International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I3P107

Volume 72 Issue 3, 46-52, March 2024

Review Article

Application of Machine Learning Techniques in Fintech Integrations in the fields of Fraud Detection and AML

Prasenjit Banerjee[1], Rajarshi Roy[2]

[1]*Director of Technical Architecture, Salesforce, Chicago, USA.*
[2]*Sr. Engineering Manager, Discover Financial Services, Chicago, USA.*

[1]*Corresponding Author : prasenjit.banerjee@ieee.org*

Received: 22 January 2024 Revised: 28 February 2024 Accepted: 14 March 2024 Published: 29 March 2024

Abstract *- Fintech Systems have evolved rapidly in the last ten years, first with the commercialization of Big data systems and then with the abundance of Machine learning models that have been trained on a large volume of data sets. However, it is important to understand the challenges faced by the organizations related to the adoption of this machine learning algorithm in the financial technology space. Finance is very well regulated, and most of the data in financial transactions involves personally identifiable information that cannot be made available to Machine learning models because of regulatory requirements. In this paper, we will examine some of the real-world challenges and the solutions offered. We are evaluating novel techniques that examine the application Bayesian perspective and allow us to model machine learning algorithms with simulated data that mask sensitive information. Our approach follows an iterative performance of classification in a diagnostic setting. These novel techniques allow for simplification without negatively impacting the efficacy of the model. We will also look at some of the other aspects of data preparation that allow for speed evaluation and rapid prototyping. Last but not least, we will focus on real-world applications of machine learning algorithms in categorization and anomaly detection.*

Keywords *- Machine Learning, Big Data, Artificial Intelligence, FinTech, Data classification, Data*.

1. Introduction

In the rapidly evolving landscape of Financial Technology (Fintech), the challenge of ensuring that models perform accurately on new, previously unseen data is critical. This challenge is not unique to Fintech but is a fundamental aspect of all Machine Learning (ML) endeavors. One of the primary concerns in this regard is the risk of overfitting, where models are so finely tuned to training data that they fail to generalize well to new data. This issue is exacerbated by the curse of dimensionality, a phenomenon where the complexity of the model increases exponentially with each additional feature, making the model less effective on new data sets. In Fintech, where accurate predictions can significantly impact financial decisions and outcomes, it is essential to employ rigorous model validation techniques [1]. This often involves dividing the dataset into multiple subsets to perform various rounds of training and testing. This iterative process helps in identifying a model that not only fits the training data well but also shows strong generalization capabilities on unseen data. The performance of the model on these tests serves as a proxy for its ability to generalize, which is crucial for its application in real-world financial scenarios. Machine learning, a subset of Artificial Intelligence (AI), focuses on creating systems that can learn and improve from experience with minimal human intervention. By analyzing historical data, ML models in Fintech can predict future trends, behaviors, and outcomes, becoming increasingly accurate over time. This capability is particularly valuable in Fintech, where predicting market trends, consumer behavior, and risk assessment is central to innovation and service improvement. The adoption of ML in Fintech has been transformative, enabling companies to offer personalized financial services, enhance security through fraud detection algorithms, and optimize operations [2]. Leading tech companies, as well as specialized Fintech startups, are leveraging ML to gain a competitive edge, making it a cornerstone of modern financial services. As technology continues to advance, the role of ML in Fintech is set to grow, underscoring the need for models that are not only powerful but also robust and generalizable.

2. Review of Existing Literature

Tuyls et al. highlights multiple obstacles encountered in Fraud Detection efforts. A primary challenge is the significantly unbalanced nature of datasets, where fraudulent instances constitute only a minor fraction of the total data, complicating the training of effective models. Additional issues stem from the presence of noisy data and patterns that overlap, making it hard to distinguish between legitimate and fraudulent activities. A critical concern is the ever-evolving nature of fraudulent behaviors, necessitating adaptive

classification models capable of identifying and adjusting to these changes. Following this introduction, we delve into an examination of key research employing machine learning and deep learning techniques specifically for fraud detection.

1.1. Comparative Study on KNN and SVM

Zareapoor and Shamsolmoali conducted a study on Fraud Detection, employing several analytical techniques such as Naïve Bayes, KNN, SVM, and the Bagging Ensemble Classifier. Their research paper highlights key challenges in fraud detection, notably the scarcity of real-world data.

This scarcity is primarily due to financial institutions like banks safeguarding their data for privacy reasons, forcing researchers to rely on simulated datasets [3]. They also point out the issue of data imbalance, with fraudulent transactions typically representing only 2% of the total, leaving 98% as legitimate. The study discusses the complexities of handling large datasets and the extensive computational resources required.

Additionally, it addresses the evolving nature of fraud, underscoring the necessity for continually updating machine learning models to detect new fraudulent tactics effectively. For their analysis, Zareapoor and Shamsolmoali utilized a dataset from the UCSD-FICO competition, which comprised 100,000 credit card transactions from an e-commerce platform, including 2,293 fraudulent cases. This dataset established a fraud-to-legitimate transaction ratio of approximately 100:3. They devised an experimental approach that segmented the dataset into four parts, with fraud percentages at 20%, 15%, 10%, and 3%, respectively. Through their experimentation, they determined that traditional metrics like accuracy or error rate were inadequate for assessing model performance in this context [4]. Instead, they opted for metrics such as the True Positive Rate, True Negative Rate, False Positive Rate, and False Negative Rate to provide a more accurate reflection of model effectiveness. Employing a 10-fold cross-validation technique, their findings indicated that the KNN algorithm outperformed the SVM and Naïve Bayes Classifier in detecting fraud, demonstrating a lower false alarm rate and a higher fraud detection rate across all subsets of the dataset.

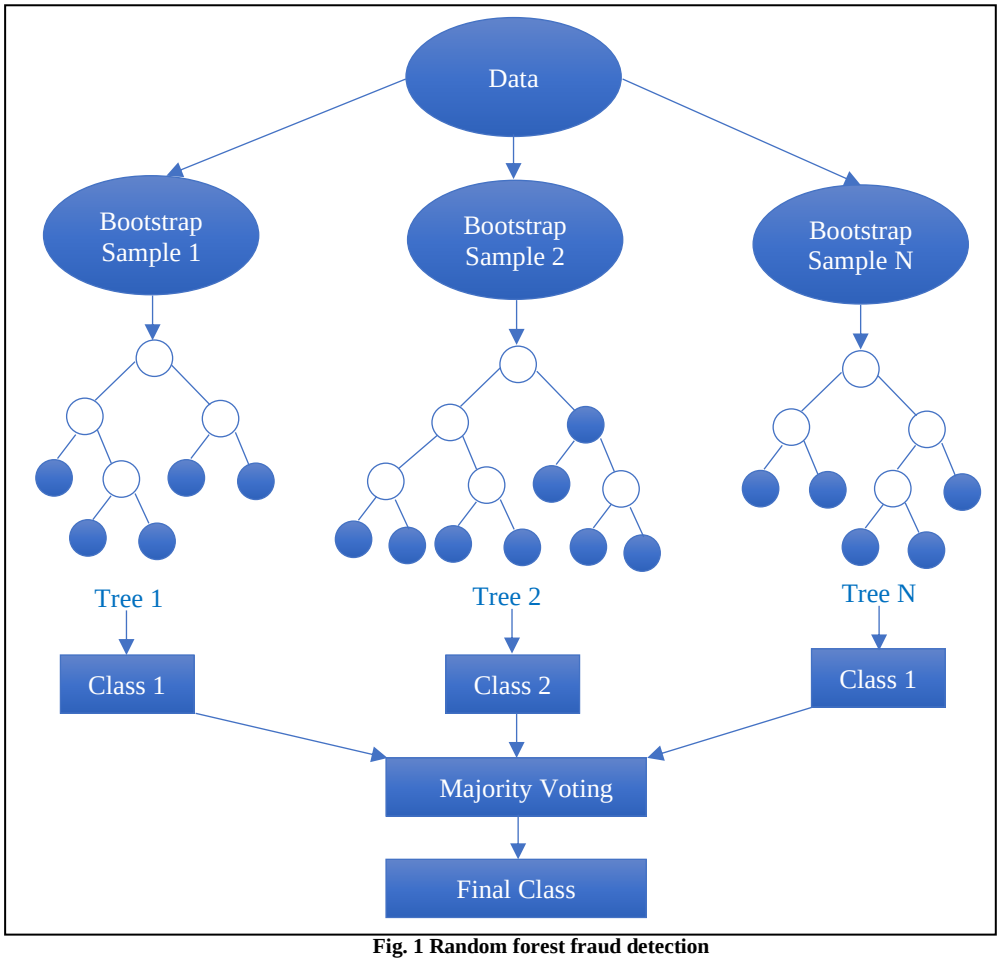

Fig. 1 Random forest fraud detection

1.2. Random Forest in Fraud Detection

In the context of Financial Fraud detection, Randhawa and colleagues' research focuses on detecting credit card fraud through the use of machine learning techniques, including AdaBoost and majority voting, among others like Naïve Bayes and Random Forest. Their approach employs "Majority Voting" to synergize the capabilities of multiple algorithms. Specifically, they examine the performance of the AdaBoost ensemble model and highlight its sensitivity to data anomalies and outliers.

The study utilizes RapidMiner software for implementation and tests the model using a dataset from the Southeast Asia region, characterized by a highly imbalanced composition with less than 1% of transactions being fraudulent [5]. To ensure robustness in their evaluation, a 10-fold cross-validation method is applied, with the Matthews Correlation Coefficient (MCC) serving as the primary metric for assessing classifier efficiency, incorporating true positive, true negative, false positive, and false negative rates.

Their findings indicate that the Random Forest classifier excels, achieving an MCC rate of 0.990, outperforming other methods such as SVM and gradient-boosted trees. Remarkably, integrating AdaBoost with Random Forest reached 100% accuracy and an MCC rate of 1. However, the study cautions about the generalization capabilities of the developed model, noting the need for further validation of unseen data [6].

The research concludes that hybrid models, which combine multiple machine learning techniques, tend to yield more reliable and effective results in fraud detection compared to single-algorithm approaches, offering significant implications for the advancement of fraud detection strategies within the Fintech sector.

3. Research and Methodology

The advancement of learning systems has been significantly propelled by probabilistic methods, with rapid developments in both techniques and their applications. This discussion will focus on the foundational principles of several key statistical methods rather than their detailed applications. At their core, statistical methods aim to define the probability distribution among a set of random variables, viewing data as manifestations of these variables' interactions. Statisticians rely on historical data to estimate these probability distributions, enabling them to address challenges such as identifying incomplete data examples or deducing the processes behind the data creation [7].

A crucial aspect of statistical modeling involves differentiating between parametric and non-parametric models based on how they conceptualize probability distributions. Parametric models presuppose a specific structure and form for the distribution, requiring only the estimation of a limited number of parameters from the data to define the distribution's exact characteristics.

3.1. Naive Bayes

The Naïve Bayes method in Artificial Intelligence (AI) is a simple yet powerful algorithm for predictive modeling and machine learning. Based on Bayes' Theorem, it is used for classification tasks, where the goal is to predict the category or class of a given sample [8]. The "naïve" aspect of the algorithm comes from its assumption that the features (variables) used to predict the class are independent of each other given the class. Bayes' Theorem provides a way to calculate the probability of a hypothesis (e.g., a class label) given some evidence (e.g., features of a sample). It is expressed as:

$$P(A|B)=P(B|A).P(A)/\ P(B)$$

Where:

- P(A|B) is the posterior probability of class A given the evidence B
- P(B|A) is the likelihood, the probability of evidence B given that class A is true.
- P(A) is the prior probability of class A
- P(B) is the probability of evidence B

To apply Naïve Bayes, a dataset is prepared where the class of each instance is known. This dataset is used to calculate the prior probabilities of the classes (P(A)) and the likelihoods (P(B|A)) for each feature. The algorithm assumes that each feature contributes independently to the probability of the class, simplifying the calculation of the likelihood [9]. This assumption is often not true in real life (hence "naïve"), but in practice, the algorithm still works well. When predicting the class of a new sample, Naïve Bayes calculates the posterior probability for each class based on the sample's features. The class with the highest posterior probability is chosen as the prediction.

Naïve Bayes is widely used in various AI applications, including Spam Detection, in which emails have been classified as spam or not spam based on word frequencies within the emails. Alternatively, in the case of Document Classification, in which documents could be categorized into topics based on the presence of specific words.

3.2. Using Convolutional Neural Networks (CNNs)

Using Convolutional Neural Networks (CNNs) for identifying suspicious activity in Fintech transactions involves transforming financial data into a format suitable for deep learning analysis. This innovative approach seeks to uncover fraudulent patterns within large volumes of transactions. A streamlined overview of the process has been provided below for observations.

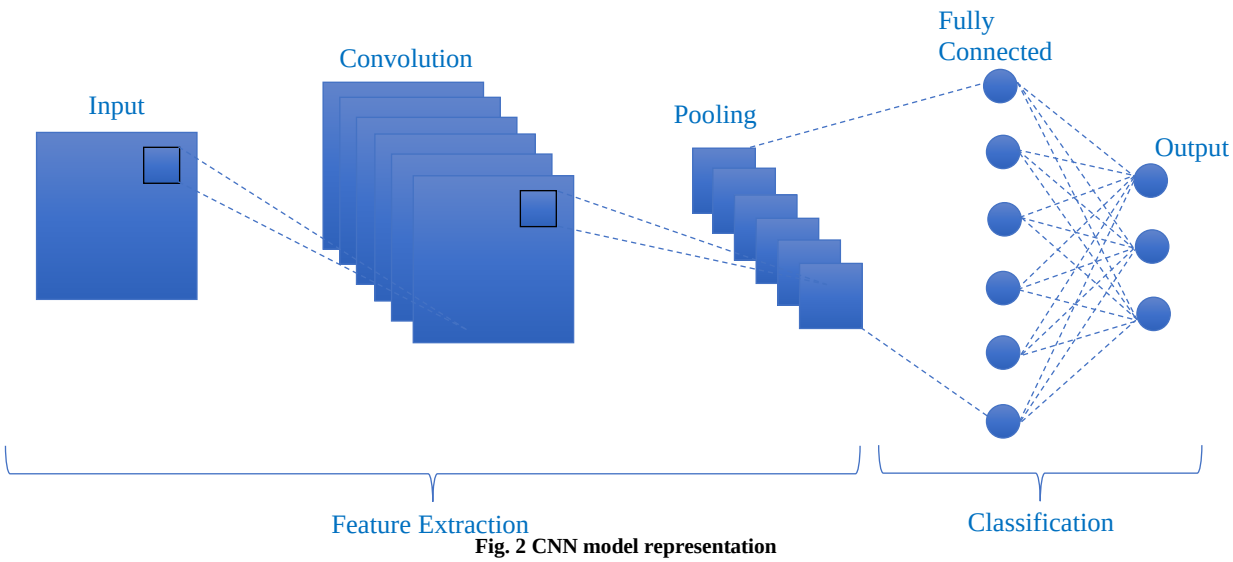

Fig. 2 CNN model representation

3.2.1. Data Preparation

Collect a dataset of Fintech transactions, including both legitimate and suspicious cases, and label them accordingly. Transform the transaction data into a CNN-compatible format, such as one-dimensional sequences or image-like representations. Tailor the CNN architecture to handle the structured nature of transaction data. This might involve one-dimensional convolutions for sequence data or two-dimensional convolutions for image-based data representations. Train the CNN on the prepared dataset, using regularization and dropout to prevent overfitting and ensure the model learns to generalize well to unseen transactions. Assess the model's effectiveness using a separate validation set and metrics like precision, recall, and AUROC [10]. Fine-tune the model as necessary to improve its detection capabilities. Integrate the trained model into the transaction processing system for real-time or batch analysis, flagging suspicious transactions for further review.

3.2.2. Considerations for Deployment

Processing Strategy: Choose between real-time or batch processing based on operational needs and system capabilities. Regulatory Compliance: Ensure the model's use complies with relevant financial regulations and privacy standards. Domain-Specific Customization: Adapt the model for different financial contexts to accurately detect suspicious activities unique to each domain. This concise approach leverages the advanced pattern recognition capabilities of CNNs to enhance fraud detection in financial transactions, demonstrating deep learning's potential in Fintech security applications.

3.3. Using Anomaly Detection (AD)

The Anomaly Detection Model is a model where techniques are used to uncover subtle anomalies indicative of suspicious transactions. The method relies on identifying the deviant patterns as a Baseline to identify fraudulent patterns. Each dataset (Transaction, Balance, Inquiry, Customer Data Changes) is associated with a scoring system that scores the bulk of the similar properties close to each other and gradually depicts the outliers.

There are various Statistical and Deep Learning Techniques like Z-Score, Mahalanobis distance, Modified Encoders etc., available for Anomaly Detection. However, a Machine Learning technique popularly used is called Isolation Forest [11].

The Isolation Forest algorithms, which isolate anomalies by constructing random decision trees, prove effective in isolating unusual patterns in financial transactions.

4. Implementation and Analysis

When it comes to implementing a Fraud Detection or AML Process for a Financial Organization there are multitudinous teams and skill sets required for it. However, predominantly, there are 2 aspects of such a program.

4.1. Model Selection, Customization and Training

The results and discussion may be presented separately or in one combined section and may optionally be divided into headed subsections.

4.1.1. Problem Definition

The first step is to define the problem clearly. This involves understanding the regulatory requirements for AML as well as Fraudulent Patterns itself. Although not a Rule-Based model, it is injudicious not to apply the years of experience gathered by the Rule/Decision engines. Also, any Machine Learning model is a construct of objective and having Clarity is the first step. Guidelines established by the

Banking Secrecy Act, Anti-Money Laundering Act, Finsen Guidelines, etc., for the US and Similar Guidelines depending on the geographical location are important. Build a profile of what constitutes normal behavior for legitimate users. This involves analyzing historical data to establish patterns, trends, and statistical distributions. This also includes Building a profile of what constitutes normal behavior for legitimate users [12]. This involves analyzing historical data to establish patterns, trends, and statistical distributions as well as Feature Engineering, which means extracting the relevant features from the data that are the key indicators that would regulate the model either by Supervised or Unsupervised Learning.

4.1.2. Model Selection

There are several models available for Fraud detection, as mentioned above, but it primarily depends on your organization's architecture and Fraud Detection Business process. If there is an appetite or need for BTL (Below the Line Testing), then probably a score based probabilistic Supervised Learning model of Classification or Regression type is your best bet.

4.1.3. Data Collection and Preprocessing

Data is the backbone of any ML model. For AML and fraud detection, this includes transaction data, user behavior data, and historical fraud data. More so the source of Truth needs to be a very reliable service specializing in collecting such data. It is also crucial to ensure that the data is free from Bias. A few big organizations who build an In-House AML solution prefer to buy the data from 3rd Party providers who specialize in Data Collection [12]. But once the data is obtained it is very important to put an automated filtration system to keep your model from performing at Optimum level. Preprocess the data to remove noise, handle missing values, and normalize features. Data preprocessing involves cleaning the data (handling missing values and outliers), feature engineering (creating new features from existing ones), and data transformation (normalizing or scaling data).

4.1.4. Model Training

The selected model is trained using the preprocessed data. This is the most important step in the whole process which is directly proportional to the behavior of the model post-implementation. It is recommended to use historical data with known fraudulent cases to train the anomaly detection model [13]. There should also be Volume Testing involved here, which compares the output of the MUT (Model Under Test) - β with the output of the legacy Fraud Detection system or the Benchmark Performance output - α.

4.1.5. Model Evaluation (Feedback Loop)

The model's performance is evaluated, and the results are used to update the model using appropriate metrics such as accuracy, precision, recall, and the Area Under the ROC Curve (AUC-ROC). This is where the evaluation of the BTL testing is needed based on the appetite. It is recommended to continuously update the model with new data to adapt to changing patterns and emerging fraud techniques. Re-evaluate and retrain the model periodically.

4.1.6. Memory Optimization

Finally, the Models, once Customized, take a long time, a lot of Memory and tremendous computational power to run. Some memory optimization techniques should be used to optimize the performance and evaluate it.

4.2. Infrastructure Setup

The Infrastructural Landscape of a Model depends upon the Data Intake Mechanism as well as the technological environments of the peripheral systems that would be consuming the output of the models. Models can be deployed and executed On-Prem Servers to any Cloud-based Environment like AWS, GCP, etc. For ease of updates to the model for both under Supervised and Unsupervised setup, a robust pipeline should be set up which takes care of automating the tasks as much as possible. Some of the aspects of the infrastructure setup are presented:

4.2.1. Hardware

Deploying an LLM (Large Language model) requires robust hardware. LLMs demand significant computational power. High-end GPUs (Graphics Processing Units) or TPUs (Tensor Processing Units) are essential for training and inference [14]. LLMs often require large storage capacities for storing model weights, training data, and results. Fast and reliable storage systems are crucial.

4.2.2. Deployment

This is the most important aspect of any LLM Model Operationally.

Model Serialization

The model needs to be in a format suitable for deployment, such as TensorFlow Saved Model or Porch's Torch Script.

API Development

An API or a service is a preferred approach that allows users or other applications to interact with your LLM. Restful APIs or group endpoints are common choices. This reduces the dependency on Batch processes, which take longer and also have an in-frequent interval of time.

Scaling

Depending on your application's requirements, scale your deployment horizontally or vertically to handle increased load. A lot of this is offered out of the box in modern Cloud based platforms.

Monitoring and Maintenance

Continuous monitoring of your deployed model for performance is needed, and Smart Alerting or Revival Capabilities should be built in.

5. Results and Observation

Here, the effect of AI on some of the most common types of financially perfidious activities is discussed.

5.1. Unauthorized Transactions

AI detects anomalies in the card owner's spending patterns and flags them in real time. By building predictive models of the user's future expenditure, it immediately sends notifications in case of suspicious behavior. The legitimate card owner can then block the card and contain damages.

5.2. Real-life Application

Mastercard's Decision Intelligence employs AI to analyze cardholder spending behavior (historical shopping), set a behavioral baseline against which it compares each new transaction, and evaluate the risk of fraud in real-time, enabling it to block suspicious transactions before they are authorized.

5.3. Identity Thefts

Cybercriminals steal a customer's identity by hacking into their account and changing crucial account user credentials like passwords. Because these models recognize the customer's behavior patterns, they may detect unexpected activities such as password changes and contact information updates violating frequency patterns. To avoid identity theft, it warns the customer and employs measures such as multi-factor authentication.

5.4. Document Forgery

Forged signatures, fake IDs, and fake credit card and loan applications, which are common issues in banking. ML algorithm-based models learn the patterns of a signature and an ID and detect minor flaws imperceptible to the human eye. They can also reduce the possibility of someone cashing a check with a phony ID.

5.5. Reduce False Positives

ML based model made a significant dent in false positives to already burdened FinCen by analyzing the patterns or suspicious activities in general. This resulted in more accurate actions and kept the cost of Regulatory Reporting Low.

6. Conclusion

Financial Institutions that have invested in building a robust ML based Fraud Detection or Suspicious Activity Detection Infrastructure have observed significant improvements in accuracy and reduction in cost. Compared to the earlier rule-based systems, AI/ML-based models made exemplary advancements in the areas of Pattern Recognition, Behavioral Biometrics, Predictive Analytics, and Real-time Monitoring. But most interestingly, the adaptability of these models helps Financial Institutions remain in step with the Fraudsters compared to the rule-based model, which requires time and effort to be updated. Coupled with the power of Cloud Engineering techniques, these are completely scalable in nature as well as coping with the rising tide of data. In recent years, the number of SAR filings has surged. In 2022, financial institutions filed over 3.6 million SARs, representing a 57% increase from pre-pandemic levels in 2019 [15]. This surge reflects heightened vigilance and regulatory compliance efforts. Technologies like generative AI and NLP can revolutionize AML by analyzing a vast amount of unstructured data and finding hidden patterns and anomalies.

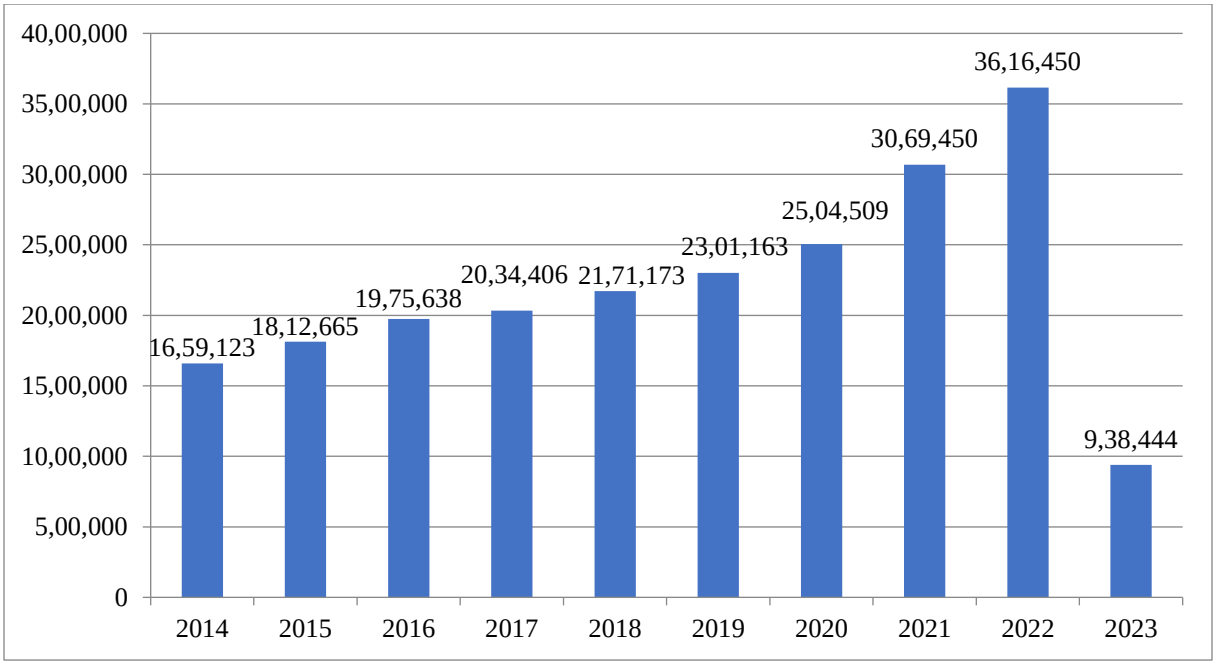

Fig. 4 Suspicious activity report all industries up to march 2023

Even with these advancements, there are still unexplored territories for AML beyond AI/ML-based capabilities which are being explored now by the financial institutions. Analysis of transaction networks, connections between entities, and behavioral patterns can reveal complex money-laundering schemes. Graph analytics can identify suspicious clusters and relationships that traditional models might miss. There is also an effort to make ML more transparent, especially for unsupervised learning models, by using Explainable AI (XAI), which tries to explain its outcomes and hence makes it easier to train the model as well as remove biases. To conclude, it does look like the most exciting days of Marriage between AI/ML and Fraud detection/AML lie ahead of us.

References

[1] Bruno Miranda Henrique, Vinicius Amorim Sobreiro, and Herbert Kimura, "Literature Review: Machine Learning Techniques Applied to Financial Market Prediction," *Expert Systems with Applications*, vol. 124, pp. 226-251, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[2] Francesco Rundo et al., "Machine Learning for Quantitative Finance Applications: A survey," *Applied Sciences*, vol. 9, no. 24, pp. 1-20, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Jaydip Sen, Rajdeep Sen, and Abhishek Dutta, *Introductory Chapter: Machine Learning in Finance-Emerging Trends and Challenges*, Algorithms, Models and Applications, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Saqib Aziz et al., "Machine Learning in Finance: A Topic Modeling Approach," *European Financial Management*, vol. 28, no. 3, pp. 744-770. 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Daniel Hoang, and Kevin Wiegratz, "Machine Learning Methods in Finance: Recent Applications and Prospects," *European Financial Management*, vol. 29, no. 5, pp. 1657-1701, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Prakhar Vats, and Krishna Samdani, "Study on Machine Learning Techniques in Financial Markets," *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, pp. 1-5, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Rudra Kalyan Nayak, Debahuti Mishra, and Amiya Kumar Rath, "A Naïve SVM-KNN Based Stock Market Trend Reversal Analysis for Indian Benchmark Indices," *Applied Soft Computing*, vol. 35, pp. 670-680, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[8] Yingjun Chen, and Yongtao Hao, "A Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for Stock Market Indices Prediction," *Expert Systems with* Applications, vol. 80, pp. 340-355, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[9] Mohiuddin Ahmed, Abdun Naser Mahmood, and Rafiqul Islam, "A Survey of Anomaly Detection Techniques in the Financial Domain," *Future Generation Computer Systems*, vol. 55, pp. 278-288, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[10] Jinbo Li et al., "Clustering-Based Anomaly Detection in Multivariate Time Series Data," *Applied Soft Computing*, vol. 100, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Alexander Bakumenko, and Ahmed Elragal, "Detecting Anomalies in Financial Data Using Machine Learning Algorithms," *Systems*, vol. 10, no. 5, pp. 1-29, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Parth Kulkarni, and Kalpesh Barde, "Applications of Generative AI in Fintech," *AIML Systems 2023 Conference*, pp. 1-8, 2023. [CrossRef] [Google Scholar]

[13] I. de Zarzà et al., "Optimized Financial Planning: Integrating Individual and Cooperative Budgeting Models with LLM Recommendations," *AI*, vol. 5, no. 1, pp. 91-114, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Rainer Alt, Gilbert Fridgen, and Younghoon Chang, "The Future of Fintech-Towards Ubiquitous Financial Services," *Electronic Markets*, vol. 34, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] *Special Report: Suspicious Activity Reports Surge; 2023 Filings On Pace For Another Record*, Thomson Reuters Institute, pp. 1-42, 2023. [Publisher Link]