# American Journal of IR 4.0 and Beyond (AJIRB)

ISSN: 2837-4738 (ONLINE)

VOLUME 4 ISSUE 1 (2025)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of IR 4.0 and Beyond(AJIRB)

Volume 4 Issue 1, Year 2025
ISSN: 2837-4738 (Online)
DOI: https://doi.org/10.54536/ajirb.v4i1.3833
https://journals.e-palli.com/home/index.php/ajirb

# Developing Machine Learning Models for Real-Time Fraud Detection in Online Transactions

Mohammad Prince[1*]

**Article Information**

**Received:** September 02, 2024

**Accepted:** October 30, 2024

**Published:** April 08, 2025

**Keywords**

*Algorithm, Detection, Fraud, Machine Learning*

## ABSTRACT

This paper offers a detailed discussion of a large–scale, real-time architecture for fraud detection specifically for use in financial organizations to combat fraudulent activities in online transactions. The proposed system in this paper uses big data capabilities and a multi-stage fraud detection pipeline to detect and combat fraudulent activities efficiently. The implemented technologies include Apache Kafka, KSQL, and Spark alongside Isolation Forest algorithm for behavioral analysis of customer transactions. The presentation of the fraud detection pipeline as a series of layers exemplifies how a transaction goes through an exacting sequence of detection algorithms with very little delay and maximum precision. Verification by simulation uses the dataset of more than one hundred million Internet transactions, the performance indicators of which are a rather high F1-score of 91% and a recall rate of 97%. The results stress the advantage of the proposed methodology over conventional techniques, suggesting the possibility of real-time fraud identification. Furthermore, the paper outlines research directions where future work should focus, such as reducing computational complexity and applying deep learning solutions to enhance the detection of new types of fraud.

## INTRODUCTION

The unprecedented rise of e-commerce and use of online finances has changed the way consumers and companies buy goods and services for increased convenience with efficiency and this digitization has also created large problems especially in the domain of fraud. Cyber criminals are using superior techniques to launch diverse paltering scams, consisting of identity theft, attempt to emulate another in order to gain credit, credit card frauds and other digital payment system complications (Bin Sulaiman *et al.*, 2022). Internet fraud statistics as captured by the Internet Crime Complaint Center (IC3) for 2022 estimates this specialty at billions of shilling and highlights the need for implementing robust mechanisms for fraud detection (Bin Sulaiman *et al.*, 2022). The criminals are also coming up with better strategies to avoid the conventional methods employed in firms today hence the need for companies to come up with ways to handle such issues better instead of the old fraud detection systems which prior thresholds as well as static patterns which were used to detect fraudulent transactions (Potla, 2023). Although these two approaches offered a certain form of protection they were able to hold fraudsters at bay, they lacked the prowess to counter the new strategies that were being developed in the market. This is the classic case of why rule-based systems are always inadequate because while they were used here to curb gains rip offs will always find a way around them which results to a movement towards using machine learning technologies which can more likely provide more flexible and adaptive fraud detection (Nakra *et al.*, 2024). Using algorithms that are capable of studying data and identifying patterns, the prospects of fraud detection in real time are greatly improved by such insights using artificial intelligence, with its subset machine learning which enables machines to learn from data, weigh it and make some decision on their own (Khan *et al.*, 2022). This capability is useful mostly in fraud detection which requires vast amounts of transactional data to be analyzed and identified within a short span of time by providing techniques based on machine learning including for example the decision tree techniques, support vector machine, and the neural network techniques, perform better than the traditional techniques in terms of detecting the fraudulent transactions (Khan *et al.*, 2022). Even with the ability of machine learning to improve the efficiency of fraud detection, the following factors hold true as the most challenging problem is the imbalance of classes which is characteristic of most fraud datasets where the number of normal or legitimate transactions is orders of magnitude larger than the number of fraudulent ones (Khan *et al.*, 2022). This imbalance if left can lead to skewed models of the entire fraud detection system which results in poor performance of the detection system with online purchases carrying out multiple transactions per second, models, need to be precise but also swift and must incorporate thousands of transactions per second. Therefore, more work needs to be done in order to adapt methods in machine learning to overcome these challenges and create better fraud identification systems. The development of the machine learning fraud detection requires understanding the necessary steps from Initiation to deployment.

### Purpose

This research work is concerned with the development of machine learning models for fraud detection for real-

[1] Trine University, Xpres Trip & Steadfast International Services, LLC, USA
[*] Corresponding author's e-mail: msprince23@my.trine.edu, mohammad@xprestrip.com

time transactions. The aim is to explore several machine learning techniques, their efficiency in detecting fraud and the problems with deploying these models in production. The paper will finally develop a model that will utilize machine learning techniques for real-time fraud detection.

## LITERATURE REVIEW

This section aims to provide an analysis of past research on the development of fraud detection models. The article for this research were sources from science articles data bases such as google scholar. The literature review is as follows;

### Overview of Fraud Detection in Online Transactions

According to Potla (2023), currently online transaction fraud is becoming common than before due to the growing use of digital business through online payment systems, transactions through credit cards, and banking operations expose such typology of fraud. Existing research suggest that fraudsters are able to perpetrate unauthorized access, phishing and or/ data intrusions and has become fundamental for financial organizations and e-commerce sites to identify fraudulent measures in real time to avoid incurring additional losses and losing customers' information. Potla (2023) futhers argued that the daily usage of financial systems on the internet due to the enhanced e-commerce businesses has enhanced the importance of adopting high-level security mechanisms to detect fraud. There were earlier types of detecting fraud including the rule-based systems that work under set thresholds or regularities but due to the very, complex crafted ways of the cheating subject the traditional rule-based systems could often fail to capture new and intricate as well as previously unobserved fraud schemes.

Potla (2023) also followed more of static techniques since there is a need for more dynamic techniques that will easily prevent fraud changes though functional fraud detection models have troublesome concerns related to False Positives (FPs), that is when legitimate credit card transaction is detected as a potential fraud, and False Negatives (FNs), the sequence where a potential fraud goes undetected, which is not easy for the financial institutions. Mir (2024) argued that over the last couple of years new trends in machine learning have greatly influenced the approach to fraud detection that currently involves both supervised and unsupervised models of learning in areas of transaction data that rule-based systems cannot see, but that machine learning is capable of seeing. For example, the usage of decision trees, random forest, and neural network has shown that the means having a better detection rate since such methods rely on past data of fraudulent and non-fraudulent transactions in an attempt to improve performance in fraud detection.

### Machine Learning Algorithms for Fraud Detection

Nakra *et al.* (2024) also found that the use of machine learning in fraud detection covers many techniques, which have their advantages and disadvantages according to classification techniques are one of the most prevalent used in industries, where models from the supervised learning group are trained on datasets containing both fraud and authentic transactions. Nakra *et al.* (2024) continues to argue that logistic regression and support vector machines and decision trees have been implemented most often due to their proficiency in the classification of transactions based on pattern that has been derived which indicated that decision trees as well as random forest perform well to detect and differentiate fraud as they are more accurate and easier to explain. Due to the nature of random forests as decision tree models, over-fitting and understanding of the general framework are addressed.

Chen and Lai (2021) believed that in the current viewpoint, unsupervised learning becomes popular in particular for a new or previously unlearned fraud pattern detection that is usually limited in quantity or costly to obtain, there are methods like clustering and novelty detection applied on the transaction data in attempts to find out the anomalous ones. These algorithms are especially useful to detect new crime methods when the model was not applied on training data with fraud examples which in most cases, unsupervised models have considerably higher number of False Positives than the supervised ones which can cause inconvenience to the customers. According to Chen and Lai (2021), other approach with big data is based on deep learning algorithms that have been also considered in the car fraud detection because of its ability to process large amount of input data using the application of autoencoders and generative adversarial networks for fraud detection, which can take advantage of latent features of transaction data with labels. Despite higher performance, deep learning models tend to make higher computational demands compared with the more straightforward and thus more transparent methods of machine learning are becoming appropriate.

### Real-Time Processing and Challenges in Fraud Detection

Zhou *et al.* (2020) revealed that online fraud prevention is an important essential for current economic systems as it reduces the time that fraudsters have in their hands even if the practices of running transaction analysis in batches are inadequate to fight the speed at which fraud operates. In real-time systems, fraud detection models analyze transactions in real-time with the outcome to be produced in millisecond since it prevents execution of fraud transactions in real-time processing frameworks such as Apache Kafka, Apache Flink and Apache Spark which have been used by researchers in processing entire transaction data streams efficiently (Zhou *et al.*, 2020). The key managing issue in real-time fraud detection is always the question of speed versus accuracy which is essential to prevent the postponement of genuine transactions, the efficiency of the model in flagging fraud cannot be negatively influenced. It should thus be noted that fraud detection systems need to be low latency compliant so that

for instance they can complete the analysis of thousands of transactions per second methods that have been well investigated on dealing continuously with transactions and other associated data streams in real time.

Streaming analytics also enable the models to learn as new data is received hence the ability to provide a high degree detection efficiency because the models adapt to the newer and more accurate data (Khan *et al.*, 2022). However, there is another problem that real time fraud detection faces, and that is scalability due to the increasing, then the outcomes of fraud detection will need to increase in capability too. The study on distributed computing frameworks, which assumes that a large amount of data is processed on several nodes with different tasks and in parallel which means, no decrease in the quality of the real-time fraud detection models employed in the backend when there are a lot of transactions during shopping festival or other important financial events (Khan *et al.*, 2022). Still, managing to keep the models as simple and still as accurate is not a trivial task and becomes an object of current research.

**Addressing Class Imbalance and Model Evaluation**

One of the biggest issues which arise when attempting to create machine learning models for fraud detection is the presence of the class imbalance issue which usually make up a small proportion of all transactions, which results in a case where models can start favoring fake transactions (Bin Sulaiman *et al.*, 2022). In such cases, while the prediction accuracy is high, most of the machine learning models do not find the least of the fraudulent cases with a number of approaches suggested to rectify this problem by increasing the sample size of the minority class samples (fraudulent transactions) and reducing that of the majority class samples (legitimate transactions). The oversampling of the minority class has therefore been efficiently done by cultivating synthetic examples through tools such as SMOTE (Synthetic Minority Over-sampling Technique) with other sampling techniques have been used to address the problem of class imbalance through cost sensitive learning (Bin Sulaiman *et al.*, 2022). Cost-sensitive learning involves the optimization of two costs, which are false positives and false negatives depending on the costliest one allowing the current algorithm to determine fraud instances in spite of few instances and can be seen to outcompete existence of single classifiers in terms of identification of fraudulent transactions as they have the competency to identify interactions between some features. Evaluation measures are used alongside when comparing the performance of fraud detection models chiefly in the imbalanced data environment with such as accuracy can be very misleading in the context of fraud because they do not adequately characterize the ability of a model to find the rare event[s] (fraud). Khan *et al.* (2022) suggested that human assessors subsume the performance by reporting metrics such as Precision, Recall, F1-Score, and AU-ROC in addition to the above, due to the real-time nature of the evaluation also involves considerations of time required for the processing and the effects of wrongly indicating a site as malicious to the customer.

**Step by Step Model Development**

The following are the steps for developing the model.

**Problem Identification and Understanding**

The initial stage in the process of creating a machine learning model for fraud detection is the formulation of the problem which includes an understanding of what is considered fraud in relation to the specific application in instances such as purchase of products online (Khan *et al.*, 2022). The parties involved should come to a consensus on what the fraudulent actions are, which may be patterns of users, amount-related irregularities, or spatial incongruities in describing the goals, it also describes what specific results the team wants to obtain, for example, in case of fraud detection it can be reducing number of false positives or increasing number of identified fraudulent transactions (Khan *et al.*, 2022). Further, historical records concerning fraud incidences offer important information that contributes to creating the focus of the model.

**Data Collection and Preparation**

The next step is to gather and acquire all the necessary data and information once the problem has been established and involves data collection where one collects relevant data that is useful for the model, which may include transaction logs, user profiles, or user behavior data (Khan *et al.*, 2022). Data collection can involve both numerical data (such as the quantities and the time of transactions) and textual data (such as the comments and feedback). Once the data has been gathered it needs to be cleaned or preprocessed so that it can be used in the analysis and the process of cleaning the data set, which may entail steps such as eliminating redundant features, dealing with features that may be missing values and even normalizing the data set. Feature engineering is also important at this stage where new features are derived from the data to improve the predictive capability of the model by calculating the number of transactions in a given time which can be important in understanding how users interact with the platform.

**Exploratory Data Analysis (EDA)**

Exploratory data analysis is performed to understand the patterns and characteristics of the data when the dataset is clean by dealing with the graphical representation and statistical assessment of the data to detect the patterns, associations and outliers (Bin Sulaiman *et al.*, 2022). Practitioners are able to understand the difference between fraudulent transactions and legitimate ones by creating distributions of transaction amounts or user activity levels and assists in identifying class imbalance whereby cases of fraud are limited as compared to legitimate cases, hence the need for measures to address

this issue during modeling (Bin Sulaiman *et al.*, 2022). Knowledge obtained from EDA also helps in deciding on which features to use and which kind of models to employ to cover different types of fraud.

**Choosing and Developing Models**

The subsequent process to be performed after EDA is the selection of proper machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, and Neural Networks which are widely applied algorithms for fraud detection (Bin Sulaiman *et al.*, 2022). The process of developing a model entails the consideration of the data set splitting into training and test data to assess the model's performance. In training, the selected algorithms optimize their parameters from the data to minimize the prediction error and the hyperparameter tuning which is also important at this stage because it involves working on different parameters to further the performance of the model (Bin Sulaiman *et al.*, 2022). To improve the model, cross-validation techniques can be used among other tools to make sure it performs well on unseen data and provide the necessary information and allow for automation.

**Model Evaluation and Validation**

After training is complete, the model requires validation for measuring its efficiency in identifying fraud by feeding the model to the reserved testing dataset and calculating the efficiency indicators, including accuracy, precision, recall, F1 score, and AUC-ROC. These metrics give a clear impression on the effectiveness of the model, focusing on its ability to prevent fraudulent transactions and endorse legitimate ones which is important to emphasize on recall (sensitivity) to ensure that it correctly predicts fraudulent transactions. Even with a good accuracy, confusion matrices and ROC curves assist in the visualization of the model and further improvement upon it if needed.

**Deployment of the Model and Monitoring**

The next step is taking the model from the validation and fine-tuning phase and implementing it into a live environment by focusing on how the model is going to be incorporated in the existing transaction processing system to enable it to approve transactions in real-time (Bin Sulaiman *et al.*, 2022). This is commonly handled during deployment so that the model can receive new data, make predictions, and feedback the predictions to the system and provide mechanisms set to assess the behavior of the model in live environment and then update the training or even carry out retraining after certain intervals (Bin Sulaiman *et al.*, 2022). However, to make the model more realistic, the scalability and latency aspects also need to be considered to allow the system to handle different transaction volumes without degrading the performance significantly.

## MATERIALS AND METHODS

The model development uses real data from transactions from one of the American banks, taking data from over 100 million transactions. This paper aims to develop a framework for determining the overall architecture of the end-to-end fraud detection solution is divided into the fraud detection pipeline which utilizes big data technologies to increase the efficiency of identifying fraudulent transactions in online banking systems in real time. These include the following: The first component involves the ingestion of the transaction data into the architecture using Real-Time Fraud Detection System taking screenshots of transactions and any relevant contextual information, and then the subsequent rule-based checks being conducted within the transaction environment. These checks are designed to be real-time since they should run within milliseconds to not slow down the user interface while helping filter out potentially fraudulent activities.

The next stage in reviewing the transactions involves more complex analytical examination after going through the preliminary checks using real-time analysis of customer behavior is achieved through the use of complex algorithms, including the Isolation Forest predictive model. This model evaluates many factors such as when and where the transaction took place to come up with risk scores for each transaction if the transaction goes beyond a specific risk level, it is detected and sent to a fraud checking system for additional analysis. The outcomes of these analyses are documented, which helps facilitate a cycle of feedback for enhancement of the model's accuracy using layered architecture also includes Apache Kafka and KSQL for real-time stream processing, enabling the immediate rejection of transactions based on certain characteristics, such as inexplicable transaction patterns.

To achieve this architecture, the deployment is run on a cluster that is devoted to the training of the model and the processing of the streaming data through Spark which is beneficial for maintainability and extensibility because Kafka and the monitoring applications are hosted on different machines. The infrastructure is built to enable scalability for the transactions and also to ensure the high availability and non-singularity. IT facilitates both real-time and batch data processing and incorporates tools such as Spark and Sparkling Water for the efficient deployment of ML models and can be integrated to perform real time analysis on big data sets so that the system is capable of adjusting to new fraud trends.

## RESULTS AND DISCUSSION

**End-To-End Fraud Detection Solution Architecture**

The purpose of this section is to provide a case study analysis of a scalable real-time architecture which can help transaction systems such as banks to detect and prevent fraud in online transactions by building upon the big data capabilities of the system to improve its capability to address complex fraud related scenarios. The idea is based on the fraud detection pipeline to prevent fraud for each transaction by passing it through multiple stages

and point to the architecture of the design as well as the technological stack that will be used in implementing the architecture. This will be clear using the data pipeline.

**Fraud Detection Pipeline**

Suppose there is a request of authorization for a certain transaction sent to a bank in this case an America. The process starts when the Real-Time Fraud Detection System captures the transaction details and related context data starting with the first line of control in minimizing fraud risk is a checklist, which contains a series of requirements that have to be assessed before approving the transaction. These rules are usually integrated within the transactional environment, which means that only transactions that satisfy specified criteria since customer experience is the primary focus, these rule-based checks are optimized to run in milliseconds to prevent any visible latency during customer engagements with the bank's systems.

After passing through the above checks, the transaction proceeds further to the next level of fraud detection using sophisticated analytical algorithms are applied to identify any potential malicious activity using the customer's behavior as a reference point with regards to the bank's IT systems. Customer data of the respective transaction is analyzed in real-time and passed through a pre-built Isolation Forest predictive model to determine the level of fraud risk using scores that are pre-defined and if they are exceeded by any transaction then it is marked and brought into a fraud monitoring system where it is reviewed by analysts. Supervisors may then take corrective actions, for instance, sending account holder alerts to the mobile App, Email or SMS of the suspicious activity and all flagged transactions are recorded into a database and labeled as suspicious by the fraud investigation team. This fosters the creation of a feedback loop that makes it easier to train and optimize the fraud detection model where each transaction passes through a fraud detection pipeline, and is systematically evaluated as described as shown in figure 1.

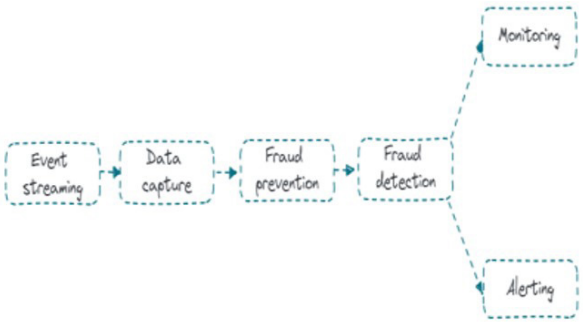


**Figure 1:** Customer Transaction Pipeline

The prevention layer will be built using of Apache Kafka and KSQL for real-time stream processing. Kafka is a versatile and popular platform well-suited for handling real-time data, while KSQL acts as a language for continuous queries to construct stream processing applications. This architecture captures large data transactions from various sources such as websites, banks, and social media in real time using Kafka and KSQL Streams especially when the system notices that the same account number was used at a different place within a short time, say ten minutes, the transaction will be immediately rejected. The real-time fraud detection system works in a layered manner which are the ingestion layer for real-time transaction data ingestion, the processing layer for handling big data with high availability and no single point of failure, and the alerting layer for the visualization of the fraud alerts. The first process is the consumption of a massive amount of online transactions, and the second process is processing using tools using two techniques which includes Spark Streaming to deploy the models and interact within the Spark framework, and Sparkling Water for large-scale analytics. To efficiently and effectively determine the risk of fraud, the Isolation Forest algorithm is used to train on the account holder's behavior and weighs parameters such as the location, time gaps between transactions, the number of transactions against historical data and is kept its usage is accomplished in a frontend application, which provides respective visualizations and allows for corrective actions through the backend APIs.

**The Infrastructure Deployment**

When it comes to the implementation of the architecture, was done separately on a distributed cluster for model training and Spark streaming data processing. In the same way, the Kafka and fraud monitoring applications were deployed using different and separate servers outside the systems as shown in figure 2.

| Component | Servers / Characteristics |
|---|---|
| Spark streaming / H2O | Driver :<br>CPU: 1 core<br>RAM: 4 Go<br>Storage: 50 Go |
| | Worker 1 :<br>CPU: 2 cores<br>RAM: 8 Go<br>Storage: 50 Go |
| | Worker 2 :<br>CPU : 2 cores<br>RAM : 8 Go<br>Storage : 50 Go |
| | Worker 3 :<br>CPU : 2 cores<br>RAM : 8 Go<br>Storage : 50 Go |
| Kafka | Broker 1 :<br>CPU : 2 cores<br>RAM : 8 Go<br>Storage : 50 Go |
| | Broker 2 :<br>CPU : 2 cores<br>RAM : 8 Go<br>Storage : 50 Go |
| | Broker 3 :<br>CPU : 2 cores<br>RAM : 8 Go<br>Storage : 50 Go |
| Monitoring application | Application server / Database:<br>CPU : 2 cores<br>RAM : 8 Go<br>Storage : 50 Go |

**Figure 2:** Used Servers

**Simulation and Results**

Over 100 million online transactions simulated data set

which represents real customer behavior in digital banking platforms for this cooperative research were used. The dataset consists of a variety of columns such as: user_id, account_no, event type, device_id, and timestamps. The events in the dataset are differences by typical activities that a user can do on a banking platform, such as login attempts, successful login or unsuccessful logins, account balance and history viewing, money transfers, bill payments provision card, view contract. Training and testing of the developed approach this deep dataset proved to be invaluable in understanding fraud patterns.

### Experimental Criteria

In this model, data splitting was done and grouped into five partitions where 80% of the data was used to train the Isolation Forest model for fraud identification focusing on accuracy, precision, recall, and the F1-score were used to assess the performance of the model. Accuracy is a general measure of how well the model is performing in predicting the probability of a transaction being fraudulent while precision is a measure of how many of the cases flagged by the model are actually fraudulent and sensitivity, also called the true positive rate, reflects the capability of the model to identify actual cases of fraud. The F1-score considers both precision and recall with emphasis placed on the latter since fraudulent cases can be rare and outcomes included True Positive (TP) which referred to actual fraudulent transactions, False Positive (FP), where normal transactions were incorrectly labelled as fraudulent, True Negative (TN), where actual non-fraudulent transactions were rightly screened out and False Negative (FN), where actual fraudulent transactions were screened out as normal ones. The evaluations metrics are as outlined in figure 3.

| Performance metrics | Formulas | |
|---|---|---|
| Precision: | $TP/TP + FP$ | (3) |
| Recall: | $TP/(TP + FN)$ | (4) |
| Accuracy: | $((TP + TN)/(TP + TN + FP + FN))$ | (5) |
| F1 score: | $2 \times \frac{(precision \times recall)}{(precision + recall)}$ | (6) |

**Figure 3:** Evaluation metrics

### The Experiments Results

The model was experimented using the isolation forest with help of python and also sparkling water engine to help in detecting fraudulent activities on real time basis using different iteration recording data on the events, transaction and labelled fraud attempts as shown in figure 4.

After testing the model basing on metrics such as accuracy, precision, recall and F1-score, the results were as shown in figure 5 below.

| | Events | transactions | Labeled Fraud attempts |
|---|---|---|---|
| Iteration 1 | 20000000 | 187234 | 151 |
| Iteration 2 | 20000000 | 234567 | 213 |
| Iteration 3 | 20000000 | 198654 | 195 |
| Iteration 4 | 20000000 | 272647 | 286 |
| Iteration 5 | 20000000 | 324546 | 323 |

**Figure 4:** Results of the Training Iteration

| Metrics | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| | 0,99 | 0,87 | 0,97 | 0,91 |

**Figure 5:** Metrics efficiency

### Discussion

In this article, the model described the sequence of actions required to create a real-time fraud detection system, which should be applied to online transactions, with the help of specific tools such as Spark, Kafka, and $H_2O$ together with the classification of the fraudulent transactions, a pre-built machine learning model known as the Isolation Forest was employed on the developed system. This system proved useful because it integrated both real-time and batch data processing for good results after evaluating the performance of the Isolation Forest model using four key metrics: accuracy, precision, recall, F1-score that depicted the performance of the model. The model also compared compared the results with other methods for fraud detection. For instance, in one research, SVM, the Apriori algorithm, and SVMIG were applied to identify fraud and the results showed 94% accuracy. In another study, six machine learning algorithms, namely Logistic Regression, XGBoost, Decision Trees, Random Forest, AdaBoosted Decision Trees, and AdaBoosted Random Forest, were trained on the data and achieved accuracy of over 98%. Another approach involved using a model known as AED-LGB to achieve a success rate of 98 percent yet another study gave Naïve Bayes a try since it performed better with a success rate of 97 percent. As for the results of the studies above, our approach based on big data tools and the Isolation Forest model demonstrated higher accuracy rates (99%) which aptly captures the efficiency of our real-time fraud detection system that identifies any transaction that may be fraudulent.

## CONCLUSION

The prevention of fraud in online transactions is important since fraud results in high monetary losses and due to the real-time nature of digital transactions, financial institutions need a solution that can efficiently and effectively detect fraud within the organization's workflow. The first one is the most important – the design of the system that would be able to operate with large volumes of data in real time and at the same time, provide high accuracy in terms of detecting suspicious activity and minimizing false positives and false negatives. The absence of timely fraud detection can lead to significant adverse financial and reputation implications for organizations and this paper outlines a new real-time

fraud detection system incorporating the Isolation Forest algorithm with big data analysis tools based on users' behavioral analysis. This approach enables the system to optimize speed, reliability, and accuracy achieved by the batch processing of the models and real-time analysis to detect suspicious transactions as they occur. Another significant feature of this architecture is that the models are constantly updated, and the system is capable of detecting fraud within real-time data feeds and was found that the F1-score of the system was 91% and the recall was 97% indicating the high reliability of the system. The following set of metrics suggests the efficiency of the Isolation Forest algorithm in the identification of fraudulent transactions with the high recall indicating the system's ability to detect fraudulent activity as it successfully captures a vast number of fraudulent transactions. Specifically, the F-measure, also known as the F1-score that combines precision and recall coefficients, evaluates the system's capacity to avoid false positives, which is critical for building trust and ensuring its effective functioning.

**Future Work and Considerations**

As with all studies, there are several limitations to consider for future research and discussion which will focus on two main objectives. Firstly, since real-time fraud detection systems work with sizeable volumes of data and respond within milliseconds, the appropriate use of computational means is crucial. Subsequent studies should focus on the strategies to enhance the performance of the system by determining the computing needs for real-time fraud identification and includes assessing the capability of infrastructure for different levels of transactions, optimizing the speed of transactions, and improving the cost effectiveness without a decline in accuracy of results. Other scenarios include cloud-based and distributed computing frameworks that will also be used in order to enable dynamic resource allocation depending on the transaction traffic. Secondly although the current research contributes the Isolation Forest algorithm to identify fraudulent behavior, future research will explore more complex machine learning and deep learning approaches for enhanced fraud detection. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are more effective in time-series data modeling and thus, using deep learning may enhance the detection of more sophisticated fraud patterns.

**REFERENCES**

Bin Sulaiman, R., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems, 2*(1), 55-68. https://link.springer.com/content/pdf/10.1007/s44230-022-00004-0.pdf

Chen, J. I. Z., & Lai, K. L. (2021). Deep convolution neural network model for credit-card fraud detection and alert. *Journal of Artificial Intelligence, 3*(02), 101-112. https://doi.org/10.36548/jaicn.2021.2.003

Khan, S., Alourani, A., Mishra, B., Ali, A., & Kamal, M. (2022). Developing a credit card fraud detection model using machine learning approaches. *International Journal of Advanced Computer Science and Applications, 13*(3), 1-76. https://dx.doi.org/10.14569/IJACSA.2022.0130350

Mir, A. A. (2024). Adaptive Fraud Detection Systems: Real-Time Learning from Credit Card Transaction Data. *Advances in Computer Sciences, 7*(1), 1-55. https://academicpinnacle.com/index.php/acs/article/download/229/252

Nakra, V., Pandian, P. K. G., Paripati, L., Choppadandi, A., & Chanchela, P. (2024). Leveraging machine learning algorithms for real-time fraud detection in digital payment systems. *International Journal of Multidisciplinary Innovation and Research Methodology, 3*(2), 165–175. https://ijmirm.com/index.php/ijmirm/article/download/97/92

Potla, R. T. (2023). AI in Fraud Detection: Leveraging Real-Time Machine Learning for Financial Security. *Journal of Artificial Intelligence Research and Applications, 3*(2), 534-549. https://aimlstudies.co.uk/index.php/jaira/article/download/189/179

Zhou, H., Sun, G., Fu, S., Jiang, W., & Xue, J. (2020). A Scalable Approach for Fraud Detection in Online E-Commerce Transactions with Big Data Analytics. *Computers, Materials & Continua, 60*(1), 1-54. https://cdn.techscience.cn/files/cmc/2019/v60n1/20190627024003_27050.pdf