[1]Surbhi Bansal,
[2]Reena Hooda,
[3]Alpa Yadav

Feature Selection and Optimization of Classifiers for Detection of Credit Card Frauds

JES Journal of Electrical Systems

Abstract: - Nowadays, in the digital era, the use of credit cards is widespread. People buy and sell goods from home. E-commerce brings a boom to the lifestyle of the people, for the working family, it is a preferable tool to purchase and pay through online mode in order to save the time and visiting hours. However, with the increasing use of credit cards, there have also been a lot of cases of credit card frauds that augmented much in the corona period. Fake credit card transactions significantly affect financial institutions, banks, and clients, whereas fraudsters try to develop new ways of committing fraud daily. To avoid credit card fraud and maintain the credibility of our clients, we need to build a model for credit card detection. There are the various machine learning tools that can be applied in detection of such frauds, these tools can be used to design an enhanced model to detect credit card frauds efficiently. In this view, the study presents a feature selection technique approach of machine learning with different classifiers. The study proposed an optimization algorithm using 'Firefly' for feature selection with four different classifiers: neural network, k-nearest neighbors, Support Vector Machine, and decision tree. The performance of all these four classifiers is compared on the basis of accuracy parameter.

Keywords: Credit card, Fraud detection, Neural network, k-nearest neighbors, Support vector machine, Decision tree.

1. INTRODUCTION

The use of digital financial services is increasing very rapidly. E-commerce has made our lives very simple. We can sell or purchase goods at home using credit or debit cards worldwide. A credit card is a small metallic card on which some unique numbers are imposed and issued by banks for electronic payments[1]. But with the increase in digital services, the speed of fraud is also growing, especially after Corona; it has become more prevalent, causing considerable losses to banks, financial companies, and customers. Credit fraud has profoundly affected our daily lives. By financial fraud, we mean activities that involve diverting the holder's money without their knowledge, which is not only eroding the credibility of the finance industry but also affecting people's cost of living. Credit card fraud is a serious issue. A research paper study revealed that out of the monthly active money, only 0.05% was fraudulent, meaning that out of every 10,000 monthly active money, five were fake, resulting in a considerable loss that had to endure[2]. Fraud

[1]Research Scholar ,Department of Computer Science & Engineering, Indira Gandhi University, Meerpur, Rewari, Haryana, India-12250211
[2]Assistant Professor ,Department of Computer Science & Engineering ,Indira Gandhi University, Meerpur, Rewari, Haryana, India-12250222
[3]Assistant Professor ,Department of Botany, Indira Gandhi University, Meerpur, Rewari, Haryana, India-12250223
1surbhibansal2011@gmail.com
2reenah2013@gmail.com
3alpa.yadav21@gmail.com

detection is a process through which we can understand whether a transaction is fraud. When a credit card is used, the service provider database stores transaction data with different features like; receiver, amount, date, and credit card identification, location, time, age and many more[3]. Credit transaction fraud and application fraud are the two most prevalent forms of credit card fraud. Moreover, card-not-present (CNP), lost or stolen cards, counterfeit cards, and identity fraud are all considered forms of credit card fraud. The percentage of fraudsters opening new bank accounts under the name of a victim using stolen identity information increased by 32% in 2022[4]. The FTC reports that about 111,000 Americans reported new account bank fraud in 2022, up from approximately 84,000 in 2021. Many rule based fraud detection techniques are used to detect the fraud in credit card, but fraudsters develop many advance techniques to do so. The abundance of data available to businesses and the increasing capability of hardware have made machine learning approaches more powerful and affordable for solving more complex challenges in our society. Machine learning has become a significant tool in the detection of credit card fraud. With the rise of technology and e-commerce, fraudulent activities have also increased, making it difficult for traditional methods to keep up. Machine learning algorithms can investigate huge amounts of data and identify patterns of fraudulent activities that may not be immediately apparent to human analysts. Machine learning models can process large amounts of data in real-time and make accurate predictions on whether a transaction is fraudulent or not. This ability to detect fraudulent transactions quickly and accurately can save financial institutions and consumers millions of dollars each year. Moreover, machine learning models can be updated and trained continuously, making them more effective at detecting new fraud techniques. The significance of machine learning in credit card fraud detection lies in its ability to improve the safety and security of electronic transactions, ultimately benefiting both financial institutions and consumers. The present paper suggests a hybrid approach to machine learning techniques and the feature selection (FS) method named the firefly algorithm, a bio-inspired approach. Data mining and machine learning methods are ineffective in handling classification and misclassification difficulties. Moreover, transactions with minimal values should not be treated lightly, whereas transactions with maximum values carry greater significance. This research seeks to reduce the frequency of these kinds of issues through bio-inspired approaches. firefly algorithm is a bio inspired algorithm used to select co-related variables.

1.1. Significance of feature selection algorithm

One of the main challenges is the enslavement of dimensionality. Extensive, complex, comprehensive data sets with many features are frequently encountered in the domain of fraud detection. Due to the vast number of dimensions, algorithms may exhibit reduced effectiveness and speed in theory, while performing badly in practice due to computational inefficiencies. Overfitting arises when an algorithm excessively fixates on the information it is already familiar with, hence diminishing its ability to learn from new and unfamiliar information.

Furthermore, there is the issue of superfluous components being included. Excessive inclusion of characteristics in a model might lead to a gradual loss of focus on the key elements. Noise is the underlying cause of both false positives and false negatives, which are both troublesome when it comes to fraud detection. Feature selection is a crucial method for addressing these problems. In order to construct a more focused and efficient model, the feature selection process entails selecting a subset of features that possess the highest level of significance. To significantly improve the model's performance, excluding all characteristics except for the most informative ones is advisable. The computational load is alleviated, and the problem of high dimensionality is diminished as a consequence. Feature selection acts as a filtering process, enabling the algorithm to focus on the most pertinent features of the input. India, a nation that has experienced a swift digital transformation, is greatly concerned about the prevalence of credit card theft. The rapid growth of the online ecosystem and the expanding array of digital payment systems have led to an augmented potential for fraudulent activities. India is ranked among the top countries globally in terms of credit card fraud, which is a significant issue. In 2020, a significant number of credit card theft incidents were reported nationwide, resulting in financial losses amounting to millions of dollars for both people and companies. Ultimately, the exponential growth of online transactions and digital payments has propelled the transformation of credit card theft into a pervasive issue. Traditional fraud detection systems are rendered useless by the sophisticated techniques employed by fraudsters, necessitating the need for more advanced solutions. Machine learning algorithms have significant potential, but their performance can be hindered by problems like dimensionality constraints and the inclusion of irrelevant information. Feature selection has become a crucial strategy for improving the effectiveness of fraud detection models and

overcoming these difficulties. The prevalence of credit card fraud in India is on the rise due to the escalating adoption of digital payment methods by the country's population. Firefly Swarm Intelligence (FSI), an optimisation system, was developed based on the mating habits of fireflies. The firefly's bioluminescence is seen as an optimisation process to solve the problem of selecting the most desirable traits, which has proven to be highly effective in attracting potential mates. FSI, or Firefly Algorithm, is a method that imitates the flight behaviour of fireflies towards brighter individuals. This algorithm is highly effective in finding optimal solutions, especially when dealing with optimisation problems.

2. RELATED WORK

Singh et al. present an innovative method to find credit card fraud termed FFSVM, which integrates a support vector machine with a bio-inspired algorithm. FFSVM has two stages that happen one after the other. In the first level, the firefly algorithm (FFA) and the CfsSubsetEval feature section method were used to improve the subset of features. In the second level, the support vector machine classifier was used to build a training model for finding cases of credit card fraud [5].

PreetiRathi and Nipur Singh Presents an innovative approach to spot growing scams and data mining methods, mainly in credit card and web-based situations. A custom design for finding fraud is made, data is gathered, and information that isn't needed is removed during preprocessing. Search time, memory usage, accuracy, and mistake rate are critical factors to judge efficiency. The study looks at different types of fraudsters and stresses how practical web mining can be, especially when using phish tank datasets. The study says that more research should be done on improving performance by combining clustering and classification algorithms. Using both methods together can speed up pattern recognition and reduce search time[6].

Awoyemi et al. study presents at how well three different models work with a highly skewed CCF dataset. There is a set of credit card transactions from European users in a credit card transaction dataset. The skewed data were used with a hybrid method that combined under-sampling and over-sampling. Additionally, the collected data had been a focus to group using various methods. The simulation work was done in Python environment, and the Matthews function was used to find performance measures like neutrality, coefficient correlation, and correlation factor. Based on the results, these models work best when they get about 98% accuracy with Naïve Bayes, 98% accuracy with K-nearest neighbour, and 55% accuracy with Logistic Regression. The results show that the k-nearest neighbour algorithm works better than the other two [7].

Bagga et al. presents several algorithms to find the best classification tool to identify credit card fraud; It is very challenging, especially in credit card transactions. It critically influences data mining algorithms in identifying fraudulent activity compared, including logistic regression, K-nearest neighbors, random forest, naive Bayes, multilayer perceptron, ada boost, quadrant discriminative analysis, pipelining, and ensemble learning. The choice of address and variable selection have a major impact on detection performance [8].

Rtayli et al demonstrated the importance of feature selection and used the recursive feature reduction method, which picks features one at a time and checks the accuracy of each one. Using SVM for hyper-plane-oriented feature optimisation was also used . Since there are more than two classes in the suggested algorithm design, SVM is not a good choice for multi-class classification[9].

Asha et al. designed a model using classifiers like ANN, SVM, and K-NN to predict occurrence of fraud . Also, a differentiation of using the classifiers had been done to tell the difference between activities that were not fraudulent and those that were. The process of difference makes sure that fraud will happen less often in the future[10].

Alharbi et al. Suggested deep learning (DL)--based method to use the Kaggle dataset to solve the text data problem. A new text2IMG conversion method is introduced that makes small pictures that can be used to identify credit card fraud. For the class imbalance problem, the pictures are fed into a CNN design that gives each class a certain amount of weight. Data mining and natural language processing (ML) methods ensured the proposed system was valid and strong. The suggested CNN's deep features were used by Coarse-KNN, which has a 99.87% performance rate[11].

Rajab Mohsen et al. Artificial neural networks, random forests, logistic regression, and support vector machines are used in the study to show a way to find credit card theft. It utilizes feature selection to figure out the most critical factors that affect the different kinds of deals. Then, the machine learning model is used and evaluated using the

confusion matrix, memory, precision, f-measure, and accuracy. A sample of 284,807 transactions was used to evaluate the performance in detecting and preventing fraud in credit cards[12].

3. PROPOSED WORK

The main target of the research paper is to identify that a transaction is fraudulent or not for this we use machine learning algorithms to train our model since it can analyse the large volume data with high accuracy and odd data pattern of the dataset. Before passing the data to train our model we firstly pre-process our data then pass the processed data to select the related attributes from huge data set by imposing firefly algorithm in the last we evaluate the machine learning algorithm performance Following steps describe about the complete process of the model;

3.1 Dataset

Dataset is downloaded from the kaggle repository[13]There were deals made with credit cards in September 2013 by people in Europe in this dataset. This dataset shows trades that happened in the last two days. Out of the 284,807 transactions, 492 were frauds. The dataset is very unbalanced; 0.172% of all trades are in the positive class, which means they are frauds. Since there are concerns about privacy, the original features and more background details about the data are not given.The main components that PCA found are features V1, V2,... V28. Time and Amount are the only features that haven't been changed by PCA. The Time is calculated by taking into account the seconds between the last transaction in the dataset and the first transaction in dataset . The last column is our target classes which have only two variables either is 0 or 1; 0 represent that our transaction is genuine or 1 represents that the transaction is fraud which was represented in figure no 1.

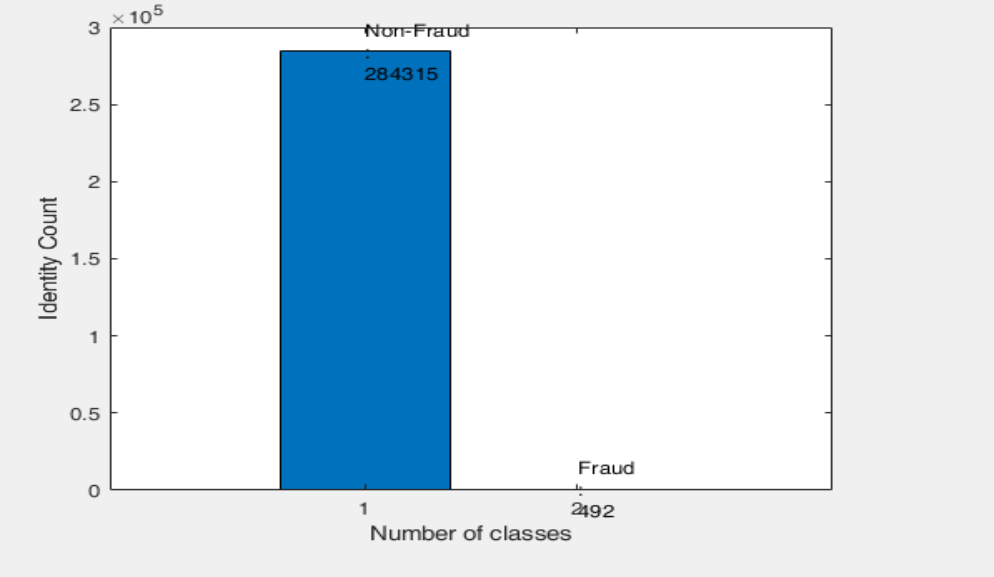

Figure1. Class Distribtion of transections

3.2 Pre- processing

After collecting the data from the dataset, in next step we preprocess our dataset. For pre-processing the data ,and to remove the outliers from dataset. Null values and empty values are replaced with 0 in the transaction to ensure data is correctly pre processed.

3.3 Feature Selection

After preprocessing the data we use a bio inspired firefly algorithm to select the feature from the dataset. Firstly we split our data in the ratio of 70:30, 70 percent of data is used to train the data and 30 percent of data is used to test the

model. Firefly algorithm(FA)[5], [14] was firstly come forth in 2008 by Xin-She Yang, which was based on the lighting pattern and activities shows by the fireflies Essentially, employs the following three idealized principles :

a) fireflies are unisexual; they attract to each other regardless of their sex.

b) Attractiveness is proportional to the light intensity of the firefly and inversely proportional to the distance between them. Thus a less brighter fly attract towards the brighter one. The behavior of a firefly is completely random if it detects no other firefly that is brighter than it.

c) The geography of the goal function determines the brightness of a firefly. When one firefly i is drawn to another, brighter firefly j.

Algorithm 1. Feature selection using firefly

Input: Data matrix data_values, max generations max_gen, features N, classes total_classes

Output: Selected features

1: gen ← 0

2: [rows, cols] ← size(data values)

3: N ← cols

4: AI ← $0_{1\times N}$

5: max_features←round(N× 0.70)

6: max_gen (maximum generations)

7: while gen <max_gen do

8: P ← round(N × 0.70)

9: Sp ← round(rand(1, P) × N) {Randomly select population}

10: Sg ← 5 {Swarm group size}

11: for i ← 1 to |Sp| do

12: Sx_i ← round (rand(1, Sg) × |Sp|) {Randomly pick swarm features}

13: $X_{i,t}$ ← data values(:, Sp(Sx_i)) {Extract features for x_i}

14: for j ← 1 to |Sp| do

15: if i != j then

16: Sx_j ← round(rand(1, Sg)×|Sp|) {Randomly pick swarm features}

17: $X_{j,t}$ ← data_values(:,Sp(Sx_j)) {Extract features for xj}

18: $Mdl_{xi,t}$ ← fitcknn($X_{i,t}$, total_classes,′ NumNeighbors′, 5,′ Standardize′, 1)

19: $Mdl_{xj,t}$ ← fitcknn($X_{j,t}$, total_classes,′ NumNeighbors′, 5,′ Standardize′, 1)

20: F 1 ← predict($Mdl_{xi,t}$, $X_{i,t}$)

21: F 2 ← predict($Mdl_{xj,t}$,$X_{j,t}$)

22: [TPR1, FPR1] ← calculatetprfpr(total_classes,F1)

23: [TPR2, FPR2] ← calculatetprfpr(total_classes, F2)

24: ξ ← mean(TPR1)

25: A ← mean(TPR2)

26: α ← 5

```
27:          β          rand
28:              γ  ←   1
29:              if ξ > then
30:                r ← ξ−
31:                  current AI ← β × exp(−γ × r2) × r + α
32:                Si  ←Sp(Sxi)
33:                AI(Si)  ←  AI(Si) + current AI
34:              else
35:                r ← A
36:    current_AI ← β × exp(−γ_× r2) × r + α
37:                Sj  ←  Sp(Sxj)
38:                AI(Sj)  ←  AI(Sj) + current_ AI
39:                end if
40:                end if
41:                end for
42:                end for
43:          gen  ←gen + 1
44:            end while
45:    [sorted _value, sorted _index] ← sort(AI, descend)
```

Above algorithm is pseudo code for the firefly algorithm used to select the feature from the dataset. The data is processed for two generation; generation 0 and generation 1. Here, AI denotes attractive index and 'n' is the number of columns, where x_i and x_j is the position of two fireflies at time t, now on calculating the value of true positive rate and false positive rate for x_i and x_j by fitting the model using k- nearest neighbor and their mean are stored in variable ξ and A.

$$\text{current AI} \leftarrow \beta \times \exp(-\gamma \times r2) \times r + \alpha \qquad [14]$$

$$Si \leftarrow Sp(Sxi)$$

$$AI(Si) \leftarrow AI(Si) + \text{current AI}_{-}$$

The above formula is used to calculate the attractive index value of the different swarm, β and γ are algorithm-specific parameters and α controls the step size.

3.4 Model Training

After selecting the 70 percent of features from the dataset the data is processed for the next step that is training of the model. To train our model we use four classifier named as neural network, support vector machine, k nearest neighbor and decision tree. In the model training section we train our model that to classify weather the transaction is fraud or not. We pass the 70% of data to train the model after that to check the performance by passing remaining 30% data. The following block diagram (figure 2) shows the steps that our model follows;

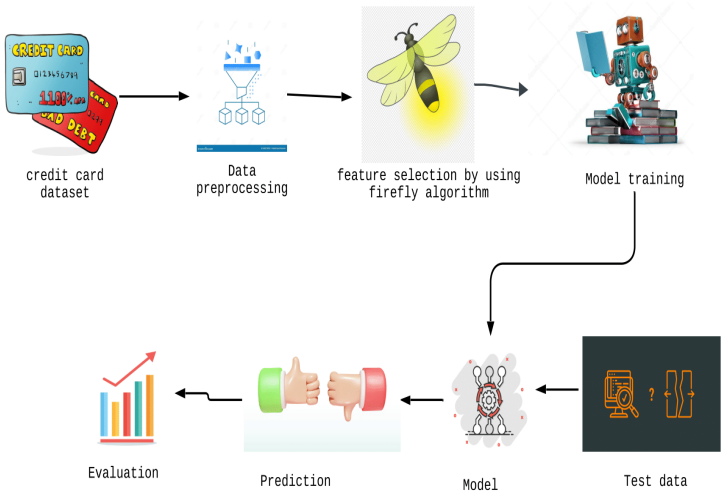

Figure 2. Flow Diagram

3.5 Evaluation

This section provides a discussion on performance evaluation measures in order to strengthen the proposed approach. Accuracy of different classifiers is calculated by drawing confusion matrix of all classifiers.

$$\text{Accuracy} = (\text{number of correct prediction}) \div (\text{total number of prediction})$$

4. EXPERIMENTAL RESULTS AND DISCUSSION

This section analyses the result of the proposed model on the basis of accuracy rate. first we pre-process the data ; to conduct our result we uses data set of European cardholders in which total number of instances are 284807 and 30 number of columns with only two classes 0 and 1.

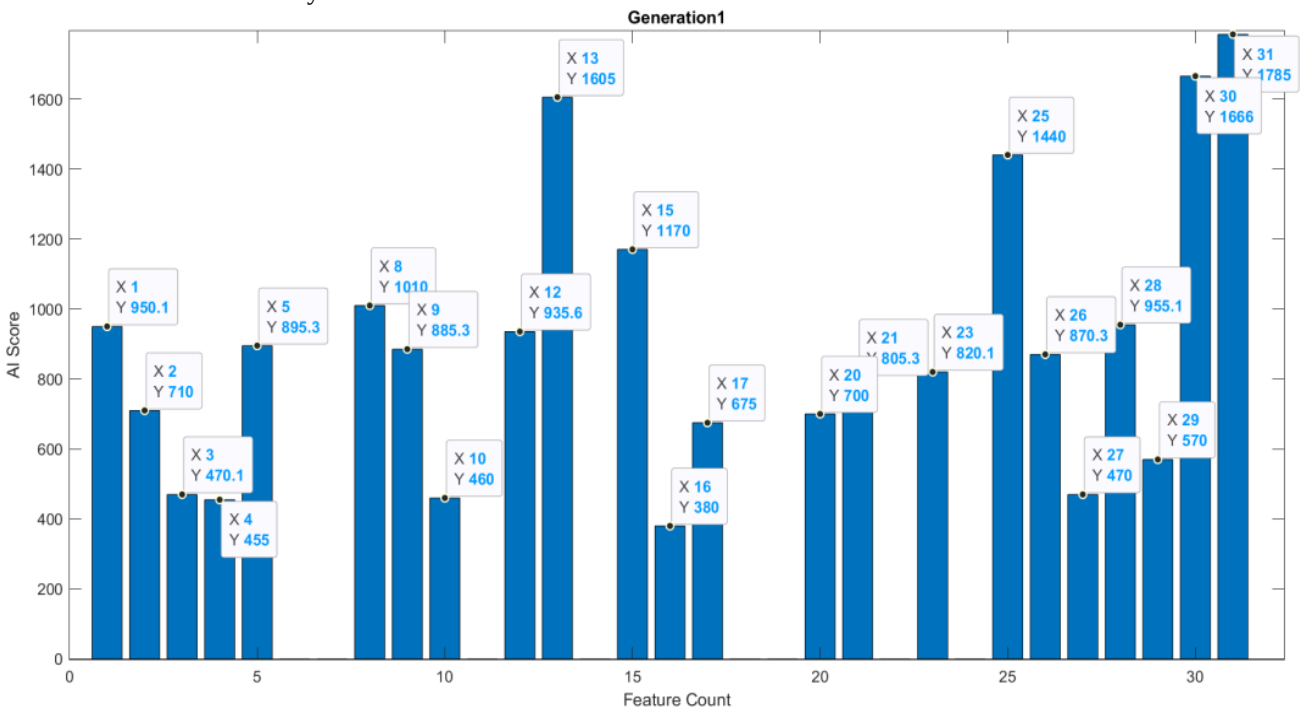

Figure 3. Feature selection using firefly algorithm

After preprocessing of the data ,a bio-inspired feature selection algorithm is imposed on the dataset to select a highly co-related attribute to improve the accuracy of the model as shown in figure 3; a feature with high attraction index is highly correlated to our target class, since every attribute is not significant . Feature selection acts as a filtering process, enabling the algorithmto focus on the most pertinent features of the input.
Table 1 shows the result of accuracy rate of the different classifiers:

Name of classifier	Accuracy rate
Neural network	99.98%
Knn	99%
Decision Tree	99.97%
SVM	21%

Table 1. Accuracy of classifiers

As shown in table 1 the accuracy rate of neural network(NN) , decision tree and knn is high where as Support vector machine shows very low, in terms of fraud detection rate, this shows that the for highly imbalanced dataset which have two classes NN, decision tree and knn shows good result as compare to the SVM. However, several performance metrics reveal different insights into a classification model's performance, such as precision, recall, and F1 score metrics.

5. CONCLUSION

The purpose of the study is to detect credit card fraud. Financial crime is an act in which someone take money for their own use without knowledge of owner. This study focused on the detection of fraud in credit card data set of European bank institute. A bio–inspired feature section algorithm to select the feature to optimize the performance of the credit card fraud detection model. To train the model for the prediction of the transaction whether it is fake or not; four different classifier named neural network, decision tree, k-nn and SVM are used. All the classifiers show the good performance in case of accuracy matrix except for SVM. In future we use new bio-inspired algorithms and deep learning to make the credit card scam detection system better at classifying things and faster at testing them in real time. The suggested method can be used to find instances of theft in real-time theft.

6. REFERENCE

[1] "Credit Card: What It Is, How It Works, and How to Get One," Investopedia. Accessed: Mar. 22, 2024. [Online]. Available: https://www.investopedia.com/terms/c/creditcard.asp

[2] R. Saini and Prof. B. Pandey, "Credit Card Fraud Detection project," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 4, pp. 2113–2118, Apr. 2022, doi: 10.22214/ijraset.2022.41704.

[3] J. K. Afriyie *et al.*, "A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions," *Decis. Anal. J.*, vol. 6, p. 100163, Mar. 2023, doi: 10.1016/j.dajour.2023.100163.

[4] "New Account Fraud: How it Works, Detection & Prevention." Accessed: Mar. 22, 2024. [Online]. Available: https://www.pingidentity.com/en/resources/blog/post/what-is-account-creation-fraud.html

[5] A. Singh, A. Jain, and S. E. Biable, "Financial Fraud Detection Approach Based on Firefly Optimization Algorithm and Support Vector Machine," *Appl. Comput. Intell. Soft Comput.*, vol. 2022, pp. 1–10, Jun. 2022, doi: 10.1155/2022/1468015.

[6] P. Rathi and N. Singh, "A NOVEL APPROACH TO DETECTION OF EMERGING FRAUD USING MINING TECHNIQUES," *ICTACT J. SOFT Comput.*, vol. 11, no. 01, 2020.

[7] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, “Credit card fraud detection using machine learning techniques: A comparative analysis,” in *2017 International Conference on Computing Networking and Informatics (ICCNI)*, Oct. 2017, pp. 1–9. doi: 10.1109/ICCNI.2017.8123782.

[8] S. Bagga, A. Goyal, N. Gupta, and A. Goyal, “Credit Card Fraud Detection using Pipeling and Ensemble Learning,” *Procedia Comput. Sci.*, vol. 173, pp. 104–112, Jan. 2020, doi: 10.1016/j.procs.2020.06.014.

[9] N. Rtayli and N. Enneya, “Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization,” *J. Inf. Secur. Appl.*, vol. 55, p. 102596, Dec. 2020, doi: 10.1016/j.jisa.2020.102596.

[10] A. Rb and S. K. Kr, “Credit card fraud detection using artificial neural network,” *Glob. Transit. Proc.*, vol. 2, no. 1, pp. 35–41, Jun. 2021, doi: 10.1016/j.gltp.2021.01.006.

[11] A. Alharbi *et al.*, “A Novel text2IMG Mechanism of Credit Card Fraud Detection: A Deep Learning Approach,” *Electronics*, vol. 11, no. 5, Art. no. 5, Jan. 2022, doi: 10.3390/electronics11050756.

[12] O. R. Mohsen, G. Nassreddine, and M. Massoud, “Credit Card Fraud Detector Based on Machine Learning Techniques,” *J. Comput. Sci. Technol. Stud.*, vol. 5, no. 2, Art. no. 2, Jul. 2023, doi: 10.32996/jcsts.2023.5.2.2.

[13] G. Preda, “Credit Card Fraud Detection Predictive Models.” Accessed: Mar. 26, 2024. [Online]. Available: https://kaggle.com/code/gpreda/credit-card-fraud-detection-predictive-models

[14] X.-S. Yang and M. Karamanoglu, “Swarm Intelligence and Bio-Inspired Computation,” in *Swarm Intelligence and Bio-Inspired Computation*, Elsevier, 2013, pp. 3–23. doi: 10.1016/B978-0-12-405163-8.00001-6.