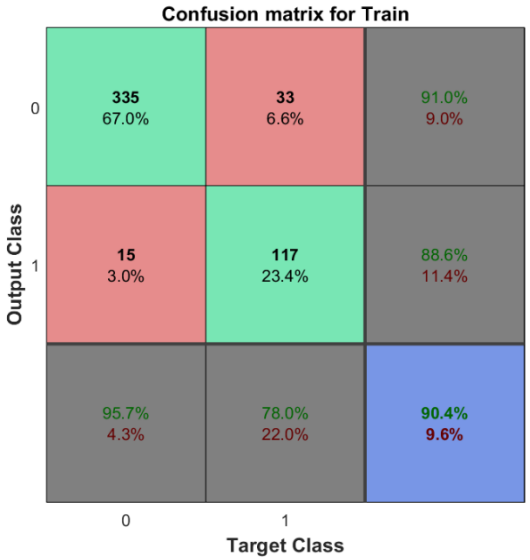

Figure 3: Disturbance matrix on training data and decision tree classification

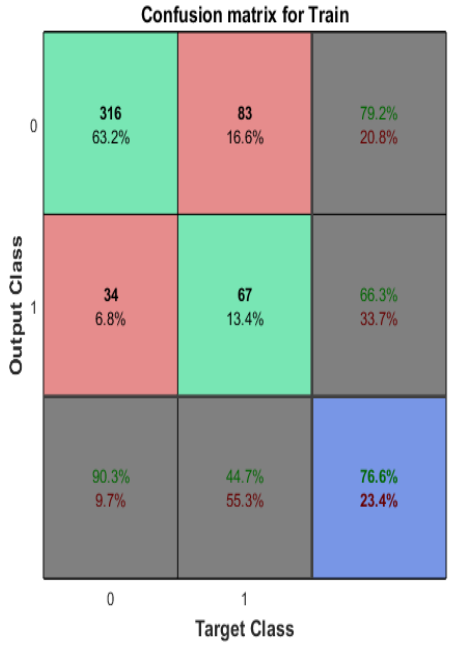

Figure 4: Disturbance matrix on training data and SVM classification

Doi: https://doi.org/10.54216/FPA.140102
Received: September 29, 2023 Revised: November 15, 2023 Accepted: December 29, 2023

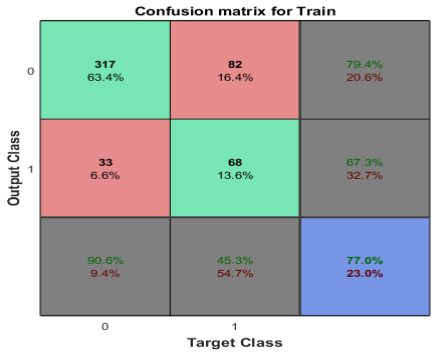

Figure 5: Disturbance matrix on training data and AdDT classification

3-2- Results obtained on validation data

In this section, the results obtained on the validation data based on the used classifications will be presented. Accordingly, the disturbance matrix will also be given.

Table 4: Results obtained on validation data

	Classified tree		SVM		AdDT	
	before	after	before	after	before	after
precision	68.53	69.00	75.00	77.50	73.50	74.00
recall	83.12	83.57	90.71	90.00	91.43	89.29
accuracy	74.45	75.00	77.44	80.25	75.74	77.16
score criterion	78.73	79.05	83.55	84.85	82.85	82.78

Following are the disturbance matrices for the three classifiers used after feature selection on the validation data:

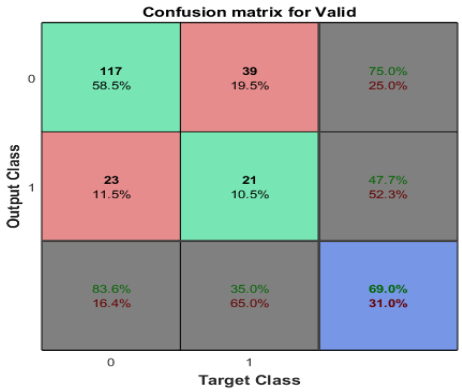

Figure 6: Disturbance matrix on validation data and decision tree classification

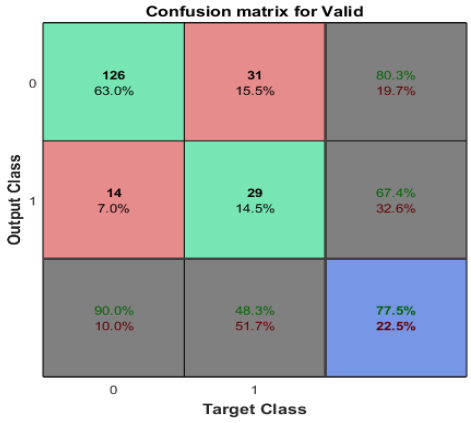

Figure 7: Disturbance matrix on validation data and SVM classification

Doi: https://doi.org/10.54216/FPA.140102
Received: September 29, 2023 Revised: November 15, 2023 Accepted: December 29, 2023

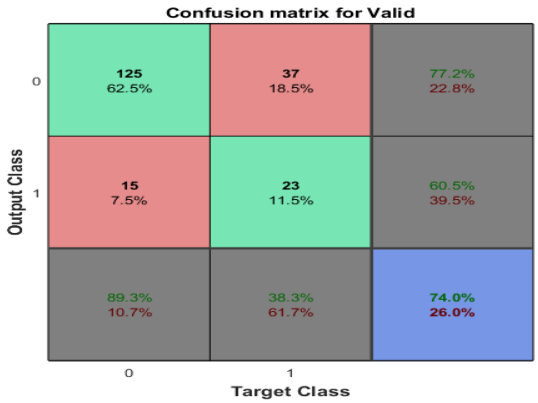

Figure 8: Disturbance matrix on validation data and AdDT classification

3-3- Results obtained on test data

In this section, the results obtained on the test data will be given based on the classifications used. Accordingly, the disturbance matrix will also be given.

Table 5: Results obtained on test data

	Classified tree		SVM		AdDT	
	before	after	before	after	before	after
precision	71.67	71.67	78.67	79.33	75.67	76.00
recall	83.33	83.33	90.00	91.43	93.33	90.00
accuracy	77.78	77.78	81.47	81.36	76.86	78.75
score criterion	80.46	80.46	85.52	86.10	84.30	84.00

Following are the disturbance matrices for the three classifiers used after feature selection on the test data:

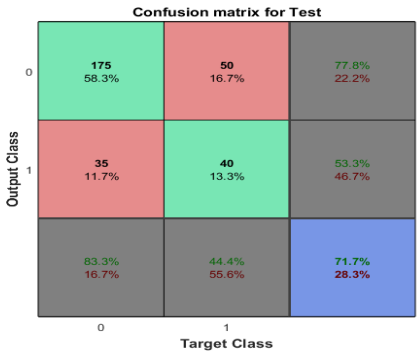

Figure 9: Disturbance matrix on test data and decision tree classification

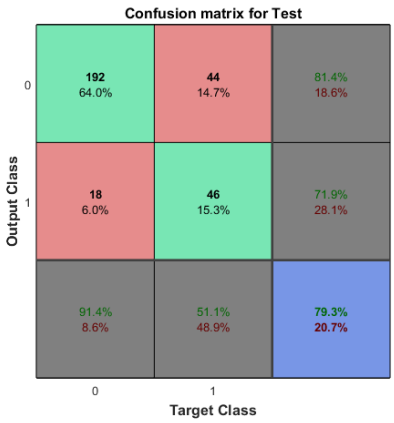

Figure 10: Disturbance matrix on test data and SVM classification

Doi: https://doi.org/10.54216/FPA.140102
Received: September 29, 2023 Revised: November 15, 2023 Accepted: December 29, 2023

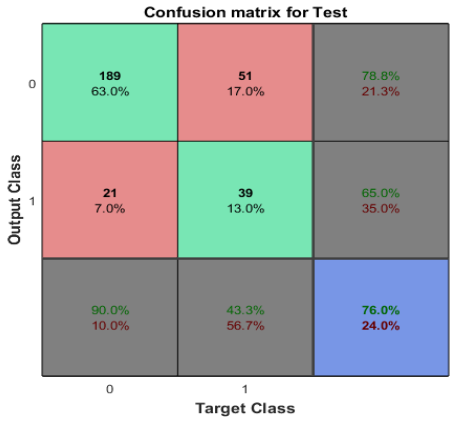

Figure 11: Disturbance matrix on test data and AdDT classification

In this section, the obtained results are analyzed based on the tables and figures given, it can be said:

- Feature selection can improve the efficiency of classification algorithms.
- The efficiency of the models on the training, validation and test data was almost close to each other only in one case which was related to the decision tree.
- In the decision tree, due to the better performance on training data compared to other data, the model has more fit.

4. Conclusion

With the expansion of financial and monetary institutions, they are strongly looking for acceleration and speed of action in recognizing the activities of fraudsters and fraudsters. This is due to its direct effect on serving the customers of these institutions, reducing operational costs, and remaining as a reliable financial service provider. Therefore, it is inevitable to use fraud detection techniques in the information security mechanism and cyber defense to prevent fraudulent actions in banking systems, especially electronic banking systems. With the advent of modern technologies, many changes have been made in the banking business. For this reason, traditional banks must either update themselves or focus on new technologies. In the last decade, the ease of online payment has created great opportunities in e-commerce. Although e-commerce has boomed in recent years, it is still a playground for fraudsters who try to exploit the transparency of online purchases and transfers through credit cards. Based on the obtained results, it can be said that feature selection can improve the efficiency of artificial intelligence and machine learning models. The results show the effectiveness of the genetic optimization method in detecting fraud in banking systems.

References

[1] Mohammed G. Fathi Al-Obaidi, Intelligent Classification for Credit Scoring Based on a Data Mining algorithm, Journal of Intelligent Systems and Internet of Things, Vol. 9 , No. 2 , (2023) : 149-161 (Doi : https://doi.org/10.54216/JISIoT.090211)

[2] Darwish, Saad M. "A bio-inspired credit card fraud detection model based on user behavior analysis suitable for business management in electronic banking." *Journal of Ambient Intelligence and Humanized Computing* 11, no. 11 (2020): 4873-4887.

[3] Bagga, Siddhant, Anish Goyal, Namita Gupta, and Arvind Goyal. "Credit card fraud detection using pipeling and ensemble learning." *Procedia Computer Science* 173 (2020): 104-112.

[4] Rai, Arun Kumar, and Rajendra Kumar Dwivedi. "Fraud detection in credit card data using unsupervised machine learning based scheme." In *2020 international conference on electronics and sustainable communication systems (ICESC)*, pp. 421-426. IEEE, 2020.

[5] Darwish, Saad M. "An intelligent credit card fraud detection approach based on semantic fusion of two classifiers." *Soft Computing* 24, no. 2 (2020): 1243-1253.

[6] Sadgali, Imane, S. A. E. L. Nawal, and Fouzia Benabbou. "Fraud detection in credit card transaction using machine learning techniques." In *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, pp. 1-4. IEEE, 2019.

Doi: https://doi.org/10.54216/FPA.140102
Received: September 29, 2023 Revised: November 15, 2023 Accepted: December 29, 2023

[7] Noor Hanoon Haroon, Hanan Burhan Saadon, Ansam Mohammed Abed, Ahmed Taha, Maryam Ghassan Majeed, Marwan Qaid Mohammed, Salem Saleh Bafjaish, Developing a Smart Economy Using Statistical Framework-Based Business Models in Smart Cities, Journal of Intelligent Systems and Internet of Things, Vol. 9 , No. 2 , (2023) : 194-205 (Doi : https://doi.org/10.54216/JISIoT.090214

[8] Reem Atassi, Fuad Alhosban, Predictive Maintenance in IoT: Early Fault Detection and Failure Prediction in Industrial Equipment, Journal of Intelligent Systems and Internet of Things, Vol. 9 , No. 2 , (2023) : 231-238 (Doi : https://doi.org/10.54216/JISIoT.090217)

[9] Al-Hashedi, Khaled Gubran, and Pritheega Magalingam. "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019." *Computer Science Review* 40 (2021): 100402.

[10] Bagga, Siddhant, Anish Goyal, Namita Gupta, and Arvind Goyal. "Credit card fraud detection using pipeling and ensemble learning." *Procedia Computer Science* 173 (2020): 104-112.

[11] Sahu, Aanchal, G. M. Harshvardhan, and Mahendra Kumar Gourisaria. "A dual approach for credit card fraud detection using neural network and data mining techniques." In *2020 IEEE 17th India council international conference (INDICON)*, pp. 1-7. IEEE, 2020.

[12] Darwish, Saad M. "A bio-inspired credit card fraud detection model based on user behavior analysis suitable for business management in electronic banking." *Journal of Ambient Intelligence and Humanized Computing* 11, no. 11 (2020): 4873-4887.

[13] Wang, Chunzhi, Yichao Wang, Zhiwei Ye, Lingyu Yan, Wencheng Cai, and Shang Pan. "Credit card fraud detection based on whale algorithm optimized BP neural network." In *2018 13th international conference on computer science & education (ICCSE)*, pp. 1-4. IEEE, 2018.

[14] Hussein, Ameer Saleh, Rihab Salah Khairy, Shaima Miqdad Mohamed Najeeb, and Haider Th ALRikabi. "Credit Card Fraud Detection Using Fuzzy Rough Nearest Neighbor and Sequential Minimal Optimization with Logistic Regression." *International Journal of Interactive Mobile Technologies* 15, no. 5 (2021).

[15] Aldo Tenis,Santhosh R., Modelling of an Adaptive Network Model for Phishing Website Detection Using Learning Approaches, Fusion: Practice and Applications, Vol. 12 , No. 2 , (2023) : 159-171 (Doi : https://doi.org/10.54216/FPA.120213)

[16] Katoch, Sourabh, Sumit Singh Chauhan, and Vijay Kumar. "A review on genetic algorithm: past, present, and future." *Multimedia Tools and Applications* 80, no. 5 (2021): 8091-8126.

[17] Nadweh, R., " On The Fusion of Neural Networks and Fuzzy Logic, Membership Functions and Weights", Galoitica Journal Of Mathematical Structures and Applications, Vol 7, 2023.

[18] Charchekhandra, B., " The Reading and Analyzing Of The Brain Electrical Signals To Execute a Control Command and Move an Automatic Arm", Pure Mathematics for Theoretical Computer Science, Vol 1, 2023.

Doi: https://doi.org/10.54216/FPA.140102
Received: September 29, 2023 Revised: November 15, 2023 Accepted: December 29, 2023