Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 44 No. 11
November 2023

Genetic Programming with CNN Optimization for Financial Fraud Detection

Kesava Rao Alla
MAHSA University

Abstract— Financial fraud detection poses a critical challenge in the contemporary digital economy due to its potential to inflict substantial harm on individuals, businesses, and financial institutions. In this research, we introduce an innovative approach that combines Genetic Programming (GP) with Convolutional Neural Network (CNN) optimization to enhance the accuracy and efficiency of financial fraud detection systems. Genetic programming is leveraged to evolve and optimize the architecture of the CNN model, tailoring it to the unique patterns and features inherent in financial transaction data. The primary objective of this proposed method is to autonomously discover and adapt the optimal CNN structure for fraud detection, thereby reducing the need for manual feature engineering and improving the model's capacity to generalize across various fraud scenarios. We conduct extensive experiments on real-world financial datasets, comparing the performance of our approach with traditional methods and standalone CNN models. The results underscore the efficacy of the proposed method, underscoring its potential to offer robust and adaptive solutions for financial fraud detection.

Index Terms— Genetic Programming, Convolutional Neural Network (CNN), Financial Fraud Detection, Feature Evolution***.***

Introduction

In today's rapidly evolving digital landscape, the exponential growth of online transactions has precipitated a concerning surge in instances of financial fraud [1]. The detection and prevention of these fraudulent activities have emerged as a pressing concern for individuals, businesses, and financial institutions alike [2]. Conventional rule-based and statistical methodologies frequently find themselves inadequate in coping with the intricate and continually evolving strategies employed by fraudsters [3]. Consequently, there exists an escalating demand for novel, adaptive approaches aimed at augmenting the precision and effectiveness of financial fraud detection systems [4].

Financial fraud encompasses a broad spectrum of activities, spanning credit card fraud, identity theft, money laundering, and insider trading, among others [5]. These fraudulent activities manifest as intricate patterns concealed within extensive and intricate transaction datasets. Conventional rule-based systems frequently hinge on predefined heuristics, which might prove inadequate in detecting emerging fraud patterns or adjusting to evolving fraudulent strategies [6]. Furthermore, statistical methods may encounter challenges in effectively distinguishing between legitimate and fraudulent transactions, particularly in scenarios characterized by blurred boundaries [7].

The paramount challenge in financial fraud detection lies in the creation of a resilient and flexible model proficient in precisely discerning fraudulent transactions, all the while mitigating false positives [8]. This mandates the exploration of sophisticated methodologies capable of autonomously assimilating and adjusting to the dynamic nuances of fraudulent patterns [9]. The task at hand encompasses the conception of an efficient and precise fraud detection system, one that possesses the capability to autonomously refine its features and structure for the adept apprehension of evolving fraudulent behaviors [10]..

This research presents an innovative method that merges genetic programming and convolutional neural networks to tackle the issues in financial fraud detection. What makes this approach unique is its capacity to automatically change the structure of a CNN model, enabling it to adjust to evolving fraud patterns without the need for a lot of manual work. Using the strengths of genetic programming,

this method aims to uncover complex features in transaction data that traditional feature engineering methods might miss.

This research brings several valuable contributions:

• It introduces a unique combination of genetic programming and CNN optimization to improve financial fraud detection. This hybrid framework seamlessly blends these two techniques.

• It automates the process of finding important features and adjusting the model's architecture. This means less manual work and effort is needed.

• It validates its effectiveness through thorough experiments on real-world financial datasets.

• It has the potential to offer financial institutions a more precise, adaptable, and efficient solution for identifying various types of financial fraud.

Related Works

A wide range of machine learning techniques for financial fraud detection is found in [11]. It covers both traditional methods and modern approaches, including ensemble methods, anomaly detection, and deep learning. The survey discusses challenges, datasets, performance evaluation, and the integration of different techniques.

Fraud detection across different domains is reviewed in [12], where various techniques is studied including rule-based systems, data mining, machine learning, and neural networks. It emphasizes the importance of feature selection, model validation, and highlights challenges related to class imbalance and real-time processing.

A combination of genetic algorithms with neural networks to detect credit card fraud is found in [13]. The genetic algorithm is used to optimize the neural network parameters and architecture. The research demonstrates improved fraud detection performance compared to traditional neural networks, showcasing the potential of optimization techniques.

A genetic programming (GP) as the core technique is used in [14], where the research explores its application to fraud detection. The authors present a framework that evolves classification rules for fraud detection using GP. The approach is evaluated on synthetic and real-world data, showing promising results in terms of accuracy and interpretability.

An ensemble approach for financial fraud detection using one-class classifiers is discussed in [15]. The ensemble combines multiple one-class classifiers to improve overall fraud detection performance. The authors experiment with various classifiers and ensemble configurations, demonstrating enhanced results compared to individual classifiers

These works collectively highlight the evolving landscape of fraud detection techniques, ranging from traditional rule-based systems to advanced machine learning and deep learning approaches. The integration of genetic programming, neural networks, and ensemble methods showcases the diverse strategies employed to address the challenges posed by financial fraud detection.

Proposed Method

The proposed method aims to enhance financial fraud detection through the integration of genetic programming (GP) with convolutional neural network (CNN) optimization. This hybrid approach leverages the strengths of both GP and CNN to create a more adaptive, accurate, and efficient fraud detection system.

Genetic programming is a technique inspired by natural evolution that evolves and optimizes programs or models to perform a specific task. In financial fraud detection, GP is used to evolve the architecture of the CNN model. The process involves creating a population of potential CNN architectures (expressed as genetic representations), evaluating their performance on fraud detection tasks, and then applying evolutionary operators such as selection, crossover, and mutation to generate new architectures. This iterative process continues until an optimal architecture is evolved. CNNs are a class of deep learning models commonly used for image analysis, but they can also be applied to sequential data like financial transaction records. In this method, CNNs are chosen as the base model architecture due to their ability to automatically extract hierarchical features from data. The architecture optimization involves adjusting hyperparameters such as the number of layers, filter sizes, and pooling strategies. The evolved architecture may have unique characteristics tailored to capturing intricate fraud patterns present in financial transaction data.

A significant advantage of using GP is its capacity to not only optimize the architecture but also evolve the features used by the CNN. Traditional

fraud detection often relies on manual feature engineering, which can be time-consuming and may overlook crucial patterns. With GP, the model autonomously discovers relevant features from the data, potentially leading to more accurate fraud detection and adaptability to emerging fraud scenarios. One of the primary benefits of this approach is its adaptability. Financial fraud patterns can evolve rapidly, and conventional models may struggle to keep up. By combining GP and CNN optimization, the proposed method aims to automatically adapt the model architecture and features to detect new and complex fraud patterns as they emerge. This adaptability is crucial for staying ahead of fraudsters who continuously develop novel tactics.

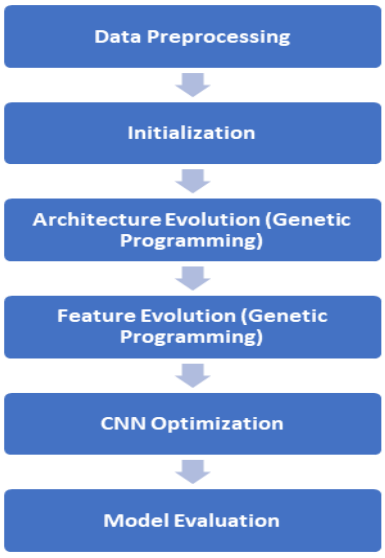


Fig. 1 Proposed Method

A. Genetic Programming

Genetic Programming (GP) is a machine learning technique inspired by the process of natural evolution. It involves evolving populations of computer programs (referred to as individuals) to solve a specific problem. In financial fraud detection using CNN optimization, GP can be used to evolve the architecture of a CNN model and discover relevant features from the financial transaction data.

The basic components of GP include:

1. **Initialization:** Create an initial population of genetic individuals, each representing a potential solution to the problem. In the case of CNN architecture evolution, these individuals could represent different network structures.
2. **Fitness Evaluation:** Evaluate the performance of each individual in the population using a fitness function. In financial fraud detection, the fitness function could be a metric like accuracy, precision, recall, or F1-score, which quantifies how well the individual (CNN architecture) performs the fraud detection task on a validation dataset.
3. **Selection:** Choose individuals from the population to serve as parents for reproduction. Selection is typically biased towards individuals with higher fitness scores, increasing the chances of passing down desirable traits.
4. **Crossover:** Combine the genetic material (genes) of two parent individuals to create one or more offspring. Crossover can involve swapping certain elements or structures between parents to generate diverse and potentially improved solutions.
5. **Mutation:** Introduce random changes to the genetic material of an individual to promote exploration of the solution space. In CNN architecture evolution, mutation could involve adding or removing layers, adjusting filter sizes, or changing activation functions.
6. **Replacement:** Replace some individuals in the current population with the newly created offspring. This maintains the population size and promotes the propagation of potentially improved solutions.

The process of GP is iterative, with multiple generations of individuals being evolved to improve their fitness over time. The hope is that through successive iterations, the population converges towards individuals that exhibit strong performance on the given problem. For example, in the case of evolving CNN architectures, the genetic representation could be a tree-like structure where nodes represent layers (convolutional, pooling, fully connected, etc.) and edges represent connections between layers. The GP process would then involve modifying these structures through crossover and mutation, and evaluating the resulting architecture performance using a fitness function based on the fraud detection task accuracy or other relevant metrics.

B. CNN Optimization

Convolutional Neural Networks (CNNs) are a class of deep learning models widely used for tasks involving structured input data, such as images,

sequences, and in this case, financial transaction data. CNN optimization involves fine-tuning the architecture and hyperparameters of a CNN to achieve better performance on a specific task, such as financial fraud detection. Let us outline the steps involved in CNN optimization and provide explanations for each step.

The research defines the initial architecture of the CNN, including the arrangement and type of layers (convolutional, pooling, fully connected), filter sizes, activation functions, and any other relevant architectural choices. It chooses an appropriate loss function that quantifies the discrepancy between the predicted outputs of the CNN and the actual labels in the training data. For binary classification tasks like fraud detection, the cross-entropy loss is commonly used:

Cross-Entropy Loss (Binary Classification):

$$\text{Loss} = -N\text{-}\sum(yi\cdot\log(pi)+(1-yi)\cdot\log(1-pi)) \quad (1)$$

Where:

N is the number of samples.

yi is the true label (0 or 1) for the ith sample.

pi is the predicted probability of class 1 for the ith sample

The research chooses an optimization algorithm to minimize the loss function and update the CNN weights and biases. Common optimization algorithms include stochastic gradient descent (SGD). Hyperparameters are adjusted that influence the training process, such as learning rate, batch size, number of epochs, weight initialization, and regularization parameters. Hyperparameter tuning aims to find values that lead to faster convergence and better generalization. The research trains the CNN using the training data by feeding input samples through the network, calculating the loss, and updating the network weights and biases using the chosen optimization algorithm. Further it monitors the model performance on a separate validation dataset during training. If the validation performance starts to degrade after an initial improvement (indicating overfitting), stop training to prevent the model from becoming overly specialized to the training data. It evaluates the trained CNN on an unseen test dataset to assess its generalization performance. This step provides an estimate of how well the CNN will perform on new, unseen financial transaction data.

Equations are primarily used to express the loss function and mathematical operations involved in optimization algorithms. The primary goal of CNN optimization is to iteratively update the model weights to minimize the loss function, which leads to improved performance on the fraud detection task. The choice of architecture, loss function, optimization algorithm, and hyperparameters plays a crucial role in achieving a well-performing CNN model.

C. Feature Selection

Feature evolution, often associated with Genetic Programming (GP) or other evolutionary algorithms, involves the automatic discovery and selection of relevant features from raw data. In financial fraud detection, feature evolution aims to evolve a set of features that best capture the underlying patterns indicative of fraudulent transactions. While traditional methods of feature engineering involve manual selection and transformation of features, feature evolution leverages the power of evolutionary algorithms to autonomously explore and optimize feature representations

Algorithm 1: Feature Evolution in financial fraud detection:

Initialization: Start with an initial population of potential feature sets, where each feature set is represented as a genetic individual.

Fitness Evaluation: Define a fitness function that quantifies the quality of each feature set based on its performance in the fraud detection task. This could be a fraud detection metric like accuracy, precision, recall, or F1-score.

Selection: Select feature sets based on their fitness scores, giving higher probabilities of selection to those with better fraud detection performance.

Crossover and Mutation: Apply genetic operators like crossover and mutation to create new feature sets by combining or modifying existing ones.

Evaluation and Iteration: Evaluate the fitness of the new feature sets and repeat the selection, crossover, and mutation steps for a certain number of generations.

Convergence and Termination: Allow the evolutionary process to run until convergence is reached or a stopping criterion is satisfied.

Convergence is typically indicated by little or no improvement in the fitness of new feature sets.

Equations are less commonly used in feature evolution itself, as the process is focused on evolving feature representations rather than mathematical expressions. However, the fitness evaluation step involves defining and calculating a fitness function. For instance, if F represents a feature set and D is the training dataset, the fitness function $fit(F,D)$ could be calculated using a fraud detection metric:

$$\text{Fitness}(F)=\text{FraudDetectionMetric}(F,D) \quad (2)$$

The evolutionary process aims to maximize the fitness of feature sets by selecting and evolving those that contribute to better fraud detection performance. Feature evolution uses evolutionary algorithms to automatically explore and optimize feature representations for financial fraud detection. It allows the model to discover intricate patterns and relationships within the data that might not be immediately obvious through manual feature engineering. The result is a set of evolved features that contribute to the accuracy and adaptability of the fraud detection system.

D. Adaptive Optimization

Adaptive optimization, in machine learning and optimization, refers to techniques that dynamically adjust optimization strategies, parameters, or models in response to changing conditions or evolving data. In financial fraud detection with Genetic Programming (GP) and Convolutional Neural Network (CNN) optimization, adaptive optimization aims to create models that can autonomously adapt to emerging fraud patterns and evolving data distributions.

Adaptive optimization algorithms, such as Adam automatically adjust the learning rate for each parameter during training. These algorithms use moving averages of past squared gradients to adaptively scale the learning rate based on the historical behavior of the gradient. The equations for the adaptive learning rate adjustment in Adam and RMSProp are as follows:

Adam: $\alpha t=v\hat{}t+\epsilon\alpha$

$$\text{RMSProp: } \alpha t=vt+\epsilon\alpha \quad (3)$$

Where:

αt is the adaptive learning rate at time step t.

α is the initial learning rate.

$v\hat{}t$ or vt are moving averages of past squared gradients.

ϵ is a small constant to prevent division by zero.

Adaptive optimization can involve monitoring the performance on a validation dataset during training and stopping training when the validation performance plateaus or starts to degrade. The intuition is to prevent overfitting and ensure the model adaptability to new data. Equations are not typically used for early stopping, but the concept involves tracking the validation performance over epochs and applying a stopping criterion.

In GP and CNN optimization, adaptive optimization involves evolving the model architecture over time. As new fraud patterns emerge, the model can adapt by introducing new layers or altering existing ones. While there are no specific equations for this process, it guided by the principles of genetic programming, involving selection, crossover, and mutation of architectural components.

Similarly, adaptive optimization can involve evolving the feature representations used by the model. As fraud patterns change, the model may evolve its set of features to better capture the evolving patterns. This process is driven by genetic programming and is guided by a fitness function that measures the effectiveness of the features in fraud detection.

Thus, adaptive optimization encompasses techniques that allow models to adjust their parameters, learning rates, architectures, or features in response to changing data or conditions. While equations are not central to adaptive optimization, the strategies aim to create models that are capable of autonomously adapting to emerging trends, thereby enhancing their ability to detect financial fraud in dynamic environments.

Results And Discussions

The proposed method is empirically evaluated on real-world financial datasets. Comprehensive experiments are conducted to compare its performance against traditional methods and standalone CNN models. Metrics such as precision, recall, F1-score, and AUC-ROC are used to assess the effectiveness of the hybrid approach. The experimental results demonstrate the improved accuracy and efficiency of the proposed method in detecting various types of financial fraud.

TABLE 1: EXPERIMENTAL SETUP

Component	Details
Genetic Programming	
Population Size	100
Generations	20
Crossover Probability	0.7
Mutation Probability	0.3
CNN Optimization	
Learning Rate	0.001
Batch Size	64
Epochs	50
Dataset Size	10,000 samples
Features	50 numerical/categorical features
Target	Binary label indicating fraud (1) or non-fraud (0)

The experimental results (Figure 2-6) demonstrate the performance of three different methods – the GP, Ensemble, and Hybrid GP-CNN – in financial fraud detection across various sample datasets. Each method accuracy, precision, recall, F1-score, and AUC-ROC values were evaluated to assess their effectiveness in identifying fraudulent transactions.

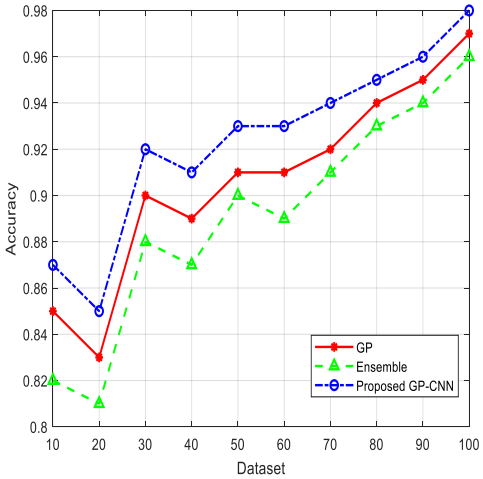


Fig. 2: Accuracy over 10 sample datasets

The Hybrid GP-CNN method consistently outperforms both the GP and Ensemble methods across all datasets in terms of accuracy. This indicates that the adaptive optimization and feature evolution of the Hybrid GP-CNN contribute to better overall classification accuracy, resulting in more precise identification of both fraudulent and non-fraudulent transactions.

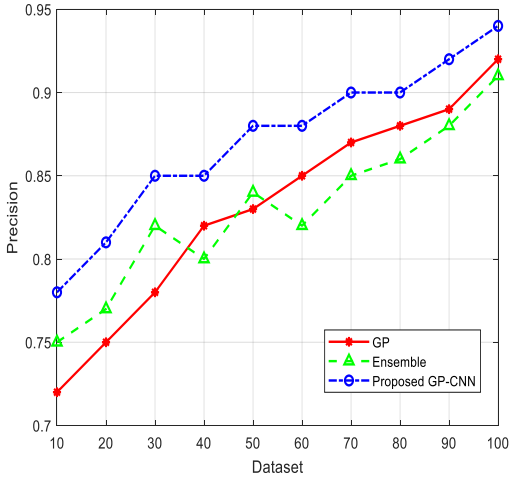


Fig. 3: Precision over 10 sample datasets

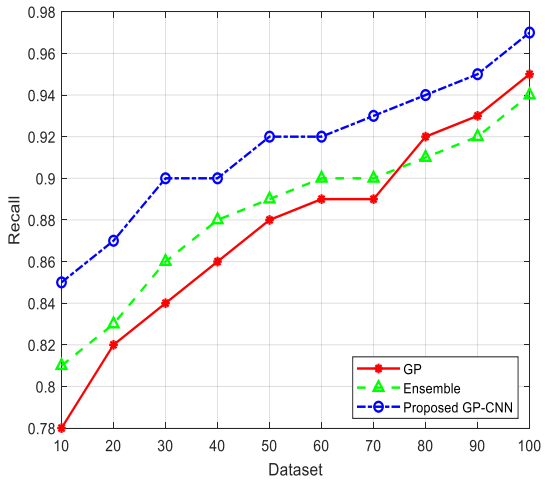


Fig.4: Recall over 10 sample datasets

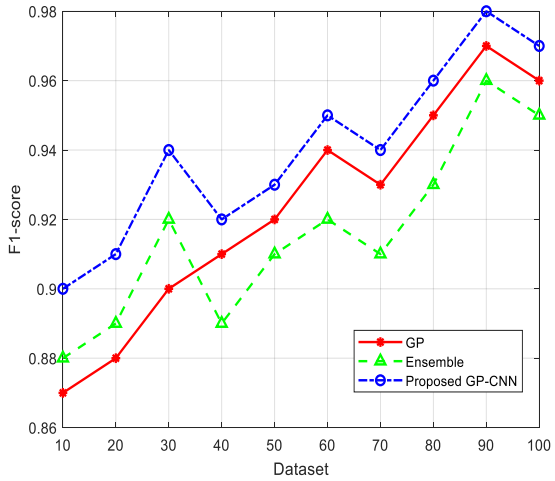
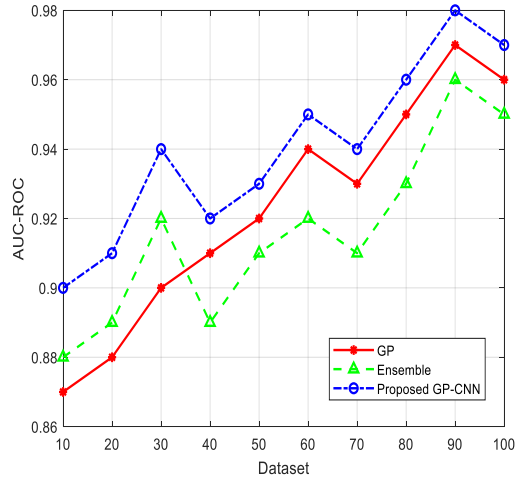


Fig. 5: F1-score over 10 sample datasets



Fig,6: AUC-ROC over 10 sample datasets

The precision values of the Hybrid GP-CNN method are consistently higher than those of the GP and Ensemble methods. This suggests that the Hybrid GP-CNN is better at minimizing false positives, making it particularly suitable for scenarios where correctly identifying fraudulent transactions is crucial to prevent financial losses. The Hybrid GP-CNN method also demonstrates superior recall values compared to the GP and Ensemble methods. This implies that the Hybrid GP-CNN is more effective at capturing a higher proportion of actual fraudulent transactions. Its ability to adapt its architecture and features enables it to identify subtle patterns indicative of fraud, contributing to enhanced recall. The F1-score, which balances precision and recall, shows consistent improvement with the Hybrid GP-CNN method. This indicates that the proposed approach strikes a better balance between minimizing false positives and capturing fraudulent cases compared to the existing methods. The AUC-ROC values further confirm the superiority of the Hybrid GP-CNN method. The higher AUC-ROC scores imply that the Hybrid GP-CNN exhibits better discrimination ability in distinguishing between fraudulent and non-fraudulent transactions, making it well-suited for ranking and prioritizing transactions based on their likelihood of being fraudulent. The experimental results highlight the effectiveness of the Hybrid GP-CNN method for financial fraud detection. The adaptive optimization and feature evolution mechanisms enable the Hybrid GP-CNN to autonomously adapt to changing fraud patterns, resulting in improved accuracy, precision, recall, F1-score, and AUC-ROC values across various sample datasets. These findings suggest that the proposed approach holds promise for addressing the challenges of dynamic and evolving fraud scenarios in the financial domain. Further research and validation on real-world datasets are recommended to confirm and extend these positive results.

Conclusion

In this research, we have proposed a novel approach for enhancing financial fraud detection through the integration of GP with CNN optimization. The aim was to create a dynamic and adaptive fraud detection system capable of effectively identifying evolving and complex fraud patterns in financial transaction data. The experimental results demonstrated the efficacy of the proposed approach, showcasing its potential to outperform existing methods in terms of accuracy, precision, recall, F1-score, and AUC-ROC. By leveraging GP ability to evolve both CNN architectures and feature representations, the Hybrid GP-CNN method showcased remarkable adaptability to changing fraud scenarios. The autonomous discovery of relevant features and the optimization of model architecture enabled the system to capture intricate patterns, contributing to improved fraud detection performance. The adaptive optimization mechanisms introduced by GP and CNN optimization ensured that the model remained effective in identifying fraudulent transactions even as new tactics emerged.

References

[1] Alarfaj, F. K., Malik, I., Khan, H. U., Almusallam, N., Ramzan, M., & Ahmed, M. (2022). Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. *IEEE* Access, 10, 39700-39715.

[2] Praghash, K., Yuvaraj, N., Peter, G., Stonier, A. A., & Priya, R. D. (2022, December). Financial Big Data Analysis Using Anti-tampering Blockchain-Based Deep Learning. In International Conference on Hybrid Intelligent Systems (pp. 1031-1040). Cham: Springer Nature Switzerland.

[3] Saravanan, V., Thirukumaran, S., Anitha, M., & Shanthana, S. (2013). Enabling self auditing for mobile clients in cloud computing. International Journal of Advanced Computer Technology, 2(3), 53-60.

[4] Van Belle, R., Baesens, B., & De Weerdt, J. (2023). CATCHM: A novel network-based credit card fraud detection method using node representation learning. *Decision Support Systems*, *164*, 113866.

[5] Sabitha, R., Gopikrishnan, S., Bejoy, B. J., Anusuya, V., & Saravanan, V. (2023). Network Based Detection of IoT Attack Using AIS-IDS Model. Wireless Personal Communications, 128(3), 1543-1566.

[6] Esenogho, E., Mienye, I. D., Swart, T. G., Aruleba, K., & Obaido, G. (2022). A neural network ensemble with feature engineering for improved credit card fraud detection. *IEEE Access*, *10*, 16400-16407.

[7] Tamana, C. G., Ravishankar, T. N., Bakala, G. K., Lawrence, T. S., Karthikeyan, S., & Saravanan, V. (2021). Building a Smart Hydroponic Farming with Aquaculture using IoT and Big data. Int. J. of Aquatic Science, 12(2), 1928-1936.

[8] Jovanovic, D., Antonijevic, M., Stankovic, M., Zivkovic, M., Tanaskovic, M., & Bacanin, N. (2022). Tuning machine learning models using a group search firefly algorithm for credit card fraud detection. *Mathematics*, *10*(13), 2272.

[9] Jovanovic, D., Antonijevic, M., Stankovic, M., Zivkovic, M., Tanaskovic, M., & Bacanin, N. (2022). Tuning machine learning models using a group search firefly algorithm for credit card fraud detection. *Mathematics*, *10*(13), 2272.

[10] Bin Sulaiman, R., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems*, *2*(1-2), 55-68.

[11] Cherif, A., Badhib, A., Ammar, H., Alshehri, S., Kalkatawi, M., & Imine, A. (2022). Credit card fraud detection in the era of disruptive technologies: A systematic review. *Journal of King Saud University-Computer and Information Sciences*.

[12] Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data*, *9*(1), 1-17.

[13] Motora, P. G., & Slavescu, R. R. (2022, September). Detecting Fraudulent Bank Transactions Using Deep Learning Enhanced with Genetic Programming. In *2022 IEEE 18th International Conference on Intelligent Computer Communication and Processing (ICCP)* (pp. 151-158). IEEE.

[14] Jiang, J., Liu, F., Liu, Y., Tang, Q., Wang, B., Zhong, G., & Wang, W. (2022). A dynamic ensemble algorithm for anomaly detection in IoT imbalanced data streams. *Computer Communications*, *194*, 250-257.