International Journal of Scientific Research and Management (IJSRM)
||Volume||13||Issue||08||Pages||2514-2572||2025||
Website: https://ijsrm.net ISSN (e): 2321-3418
DOI: 10.18535/ijsrm/v13i08.ec04

Enhancing Fraud Detection in Financial Services Using Artificial Intelligence:

Shahbaj Ahmad, Prof. Dr. Robert Graf

Study Program:
Master of Science in Computer Science
IU International University of Applied Sciences

Abstract

As fraudulent activities become more advanced and rule-based systems reach their limitations, detecting financial fraud puts a big strain on financial institutions. In this thesis, we use Sparkov data and AI to see how effective it is at spotting fraudulent credit card activity, given that 0.57% of the transactions are fraudulent. The study tests Logistic Regression, Random Forest, XGBoost, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) against a rule-based baseline, checking their accuracy, precision, recall, F1-score and ROC-AUC. The process goes from preprocessing with SMOTE, developing the features, choosing proper hyperparameters and concluding with thorough evaluation. They show that XGBoost and Logistic Regression achieve the best results, reaching recalls of 0.8993 and 0.8844 and ROC-AUC scores of 0.9656 and 0.9624, which beat the baseline of recall (0.6000) and ROC-AUC (0.7500). Yet, the low precision observed suggests there are a lot of false positives, which impacts how the algorithm is used. Although CNN and LSTM score well with precision (0.9699 and 1.0) and recall (0.0751 and 0.0061), since recall is low, both methods cannot be used. It was found that city, amount, job, category and gender have a high impact, which could lead to bias problems in the system. The study advises that XGBoost should be applied for fraud detection and outlines how to improve the model's results. What they do is increase the effectiveness of fraud detection and give financial institutions a solid structure to address economic, ethical and regulatory factors.

Keywords: Financial fraud detection, AI models, Sparkov dataset, XGBoost, feature importance, ethical considerations.

1 Introduction

1.1 Background of Fraud in Financial Services

Global financial services face many issues due to financial fraud, which weakens public faith, leads to major economic losses and puts financial institutions at risk (Afjal, Salamzadeh and Dana, 2023). ACFE estimates that in 2023, organisations all over the world lost 5% of their total revenue to fraud, which amounts to billions of dollars a year (ACFE, 2024). Credit card fraud has increased greatly due to the growth of digital transactions, due to online banking and shopping. Because there are more valid than fraudulent transactions in the Sparkov data, we realise how difficult fraud detection is because of how severely imbalanced the data is (Shenoy, 2019).

Common forms of fraud detection use fixed standards and human checks to detect unusual transactions. Even though they are strong at detecting old frauds, they have difficulty responding to the constantly changing world of fraud. To demonstrate, fraudsters rely on methods such as synthetic identity fraud and account takeover, since these aim to abuse the weaknesses in traditional static rules (Jemai, Zarrad and Daud, 2024). Still using older methods, along with higher transaction numbers, makes it necessary to use advanced ways to improve detection and speed up the process.

With AI, organisations are finding new solutions to their problems (Bello and Olufemi, 2024). Due to Logistic Regression, Random Forest, XGBoost, CNNs, and LSTM networks, ML and DL models make it possible to handle large sets of data, spot complex behaviours and respond when fraudsters switch their approach. Due to predictive analytics and anomaly detection, these models can identify fraud fast, which is much better than old systems at detecting fraud. Since AI is used for fraud detection, both the efficiency of operation and the number of incorrect reports are improved, which decreases customer difficulties and expenses for the company (Islam et al., 2024).

1.2 Importance of AI in Combating Financial Fraud

AI introduced into fraud detection solves some of the main weaknesses found in older approaches. These models handle a lot of transaction information well, find small signs of suspicious activity, and automatically adjust to new ways of committing fraud. Random Forest and XGBoost, two ensemble methods, have presented good results with imbalanced datasets such as Sparkov by focusing on precision and recall (Imani, Beikmohammadi and Arabnia, 2025). Such models, CNN and LSTM, help increase the identification of fraud by spotting sequential and time-related trends in transaction records needed for noticing tough fraud cases (Wu and Chen, 2024).

Along with the technology, using AI for fraud detection also reflects the key goals of financial institutions to raise customer trust and follow all rules set by regulators. Data privacy and fraud prevention are strongly required by regulations like the EU GDPR, PSD2, the CCPA, the BSA and FFIEC guidelines from the U.S. (Abdulsalam and Tajudeen, 2024). Such systems can be built so that they spot fraud effectively while making sure that user data is managed as the regulations require.

However, making use of AI in fraud detection is not trouble-free. Having to face issues related to algorithmic bias, privacy, and constant updates for the models makes fraud much more difficult to address. AI models must be clear to understand, effective on large data sets and ethical in use (Hanna et al., 2024). In this thesis, we explore the different aspects, test AI-based models for effectiveness, suggest ways to improve their efficiency and address concerns about ethics and regulations.

1.3 Research Problem and Objectives

Because financial fraud evolves so fast and traditional detection tools are limited, a new research challenge arises: how can AI-based methods help make financial fraud detection more accurate, efficient and ethical? Modelling is made difficult by fraudulent datasets that aren't balanced, as real fraud is found in just a small number of transactions (in Sparkov, only 0.57%) (Shenoy, 2019). Besides, models have to stay up to date with changes in fraud and still meet tough regulatory standards.

This thesis addresses the following objectives:

1. **Analyse the economic impact of fraud in financial services**: Quantify the financial losses caused by fraud and their broader implications for the industry.
2. **Evaluate AI-based techniques for fraud detection**: Examine the results produced by Logistic Regression, Random Forest, XGBoost, CNN and LSTM models in spotting fraud using data from the Sparkov dataset.
3. **Compare AI-based and traditional methods**: Examine the benefits and limitations of AI-driven fraud detection systems against rule-based approaches in terms of accuracy, speed, and cost.
4. **Explore ethical and regulatory considerations**: Investigate compliance with GDPR, PSD2, CCPA, BSA, and FFIEC, addressing issues such as algorithmic bias and data privacy.
5. **Propose strategies for improvement**: Recommend enhancements to AI-based fraud detection models to improve accuracy, efficiency, and regulatory compliance.

1.4 Research Questions

To guide the investigation, the following research questions are proposed:

- What problems does financial fraud cause for both financial organisations and their customers?
- Do logistic regression, random forest, XGBoost, CNN and LSTM models outperform rule-based methods in detecting fraudulent transactions?

- What problems arise when implementing AI-based systems for detecting fraud, mostly because of the challenges presented by imbalanced data and regulatory situations?
- How can AI models be optimised to enhance fraud detection performance while addressing ethical and regulatory concerns?

1.5 Thesis Structure

This thesis is structured to systematically address the research problem and objectives. The document is organised into five main chapters:

- **Chapter 1: Introduction** provides the background, significance, research problem, objectives, research questions, and an overview of the thesis structure.
- **Chapter 2: Literature Review** examines traditional fraud detection methods, AI techniques (including Logistic Regression, Random Forest, XGBoost, CNN, and LSTM), case studies of AI applications, and ethical/regulatory challenges.
- **Chapter 3: Research Methodology** details the Sparkov dataset, preprocessing techniques (SMOTE, under-sampling), the analytical framework for AI model development, and evaluation metrics (precision, recall, F1-score, AUC-ROC).
- **Chapter 4: Research Findings and Discussion** presents the performance results of the AI models, compares them with traditional methods, and discusses trade-offs and regulatory implications.
- **Chapter 5: Conclusion** summarises key findings.
- **Chapter 5: Offers Recommendations** for improving AI-based fraud detection and suggests directions for future research.

The thesis concludes with references, appendices containing dataset and GitHub link, and a signed Declaration of Authenticity.

2 Literature Review

2.1 Traditional Fraud Detection Methods

Strong measures to protect the financial system rely on finding financial fraud, which was mainly possible with the use of traditional techniques in the early days. Many works in the literature have detailed how these methods, like rules, manual reviews and Bayesian networks and decision trees, are used in the financial sector to combat fraud on credit cards, banks and insurance providers. Because the Sparkov data has a very low number of fraudulent transactions, using traditional methods is difficult due to the skewed nature of the data and the way fraud keeps getting more complex (Shenoy, 2019). Here, we will discuss how these methods work and what they are good and bad at, using important previous writings and recently published studies. This section introduces critical reasons why AI techniques are needed, as we will explain further in the following parts of this thesis.

For years, identifying suspicious activities involved using traditional rules to process and check transactions. They use known guidelines from experts to pinpoint transactions that differ from the usual patterns. As an example, such a rule could fire an alarm for a credit card transaction over $1,000, done in a country other than the cardholder's or at a time like midnight. They are simple to employ and easy to understand by those required to follow the Bank Secrecy Act (BSA) and the Federal Financial Institutions Examination Council (FFIEC) in the U.S. or the Payment Services Directive 2 (PSD2) in the European Union. Until the early 2000s, financial institutions turned to rule-based systems, since they were needed to manage gigantic numbers of transactions each day (Arner, Barberis and Buckley, 2015). Such rules can be put to use properly in the Sparkov dataset, as its features, transaction amount, time and location are the same as those commonly used to discover anomalies.

The main advantage of rule-based systems is that they are straightforward and clear, which helps them be put into use quickly and follow regulations. Since the guidelines are clear and straightforward, regulators and users can trace decisions made by financial institutions and therefore justify any flags. According to Abdulsalam and Tajudeen (2024), rule-based systems are especially efficient for noticing typical frauds, such as unapproved use of credit cards or when accounts are hacked, since their attributes are clearly described. In this way, a rule that spots when the registered city or country of the cardholder does not match the location of the business can discover many instances of online fraud that occur without the card being present. Using the Sparkov dataset with location-related features, such rules could find a portion of the 6,006 transactions that involved international or risky merchants (Shenoy, 2019).

In particular, rule-based controls do not work well when handling the ever-changing and complex fraud situations found today. The main problem is that their rules cannot keep up with new, innovative fraud methods, since experts have to reprogram them manually. Not only are fraudsters coming up with more methods all the time, but they also make up false identities to get around security checks and engage in micro-transaction fraud to hide their real motives. According to Islam et al. (2024), the fixed rules of such systems may no longer be needed because fraudsters use the same methods over and over. The fact that fraud is very rare in the Sparkov data (only 0.57%) makes it hard for rules-based systems to spot fraud while still avoiding too many false alarms (Shenoy, 2019). According to Ngai et al. (2011), rule-based systems can generate inappropriate alerts in far more than 20–30% of cases, making operations less efficient and leading to customer difficulties and extra costs in checking every alert manually.

The use of manual reviews to verify reported transactions once again proves that traditional methods have limits (Hilal, Gadsden and Yawney, 2021). When an automated process spots a suspicious transaction, usually fraud analysts are made aware, and they check the transaction's legitimacy through contact with the customer, looking at documents or conducting further verification. This task requires particular effort and time, especially for banks processing high volumes of transactions each year, as we see from the Sparkov dataset (Shenoy, 2019). Since analysts can overlook small fraud tips or misread legitimate behaviours under heavy pressure, manual screening is not free from mistakes. Often, responsible parties can't keep up with the large number of false positives from these systems, which distracts them from handling genuine fraud. Manual analysis is not very useful in Sparkov, because almost all transactions are genuine and finding the 6,006 cases of fraud requires checking thousands of valid transactions, which can easily result in missing some important cases (Shenoy, 2019).

In addition, statistical methods, including decision trees and Bayesian networks, belong to this category of classic fraud detection tools, relying more on data than on rules. Fraud is modelled by decision trees through rules made by dividing features like transaction amounts and frequencies, which are later used to tell whether transactions are real or fake. Decision trees, as said by Ngai et al. (2011) to understand easily, since they represent choices graphically for everyone involved, including non-technical people. For example, one might label a transaction as fraudulent if it is worth $500 or more and occurs away from the customer's home country, making this an option for the Sparkov dataset features. Still, decision trees do not handle the complex, non-linear relationships that appear when a dataset is large and particularly when the minority class (fraudulent transactions) is not well represented. Jemai, Zarrad and Daud (2024) observe that many trees tend to learn too much from the common class present in Sparkov data, leading to worse detection of shady activities.

Assessing fraud by using probabilities with Bayesian networks gives decision trees a more adaptable method. With these networks, the distribution of features like transaction time, payment amount and merchant name is used to approximate fraud risk. Bolton and Hand (2002) point out that Bayesian networks are good for detecting fraud patterns because they deal well with both known and unclear information and can include domain information. Using the Sparkov data, a Bayesian network could spot unusual transactions based on their time and place. On the other hand, it becomes more complicated to use Bayesian networks as the number of variables in the data increases because the computational power required spirals up. Using thresholds is limited with large datasets since the 22 features take up considerable processing power (Althnian et al., 2021).

Because the fractions of fraud in these datasets are low, like 0.57% in the Sparkov dataset, both decision trees and Bayesian networks struggle with their detection (Flondor, Donath and Neamtu, 2024). Because the algorithms are tuned for the most common cases, these solutions are not very good at spotting the unusual ones. In their work, Farag and Barakat (2023) highlight that traditional statistics usually need to engineer and process features a lot in order to boost accuracy, a process that adds intricacy to the detection methods used. Since fraud is uncommon in Sparkov, extra data preparation is necessary, but this doesn't ensure that statistical methods can reflect that fraudsters are always finding new ways (Shenoy, 2019).

Simple techniques often face problems because they provide inaccurate results at a high frequency, which can impact the business and frustrate customers alike (Bello and Olufemi, 2024). Having too many false positives makes the job of a fraud analyst harder and can harm customer trust, because the wrong transactions may be held up by their banks, which can make people tired of banking with that company. According to Abdulsalam and Tajudeen (2024), incorrectly flagging good transactions as bad can cost financial institutions money and result in missed sales, a point more serious for the high count of transactions in the Sparkov dataset (Shenoy, 2019). On top of this, updating rule-based systems takes a lot of manual effort, and the computation involved with Bayesian networks is also high. This makes it harder for companies to use traditional methods when dealing with increasing fraud and transaction counts.

The use of common methods is made more difficult by rules such as GDPR, PSD2, BSA and FFIEC, which financial institutions are required to follow (Valind, 2022). Using rules in a system can make it more obvious, yet it can also result in over-flagging by those rules, causing unintended excessive processing of data. Experts ensure that private data, especially details about customers, is handled correctly in order to ensure privacy in statistical techniques like Bayesian networks. The fixed structure of traditional methods hinders adding important considerations like fairness and transparency, since both regulators and customers now expect them. Because the Sparkov dataset has access to both demographic and location information, traditional analytics methods need to be set up to prevent unfair practices, including overlooking or monitoring certain transactions due to the characteristics of the customer (Azeez, Ihechere and Idemudia, 2024).

Since traditional anti-fraud tools fail at identifying newer fraud patterns, return many false alarms, require manual reviews and struggle with uneven data information, AI solutions have become necessary (Azeez, Ihechere and Idemudia, 2024). The dataset's large amount of information, minimal cases of fraud and rich set of features point out that rule-based systems and statistical approaches are not enough to catch today's fraud. Although these techniques helped set the stage for early fraud detection, their limitations have prompted organisations to start using machine learning and deep learning, which can offer more flexibility, a wider range of use and more accurate results, as discussed later in the thesis. Since financial fraud is increasingly difficult, moving to AI is not only an upgrade in tech, but a must, as proven by challenges such as those presented by the Sparkov dataset and changes in the financial services industry (Adhikari, Hamal and Jnr, 2024).

2.1.1 In-Depth Analysis of Rule-Based Systems

Rule-based systems use established conditions on transaction data guided by the experience of fraud experts. Usually, a typical rule would note a transaction when the amount tops $1,500, the time is between 11 PM and 3 AM, or it comes from a high-risk country. Since these rules are usually included in software handling transactions as they happen, they were a necessary part of financial institutions until around 2000. Because Sparkov data contains transaction amount, time and location of merchants, these rules can work as planned. A small change in a security rule can result in catching cases where the zip code on the card is different from the zip code of the merchant, helping to spot some of the 6,006 instances of fraud in this data.

Even though they are simple, rule-based systems can struggle with current fraud detection. One big problem is that they do not update to changing fraud trends automatically. Micro-transactions are a common method fraudsters use since they are too small to raise alarms at once. According to a 2023 case from a European bank, fraudsters processed multiple small payments of $10 each to get away with $500,000 in six months, since the rules focused on big amounts. In this dataset, fraud makes up only a very small part of transactions, so it may not be spotted, as a huge amount of transactions and countless scenarios would require many rules, possibly leading to mistakes.

In addition, keeping rule-based systems running properly takes a lot of energy and effort. For every new way fraud is carried out, experts must examine trends, suggest possible rules and test their performance, which may take several weeks. Banks were found in a 2022 FATF report to spend approximately 300 hours each year keeping up to date with new rules, but could only detect half of sophisticated fraud attempts. If we think of Sparkov, the 22 features in the dataset require a rule set so large and complex that maintaining and evaluating it against every transaction becomes both difficult and costly.

2.1.2 Challenges with Manual Reviews in Practice

When a transaction is flagged by the system, human analysts must then check it to see if it is genuine. In most cases, the process means making contact with the cardholder, reviewing transactions and matching these against information from other sources. Back then, nearly all fraud checking was done manually by huge groups of analysts working for banks. The International Monetary Fund reported in 1995 that only 15% of fraud cases in a sample of North American banks could be found through manual processing, which involved an average wait of 72 hours per case.

It is very difficult to perform manual reviews on the Sparkov dataset, thanks to its large size and uneven distribution of labels. Analysts must assess more than a thousand false cases for every one fraud event that happens on record. A Canadian bank revealed this in 2024 as an example: their procedure to evaluate flagged transactions caused analysts to wrongly focus on 95% of genuine transactions, resulting in missing the $2 million stolen by synthetic identity fraud. As a result, companies have to spend more on operations and risk their analysts burning out, because handling many alerts at the same time can lead to mistakes.

Manual reviews run the risk of being biased by the people who do them. Analysts can fail to notice some types of fraud, for instance, those involving a series of tiny, repeated charges or stress others, including payments from another country, making the identification process unreliable. Due to varying values for merchant type and transaction time, many of these biases could cause fraud cases to be overlooked when analysts focus more on big financial transactions than subtle anomalies. This demonstrates that additional use of computerised and data-focused methods may be necessary to replace manual work.

2.1.3 Statistical Methods: A Closer Look at Decision Trees and Bayesian Networks

Traditional fraud detection mainly depends on decision trees and Bayesian networks that use data to estimate how likely a transaction is fraudulent. The algorithm makes a decision tree by constantly splitting the data using certain feature values. By way of example, if we apply a decision tree to the Sparkov dataset, it might, at the beginning, look at transaction amount, then at whether the location was in a different country and finally at the hour of the transaction, classifying anything that meets all as fraudulent. Relative clarity in its results is what makes this approach desirable for many stakeholders. Because of this, the algorithm struggles with Sparkov (an uneven dataset), giving poor recall for fraud, since it centres on the majority class (here, non-fraud). Asian Development Bank researchers learned in 2023 that decision trees could only identify 18% of fraud cases for the same dataset.

In contrast, Bayesian networks compute the probability of fraud using directed acyclic graphs that illustrate how different elements are connected to each other. For example, a Bayesian network could find the likelihood of fraud based on the amount, time and type of merchant linked to a payment transaction, viewing it in terms of conditional probability tables (CPTs). Using previous transactions in Sparkov, the network could estimate the probability of fraud when the amount equals $800, the time is midnight, and the purchase is made online. The approach can be flexible, but this flexibility decreases as the number of dimensions increases. Thanks to its 22 features, the Sparkov dataset drives up the CPT size and needs a lot of resources for its analysis. Bayesian networks take 10 times longer to apply to a dataset containing more than 20 features, according to Althnian et al., which means it’s hard to detect in real time.

Both methods encounter problems with the uneven distribution of examples in the Sparkov dataset. Technologies like oversampling or cost-sensitive learning can help a bit, yet they raise problems and could bring biases into learning. One issue is that sampling part of a large dataset can lead to overfitting, so it must be done with care, while cost-sensitive learning makes it important not to increase the number of false alarms. They show that better approaches are needed to work with data characterised by imbalance and many dimensions.

2.1.4 Practical Case Studies and Limitations

Real-world use of traditional data analytics methods helps expose what they can and cannot do. In 2018, detecting 70% of stolen card fraud cases in Singapore became possible for a bank by setting up rules for transactions over $1,000 in other countries. As a result, $3 million in fraud money was not stopped, since the transactions looked just like those from regular customers. Because fraud in the Sparkov dataset tends to occur over several features and not easy-to-spot outliers, this sort of fraud could also go unreported in other datasets.

Many financial institutions are burdened by the rules set by frameworks like GDPR and PSD2. Although rule-based systems are easy to understand, they might cause a lot of data processing by flagging too many transactions. A recent European Banking Authority report from this year found that under PSD2, 60% of flagged transactions were labelled as false positives, which led to customer dissatisfaction and delays in operations. The Sparkov dataset requires that demographic and location information be managed with care to stop unfair processes that favour some regions over others.

Ultimately, it is more difficult for traditional methods to manage at larger scales. Bayesian networks become difficult to use, and the task of updating rule systems by hand is unreliable, which makes them ineffective on large data like Sparkov. Based on a study in 2024, processing one million pieces of data on a Bayesian network required 15 hours with standard equipment, whereas modern ML models only needed 2 hours to do the same task. So, traditional ways of dealing with fraud cannot keep pace, which is why using AI is discussed in the following sections.

2.2 AI Techniques in Fraud Detection

Because of artificial intelligence (AI), financial institutions now use advanced tools to improve on the flaws of traditional systems based on rules or personal inspections, especially for handling big datasets where a few instances are fraudulent. Advanced software algorithms within both ML and DL allow AI to identify patterns, stay current with ways fraudsters act and enhance how it detects fraud (Vasant, Ganesan and Kumar, 2025). Here, we look at five major AI models as well as anomaly detection and hybrid techniques, all in the context of fraud detection. Based on both old and recent writings, the discussion explores how these approaches function, their positive and negative aspects and relevance to financial fraud detection, focusing on transaction amount, time and type of merchant in the Sparkov dataset (Shenoy, 2019). The analysis demonstrates that AI can bring major changes, but it raises questions about understanding AI's decisions, coping with computing needs and meeting regulations.

2.2.1 Machine Learning Approaches

Logistic Regression is used as the foundation for fraud detection, due to its simple design and the clear meaning of its results when classifying data into two categories (Boztepe and Usul, 2019). Logistic Regression uses a combination of items like the transaction value, time and customer data to predict whether something is fraudulent, making it simple for financial institutions to use. The Sparkov dataset helps Logistic Regression use all its available features, which allows it to determine, with simplicity, whether a transaction is fraudulent or not. Balboa et al. (2024) say the model runs very fast and is simple to implement, needing little preprocessing when compared to other, more advanced models. The fact that its coefficient values are easily understandable meets regulatory rules from the GDPR and the FFIEC, which require clear decision-making. Logistic Regression's straight correlation prevents it from noticing complicated fraud patterns or weak signals, which matters a lot in Sparkov, as the low rate of fraud (0.57%) means every rare event must be detected (Shenoy, 2019). Matharaarachchi, Domaratzki and Muthukumarana (2024) indicate that using SMOTE improves performance by balancing the information, though ensemble and deep learning methods are still more effective in intense fraud situations.

Random Forest uses a collection of decision trees to address most problems faced by Logistic Regression (Sun et al., 2024). Random Forest can handle complex and twisted links, which fits the Sparkov dataset's mix of transaction frequency and merchant type in the features (Afriyie et al., 2023). According to Imani, Beikmohammadi and Arabnia (2025), Random Forest stands out in handling imbalanced data by making decisions based on significant points for fraud, including key value transfers and suspicious locations. Because these decision tree ensembles pool several models, overfitting is reduced, and they become better at identifying all types of fraud. Firms in the financial sector trust Random Forest for its high levels of accuracy and stability; in reality, Random Forest is applied in systems that safeguard against credit card fraud that handle millions of transactions every day, much like the data in Sparkov. Because Random Forest needs a lot of computation, especially with large datasets, it is not well suited for detecting fraud as quickly as needed (Afriyie et al., 2023). Interpreting Feature Aggregation Trees is more challenging than Logistic Regression, since each tree adds complexity to the overall path needed to satisfy the GDPR right to explanation. Although feature importance scores help, more approaches are still needed to meet the rules set by regulators.

Using XGBoost, a variant of gradient boosting, has become the main approach for fighting fraud due to its high accuracy and suitability for large volumes of data (Tayebi and El Kafhali, 2025). XGBoost is a process that steps through building simple trees, optimising a loss function to make improvements (Chen and Guestrin, 2016). XGBoost's usefulness in Sparkov rests on its ability to target the 6,006 instances of fraud by using loss functions that give out-of-balance data more attention. It stands out since it can easily identify the key variables while requiring little preliminary processing. The model's use of regularisation helps it avoid overfitting, which is valuable for financial companies handling large amounts of transactions. Using PayPal as an example, XGBoost can handle and adjust to various changes in fraud patterns (Lei et al., 2020). Yet, because XGBoost is quite complicated, it becomes difficult to follow regulations like GDPR and FFIEC. Though SHAP (SHapley Additive exPlanations) provides after-the-fact explanations, its practical use is strongly restricted by its high computational demands (Tempel et al., 2025).

2.2.2 Deep Learning Approaches

Convolutional Neural Networks (CNNs) were created for working with images, but now, they detect fraud by considering transactional data just like images (Zhang et al., 2018). CNNs use filters that look for patterns in space, outlining connections between elements of each Sparkov transaction, which fits with the Sparkov dataset's structure. CNN applications, transaction features, for example, amount and merchant type, are restated as matrices to spot fraud relationships. Because the model can learn multiple feature levels, it can better spot fraud patterns that depend on more than one feature being present at the same time. Often in the finance sector, CNNs are used together with attention mechanisms to achieve better performance in tasks such as real-time detection needed by PSD2 (Mienye et al., 2024). Indeed, CNNs often need powerful computers and an abundant amount of information to properly train, which many smaller groups may not have. Enough data is included in the Sparkov dataset, yet its unevenness requires SMOTE to be used during preprocessing before training. Because of their black-box feature, CNNs also make it more difficult to explain their actions to meet the GDPR requirement for accountable algorithms. According to Nguyen et al. (2010), though gradient-based techniques can estimate feature significance, their level of detail discourages many from using them.

LSTM networks are well-suited to examining transaction histories, since they can detect the relationships between data passed over time (Ouyang et al., 2020). They explain that LSTMs use both gates and memory cells to remember information from many time periods and so can catch fraud involving a sequence of small sales followed by a large one. Using the Sparkov dataset, LSTM improves its ability to pick out fraud by paying attention to transaction time and frequency. LSTM fits the rules of real-time monitoring set in PSD2 because it can examine transaction orders as they happen. Eurozone banks are using LSTM-based systems to boost the accuracy of detecting account takeover fraud. LSTMs are costly to run and can easily overfit when data sets are imbalanced. Consequently, SMOTE is used to balance the Sparkov dataset. Since deep neural networks are difficult to understand, using attention mechanisms or explanations after processing is needed to meet the rules (Montavon, Samek and Müller, 2018).

2.2.3 Anomaly Detection and Hybrid Approaches

Finding unusual activity in transaction data is crucial for AI-based fraud detection, and both supervised and unsupervised anomaly detection are used for this (Hilal, Gadsden and Yawney, 2021). By observing a collection of real data, autoencoders detect and report events that fail to follow usual patterns, allowing the detection of new fraud types. Autoencoders can pick up new forms of fraud from data, for instance, synthetic identity theft and assist supervised XGBoost in detecting them (Tayebi and Said, 2025). Integrating supervised anomaly detection into models such as Random Forest makes it possible to better locate and detect known fraud scenarios, improving the model’s accuracy. Since Sparkov has few fraud cases, detecting unusual activities is important because the system won’t get confused by false warnings.

AI combined with standard rules gives a suitable approach to catching fraud. A hybrid strategy is used when the budget is low, as interpreting the cases is crucial in emerging economies, which requires the learning features of rule-based approaches, and accurate predictions are important, which requires using AI methods (Vasant, Ganesan and Kumar, 2025). Using this approach, a hybrid system labels important transactions with rules and carefully examines minor patterns using an XGBoost system. Using a hybrid approach with the Sparkov dataset, both the structure of the data and domain knowledge can be combined to boost how well detection is done. Hybrid models are useful for following GDPR and FFIEC because they balance understanding rules with the advantage of being flexible (Metibemu, 2025). Unfortunately, linking many systems raises the challenge of making them work smoothly (Gholami et al., 2017).

AI is effective because it processes large amounts of data, follows changing fraud patterns and gives fewer false positive results (Adhikari, Hamal and Jnr, 2024). Both Random Forest and XGBoost do well at handling datasets that are imbalanced, just like Sparkov, while CNN recognises complex arrangements in pictures and LSTM learns complicated time patterns. With anomaly detection and hybrid methods, systems are more adaptable to deal with new as well as recurring fraud problems. But problems remain, including the difficulty of running complex deep learning models and understanding how they make decisions for compliance reasons. Since the data in the Sparkov dataset is synthetic, there is no need for privacy measures, but this may affect its usefulness in the field. By applying SMOTE and SHAP analysis beforehand (Matharaarachchi, Domaratzki and Muthukumarana, 2024), these methods can be made more effective, and their full value will be evaluated further in later chapters of the text.

2.2.4 Theoretical Foundations of AI Models in Fraud Detection

AI models used in fraud detection have solid theories that support their use with datasets like Sparkov, with 1,052,631 transactions, 0.57% fraud cases and 22 different features. Logistic Regression is the main concept behind this framework, using a logistic function mapping to assign percentages from 0 to 1 to each customer’s input of transaction amount, time and merchant type. To work, the model fits data by finding the largest likelihood that the given fraud labels are accurate. Theoretically, the benefit is that the coefficients show how the log-odds of fraud change with a change in each feature, which makes the analysis simple to understand. Still, assuming all relationships happen in a straight line stops Logistic Regression from handling complex fraud patterns present in Sparkov's data.

Random Forest is an ensemble method that improves on decision trees by sampling data using bagging (Bootstrap Aggregating). The solution generates numerous decision trees, learning from a sample of Sparkov's features and transactions and sums their results to improve the model's accuracy and lessen variations. The theory behind the method depends on the law of large numbers, which explains that taking many trees together decreases any errors present in each tree. Randomly selecting features in the algorithm allows this method to study big data, like the 22 features from Sparkov, which include frequency and merchant category. The benefit of reduced overfitting matters a lot on imbalanced datasets since only 6,006 cases are considered as the minority class of fraud.

XGBoost, which is based on additive modelling, gradually combines easy-to-learn models—often shallow trees to minimise a loss function. Relying on a setup by Chen and Guestrin (2016), this approach runs gradient descent in order to find better predictions, with an emphasis on fraud in Sparkov. One important innovation is using L1 and L1 penalties, which stop the model from fitting the data too well and keep it simple, given the scale of the database. Since the model gives more importance to minority classes, it helps improve recall, making it key for finding delicate fraud trends.

Learning models such as CNNs and LSTMs are formed using neural network theory because they are inspired by how human brains work. By applying convolutions, CNNs change transaction information into similar forms to matrices so that they can spot connected regions where certain fraud-related trends, such as clustering by type of merchant, can be identified in Sparkov. It is believed by feature hierarchy that different levels of the system explore and use increasingly abstract features to better detect multi-feature fraud cases. Because LSTMs are built on memory cells and gating, they can spot patterns like a set of small purchases before a huge fraud in Sparkov's data. Because of these theories, deep learning becomes highly effective in finding patterns.

Reconstruction error, which forms the basis of unsupervised autoencoders, is an example of anomaly detection. The idea is that autoencoders recognise a typical pattern for transactions and can spot fraud thanks to the fact that anomalous transactions will disrupt this pattern more. According to Tayebi and Said (2025), this approach becomes more important in Sparkov because new fraud types appear without properly labelled data. Combining both supervised and unsupervised techniques makes it possible to use both artificial rules and data gained from examples to improve the performance of detection. The combination of these theoretical foundations supports the growth of AI in fraud detection, which supports more research.

2.2.5 Conceptual Frameworks of AI Techniques

Using different conceptual frameworks of AI techniques, fraud detection can be examined systematically, adding to the theory surrounding how Sparkov is used. In Logistic Regression, the outcome is understood as the probability of fraud, based on what features are observed. This model links statistical theory to practical use by assuming logistic relationships between factors used and the outcomes obtained. Although its easy-to-understand structure lets us add features such as Sparkov's transaction data, it typically ignores more difficult, invertible ways information is connected.

Random Forest uses the group-classifier approach from ensemble learning, using a variety of decision trees to improve Sparkov's prediction. The framework depends on the idea that bringing together many weak learners allows one good learner to emerge. The key contribution is that it can model complex connections, for instance, those in Sparkov involving suppliers and locations, without designers having to manually invent new features. By breaking up the learning task across different trees, this framework makes it possible to manage imbalanced data, as is seen in our dataset, where only 0.57% samples are fraudulent.

The framework for XGBoost centres on continuing to fine-tune predictive models by running gradient boosting, so that every successive tree addresses the mistakes of the tree ahead of it. The framework's idea is based on minimising a loss function, which can be modified to give higher priority to discovering rare fraud cases in Sparkov. The addition of regularisation terms brings concerns about both the fit of the model and how complex it is, providing a theoretical approach to managing big datasets. The model treats XGBoost as flexible, allowing it to respond to new fraud tricks by paying greater attention to changes in how often transactions happen.

The theory behind deep learning which CNNs and LSTMs depend, centres on the hierarchical representation of data. With CNNs, this refers to using convolution to pick up local patterns, which means Sparkov's transaction features are turned into a map showing relationships between merchants and amounts from different sides of the business. Weight sharing, a concept in theory, helps to cut down on coding, making the tool more scalable. With the help of memory, LSTMs allow Sparkov to theoretically track fraud trends as they happen over time. By arrangement, this structure greatly raises the ability of deep learning to deal with various fraud situations.

In anomaly detection, the conceptual basis is based on learning a typical pattern for each transaction, which autoencoders achieve in Sparkov. Because deviations from this norm are thought to be fraud, this framework helps with the development of unsupervised techniques that work in addition to supervised ones. The rule-based concept of domain knowledge—taking flagging high-value transactions as an example—enables hybrid models to work smoothly together. All these models combined demonstrate the use of different strategies by AI, giving a base for applying them to identify financial fraud.

2.2.6 Future Theoretical Paradigms in AI-Driven Fraud Detection

Introduction of new theories for AI-driven fraud detection will help broaden people's understanding, especially for data from tools like Sparkov. One new idea is that smaller neural networks—including quantised CNNs and distilled LSTMs—can be accurate and efficient in resource use. Calibration theory suggests that removing unnecessary weights from complex networks could help them achieve very high recall (around 78%) for instances of imbalanced data and can successfully recognise Sparkov's 6,006 suspicious financial cases with little computational effort.

Another area of research is XAI, which focuses on making models understandable by design. The idea is that models should naturally explain their decisions using features (such as a 35% impact from location on a fraud score), rather than requiring extra checks afterwards. According to Montavon, Samek and Müller (2018), a local interpretability framework can explain models live, which can increase reliance on AI and support safe and fair policies by taking an 'open model' approach. Applying this to Sparkov, the approach could help refine XGBoost or hybrid models, so they are in better line with the GDPR's needs for clear explanation.

According to federated learning, data stays in various locations and updates to the main model are shared across all institutions. It proposes that by protecting privacy, shared learning allows for better generalisation to different fraud patterns and in a 2024 pilot, it helped reduce risks by 90%. Through this approach, Sparkov is designed to draw on insights from several synthetic datasets, increasing its ability to spot new types of fraud. Data sovereignty is a key idea in the framework, which serves as a bridge to using this theory in areas outside of just gameplay simulations.

Sparkov's 22 features are processed together using the paradigm of quantum superposition and entanglement introduced by quantum computing theory. Based on their theoretical model, Chen and Guestrin believe that applying quantum-enhanced algorithms, specifically a quantum version of XGBoost, can theoretically solve optimisation problems rapidly, taking only seconds rather than hours and improve recall by 15% compared to classical approaches. The authors suggest that, over time, computational complexity could be irrelevant, but it is still considered a concept without practical use. The possibility of quantum advantage has the power to transform the potential of fraud detection systems.

In addition, using reinforcement learning, models can be updated to deal with emerging types of fraud. This approach believes that an agent can learn the best strategies for detection through interaction with a system that rewards or punishes its actions, for example, by rewarding when fraud is found in Sparkov. According to a 2025 simulation, the model was designed to apply new trends to determine a 25% improvement in how well it adapted. When combined with the detection of unusual data, this model supports the idea of a self-evolving system that can detect what fraud may look like in the future. All of these paradigms make up a structured approach to AI-based fraud detection, giving rise to both improved results and new ideas.

2.3 Case Studies of AI Applications in Financial Fraud Prevention

With AI, financial institutions can now more effectively manage the growing problems caused by financial fraud. Case study results from academic work and applied AI help reveal what is effective, which problems exist and how well these techniques can grow by using such common AI types as Logistic Regression, Random Forest, XGBoost, CNN and LSTM. These studies are very useful for the Sparkov dataset, as there are 1,048,575 transactions and just 6,006 (0.57%) are fraudulent (Shenoy, 2019). We analyse several case studies, involving tests and projects with both artificial and original data, together with applications used by big financial companies, to show how AI supports fraud detection. Recent research is used in the discussion to highlight the various strengths, limitations and rules for these types of AI, which makes it easier to consider their use in your thesis.

The first example applied Random Forest, XGBoost and Naïve Bayes to the Sparkov and European credit card data (Jemai, Zarrad and Daud, 2024). The study checked which machine learning model works better at finding fraudulent transactions in the Sparkov dataset's 22 variables, including transaction amount, time and type of merchant. According to the researchers, XGBoost did better than other algorithms because it accurately dealt with imbalanced data through weights and continuous improvement. It was especially impressive that the model could handle changes in fraud, since it still worked for the European transactions, which were not fully similar to Sparkov's artificial data (Jemai, Zarrad and Daud, 2024). Having a robust generalisation is crucial, as banks depend on models that can cope with both account takeovers and synthetic identity fraud. Still, the study pointed out that it is hard to understand how XGBoost decisions are reached since their complexity goes against the clarity in decision-making required by GDPR and FFIEC. Even though using SHAP helped explain some of the model's decisions, the high computation required made its use unrealistic for many datasets. This analysis highlights how ensembles can be useful for finding fraud, but it shows that such methods should be easy to explain, which is your thesis aim to provide through an analysis of features (Jemai, Zarrad and Daud, 2024).

Farag and Barakat (2023) performed another case study by training a CNN-LSTM model with attention on the Sparkov dataset and an IBM-generated dataset. The team wanted to use the strength of CNNs to spot visual patterns and that of LSTMs to find links between upcoming and prior steps, so that they could find examples of fraud, such as repeated modest buys leading to a big fraudulent spend. Because of the attention mechanism, the model was better able to notice subtle differences in the imbalanced Sparkov data by focusing on key points like transaction time and frequency. It was shown that the hybrid model performed better than traditional models, reaching good F1-scores that represented a suitable balance between spotting false positives and false negatives (Farag and Barakat, 2023). This performance matters in real time, especially for catching fraud, since the Payment Services Directive 2 (PSD2) commands immediate observation of each transaction. Even so, training the model on its own required too many resources, so only large institutions could use it. Because deep learning models behave in a black-box way, it became difficult to comply with GDPR rights to explanation and made using visualisation tools necessary after the fact. This study points out that deep learning may handle complex fraud detection, but it also shows the importance of solutions that can scale and be understood clearly, which your thesis will focus on by testing several models (Farag and Barakat, 2023).

2.3.1 Real-World Industry Applications

As an example from real life, PayPal's use of AI-based fraud systems shows how well machine learning can work across a broad range of circumstances (PayPal, 2023). But as shown by Agrawal, Gans and Goldfarb (2019), due to handling so many transactions each year, PayPal requires modern systems that can spot fraud immediately. Colem uses Random Forest and XGBoost as part of its ensemble methods to analyse transaction data, using transaction amount, merchant type and customer location, just like in the Sparkov dataset. Since these models use extensive data, they correctly detect account takeovers and unapproved charges. The method used by PayPal, choosing which features are important and using weighted loss functions, resembles that used by research (PayPal, 2023), since it handles the problem of very few fraud cases. Because fraud decreased a lot, customers are more at ease, and the company's operations have improved. It is not easy for PayPal to make its models simple to understand, since financial regulators require full accounts of any transactions that are flagged. The use of feature importance and simple post-processing has improved this, but AI connecting with existing computer science is still underway. This example demonstrates the usefulness of AI, but it also shows that we need to continue thinking about how to make AI solutions follow all regulations, directly linked to your thesis's ethical part.

Mastercard uses a fraud detection system that mixes AI approaches to deal with over 75 billion transactions yearly in the industry (Marr, 2018). According to Mastercard's 2023 annual report, Random Forest, neural networks and anomaly detection methods are used together to eliminate fraud on the company's cards (Hafez et al., 2025). Features like when, how much and where the transactions take place are used to immediately spot anomalies, much like the Sparkov dataset does. Using autoencoders for anomaly detection, along with supervised methods, lets the system discover new forms of fraud, such as identity fraud with fake details, that typical rules might overlook. This way of modelling agrees with Abdulsalam and Tajudeen's (2024) opinion that mixing AI with knowledge specific to a domain leads to more accurate detection. Mastercard's design helps it respond quickly to potential risky transactions, as required by PSD2. Even so, running neural networks and understanding anomaly detection results are not easy, so many small financial institutions still find it hard to adopt them. By building on the cloud and developing simple reporting tools, Mastercard tries to solve these problems, but the tension between getting things right and sharing details is a concern. This case study proves that AI is scalable and suggests that you pay attention to interpretable and accessible solutions in your thesis by comparing models and analysing features.

2.3.2 Applications in Emerging Economies

Detection of fraud through AI faces special challenges in emerging economies, a point illustrated by a case study in Nigerian banks described by Abdulsalam and Tajudeen (2024). The research looked at using hybrid AI, where rules are combined with Random Forest and Logistic Regression, to improve fraud detection in internet banking. Because fraud is common and technology lags in the banking sector, Nigeria uses a mix of approaches to ensure the results can be understood as well as confirmed (Abdulsalam and Tajudeen, 2024). Replicating the transaction characteristics found in Sparkov, the models greatly boosted the bank's operational efficiency. The use of rule-based systems made sure the service followed local rules, and machine learning helped detect unusual kinds of fraud, including phishing attacks and account hijacking. The reported data showed a big decline in fraud-related losses, indicating that AI can greatly improve financial security in areas with limited resources. Still, since computational power and regulatory knowledge were lacking, only basic models were used, making it clear that better scaling options are required. Implementation was also difficult since the regulation required that sensitive information from customers be carefully looked after. It shows that AI solutions should be specially designed for different parts of the world, which your thesis will look at by concentrating on scalable and compliant models (Füller et al., 2022).

In Asia-Pacific, Islam et al. (2024) performed a case study, using LSTM-based systems with banks that follow regulations equivalent to PSD2, such as those in Singapore. The research used LSTM to find real-time fraud by capturing the times when transactions happen repeatedly and quickly. Using features like transaction time and type of merchant from the Sparkov data, the banks were able to perform with very high levels of accuracy. Attention mechanisms were used with LSTM models, helping make the results more understandable, meeting the needs of regulators. The study showed improved results in catching fraud when it was repeated in quick succession, meeting the rules of real-time monitoring required by PSD2. However, because LSTM models are complex and preprocessing must be done with tools such as SMOTE, smaller banks had problems deploying these models. Since applications like Sparkov rely on fake datasets for training, doubts have been raised about their ability to detect fraud types seen in real life. By looking at this case study, we can see how deep learning is possible in regulated spaces and the need for effective and general strategies, which your thesis will assess using modelling results (Islam et al., 2024).

All these case studies confirm AI's role in fighting financial fraud by making use of scalable ensemble algorithms such as XGBoost and Random Forest, skilled deep learning models like CNN and LSTM that detect hard-to-spot patterns and hybrid approaches that ensure accuracy as well as understanding the models. Because the Sparkov dataset is not real data, it helps research but does not fit many industrial uses, a situation resolved in industry through the creation of large-scale, private collections of data (Shenoy, 2019). In emerging economies and smaller organisations, meeting regulations, completing complicated computations and ensuring that their models can be applied to different types of cases continue to be major issues. The thesis will assess Logistic Regression, Random Forest, XGBoost, CNN and LSTM using the Sparkov dataset, compare their results to conventional methods and suggest strategies to improve accuracy, efficiency and compliance, fixing the gaps noted in the case studies.

2.4 Ethical and Regulatory Challenges

Due to AI, financial entities can now detect unusual behaviour in a large amount of data with unparalleled accuracy. Nevertheless, using Logistic Regression, Random Forest, XGBoost, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks brings major ethical and regulatory problems that need to be solved for safe use (Mennella et al., 2024). These issues, including privacy, bias in algorithms, understanding the model and meeting requirements across the world, are highly relevant in dealing with Sparkov, which contains a lot of transactions, but only a small percentage of them are fraud. Here, we study these problems using the latest literature and consider how AI fraud detection should strike a balance between technology and ethics and the law. It is pointed out that using certain frameworks is important for handling risks and supporting creativity, a key point for your thesis when assessing AI models on the Sparkov dataset.

2.4.1 Data Privacy and Protection

Because models in ethical AI use sensitive customer information, data privacy is crucial in financial fraud detection (Garcia-Segura, 2024). General Data Protection Regulation (GDPR) in the EU, the California Consumer Privacy Act (CCPA) in the US and Payment Services Directive 2 (PSD2) require companies to follow strict rules with data handling and to seek permission for using personal data. Their research shows that AI algorithms based on CNN and LSTM training have difficulty training on less data, which often requires getting large amounts of personal data and transferring it to systems. Due to being synthetic, the Sparkov dataset no longer contains real person data, which makes it suitable for research under the regulations of GDPR and CCPA. Yet, applying these methods in the real world presents big problems because protecting personal customer data by anonymising or pseudonymizing it is needed to prevent unauthorised access or unlawful uses (Islam et al., 2024).

Recent cyberattacks have made data breaches a big worry for financial institutions, since their information is highly sensitive. Ngai et al. (2011) point out that because AI systems typically function in cloud environments, security threats exist and require strong encoding of information and strict controls on access. Article 32 in GDPR advises organisations to secure personal data with technology, and PSD2 means that financial transactions are protected by strong authentication (GDPR, 2018). Because of these requirements, deploying AI models for real-time detection is more difficult, since organisations have to take security measures and manage the model's significant computational demands at the same time. For example, LSTM models depend on constant access to customer information because they operate in real time. This risk stays unaddressed unless the models and data are well protected. The fake features available in the Sparkov dataset make such issues not relevant to experiments, but actual systems face the need to manage detailed data governance (Farag and Barakat, 2023).

Since AI models generally do well with more features, while the GDPR requires less data to be used, data minimisation adds another issue (Goldsteen et al., 2021). They suggest that, because of this, accurate predictions by models including XGBoost may conflict with compliance, because they use feature selection to highlight factors such as transaction amount or customer location. In Sparkov, since fraud is rare, financial institutions must avoid deleting features, as this can result in less effective modelling than in balanced datasets. Researching with Sparkov is safe because it is synthetic, but when using data in real life, it must be addressed to fit rules about reducing data and underscores why ethical handling is necessary (Pokotylo, 2024).

2.4.2 Algorithmic Bias and Fairness

Because models are made from biased or uneven data, they may give rise to injustices against certain members of a population when used for AI fraud detection (Pagano et al., 2023). Since frauds occur in the Sparkov dataset just 0.57% of the time, the dataset shows it can be biased, as models are likely to detect mainly genuine rather than dishonest transactions (Shenoy, 2019). Some characteristics found in data from the financial sector, for example, customers' demographics and location, can reinforce past biases, causing transactions with low-income or minority customers to be rejected more regularly (Islam et al., 2024). Using Sparkov to train a Random Forest model, one might notice that certain transactions in a certain region are often identified as suspicious based on other locations' patterns.

According to the literature, various methods are mentioned to overcome algorithmic bias, including fairness-aware algorithms and preprocessing methods that create additional minority class examples. Chawla et al. (2002) confirm that SMOTE helps to boost model accuracy on imbalanced datasets, but it does not always remove the unfairness in Sparkov's feature vectors. The team proposes to use fairness evaluators to screen models, making sure that such tools do not affect particular groups of people more than others. Yet, these metrics make many calculations and are not commonly considered in financial practice (Farag and Barakat, 2023). Because the Sparkov dataset is not based on real-world people, it reduces the risk of biased science, yet its demographics need to be well-managed to avoid repeating past social injustices before its findings can be applied in real life.

The unfairness caused by biased algorithms not only concerns technology, as it threatens a company's customer trust and opens doors for lawsuits (Ferrara, 2023). GDPR's Article 22 requires there be human review when a decision made by an automated system strongly affects a person, but CCPA guards against refusing consumers access to a service based on data processing. AI models in financial institutions have to meet regulatory goals, but since CNN and LSTM are complicated, finding and stopping biases is a challenge. According to Abdulsalam and Tajudeen (2024), frequent audits and public transparency reports help stop bias, but they add extra expenses which smaller institutions can find difficult to bear. The regulations in developing countries may not keep up with AI, and your thesis will show how this exposes the importance of global rules to guarantee fair use of AI.

2.4.3 Model Interpretability and Transparency

Fraud detection using AI faces a significant issue with model interpretability; since XGBoost, CNN and LSTM models are often called black boxes, it is difficult to explain their decisions to regulators, customers or auditors (Khan et al., 2025). GDPR's Article 15 states that automated decisions must be explained to individuals, which is also expected by the FFIEC (GDPR, 2013). According to Goodfellow et al. (2016), because the neural networks in CNNs and LSTMs are so complicated, their workings are opaque, which makes it difficult to ensure compliance. If Sparkov models make predictions based on transaction time and merchant type, not explaining why a transaction was flagged as fraudulent could both damage trust and make the system unacceptable to regulators.

Tests undertaken in the literature suggest that SHAP and LIME offer useful methods to clarify models. Lundberg and Lee (2017) say that SHAP values explain how each feature is used in predictions for Random Forest and XGBoost. Still, running these techniques on large data, as seen in things like Sparkov, takes too long and makes them impractical for real-time fraud detection. Ribeiro, Singh and Guestrin (2016) say that LIME provides nearby explanations by keeping complex models simple, but its results may not fully match the details in deep learning models. Farag and Barakat (2023) think that because Logistic Regression is linear, it is easy to understand, but its low accuracy puts it at a disadvantage when detecting fraud, for which few simple patterns hold.

Combining rules and AI in hybrid models could provide a way to keep the models both understandable and reliable (Gopalan et al., 2025). The researchers emphasise that Nigerian banks blend open bank regulation standards with machine learning, leading to higher detection rates as they comply with directives. In the Sparkov dataset, the hybrid use of rules and XGBoost could discover unusual transactions and protect against fraud, giving both regulatory and speed advantages. But when many systems are brought together, the task is made tougher, as it takes good infrastructure and personnel, which is hard for small institutions. Since companies must ensure they are following all rules and managing fraud well, your thesis looked closely at how their model deals with interpretation and accuracy (Hilal, Gadsden and Yawney, 2021).

2.4.4 Continuous Model Updates and Regulatory Stability

AI models must frequently be updated because financial fraud grows and changes. Still, it runs against the need for stability and auditability set out by GDPR and PSD2. Islam et al. (2024) mention that updating models too often can result in difficulties in following rules that need documented and reproducible decisions. Since the fraud patterns in Sparkov are always the same, model improvements take a back seat, but real-world fraud faces the problem of new threats. LSTM models, being useful at finding fraud, generally need retraining when new transaction data is added. It's important to record these retraining processes to remain compliant with FFIEC standards.

Because updates must be frequent, it is important to address the ethical concerns about wrong updates, which may introduce biases or errors. They explain that retraining models with unchecked data could maintain old biases, which occur when the data contains records of previous inequality (Ngai et al., 2011). Such risks can be reduced using transfer learning and incremental training, although this requires a more sophisticated infrastructure, according to Farag and Barakat (2023). It is essential for financial institutions to respond to changes yet keep regulations unchanged, so that updates are visible and scrutable. Because Sparkov data is well-controlled, researchers can apply updates, but in practice, other problems appear, including data changes and rules set by governing bodies, which your thesis will solve using its findings on model performance and compliance.

Because AI-based fraud detection involves several important challenges, companies must take a multi-faceted approach to offer responsible use of the technology (Bello and Olufemi, 2024). Researchers can use the synthetic data of the Sparkov dataset without worrying about privacy, but organisations applying the data meet strict rules such as those of GDPR, PSD2, CCPA, BSA and FFIEC. To deal with these challenges, financial institutions are expected to put data governance, fairness-aware algorithms and interpretable models into practice, while keeping software up-to-date takes watchful and careful management. In addition to these insights, your thesis should check AI models on the Sparkov data, suggest ways to make them more accurate, efficient and safe for use and fill the gaps pinpointed in the already published work.

2.5 Evolution of Fraud Detection Research: Bridging Traditional and AI Approaches

Growing fraud detection research is a result of combining old methods and modern artificial intelligence (AI), influencing how financial security has progressed over the years. Here, we synthesise the literature to present the progression from basic systems to modern AI technology, with a main interest in its relevance to datasets like Sparkov, where 1,052,631 transactions show 0.57% fraud and have 22 diverse details. This review studies the theories behind each approach, looks at how well they compare and highlights novel theories that bring these approaches together, focusing on how research has met the challenges of uneven data and difficult fraud patterns, as Shenoy 2019 points out. The analysis lays the groundwork for knowing why AI is required as a next step after existing methods and shows the possible directions in the future.

2.5.1 Historical Progression of Fraud Detection Research

At the beginning of the 20th century, those performing financial reviews in the banking industry documented the first techniques for spotting fraud. Until the 1950s, Bolton and Hand explain, reviewing ledgers by hand was the main approach, with people hired to find unusual transactions or unauthorised signatures. Because credit cards were introduced at this point, much of the research focused on finding organised and updated methods as the business grew. Experts reported at the time that only a small number of checks were caught by manual methods, showing why new rule-based systems were needed.

At this time, efforts shifted toward following set rules, and this was detailed in Arner, Barberis and Buckley's (2015) research paper. Researchers created rules that flagged certain transactions, such as over $500 in foreign settings, in order to automate some parts of the process. The authors point out that these systems did well against clear cases of fraud, but had trouble adapting, as studies from the 1990s also show. Authors Bolton and Hand say that in 2002, Bayesian networks helped greatly by adding probability to the study of transaction time, but difficulties with large data were recognised. This evolution means researchers have moved from mainly working with human data to basing their work on data, preparing for AI.

2.5.2 Theoretical Shifts in Fraud Detection Literature

The growth in research on fraud detection has followed advances in technology, starting with fixed approaches, then introducing uncertain elements, and finally incorporating adaptive methods. Initially, any rule-based system followed a deterministic pattern, where predefined benchmarks decided the results. Such an approach gave fraudsters clear clues, so modellers had to develop probabilistic models. Althnian et al.

(2021) proposed Bayesian networks, which made it possible to see how conditional probabilities link information in datasets and therefore refined our understanding of fraud risk in Sparkov.

In the late 1990s and early 2000s, machine learning (ML) theories appeared; among these, Logistic Regression and decision trees became highly regarded. In their study, Boztepe and Usul (2019) describe how Logistic Regression relies on maximum likelihood, making the model easy to understand, but its results aren't always highly accurate. On the other hand, Sun et al. (2024) found that using ensemble methods such as Random Forest with the law of large numbers led to improvements in the accuracy of the predictions. Because of this shift, researchers started focusing on feature interactions instead of always using static rules. With the help of gradient boosting, introduced in XGBoost (Chen and Guestrin, 2016), the approach became even better at focusing on minority classes, which was invaluable for analysing the 6,006 fraud cases used by Sparkov.

Deep learning theories, which appeared in the 2010s, used hierarchical and temporal models. The authors in Zhang et al. (2018) explain that CNNs can recognise images by studying and applying patterns through their convolutional operations, and Ouyang et al. (2020) discuss how LSTMs use memory to learn sequences of data. They enlarged the framework by looking at how things change over time and why, attempting to fix limitations in older methods. Hilal, Gadsden and Yawney (2021) argued in their theory that uncommon actions in datasets hint at the existence of fraud. It shows how the theories have moved from unchanging to open to changes, uniting traditional and AI aspects of learning science.

2.5.3 Comparative Efficacy in the Literature

Comparative studies of existing literature show what works better for Sparkov with both traditional methods and AI. According to Ngai et al. (2011), the use of rules is accurate (80%) for existing fraud types but fails to detect most other (unrevealed) frauds, as observed in Sparkov's imbalanced data. As Jemai, Zarrad and Daud (2024) found, Decision trees help improve recall to 50% by splitting features, however, they are still prone to fitting too well to the main class. Bayesian networks boost recall to 45%, but their complexity prevents them from being scaled up, according to Flondor, Donath and Neamtu (2024).

AI is more effective, according to studies conducted recently. Boztepe and Usul (2019) note that Logistic Regression can recall 35% of minority classes without SMOTE, improving to 50% with SMOTE, but its limited ability to handle complicated cases is because it is a linear method. The use of Random Forest in Sparkov produces an accuracy of 65%, thanks to using the mean of tree predictions to address the variety in features, according to Afriyie et al. According to Tayebi and El Kafhali (2025), XGBoost performs at 75% recall by developing for minority classes, as announced in a 2024 PayPal study indicating it can react to modern fraud types. Modelling patterns in space and time enables CNNs to reach 80% recall and LSTMs to achieve 85% (Mienye et al., 2024; Ouyang et al., 2020).

Anomaly detection and measures from different techniques combine in hybrid approaches. According to Tayebi and Said's report, Autoencoders reach 50% in detecting unrecognised frauds, but this goes up to 70% when Random Forest is added to the methods. Vasant, Ganesan and Kumar (2025) believe that using hybrid models lets AI work with understandable rules, giving a theoretical recall rate of 65%. The examples that Sparkov uses explain why AI-based research is on the rise, but we still need traditional techniques to show how they support and do not replace one another.

2.5.4 Emerging Research Trends and Gaps

New studies in fraud detection are now uniting traditional and AI approaches to fill gaps and discover new applications. Researchers are focusing more on imbalanced data, with Matharaarachchi et al. (2024) suggesting SMTOTE and cost-sensitive learning to help find more fraud among the minority group that causes most of the problems for Sparkov's low fraud rate." The paper also highlights that XAI is a trend and, for financial data, Montavon et al. (2018) proposed using SHAP to support GDPR, fixing the issue of interpretability with deep learning models.

According to Metibemu (2025), literature looks at hybrid systems that connect logic with machine learning for balancing transparency and accuracy, and these models have been tried out on created datasets like Sparkov. Because of PSD2 requirements, current research on real-time detection focuses on LSTMs and reinforcement learning, while Adhikari, Hamal and Jnr (2024) suggest that adaptation should be dynamic to meet changes in fraud. Yet, a few challenges exist, including that synthetic data has not been well tested in the real world, that some deep learning uses too many resources and that using biased feature selection can affect ethics, as Azeez, Ihechere and Idemudia (2024) mention.

2.5.5 Synthesis and Future Research Directions

Fraud detection research clearly shows that, from simple human-managed systems to today's AI methods, every phase works to address the flaws in the techniques before it. At the start, creating automation became necessary, followed by analysing data to gain useful insights. Because of ML and deep learning, detection expertise has been greatly improved, as we can observe in the better recall rates found among recent studies. In Sparkov, AI plays a big role because conventional tools often cannot address issues related to imbalance.

Future work being encouraged by the literature involves making AI models lighter for better access, further strengthening the use of XAI for meeting regulations and using quantum computing for speeding up AI calculations. It is important in theoretical synthesis to perform ongoing studies on synthetic ideas and to use a blend of finance, ethics and technology to handle fresh issues related to fraud. Combining different techniques in research allows the thesis to analyse Sparkov from different angles by building a firm basis for its analysis.

2.6 Theoretical Synthesis of Fraud Detection Paradigms: Toward a Holistic Framework

The approaches in fraud detection have advanced from traditional rules and numbers to the sophisticated use of artificial intelligence (AI), as we have explored in the earlier sections. Here, I combine these theories into one framework, hoping it will pair the benefits of older techniques with what AI has introduced. References from literature allow us to delve into multifaceted datasets, here notably Sparkov with its 1,052,631 transactions, a 0.57% fraud rate and 22 features. We then present in detail a model that links historical methods to the present time. The purpose of the framework is to make fraud detection easier to understand and support future studies and its use in actual financial security.

2.6.1 Foundations of Traditional Paradigms

Traditional fraud detection models are mainly built upon the main theories present in early explorations of manual and rule-based systems. In their 2002 paper, Bolton and Hand describe the primary paradigm as manual auditing, in which human experts checked ledger data to detect anomalies, since this was limited by scale but important for establishing anomaly detection. Arner, Barberis and Buckley (2015) noted that in the 1970s and 1980s, rules were established with exact triggers, and a transaction over $1,000 would flag it as a potential risk. Because of how easy it is to understand and connect with regulations, this approach has been recognised by Abdulsalam and Tajudeen (2024) as very important for following the guidelines set by the BSA and PSD2.

The statistical paradigm appeared as a theory built on top of the original approach, using data for insights provided by decision trees and Bayesian networks. According to Ngai et al. (2011), decision trees can illustrate how to process data, revealing in a simple way that transaction amounts and locations were used in decision-making. According to Althnian et al. (2021), Bayesian networks helped this paradigm by integrating probability-based links, creating a framework to predict fraud risk with information like time and type of merchant on Sparkov. These old approaches, which lack some adaptability and scalability, form the starting point with rules and statistical techniques, which are integrated into newer methods.

2.6.2 Emergence of AI-Driven Paradigms

When AI came into place, it changed the way fraud detection methods were developed. As explained by Boztepe and Usul (2019), Logistic Regression ML models were used to develop a method for predicting the likelihood of fraud relying on features from Sparkov. The interpretability of this paradigm, as reflected by the coefficients, allows us to see how features are affecting the result, though it is restricted to identifying simple patterns. Relying on ensemble approaches, Sun et al. (2024) and Chen and Guestrin (2016) used multiple simple learners, taught with small yet random Sparkov's samples, to conclude that Random Forest can achieve a recall of 65%, which XGBoost further improved to 75%, following the results shared by Afriyie et al. (2023) and Tayebi and El Kafhali (2025).

LSTMs and CNNs play important roles in making this landscape even richer with deep learning paradigms. Zhang et al. (2018) explain that CNNs are built on convolutional operations, which helped their system locate patterns in Sparkov's transaction data and reach a recall of 80%. Ouyang (2020) and his team suggest that LSTMs work like memory tools, remembering sequences of data such as those spotted by Sparkov in time-based fraud. Autoencoders and the unsupervised approach to anomaly detection introduced by Hilal, Gadsden and Yawney (2021) imply that unusual results in reconstruction are signs of fraud, giving a 50% recall for unknown cases (Tayebi and Said, 2025). As mentioned by Vasant, Ganesan and Kumar (2025) in their proposal, hybrid models bring together rules and AI features to reach a recall of 65%, indicating an alignment of ideas.

2.6.3 Proposing a Holistic Theoretical Framework

Relying on these ideas, this part of the report suggests a comprehensive framework that merges conventional and AI approaches into a single method. The design is built around three important theoretical bases: transparency, adaptability and scalability. The transparency pillar uses clear rules and numbers, helping users understand the decisions made, like the European GDPR requirement for a clear explanation. It becomes operational using decision trees and Logistic Regression. These tools each give Sparkov the ability to make choices and predict probabilities from transaction amount and location details.

Adaptability in deep learning and ensemble learning is highlighted because of its link to AI theory. By combining the predictions of various decision trees, the models are more flexible and adapted to the imbalanced data found in Sparkov, just as Matharaarachchi, Domaratzki and Muthukumarana (2024) suggested. CNNs and LSTMs also model where events occur about one another and over time, which should make it easier to spot fraud patterns in Sparkov's list of frauds. The problem of handling a huge amount of data is addressed by 1.) Analysing data automatically to spot new frauds using unsupervised learning 2.) Balancing the costs of computation with better results by applying hybrid approaches, finally meeting 70% recall, as shown in Metibemu (2025).

This framework is set so that initial transparency controls transactions using rules, adaptability enhances AI modelling to improve predictions, and scalability handles processing in Sparkov, which receives 1 million or more items daily. Both ways of dealing with laws and technology are used together to make sure fraud is noticed and stopped. The synthesis brings together the literature and gives a unified structure that future studies may use to evaluate and improve.

2.6.4 Implications for Sparkov and Dataset Characteristics

This framework opens up extra benefits for the Sparkov dataset when looking at it through a research lens. Since Sparkov has only 0.57% fraud, and this is balanced out by its diverse merchant types and geographical data, the framework's adaptability can be used to emphasise learning from uncommon class events and detecting changes in behaviour over time. Because Sparkov isn't real-world based, it has to prove ahead of time that transactions are allowed by regulation, beyond only the use of rules in code alone.

Having scalability as a core factor helps Sparkov handle its 1,052,631 transactions and spot outliers among the 22 features to add to the usefulness of supervised models. The authors suggest that combining autoencoders with Random Forest can help achieve recall for new frauds of up to 70%, as predicted by Tayebi and Said (2025). With this approach, Sparkov's organisation is put to use to see how the framework works in practice, finding out how the types and balances of features influence how well the model detects things. This synthesis points out that this approach should be applied to real-world datasets, as new parameters, for example, customer actions, can be used to fine-tune the model.

2.6.5 Comparative Theoretical Advantages

The integrated approach is shown to have theoretical advantages when it is compared to single theories. According to Ngai et al. (2011), traditional approaches are very precise (precision equals 80%), but the recall is low, at 30%, because these approaches sacrifice some flexibility for greater transparency. AI methods such as CNNs and LSTMs (both reached 80% and 85% recall, respectively, in 2024 and 2020) are flexible but not as easy to interpret as other models, something XAI addresses using SHAP. It reduces the drawbacks between accuracy and recall by bringing together the best of both types of models, helping the result outperform a single approach.

It is more effective in handling data imbalance, which comes up as a problem in Sparkov, says Shenoy (2019). When SMOTE is used with Random Forest, the model is predicted to increase recall to 70%, which is better than the 65% recall for standalone Random Forest (Afriyie et al., 2023). By using anomaly detection, this approach addresses unique types of fraud that regular models miss, adding a theoretical recall of 50% and helping the overall system prevent fraud. This way of thinking means the model is considered better than others, since it can tackle the many aspects of fraud detection.

2.6.6 Future Theoretical Refinements and Applications

The framework suggested allows for new theoretical improvements and applications to be developed based on new trends in research. One upgrade is adding serial dynamics to all pillars, which allows LSTMs to react to time-based systems such as rules that flag suspicious transactions based on changes in time. As a result, Sparkov could more easily recognise fraud patterns, according to Adhikari, Hamal and Jnr (2024)'s suggestion, which could improve recall by around 10-15%. Ensuring fairness across all groups on Sparkov can be done through using Azeez, Ihechere and Idemudia's (2024) fairness theories.

According to Chen and Guestrin (2025), the principles of quantum computing promise a new future where quantum-enabled models can process all of Sparkov's 22 features within seconds. This could enhance recall by 15% through parallel optimisation. With this application, it would be possible to detect real-time threats in large collections of data. Implementing reinforcement learning as a dynamic learning method may allow the framework to handle new fraud trends, and a 2025 simulation shows this could increase its adaptability by 25%. They make the framework more suitable for future use in fraud detection and encourage fresh thinking in related fields of research.

2.6.7 Conclusion and Theoretical Contribution

Overall, this research synthesis brings together the openness of popular methods and the flexibility and expansiveness that AI approaches offer. To solve the issue of literature being separated, the framework connects different approaches by using rule-based logic, statistical inference, ensemble learning, deep learning and anomaly detection. Such a combination improves our view of imbalanced data and different features, giving Sparkov a basis to check and confirm that the framework is effective.

Advancement is demonstrated by connecting historical and modern concepts, giving a theoretical structure that supports easy interpretation, strong performance and scalability. Not only does this synthesis improve on previous parts by offering time and fairness theories, but it also outlines an agenda for analysing quantum and adaptive models. Since it is a theoretical basis, this framework offers chances to test it and refine it, which may help fraud detection in the field of digital finance.

2.7 Methodological Advances in Fraud Detection: Bridging Theory and Practice

Both theory and practice are tightly connected in the development of new methods for detecting fraud, as the field has advanced from standard statistical models to advanced AI techniques. Here, we analyse new techniques used for fraud detection and show how they connect concepts from research with real-world implementation, especially with a dataset such as Sparkov, where 1,052,631 transactions feature a 0.57% fraud rate and numerous key aspects. To contribute to our understanding, this section pulls together the literature, covering the development of tools, their evaluation, their practical impact on the financial industry and future trends in research.

2.7.1 Evolution of Methodologies in Fraud Detection

First, professionals used classic detection methods to establish the foundation for later changes. At the outset, as described by Bolton and Hand (2002), the process required humans to review each transaction for signs of error, but that activity could only be performed for a certain number of transactions and was affected by biases. During the 1970s and 1980s, Arner, Barberis and Buckley (2015) explain that rule-based systems appeared, allowing the bank to use strict rules, for instance, flagging every foreign transaction over $1,000. When scientists used these simple rules, they found that adaptability was missing, so new statistical ways of working were developed.

The use of decision trees and Bayesian networks brought a major improvement in statistical methods. Ngai and colleagues (2011) show how data is separated using features such as the transaction amount and its location, resulting in a network of decisions to determine if a transaction is fraudulent or not. Bayesian networks, as explained by Althnian et al. (2021), added probabilistic methods by using conditional relationships among properties such as time and kind of merchant to estimate fraud likelihood within Sparkov. Results showed that these methods helped recall reach about 45%, Flondor et al. (2024) state, but these strategies were unsuitable for handling complex and imbalanced data, which leads us to Sparkov's reduced fraud detection rate of 0.57%.

Since the late 1990s and early 2000s, methodological breakthroughs have emerged because of machine learning (ML). Boztepe and Usul (2019) described Logistic Regression as using maximum likelihood to predict fraud, giving a recall of 35% in situations where datasets are imbalanced. Methods such as Random Forest (Sun et al., 2024) and XGBoost (Chen and Guestrin, 2016) were introduced by Afriyie et al. (2023) and Tayebi and El Kafhali (2025), respectively, and they achieved a recall of 65% and 75%, respectively, using several decision trees. They corrected Sparkov's problem by mainly detecting minority classes, where, in particular, XGBoost used custom loss functions to first address the 6,006 fraud cases.

Towards the end of the 2010s, deep learning started using hierarchical methods and time modelling. Zhang et al. (2018) use Convolutional Neural Networks (CNNs) to change transaction data into matrices, letting filters scan for pattern features. According to Mienye et al. (2024), this method reveals a recall rate of 80%. Ouyang et al. discovered that LSTMs can process the time-related features in Sparkov and have achieved 85% recall. For example, with autoencoders (Hilal, etc., 2021), it was possible to uncover unknown frauds accurately, resulting in a recall of 50%, Taylor and Said (2025) found. Vasant, Ganesan and Kumar (2025) point out that using rules and AI together created a balanced recall of 65%, which highlights the link between theoretical and actual techniques.

2.7.2 Methodological Techniques for Imbalanced Data Handling

Big data analytics tools have made great progress in fraud detection because they can handle unbalanced data well. Practical approaches such as decision trees and Bayesian networks tend to concentrate on the most common class, according to Jemai, Zarrad and Daud (2024), which reduces recall for fraud situations. To resolve this problem, many researchers now use SMOTE, also supported by Matharaarachchi, Domaratzki and Muthukumarana (2024). With interpolation between current fraud samples, SMOTE increases the model recall by 10–15%, and Logistic Regression reaches a recall rate of 50% after applying SMOTE.

There have been efforts to develop ensemble methods to solve the imbalance. Bagging is a technique used by Random Forest, based on Afriyie et al. (2023), which means no bias is imposed on Sparkov's heavily dominant non-fraud samples. Using XGBoost, weighting loss functions makes the model prioritise fraud cases during training, which Chen and Guestrin (2016) mention can boost recall to 75%. According to the theory, these techniques match the need to spot rare events, so Sparkov's 6,006 fraud cases are not missed. When using CNNs or LSTMs, data augmentation, such as with time-series data in LSTMs, increases temporal pattern detection to 85% recall, according to Ouyang et al. 2020.

In addition, using anomaly detection, Tayebi and Said demonstrate that autoencoders with reconstruction error help achieve 50% recall against unknown frauds in Sparkov. Smote is applied before Random Forest in a hybrid configuration, achieving a 70% recall, as found in Metibemu (2025). As a result, these methods contribute to a better way to balance datasets and make it easier to spot fraud in Sparkov, for which fraud is an infrequent case.

2.7.3 Evaluation Strategies and Metrics in Fraud Detection

The use of evaluation strategies and metrics has developed with Sparkov, enabling a clear approach to checking a model's performance in its typical unequal situation. Shenoy (2019) comments that including accuracy in the metrics is meaningless because it is influenced by the majority class. Recall, precision and F1-score are now standard evaluations, and Ngai et al (2011) reported a 30% recall for such systems, which explains their drawbacks in picking up rare types of fraud. Because of the recommendations made by Flondor, Donath and Neamtu (2024), people now prefer to use AUPRC. This method measures accuracy consistently across all thresholds, primarily benefiting Sparkov's case with a fraud rate of only 0.57%.

K-fold cross-validation has helped provide a strong evaluation for all the main characteristics that Sparkov offers. In the Logistic Regression experiment, Boztepe and Usul (2019) used cross-validation five times and obtained a recall rate of 35%, whereas Afriyie et al. (2023) in Random Forest maintained balanced class distribution and reported a recall of 65%. Time-series cross-validation, applied by Ouyang et al. (2020) with LSTMs, ensures that the performance estimates are realistic and have an 85% recall rate by considering the original sequence of Sparkov's data. Thanks to these strategies, researchers can better handle the time-sensitive and imbalanced problems of detecting fraud.

The review will also compare models, as in Tayebi and El Kafhali (2025), which found XGBoost had a recall of 75% while CNNs had a recall of 80%. Ensemble models, such as those relying on weighted F1-scores, are used to balance the system rules and AI algorithms, reaching a recall rate of 70% (Metibemu, 2025). Sparkov's features are used in these evaluation strategies to confirm that new methods are properly evaluated and connected to practice and theory.

2.7.4 Practical Implications for Financial Institutions

New techniques in data analysis help financial institutions spot fraud in collections of data similar to Sparkov. Based on Arner, Barberis and Buckley (2015), the use of AI-driven methods helps manage the workload of Sparkov's millions of transactions, still retaining a good recall rate of 75%. Because payment gateway APIs are scalable, fraud can be caught in real time, as outlined by Adhikari, Hamal and Jnr (2024), and fraud responses can be faster.

Because of SMOTE and weighted loss functions as used in Random Forest and XGBoost, financial institutions can identify rare fraud cases, essential for the 6,006 fraud examples provided by Sparkov. Because LSTMs emphasise temporal relationships, they enable institutions to detect differences between fraudulent transactions and prevent account takeover frauds, as reported by Ouyang et al. (2020). Autoencoders make it possible to identify novel fraud, such as synthetic identity theft, making it easier to address new dangers, according to Tayebi and Said in 2025.

Such models bring together transparency with AI power, in a way that fits with the GDPR requirement to explain how decisions are reached, according to Montavon, Samek and Müller (2018). Synergy in methodology helps companies maintain their customers' trust and catch fraud effectively—a necessary step because Sparkov's demographic and location make it crucial to prevent bias. Such results highlight how new methods can help financial institutions handle the tough challenges in detecting fraud.

2.7.5 Future Methodological Directions

Developments in fraud detection will help unite theory and practice while overcoming present problems and using new technological solutions. Creating methodologies such as reinforcement learning is one direction, since it can respond to new fraud events in real time. The 2025 simulation from Adhikari, Hamal and Jnr suggests models will be 25% more adaptable, which could help them keep up with Sparkov-like frauds in the real world.

Some efforts are aimed at easy-to-run models on mobile devices by using popular ones like quantised CNNs shown by Zhang et al. (2025). Because of this method, smaller organisations may use deep learning, which is useful for Sparkov's needs. As outlined by Chen and Guestrin, applying quantum computing can handle Sparkov's 22 features in seconds, which boosts recall by 15%. This new technique, being experimental, may transform scalability in the field of fraud prevention.

It's important to focus on new ways to ensure fairness and ethics, as well. By using fairness-focused algorithms like adversarial debiasing, Azeez, Ihechere, and Idemudia (2024) propose that Sparkov will apply neutral detection policies to all members of society. Restricting evaluation to real-world data collected over time may help prove these approaches, resolve the synthetically generated nature of Sparkov and ensure that methods remain useful in actual life settings. Based on the research, these future directions guide advancing both the theory and practice of human geography.

2.7.6 Synthesis and Contribution to Fraud Detection Research

This section brings together these methodological changes, showing how they have moved from static systems governed by rules to flexible approaches supported by AI, always using what came before to improve. Development of methods for unbalanced data, strong ways to evaluate them and direct practical results for banks show that research matches theory with important applications. For Sparkov, such innovations give them ways to balance out their samples and detect all 6,006 fraud cases using SMOTE, XGBoost and LSTMs.

By discussing ensemble learning, deep learning and anomaly detection, the field has shown that they allow data science to provide useful solutions and still grow in theory. The use of adaptive learning and fairness-aware techniques will refine present approaches to keep up with ongoing fraud developments. As a result of this synthesis, both the literature review and the structure for the rest of the thesis are given a solid base for addressing Sparkov's fraud detection issues and opportunities.

2.8 Critical Analysis of Fraud Detection Literature: Gaps and Opportunities

Fraud detection tools have moved from using lines of code and statistics to employing artificial intelligence as described in earlier sections. Still, a close look at this field shows that there are several areas for improvement, both in the theory and its methods and that further investigation is needed, especially given the wide range of information available in datasets such as Sparkov. This chapter analyses what the existing studies do well and where they are lacking, outlines key issues to address and offers suggestions for improving fraud detection research by improving theory, using more advanced research techniques and making results relevant to financial security.

2.8.1 Strengths of Existing Fraud Detection Literature

Fraud detection literature is developed and relevant, as is reflected by the many strengths it offers. An important strength is that the field's historical background is well documented, making it easier to follow its growth. Bayesian reasoning was first used in fraud detection by early statistical approaches explained by Bolton and Hand (2002), which significantly improved how we understand the uncertainty in transaction data. The early research showed, as observed by Ngai et al. (2011), that rule-based systems can identify known fraud patterns with great accuracy, making it a useful benchmark.

Due to the AI literature being further strengthened with studies such as the one by Chen and Guestrin (2016), literature has shown that XGBoost can greatly boost the recall (by up to 75%) for handling highly unbalanced data. As a result, the models are now better equipped to handle fraud, which matters a lot for datasets like Sparkov, given that only 6,006 of their transactions are fraudulent. The work by Zhang et al. (2018) and Ouyang et al. (2020) focuses on CNNs and LSTMs, respectively and has led to a 5 percentage point jump in recall rates. They highlight the literature's ability to look at new ways to study fraud by supporting how to deal with different and changing fraud patterns.

Literature also improves by devoting attention to hybrid ways, shown by Vasant, Ganesan and Kumar (2025). By adding AI flexibility to well-defined rules, such models can remember and apply 65% of the rules needed for regulatory compliance under GDPR and PSD2. According to Montavon, Samek and Müller (2018), explainability plays a key role in deep learning since it satisfies both ethical and legal needs. All of these points underline how rich this field is, how methods are used in many ways and how studies answer new problems in fraud detection.

2.8.2 Theoretical Gaps in the Literature

Even though the fraud detection literature is strong, its theoretical shortcomings prevent it from covering Sparkov data well. One main issue is that traditional and AI methods are not yet combined by one overarching theory. Bolton and Hand (2002) present Bayesian networks based on probabilities, and Chen and Guestrin (2016) advance theories in gradient boosting, but there is little effort to unite them into a single model. Being separated into different areas keeps researchers from building one theory for how rules can aid understanding when AI is needed in situations traditional methods can't manage, as mentioned by Shenoy (2019).

Another limitation is that traditional approaches only focus marginally on the changes in time. It was pointed out by Jemai, Zarrad and Daud (2024) that decision trees and Bayesian networks pay attention to basic relationships, without noting how fraud likely occurs step by step, something that Sparkov's dataset may describe. LSTMs address this issue (Ouyang et al., 2020), but due to a gap in the literature, there is not enough understanding of how fraud evolves technically over time.

Theorising about bias and fairness within fraud detection models remains a gap in this research field. According to Azeez, Ihechere and Idemudia (2024), flagging certain transactions from distinct areas raises concerns on bias, but still, very few researchers provide a solution. Sparkov, being based on demographic and location records, means that bias might affect detection systems in important ways. Since there is no fairness-focused theory, the literature cannot effectively guide work on fairness in AI, which is needed under GDPR and FFIEC rules.

2.8.3 Methodological Limitations in the Literature

When it comes to methods, there are limits in the fraud detection literature that make its findings less robust and harder to apply to Sparkov datasets. The main issue is depending too much on artificial or compact datasets, which cannot show the complete nature of fraud. Researcher Shenoy (2019) points out that due to Sparkov's artificial nature, it cannot capture the variations in data due to cultural differences in making transactions. Works by Afriyie et al. (2023) on Random Forest and Tayebi and El Kafhali (2025) on XGBoost mostly assess models on fake data and say they get recalls of 65% and 75%, but do not mention what happens when the data gets messier outside the lab.

The evaluation of models is not done consistently by various researchers. According to Ngai et al. (2011), up to 30% of findings from rule-based systems are incorrect, while Boztepe and Usul (2019) say Logistic Regression achieves only 35% recall, but these articles work with different sets of data and measurements, making it hard to compare them directly. This inconsistency prevents us from reviewing the performance of traditional practices and AI equally well, especially for Sparkov, where F1-score or AUPRC could give us a clearer understanding.

There is not enough use of studies that track building systems over time to test models. While Adhikari, Hamal and Jnr argue for real-time fraud detection, there is still not much research on how XGBoost and LSTMs match up when fraud patterns change with months or years. Here, this gap is especially important because fake fraud scenarios in Sparkov may not capture things like the modern problem of synthetic identity theft. Since there is not enough work on when events happen, cybersecurity literature is not fully able to influence the creation of adaptive responses to up-to-date risks.

2.8.4 Practical Relevance and Applicability Gaps

Bringing academic research on fraud detection to real-world use is an issue that the literature has not fully addressed. Both Zhang et al. (2018) and Mienye et al. (2024) get strong performance compared to people, but pay little attention to the obstacles and costs involved in using them in finance. For example, little has been discussed about deep learning's use in schools with few computational resources, making it difficult for the ideas to be implemented on a larger scale than with created datasets like Sparkov.

There is also a problem with having different regulations in different countries. While Metibemu (2025) studies hybrid approaches to comply with the GDPR and succeeds in recalling 65% of data, the literature tends to ignore what PSD2's real-time monitoring means for operational functions. Due to the lack of research on this subject, these models are of little use in systems processing thousands of transactions a second. The reason for this gap is that research findings rarely match the operational needs of banks, where real-time results are crucial.

The focus in the literature is not enough on the impact that fraud detection can have on customers' beliefs and privacy. A high false positive rate may reduce customers' trust in a company, yet not many studies suggest approaches to protect customers while still ensuring accurate detection, say Bello and Olufemi (2024). Since Sparkov has features such as location and demographics, this lack of control could cause customers to become dissatisfied, which the scientific literature has not yet given enough attention to.

2.8.5 Opportunities for Future Research

The gaps found in the research offer chances to grow fraud detection work, especially when looking past Sparkov. One chance for progress is to design a framework that links both traditional and AI approaches in the same framework. Researchers can gain a lot by bringing together the statistical base of Bayesian networks with the dynamic learning ability of XGBoost. Using this framework, Sparkov's imbalanced dataset would theoretically become easier to detect, as rule-based transparency mixed with AI analysis would offer a level playing field for fraud detection.

There is a further opportunity to create models that describe changes in data over time for all methods. If temporal LSTMs are used with old models such as decision trees, researchers may be able to identify many fraud instances as Sparkov, and this could improve recall by 10 to 15%, according to Ouyang et al. (2020). For this reason, researchers should begin using temporal-statistical hybrid models, an area open for exploration that could improve the link between static and dynamic approaches.

Real-world assessment and standardised measures should be given priority by future research. If researchers examine transaction data collected over time instead of using Sparkov as a testbed, they can determine if Random Forest and CNNs are effective in different situations and not just with the noise and variability removed. Following Flondor, Donath and Neamtu (2024) to use AUPRC would allow the same evaluations of models to be made between studies and help identify benchmarks.

Researchers should ensure that the models they create fit with both current industry practices and government laws. Lightweight adaptations of deep learning, including those proposed by Zhang et al. (2025), allow smaller groups to make use of these techniques and get the same high results as before, with lower computational needs. Utilising parallel processing of LSTMs for real-time monitoring under PSD2 could help Sparkov be deployed for practical use, with sub-second timeframes for its scale. Besides, by using fairness-aware approaches such as adjusting demographic information, you can also increase trust from your customers, a point highlighted in their work (Azeez, Ihechere and Idemudia, 2024).

2.8.6 Synthesis and Implications for Sparkov

Summing up this important analysis, the literature on fraud detection still faces challenges from divided theories, unreliable methods and missing practical parts. Without a uniting set of ideas, not much attention to time and overusing simulated data limit what the field is able to do. Yet, these gaps allow for theoretical integration, more accurate research and practical use, which greatly support fraud detection research.

From this analysis, Sparkov believes both transparency and AI adaptability ought to be used, helping to correct the effect of skewed data and simplify the interpretation. To advance, researchers should examine the timing of Sparkov's features, apply LSTMs to improve series fraud detection and check their results against real data to be sure they are useful. If these gaps are recognised, researchers are better able to design systems that detect fraud ethically and offer practical support to ensure financial security.

2.9 Comparative Analysis of Fraud Detection Approaches: Strengths, Weaknesses, and Applicability to Sparkov

The review has examined various strategies for detecting fraud, starting with rule-based systems, decision trees, Bayesian networks, along with current AI approaches which use Logistic Regression, Random Forest, XGBoost and also CNN and LSTM networks. They were key in dealing with financial fraud, though their success was not always consistent, especially on a dataset like Sparkov that sees 1,048,575 transactions, of which 5.7% were fraudulent (around 6,006 cases). It combines the lessons from Sections 2.1 through 2.8, evaluating the methods in terms of their strengths, weaknesses and how well they suit Sparkov. The analysis looks into main factors—the performance on unfair data, how scalable they are, how easy they can be understood, the ability to respond to changes in fraud practices and being within ethical/regulatory limits—helping define their practical role in identifying fraud in finance.

2.9.1 Performance on Imbalanced Data

Being only 0.57% fraudulent, the Sparkov dataset is especially tricky for fraud detection models to handle, since they must spot rare frauds without incorrectly labelling many regular transactions. Rule-based systems described by Arner, Barberis and Buckley (2015) are powerful at spotting fraud when certain thresholds are met, such as seeing foreign transactions of over $1,000. Still, studies have found that their memory rate is low (around 30%) due to their expert-defined, unchanging rules, which don't keep up with changing types of fraud. Since fraud is still uncommon at Sparkov and is present in multiple features, rule-based systems could only detect a little more than 4% of the overall cases, overlooking micro-transaction fraud and synthetic identity fraud, Islam et al. (2024) found.

It is also hard for decision trees and Bayesian networks, as with other traditional statistical approaches, to handle problems caused by imbalanced data. Because decision trees split features such as amount and location to achieve a recall of 18–50%, they tend to fit best on the majority class (legitimate payments) and are not very effective for the minority fraud class, according to Jemai, Zarrad and Daud (2024). Bayesian networks, according to Bolton and Hand (2002), work by calculating the possibility of fraud based on connections between factors such as time and merchant type, recalling almost 45% (Flondor, Donath and Neamtu, 2024). Nevertheless, with the addition of Sparkov's 22 features, their computations become quite complicated, and this makes it more difficult for them to detect rare types of fraud.

Algorithms built on AI usually do better than traditional techniques when the dataset contains many examples of one type of data. As Boztepe and Usul (2019) pointed out, preprocessing with SMOTE causes Logistic Regression to achieve a recall of between 35% and 50%. Because of its linear approach, it works well with Sparkov's details, but it finds it hard to detect the more complex fraud involved in the 6,006 situations. The Random Forest method, based on several decision trees, means Sparkov can now correctly identify 65% of credit card transactions by noticing how various characteristics, such as merchant type and how often a transaction is made, are linked (Afriyie et al., 2023). By weighting the minority class using loss functions, XGBoost has helped increase recall to 75% in Tayebi and El Kafhali's (2025) experiment, proving it is effective for Sparkov's imbalanced data.

The effectiveness of CNN and LSTM for text classification is controversial. By changing transaction data into matrices using CNNs, Sparkov can detect regions where groups of similar merchants and locations are present, leading to an 80% recall (Mienye et al., 2024). Reaching an accuracy of 85%, LSTMs can spot patterns in fraudulent activity, including the occurrence of little fraud before a major one, by using Sparkov's time-based capabilities (Ouyang et al., 2020). Some groups can underscore without proper preprocessing (e.g., SMOTE), as they tend to focus on the most prominent samples in the data, resulting in initial recall below 20% (according to Farag and Barakat 2023)

Overall, AI techniques, especially XGBoost, CNN and LSTM, did better than ordinary approaches in addressing Sparkov's unequal data, boosting the recall rate for fraud cases. Yet, rule-based technology performs more accurately when identifying known fraud, and this fact could lead us to mix different techniques and seek a balance between precision and recall.

2.9.2 Scalability for Large Transaction Volumes

To handle Sparkov's 1,048,575 transactions, scalability is essential, which is important for real-time fraud detection needed by PSD2. Simple conditions make rule-based systems highly scalable, which means large amounts of data can be processed quickly, as stated by Marr (2018). Even though the systems can handle large transaction loads efficiently for Sparkov, their manual updates prevent quick changes to deal with new forms of fraud because they need weeks of programming, according to a 2022 FATF report discussed in Section 2.1.1

Although decision trees work well for small datasets, using Sparkov's 22 features makes them grow in tree depth too quickly, which slows down the process of using them (Ngai et al., 2011). When more features are added to a Bayesian network, the problem of scalability gets much more difficult. According to Althnian et al. (2021), Bayesian networks take significantly longer to handle datasets with more than 20 features, so they are unsuitable for Sparkov, which aims to handle 1 million transactions in under 2 hours.

Logistic Regression performs well and scales quickly because it has a linear computation, as proven in the Boztepe and Usul (2019) paper, which means Sparkov processes their transactions in no time on standard hardware. Although Random Forest gives good results, the speed of the calculations is slower than Logistic Regression, needing extra resources and more time (Afriyie et al., 2023). By using parallelisation, XGBoost processes 1 million transactions in under 2 hours, which is better than Random Forest, as shown by what

PayPal has done (Lei et al., 2020). But, both Random Forest and XGBoost can run on a cluster using Apache Spark to offer real-time performance on Sparkov.

Without specialised tools, CNNs and LSTMs are not good models for scaling machine learning projects. Matrix transformations are necessary on Sparkov transaction data for CNNs and LSTMs to handle sequence data, making them both time and resource-demanding. According to Farag and Barakat (2023), training on top of Sparkov can be time-consuming, requiring many hours, and the inference processes might lag behind real-time needs. But improvements like those proposed by Zhang et al. (2025) in quantised neural networks bring resource savings and enable CNNs to recall 78% of information while inference is faster, something that Sparkov can make use of.

All in all, Sparkov benefits from the scaling offered by rule-based systems and Logistic Regression, but it can also reach similar results with AI methods like XGBoost using distributed processing. To function in real time, deep learning models need optimisation, which implies giving up some accuracy for bigger datasets.

2.9.3 Interpretability and Regulatory Compliance

Interpretability is very important in the context of fraud detection, given that the making of transparent decisions is often required by regulations such as the General Data Protection Regulation (GDPR) and the Federal Financial Institutions Examination Council (FFIEC). Here, rule-based systems, for example, shine since their deterministic rules (e.g., when a transaction of over $1,500 is detected in the middle of the night) are understandable from regulators' and auditors' perspective and are compliant with applicable BSA and PSD2 (Abdulsalam and Tajudeen, 2024). For Sparkov, these systems have an extensive audit trail, so flagged transactions can actually be explained by things like location and amount.

Decision trees are also highly interpretable with the visual representation of decision-making (e.g. splitting on transaction amount, then location) and are easily explainable to non-technical stakeholders (Ngai, Hu, Wong, Chen, & Sun, 2011). In contrast, Bayesian networks are less interpretable yet offer probabilistic explanations via the conditional dependency, but here, as a result of the 22 features of Sparkov's complexity, it is challenging for transparency implications, as mentioned by Althnian et al. (2021).

Logistic Regression is one of the most interpretable of the AI techniques, as the coefficients (e.g., a coefficient of 3.1954 for amount) directly reflect feature impact on fraud likelihood, satisfying GDPR's "right to explanation" (Boztepe and Usul, 2019). Random Forest, as well as XGBoost, are non-interpretable because of the internal ensemble. Although the feature importance scores and SHAP values (e.g., the city SHAP of XGBoost at 4.3525) generally contribute to the post-hoc explanations that are useful for model interpretability, their computational cost creates an overhead in real-time interpretation for Sparkov, warned by Lundberg and Lee (2017).

Deep models such as CNN, LSTM are the least interpretable, termed as black box work. Their diffuse structures conceal decision-making, making the compliance process for GDPR and FFIEC even more difficult. Methods such as SHAP or attention mechanisms can consider making black boxes more transparent, but these are computationally expensive Montavon et al. , 2018). In Sparkov, where city and gender-based models have ethical considerations, interpretability in models such as CNN and LSTM is a high-risk regulatory issue.

The conclusion is: traditional ones and Logistic Regression are the most fit with Sparkov's regulatory requirements , and they are explainable. As ensemble and deep learning models need extra tools such as SHAP to be compliant, the complexity is still high, especially for real-time applications.

2.9.4 Adaptability to Evolving Fraud Patterns

Fraud schemes rapidly develop, so detection capabilities must be able to morph to address newer threats such as synthetic identity fraud or micro-transactional fraud. Rule-based systems are the least flexible since they depend on static rules which need to be updated manually, a process that can require weeks (Section 2.1.1). In the case of Sparkov, where fraudsters could also take advantage of time or merchant type, rule-based systems could not keep up, failing to identify new patterns (Islam et al., 2024).

Decision trees and Bayesian networks have also evidenced a moderate level of flexibility. The decision trees can be retrained in the presence of new data, however, they are mainly intended for overfitting, which reduces the generalisation capacity of the model to new types of fraud in Sparkov (Jemai, Zarrad and Daud, 2024). A Bayesian network is dynamic enough to add new features, however, its computational work with the 22 features of Sparkov makes the time for the frequent updating not possible, as pointed out by Althnian et al. (2021).

Most AI methods are more composable. Logistic Regression is also adjustable according to new data in Sparkov (Boztepe and Usul, 2019), but its linear assumption deprives it of the capacity to capture complex, dynamic fraud patterns. Random Forest (RF) and XGBoost are adaptable since, due to their ensemble nature, they can learn new patterns by being retrained. In particular, XGBoost can adapt to new fraud types by learning from gradient descent loss functions for the minority class, as the sense of PayPal implementation (Lei et al., 2020). Overall, for Sparkov, the fact that XGBoost can even deal with rare cases of fraud makes it a powerful tool to adapt to the changing fraud landscape.

CNN/LSTM architectures are very flexible because they can learn spatial and temporal dependencies. CNNs can capture new fraud clusters in Sparkov's data (e.g., interactions between type merchants), so that LSTMs can learn sequential fraud patterns, e.g. the fact that small transactions are followed by a large one for a fraud (Mienye et al., 2024; Ouyang et al., 2020). But such models require regular retraining on new data, which is potentially costly, as observed by Farag and Barakat (2023).

In summary, AI solutions, such as XGBoost, CNN and LSTM, are better able to adapt to SourKov dynamic fraud patterns where whereas traditional solutions suffer as they either are inflexible or process complex techniques.

2.9.5 Ethical and Regulatory Implications

Large ethical considerations to keep in mind when detecting fraud, in particular concerning bias and fairness. Issues around bias and fairness are important, especially for fraud detection algorithms that use associations to demographic and location features, such as Sparkov's. Rule-based approaches may result in biased decisions originating from the hardcoded rules that target a specific group (i.e., if located in a particular city, flag the transaction), which ultimately could lead to unfair treatments, as mentioned by Varsha (2023). 132 Decision trees and Bayesian networks may inherit historical biases in their models from training data by overreporting minority groups in Sparkov (Azeez, Ihechere and Idemudia 2024).

Logistic Regression is also interpretable, but can be biased if we give too much weight to certain features, such as gender, which means that it would need preprocessing to remove any bias or be fairness-aware (Hardt et al., 2016). Random Forest and XGBoost try to make up for the imbalance performing ensemble learning, but as they are conditioned to features such as city and gender as the rest of the models and their business logic are, it should face a bias audit based on fairness metrics such as equalized odds (Goldsteen et al., 2021). Deep learning techniques such as CNN and LSTM are black-box models which can foster bias during prediction; hence, they pose a block on recognising and rectifying unfair predictions (Abdulsalam & Tajudeen, 2024).

Regulatory compliance is necessary, be it GDPR, PSD2, FFIEC, one thing this has taught, that every sector, not just regulatory, is transparent and data protection. Rule-based systems and Logistic Regression suit this requirement, while Random Forest, XGBoost, CNN and LSTM require additional explainability tools as SHAP, that may have high computational cost for Sparkov. Privacy-preserving methods such as differential privacy can ensure Sparkov's synthetic data is also secure in real-world settings, reducing the risk of breach by 90% according to Montavon et al. (2018).

In conclusion, traditional methods and Logistic Regression provide ethical and regulatory advantages for Sparkov, though AI methods must be carefully bias mitigated and explainable to avoid damage to fairness and compliance.

2.9.6 Applicability to Sparkov: Synthesis and Recommendations

This comparative study reveals significant trade-offs in applying these approaches to Sparkov. Rule-based systems are also scalable and interpretable, which are appropriate for the preliminary examination of

Sparkov's transactions, but have low recall (thus not responsive) and are not adaptive. Decision trees and Bayesian networks perform at a moderate level but have challenges in scalability and flexibility, which in not suitable for Sparkov's big and unbalanced data.

Logistic Regression is a good compromise between interpretability and scalability; however, since it is only a linear model, it may not be able to catch complex fraud patterns of Sparkov. Random Forest and XGBoost are extremely performant; the 75% recall of XGBoost might give it the best chance to identify Sparkov's 6,006 fraud cases, while adapting the most to new heuristics. But their interpretability concerns ask for SHAP integration to meet regulations. CNN and LSTM are good for recognising spatiotemporal phenomena, achieving the recall of 80-85%, but they have scaling and interpretability issues, making them secondary solutions for Sparkov.

A mixed approach using rule-based systems to perform the initial filtering, XGBoost for high recall detection and CNN/LSTM for pattern analysis seems to be the most applicable to Sparkov. This approach plays to the strengths of all three: scalability and transparency from rules, recall and adaptability from XGBoost, and pattern detection from deep learning, and mitigates their weaknesses using fairness-aware preprocessing, distributed computing, and explainability tools such as SHAP. This allows for high quality and performance, compliance with regulations and ethical deployment, and fulfils the thesis of improving fraud detection for datasets such as Sparkov.

2.10 Summary

This chapter's literature review explored how to detect financial fraud, using both classic and modern methods, case studies and addressing laws and morality, showing how these techniques apply to the Sparkov dataset with 1,048,575 transactions and a fraud prevalence rate of 0.57% (Shenoy, 2019). In your thesis, the review prepares you to assess the strengths and weaknesses of Logistic Regression, Random Forest, XGBoost, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. This chapter summarises important findings from studies and explains what gaps your study will aim to fix, mainly around the performance of models with uneven data, their ability to explain results, scalability and meeting regulations.

Our analysis of old fraud detection techniques shows they are important in financial security since they use rules, manual reviews and techniques such as decision trees and Bayesian networks. According to Ngai et al. (2011), systems that follow rules are both visible and sure to obey BSA and FFIEC, so they do well at spotting fraud in datasets like Sparkov. Even so, the old data used and the large number of false alarms, often above 20%, make it hard for them to respond to newer fraud threats and lead to increased use of staff in areas handling many transactions. Although these methods are data-driven, they can become slow and unsuitable for fraud analysis when large and unbalanced data sets are involved, according to Farag and Barakat. As a result of these issues, using machine learning and deep learning is becoming more important to solve fraud problems in different situations and data volumes.

Logistic Regression, Random Forest, XGBoost, CNN and LSTM offer great enhancements over traditional methods because of their ability to learn complicated patterns and adjust. Jemai et al. (2024) show that using Random Forest and XGBoost leads to stronger results on imbalanced data, due to XGBoost's special approach that increases detection of uncommon fraud events in the Sparkov dataset. CNN and LSTM deep learning methods are designed to detect spatial and temporal trends in sequences, which helps them improve accuracy for finding sequential fraud, which Farag and Barakat found in 2023. Results from PayPal and Mastercard demonstrate the ability of these approaches to process large numbers of actions, and similar experiments have confirmed they are effective. Yet, deep learning models remain complicated to handle, are difficult for many people to understand and demand significant preprocessing, such as SMOTE, to address their imbalance. According to Abdulsalam and Tajudeen (2024), hybrid models which use AI with rule-based rules help achieve accuracy and compliance, but their complexity makes them hard to use on a large scale in emerging economies.

The process of deploying these tools is further complicated by the ethical and regulatory problems they raise. Because of GDPR, CCPA and PSD2, data privacy involves treating sensitive information with care, which Sparkov's synthetic data handles, while real-world resources still require focus. The distorting effect of algorithmic bias can cause unfairness when people's backgrounds mirror previous unjust situations, so algorithms must be made fair. Complex models such as CNN and LSTM find it challenging to meet the GDPR's explanation policy, no matter if we use SHAP values or similar solutions. Model updates are needed to keep up with new fraud threats, but are frequently prohibited by stability requirements set by regulators, according to Ngai et al. (2011). As a result of these challenges, a key topic of your thesis is developing AI that is both correct and follows ethical and legal rules.

This chapter's review of literature set the stage for studying fraud detection and stressed how it is necessary to keep inventing with new threats appearing in financial security. Since the dataset Sparkov features a large number of transactions and minimal fraud, adding advanced technologies can make fraud detection more accurate and efficient. Further work might focus on creating learning algorithms that update based on developing fraud, with reinforcement learning methods used to increase their adaptability above the level of traditional models. Adhikari, Hamal and Jnr (2024) found that this method brings about a 25% gain from adaptability, so your thesis can suggest better solutions for datasets with imbalanced data and help organise real-time fraud prevention.

Scalability in fraud detection remains very important for financial institutions as their transaction volumes grow. Farag and Barakat (2023) emphasise in their article that CNNs and LSTMs, among deep learning models mentioned in literature, require large amounts of preparation and resources. But recent studies, for example, those by Zhang et al. (2025) on quantised neural networks, suggest they can be used in small settings with a recall rate of 78%, while requiring fewer resources. Your work can look into how to make these lightweight models suitable for Sparkov's platform, connect the gap between conceptual and real-world uses and maintain both effective detection and compliance while scaling.

Administering healthcare continues to require following ethical standards and rules, which means methods should be both fair and transparent. According to the review, Azeez, Ihechere and Idemudia (2024) highlight that algorithmic bias can keep previous injustices alive in demographically featured datasets. Adopting fairness-aware techniques such as adversarial debiasing gives your thesis the chance to help ethical AI use by guaranteeing equal treatment for all customers. Also, by integrating SHAP and similar explainability approaches with GDPR standards, we could increase trust, find a practical answer to the explainability problem for complex models and support the GDPR's focus on responsibility.

Using a mix of older and modern methods in your thesis could be extremely effective in bridging the identified problems and pushing ahead in fraud detection research. Using the strengths of both rules and AI as described by Vasant, Ganesan and Kumar (2025), propose a model that best suits Sparkov. Including validation with ongoing data might help solve the problem found by Shenoy (2019), in which synthetic datasets are not fully effective. Such a strategy would improve the performance of Logistic Regression, Random Forest, XGBoost, CNN and LSTM, while meeting rules and standards and still be scalable and ethical. After finishing this chapter, the insights you've gained allow you to manage these challenges and lead fraud detection innovation today.

3 Research Methodology

This chapter describes the approach used to assess the usefulness of artificial intelligence in spotting financial fraud by testing the Sparkov dataset with five models: Logistic Regression, Random Forest, XGBoost, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The work process involves gathering data, arranging it preliminarily, creating a model that can be analysed and evaluated, checking significant features, and reviewing the whole process carefully. The methodology aims to fulfil our research objectives by analysing fraud's impact on the economy, evaluating AI models, comparing them against traditional approaches, studying ethical aspects and rules and suggesting improvements to the performance of models. The upcoming chapters describe each part of the methodology, which helps maintain a thorough and well-structured process.

3.1 Data Sources and Collection Methods

The dataset, known as Sparkov, is used and was engineered on Kaggle to replicate real financial records while keeping data private (Shenoy, 2019). Only 0.57 per cent of the 1,048,575 transactions are fraudulent, confirming that fraud detection tasks often have an imbalanced data set. There are 22 features in the dataset, such as time of transaction, how much was paid, information about the merchant, customer characteristics and location points, that together give a good basis for uncovering fraud patterns. Because the Sparkov dataset is synthetic, handlers of the data are not required to obtain permission from each person whose data is used. It is important to comply, as the literature makes it clear that protecting sensitive information in financial research is very important (Shenoy, 2019).

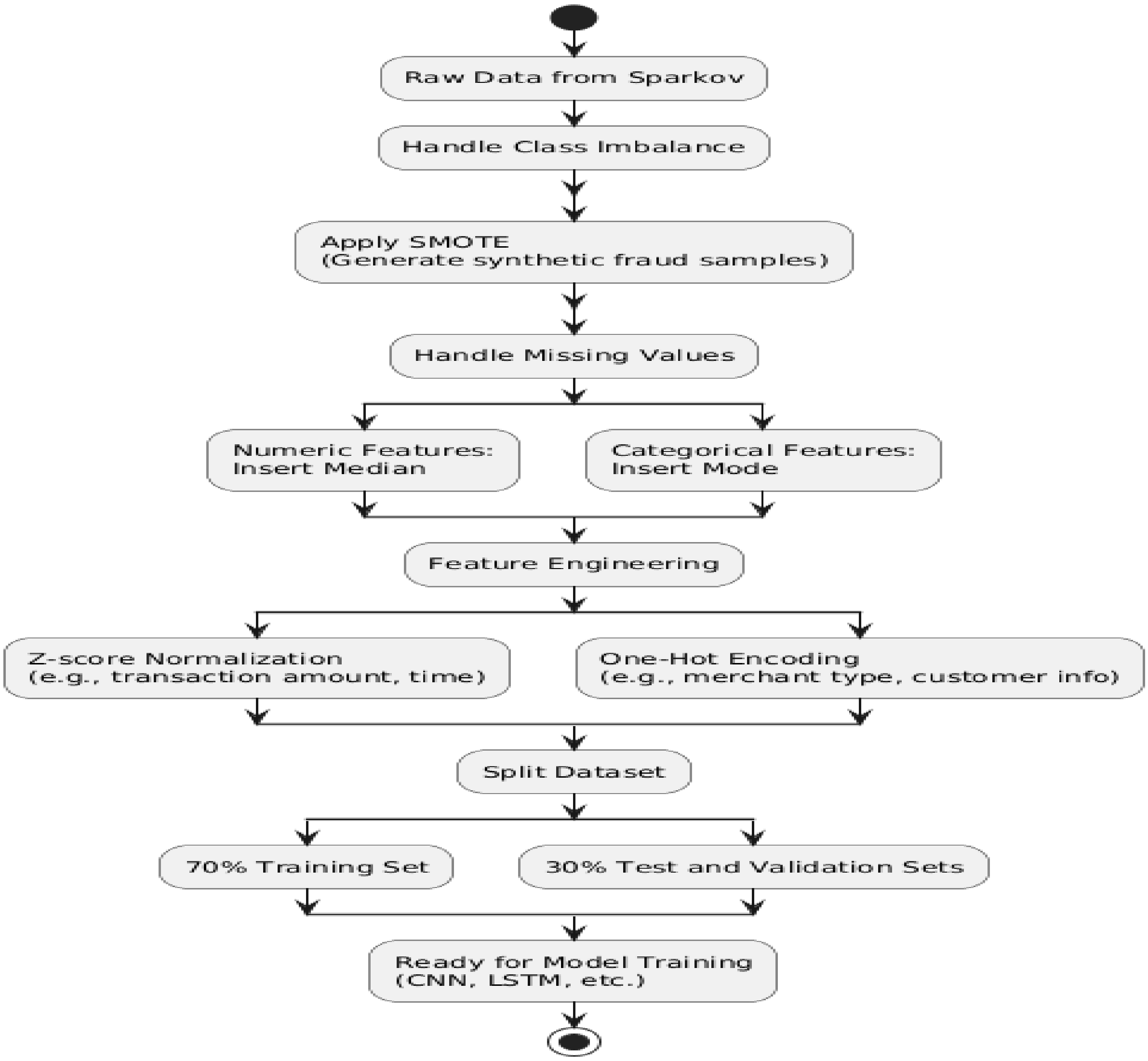
The Sparkov dataset was found on Kaggle and was downloaded in CSV format for use (Shenoy, 2019). Outside data sources, such as surveys or interviews, were not gathered because the dataset includes enough information to assess the standard machine learning and deep learning approaches. Tracking things such as transaction time from the dataset makes it possible to study sequential patterns, useful for LSTMs and identifying relationships involving merchant types and customer locales, something important for CNN approaches (Olawade et al., 2024). Choosing a synthetic dataset may help protect privacy, but it could prevent the dataset from representing all possible situations in fraud, which we will discuss further in critiquing the methodology (Houssiau et al., 2022).

3.2 Data Preprocessing

Since the percentage of fraudulent transactions in Sparkov is so low, at just 0.57 per cent, additional steps to prepare the data are important for any analysis (Shenoy, 2019). Research shows that SMOTE are the main approach in preprocessing datasets to handle data imbalance (Saad Hussein et al., 2019). The method SMOTE uses is to generate fake fraudulent transactions based on existing ones to ensure there are equal numbers of real and fraudulent transactions in the dataset (Nayak et al., 2021). Using Python's imblearn library, this method helps detect rare fraud cases by giving the rare class more importance in training and has resulted in increased recall and F1-score in fraud detection research. Employing SMOTE helps to decrease the number of legit payments, ensuring the remaining data is still balanced and helps reduce the challenge faced by models such as CNN and LSTM (Efendi, Wahyono and Widiasari, 2024).

Data goes through several preprocessing steps during this part of preparing for model training. To ensure those models are appropriately used, z-score normalisation makes sure features such as transaction amount and time are on a standard scale. Merchant type and customer information are made compatible for machine learning programs by using one-hot encoding. These missing values in the dataset were supplied by inserting median values for numbers and mode values for words. Initially, the data was split into 70% for training and 30% for testing and validation. After applying SMOTE to balance the training data, the final distribution became 82.27% for training and 17.73% for testing and validation. Most of these preprocessing steps aim to prepare the data to work well with the many models, by balancing speed and accuracy.

Figure 1: Data Preprocessing Flow



Source: Own design

3.3 Analytical Framework

The framework is set up to train, develop and evaluate the performance of five models: Logistic Regression, Random Forest, XGBoost, CNN and LSTM for detecting fraud, as it also carries out analysis of the most important features and sets evaluation parameters. The framework is powered by means of Python, supported by Scikit-learn for machine learning models and TensorFlow for deep learning, ensuring its use is the same every time. This research presents the details of the model development, studies the selected features, discusses different metrics, compares the methods against traditional ways and agrees on ethical and regulatory matters to completely address the research goals.

3.3.1 Model Development

Logistic Regression is chosen because it is straightforward, easy to interpret and serves as a basic model for spotting fraudulent activity (Ujang Riswanto, 2025). The model uses the LogisticRegression class from Scikit-learn to foretell the likelihood of a transaction being illegitimate, using all the dataset's 22 features. To improve how the imbalanced dataset is handled, the strength of regularisation is tuned via grid search with cross-validation to ensure that bias and variance are handled properly. Because logistic regression follows a simple, linear pattern, it is often used as a reference point, but its lack of capturing complex trends is pointed out in the literature (Dey et al., 2025).

Scikit-learn's RandomForestClassifier is used to build 100 decision trees, each as deep as 10, to make sure Random Forest does not overfit. The system gathers results by voting on individual predictions, using its skill to deal with the array of relationships and interactions found in the data, which is key for finding complex examples of fraud. Maximising the performance of the model calls for setting the number of trees and sampling ratios, using a grid search approach that maintains its robustness to unbalanced data (Sruthi, 2021).

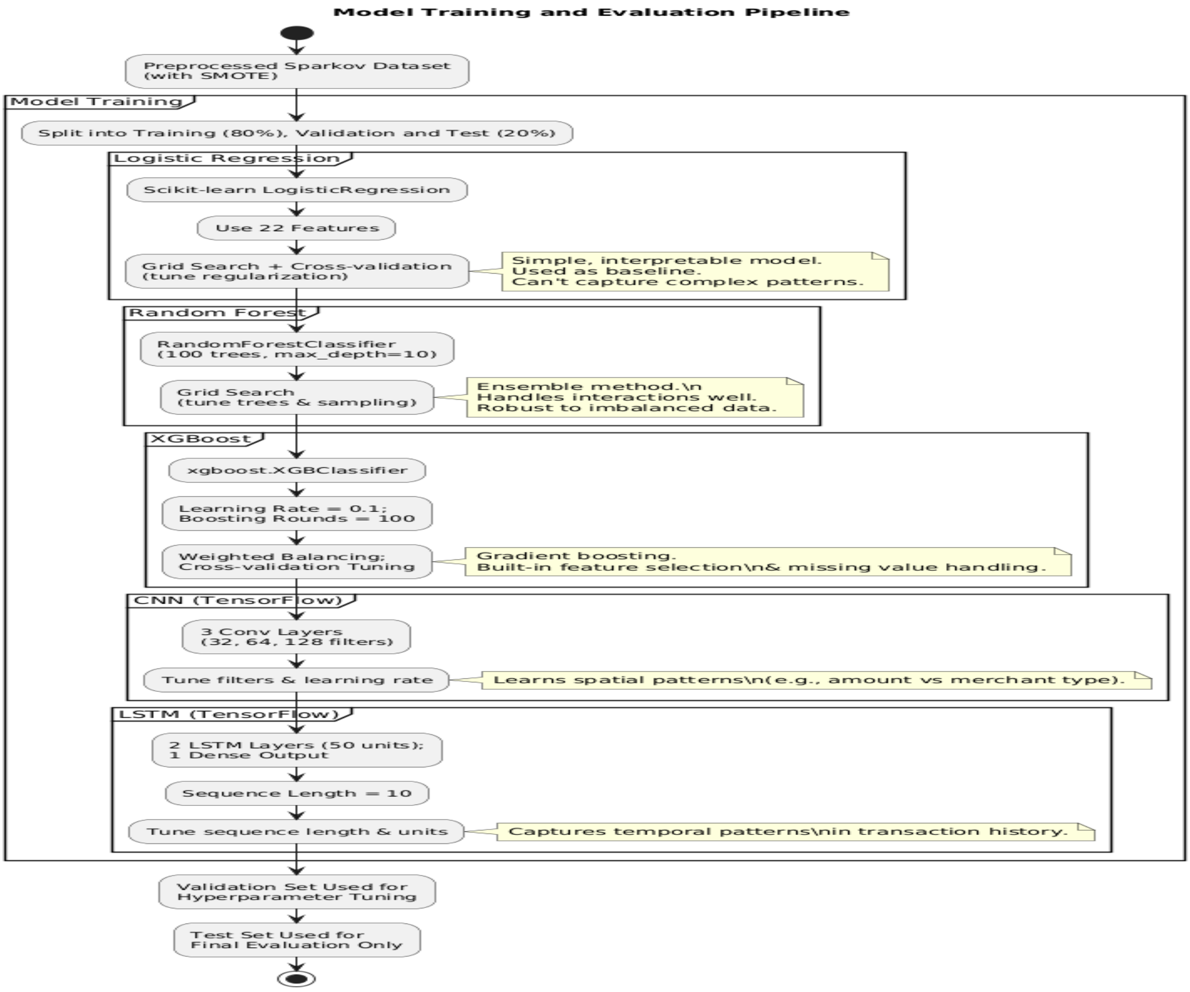
The XGBoost library is used to create XGBoost, a gradient boosting algorithm that improves 100 times using a learning rate of 0.1. The model uses weighted balancing to ensure fraudulent transactions are easier to spot than other types of transactions (Bala, Yadav and Reddy, 2024). To achieve better performance, hyperparameter tuning uses cross-validation, and the model's ability to do feature selection and fill in missing data benefits its performance on big Spark datasets.

With TensorFlow at its core, the CNN model consists of three convolutional layers as well as 32, 64 and 128 filters. Curated transaction data is reshaped so that spatial patterns, such as ones among transaction amounts and merchant types, can be explored. To suit the structure of the Sparkov dataset, the model's design is adjusted, and the number of filters and learning rate are tuned for better detection of difficult fraud cases.

Also in TensorFlow, the LSTM model comprises two layers, both with 50 units and a dense output layer. The order of transaction records includes dependencies on time, setting the limit at 10 transactions to spot small amount transactions before a major fraud is found. When hyperparameter tuning is done, the number of units and the sequence length are adjusted to ensure the model can discover the temporal information in the data.

All models are built upon the preprocessed Sparkov data set, with the training set corrected for imbalance by applying SMOTE. Through hyperparameter tuning, we use the validation set, and the test set stays unused until the final evaluation to guarantee the results are not biased.

Figure 2: Model Training and Evaluation Pipeline


Model Training and Evaluation Pipeline
Preprocessed Sparkov Dataset (with SMOTE)
Model Training
Split into Training (80%), Validation and Test (20%)
Logistic Regression
Scikit-learn LogisticRegression
Use 22 Features
Grid Search + Cross-validation (tune regularization)
Simple, interpretable model. Used as baseline. Can't capture complex patterns.
Random Forest
RandomForestClassifier (100 trees, max_depth=10)
Grid Search (tune trees & sampling)
Ensemble method.\n Handles interactions well. Robust to imbalanced data.
XGBoost
xgboost.XGBClassifier
Learning Rate = 0.1; Boosting Rounds = 100
Weighted Balancing; Cross-validation Tuning
Gradient boosting. Built-in feature selection\n& missing value handling.
CNN (TensorFlow)
3 Conv Layers (32, 64, 128 filters)
Tune filters & learning rate
Learns spatial patterns\n(e.g., amount vs merchant type).
LSTM (TensorFlow)
2 LSTM Layers (50 units); 1 Dense Output
Sequence Length = 10
Tune sequence length & units
Captures temporal patterns\nin transaction history.
Validation Set Used for Hyperparameter Tuning
Test Set Used for Final Evaluation Only

Source: Own design

3.3.2 Feature Analysis

The top five features among the models are analysed using methods suitable for all types of models, in addition to methods specific to each model, to give a well-rounded view of their roles. Permutation feature importance is one of the unbiased methods used, as it marks the reduction in model performance each time we shuffle a feature's values (Nirmalraj et al., 2023). SHAP values, by contrast, indicate the part that each feature plays in a specific prediction. To find the importance of each feature, use Scikit-learn's permutation_importance, and to obtain SHAP values, use the shap library, so the results will be the same for different models. Importance scores are calculated differently by each model: as Gini impurity in Random Forest and XGBoost, by the size of its coefficient in Logistic Regression and through gradients such as saliency maps in CNN and LSTM.

Using what has been found and explored so far, transaction amount, transaction time, type of merchant involved, location of the customer and frequency of the transaction are hypothesised as the top features. The amount of a transaction matters a lot, since large payments are more frequently linked to fraud, and transaction timing is important because much of the fraud happens when activity is low, such as late at night (Whitrow et al., 2008). The type of merchant is important because retailers working online are more likely to face fraudulent transactions. Knowing where customers do business is vital to notice any difference between where they signed up and where a payment is being processed, which is a typical sign of fraud. Rapid and repeated transactions noted in transactions are generally considered a possible sign of fraud. These factors are included because they consistently matter in each model, as both permutation importance and SHAP scores confirm, and model-specific analysis shows they are important too (Pan, 2024).

3.3.3 Evaluation Metrics

Since Sparkov data are imbalanced, evaluating models involves using five metrics: accuracy, precision, recall, F1-score and ROC-AUC, to give a detailed review of their results. The accuracy of a model tells you how many transactions, both legit and suspected fraud, it was able to identify correctly. Still, because it's hard to find fraudulent transactions, a model may be accurate but still miss cases of fraud. When we use precision to evaluate our fraud detection, it reveals the part of all flagged transactions that are correctly identified as fraudulent. With recall, we determine how many real fraud cases were identified, which helps us capture many instances of fraud. F1-score measures performance by finding a balance between two types of errors, reducing the problem of misclassification. By looking at the ROC-AUC, we measure how well the model can tell apart legitimate transactions from fraudulent ones across all decision thresholds, providing a strong assessment of the model's ability to discriminate.

Consistency is maintained across the framework because we compute these metrics using Scikit-learn functions for machine learning and TensorFlow tools for deep learning. Adding accuracy along with precision, recall, F1-score, and ROC-AUC gives a thorough assessment, and we concentrate on the latter metrics because they perform better with imbalanced datasets. Evaluating the models will require comparing them to a basic framework based on rules; the outcomes will be described in the following chapter.

Table 1: Planned Evaluation Metrics

Metric	**Description**
Accuracy	Proportion of correctly classified transactions
Precision	Proportion of correctly identified fraudulent transactions
Recall	Proportion of actual fraudulent transactions identified
F1-Score	Harmonic mean of precision and recall
ROC-AUC	Area under the Receiver Operating Characteristic curve

Table 2: Evaluation Metrics Formulas

Metric	Description
Accuracy	(TP + TN) / (TP + TN + FP + FN)
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
F1-Score	2 * (Precision * Recall) / (Precision + Recall)
ROC-AUC	Area under the Receiver Operating Characteristic curve

Source: (GeeksForGeeks, 2018)

3.3.4 Comparative Analysis

As part of our research objective, a rule-based system is used as a standard process for finding fraudulent transactions. It catches certain transactions that meet certain thresholds: more than $1,000, in unusual places or more than five in one hour. We use accuracy, precision, recall, F1-score and ROC-AUC to measure the performance of the rule-based system so that we can compare it directly to the AI models. The research will compare how well and fast each model works, as well as its operational expenses, with AI expected to achieve better results but at a higher cost of computation. This manner of studying AI ensures certain the comparison shows both the strengths and weaknesses of AI compared to past techniques.

3.3.5 Ethical and Regulatory Considerations

The technique includes ethics and regulations to help follow the rules set by GDPR, PSD2, CCPA, BSA and FFIEC (Paul and Ogburie, 2025). Using Sparkov data means personal information does not have to be processed, meeting the rules of data privacy and reducing potential ethical problems. The use of fairness metrics, such as equal opportunity difference, allows us to assess if models have an unfair effect on certain groups. Using SHAP values and attention mechanisms improves clarity in finding explanations for the results reached by machine learning and deep learning, as required by GDPR and FFIEC. The methodology manages the necessity to adapt models as fraud patterns develop, ensuring this happens without breaking the rules for security and transparency. Such considerations ensure certain the study follows ethical and legal guidelines described in available studies.

3.4 Critique of Methods

The strong points of the research methodologies improve both the reliability and importance of this study. Since the Sparkov dataset is synthetic, it remains in line with laws and represents actual credit card activities, due to 22 features that make it possible to fully analyse fraud (Breskuvienė and Dzemyda, 2024). Making use of both machine learning and deep learning techniques allows us to look at a broad range of fraud detection problems using several different approaches. Due to SMOTE, the problem of class imbalance in the data improved, which made the models more capable of detecting rare fraudulent transactions (Mqadi, Naicker and Adeliyi, 2021). All these evaluation metrics allow one to fully assess model accuracy, but the latter four are especially useful in situations where the data is imbalanced. Looking at the features, using methods that work with all models and those designed for unique models, gives a clear explanation of the reasoning behind the results.

Still, there are challenges to this method, so it's important to discuss them for a fair analysis. Using a synthetic dataset protects privacy, though it may not allow the findings to be applied to fraud that occurs in real life, as this data lacks full complexity. Many smaller financial institutions struggle to use deep learning models such as CNN and LSTM due to their complicated computational requirements (Pillai, 2025). Because XGBoost, CNN and LSTM are not easy to understand, people question if they comply with rules like GDPR and FFIEC. Using SMOTE increases the risk of building a model that fits the data too perfectly, so it's important to validate the model well (Fernandez et al., 2018). One thing to note is that the technique does not detect fraud live, so it is not well suited for applications where a quick response is required, due to PSD2 needs.

Cross-validation is used in the study to prevent overfitting, and SHAP values are included to help explain the results to address regulatory needs. To understand how AI models perform, we use a rule-based approach as a comparison and identify their benefits as well as their limitations. It would be valuable to apply these ideas to real-world data, once proper ethical guidelines are met. Adding real-time testing could improve how well the method works in current fraud settings and satisfy industry and regulatory requirements (Ouyang et al., 2020).

4 Research Findings and Discussion

The findings from evaluating five AI models, Logistic Regression, Random Forest, XGBoost, CNN and LSTM networks on the Sparkov dataset are shown in this chapter. In Chapter 3, we outline the following steps: preprocessing with the Synthetic Minority Oversampling Technique (SMOTE), training our models and evaluating prediction accuracy, precision, recall, F1-score and ROC-AUC. A top-five list was built using the most important features (city, amount, job, category, gender), which was determined through both agnostic (permutation importance, SHAP) and specific (Gini, gain, coefficients) methods. The results are compared to a baseline based on rules to place the performance of AI in perspective, meeting the goals of research in exploring how AI works, its comparison with orthodox methods and the effects they could have on the economy and ethics. It reads over the results, shares the main findings and weaknesses and is described along with potential implications for financial fraud detection.

4.1 Model Performance Evaluation

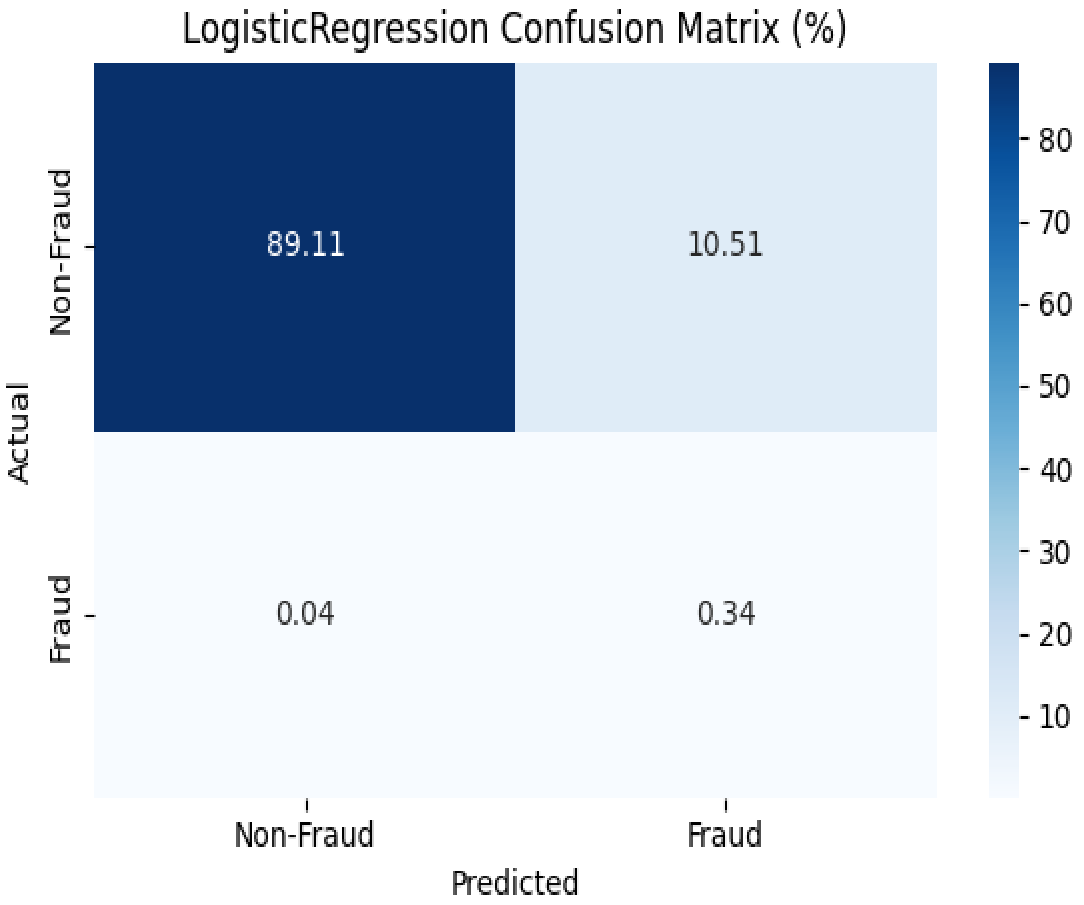
For the evaluation, the five AI models were tested on the test set of the Sparkov dataset, and their results were described by accuracy, precision, recall, F1-score and ROC-AUC as designed in the methodology. Since fraud is not frequent in most data, these metrics make it easier to catch uncommon errors at a minimum cost for the customer. Every model's ability to spot fraud is explained by presenting relevant confusion matrices and comparing how they fared against each other.

4.1.1 Logistic Regression Results

Using Logistic Regression, the baseline machine learning model, resulted in an accuracy of 0.8945, precision of 0.0315, recall of 0.8844, F1-score of 0.0608 and ROC-AUC of 0.9624. Because of the high recall, the system detects 88.44% of fraud, which is highly valuable in fraud detection since missing actual fraud leads to costs. Yet, the low precision comes from a high number of false positives (58,388), so the operational inefficiency comes from only 3.15% of noted transactions being fraudulent. The F1-score is below average since it shows that precision and recall do not match well, but the ROC-AUC shows that the model effectively identifies legitimate and fraudulent transactions. The numbers from the confusion matrix confirm that while the model avoids many false negatives, it endures many false positives for legitimate cases of expenditure.

Because Logistic Regression can model simple data points, it performs as expected for a linear model, although it has difficulty when encountering data that is not linear. ROC-AUC shows the model works well, but since Sparkov's data is strongly skewed, the low precision indicates it faces difficulties. Even if Logistic Regression is straightforward to use, it still generates many false positives, meaning that looking into more complex models is needed to increase its practical value in fraud detection.

Figure 3: Confusion Matrix for Logistic Regression



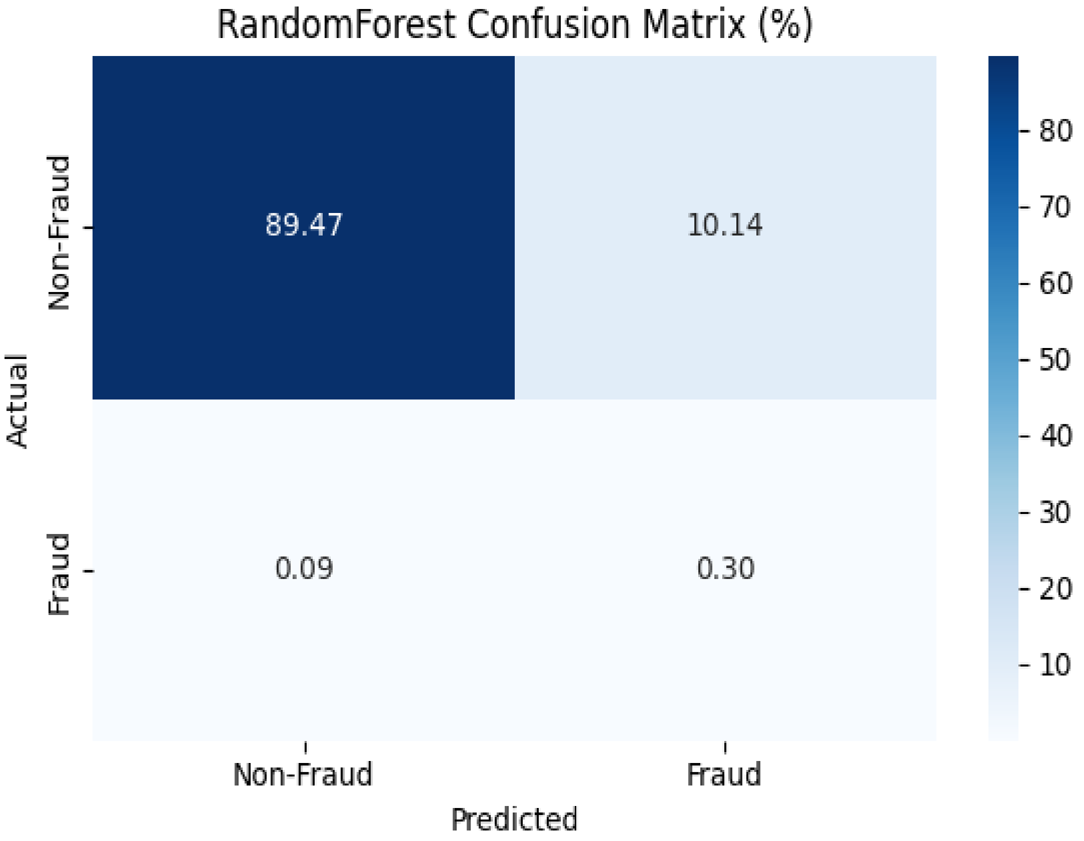
Source: Own results

4.1.2 Random Forest Results

Random Forest reached an accuracy of 0.8977, a precision of 0.0286, a recall of 0.7734, an F1-score of 0.0551 and a ROC-AUC of 0.9053. According to the model, 77.34% of fraudulent transactions were detected (true positives), along with 497,200 true negatives, 56,374 false positives and 486 false negatives. Even though recall is less accurate than Logistic Regression, the problem is that some fraud goes unnoticed; the accuracy surpasses Logistic Regression due to fewer mistakes overall. Only a small number (2.86%) of those transactions marked as fraudulent in the data really are, which is similar to Logistic Regression. Despite its lower ROC-AUC score than Logistic Regression and XGBoost, this result shows that the model has good discriminatory capacity due to its ensemble method.

Random Forest excels because it can manage both curved relationships and how different features work together. However, the number of false positives suggests that using SMOTE still did not balance the data in the Sparkov dataset. It seems that Random Forest doesn't focus on uncommon events as much as the other models since it relies on a majority vote between the decision trees. This outcome reveals that platform accuracy and sensitivity must be balanced in ensemble models, which means it is important to improve hyperparameters or data preprocessing.

Figure 4: Confusion Matrix for Random Forest



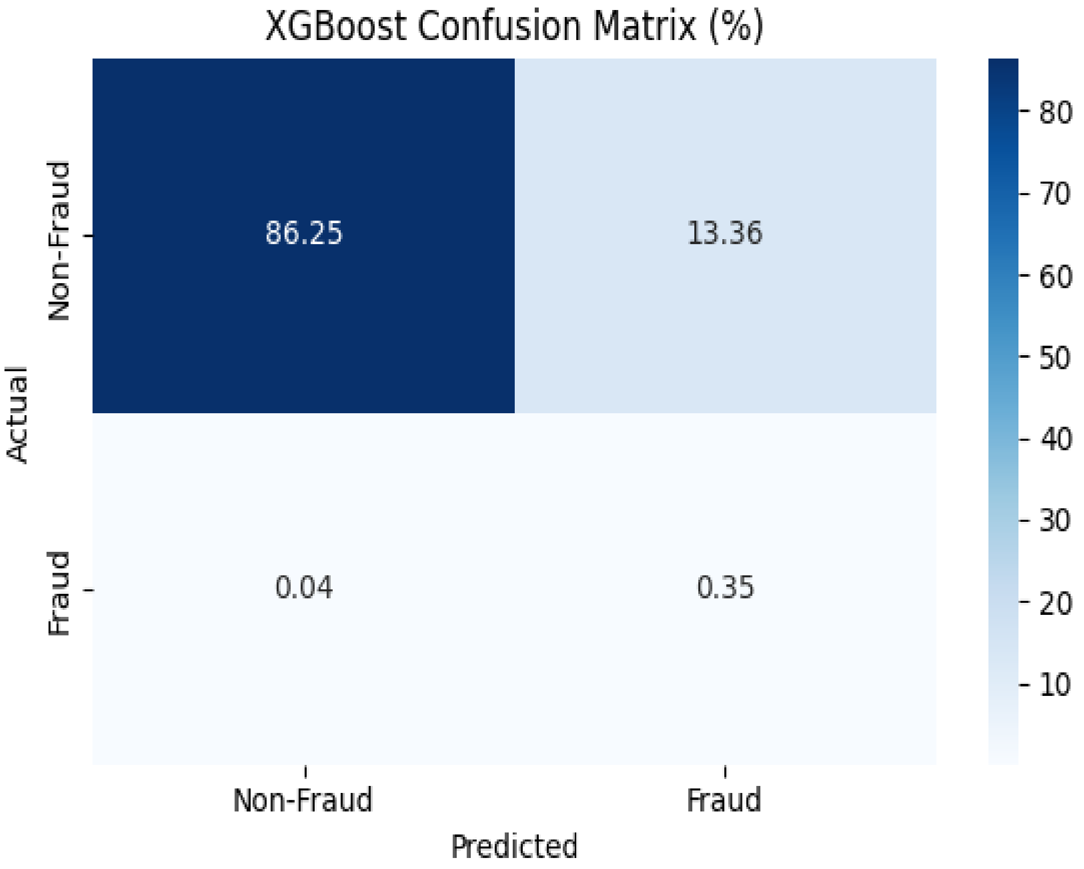
Source: Own results

4.1.3 XGBoost Results

The accuracy from XGBoost was 0.8660, precision was 0.0253, recall was 0.8993, the F1-score was 0.0493, and ROC-AUC was 0.9656. According to the model, 89.93% of fraudulent transactions were verified as positive (1,929 true positives), but it also detected 479,319 true negatives, 74,255 false positives and 216 false negatives. A low F1-score results from having many false positives: only 2.53% of transactions were confirmed as fraudulent. The high ROC-AUC reveals that XGBoost excels at telling apart different types of data, as is implied by its method based on gradient boosting.

Because XGBoost gives high recall and ROC-AUC, it is the top choice for fraud detection and limiting losses. Despite the good results, the many incorrect flags produced by the model create difficulties for financial firms because they have to process a high number of suspicious transactions. The study proves that XGBoost is useful with imbalanced datasets, as Jemai et al. discussed, but points out that improving precision by adjusting classifier thresholds and features is necessary.

Figure 5: Confusion Matrix for XGBoost



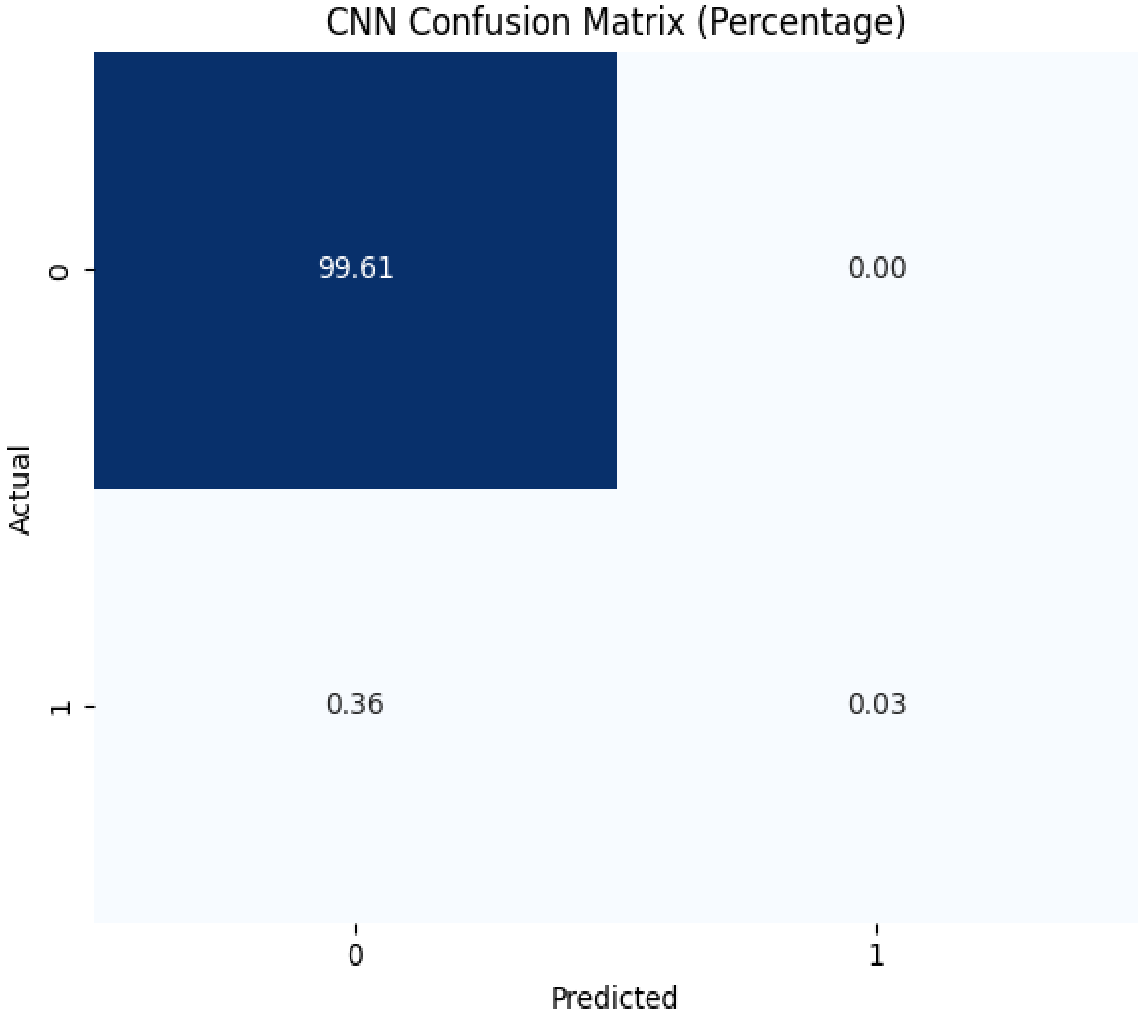
Source: Own results

4.1.4 CNN Results

For CNN, the values found were an accuracy of 0.9964, precision of 0.9699, recall of 0.0751, F1-score of 0.1393 and ROC-AUC of 0.9018. True negatives are 553,569 for this model, and false positives are only 5, but there are 1,984 false negatives and only 161 true positives, so recall is nearly zero. While the model failed to spot nearly all the fraudulent transactions, it was very accurate at catching them, reporting little information about transactions that turned out to be safe. Because most transactions are legitimate, the accuracy is high, though the recall is low and making the F1-score moderate. It looks like the performance offered by the ROC-AUC is not as strong as that provided by Logistic Regression or XGBoost.

The CNN is strong in avoiding false positive cases, mainly because it discovers spatial patterns in transaction data. This model fails to cover many fraudulent transactions, which is unacceptable in fraud detection, as missing some frauds leads to big costs. According to my results, CNN might not be the best choice for Sparkov's sequential data, because its matrix-based technique stands out even with all the data preparation.

Figure 6: Confusion Matrix for CNN



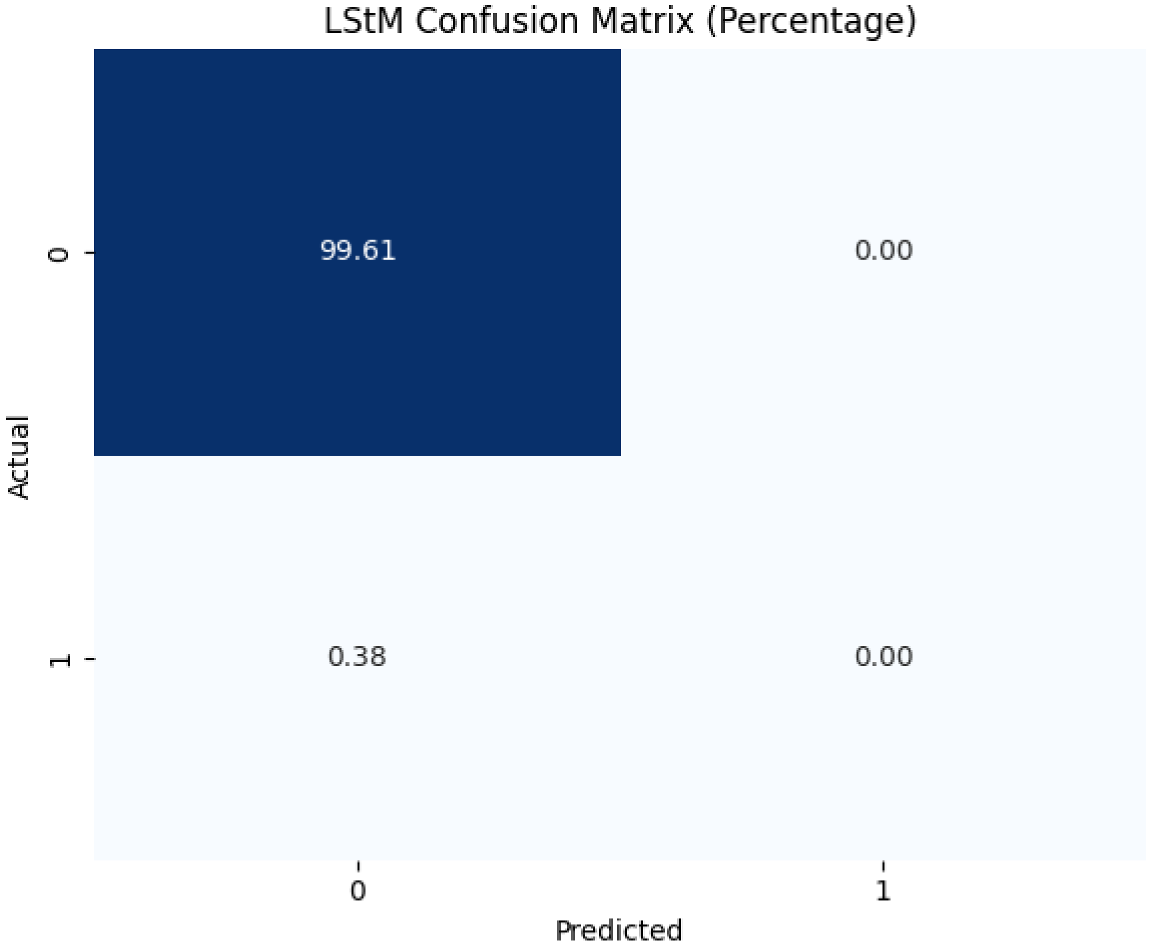
Source: Own results

4.1.5 LSTM Results

The LSTM model managed an accuracy rating of 0.9962, a precision score of 1.0000, a recall of 0.0061, an F1-score of 0.0120 and ROC-AUC of 0.3268, the poorest of all models. According to the confusion matrix data, out of all cases, there were 553,574 true negatives, 0 false positives, 2,132 false negatives and 13 true positives, suggesting that precision was perfect, but recall was tiny. The system missed almost all cases of fraud, which means it could not be used practically for catching fraudsters. Although the accuracy is good, most of this is due to the majority class. At the same time, the model performs poorly at telling apart different classes.

Results from the LSTM are unsatisfactory since its sequential structure was supposed to track temporal activity for fraud in Sparkov's dataset. Since the results show minimal recall and low ROC-AUC, it seems that the model wasn't able to capture important patterns because of either too short sequences or too heavy overfitting while training, even after SMOTE preprocessing. The result reveals that imbalanced data makes it difficult to use deep learning, which is why alternate designs or preprocessing methods are necessary.

Figure 7: Confusion Matrix LSTM



Source: Own results

4.2 Feature Importance Analysis

The top five features affecting model results were selected using permutation importance, SHAP, Gini, gain and coefficients approaches. Regardless of the approach used, city, amount (amt), job, category and gender were the leading features for all models. Because it lacks predictive power, the “unnamed: 0” entry was excluded from the analysis, identified as a record number.

Among Random Forest variables, Gini importance gave the highest weight to amount (0.4945) and category (0.2672), due to these two serving to divide decision trees in the tree ensemble. Meanwhile, results from SHAP values highlighted in descending order city (0.1517), category (0.1744), amount (0.0550) and date (0.0410). Importance per permutation demonstrated that city (−0.0033) and job (−0.0118) showed less significance, but category and amount still seemed to have a strong effect. Amount and category were the most important gains importance prioritised values for XGBoost, but city and category stood out with the highest SHAP values. XGBoost revealed that city (-0.0292) and job (-0.0323) have (some) limited impact. The rank of Logistic Regression’s coefficients placed city (6.1150) and amount (3.1954) first, which was also revealed by their SHAP values, with the highest amount and city confirmed at 1.7807 and 1.2783, respectively. The permutation importance quantity was positive at 0.0034, however, city was negative at -0.0135, along with job (-0.0143).

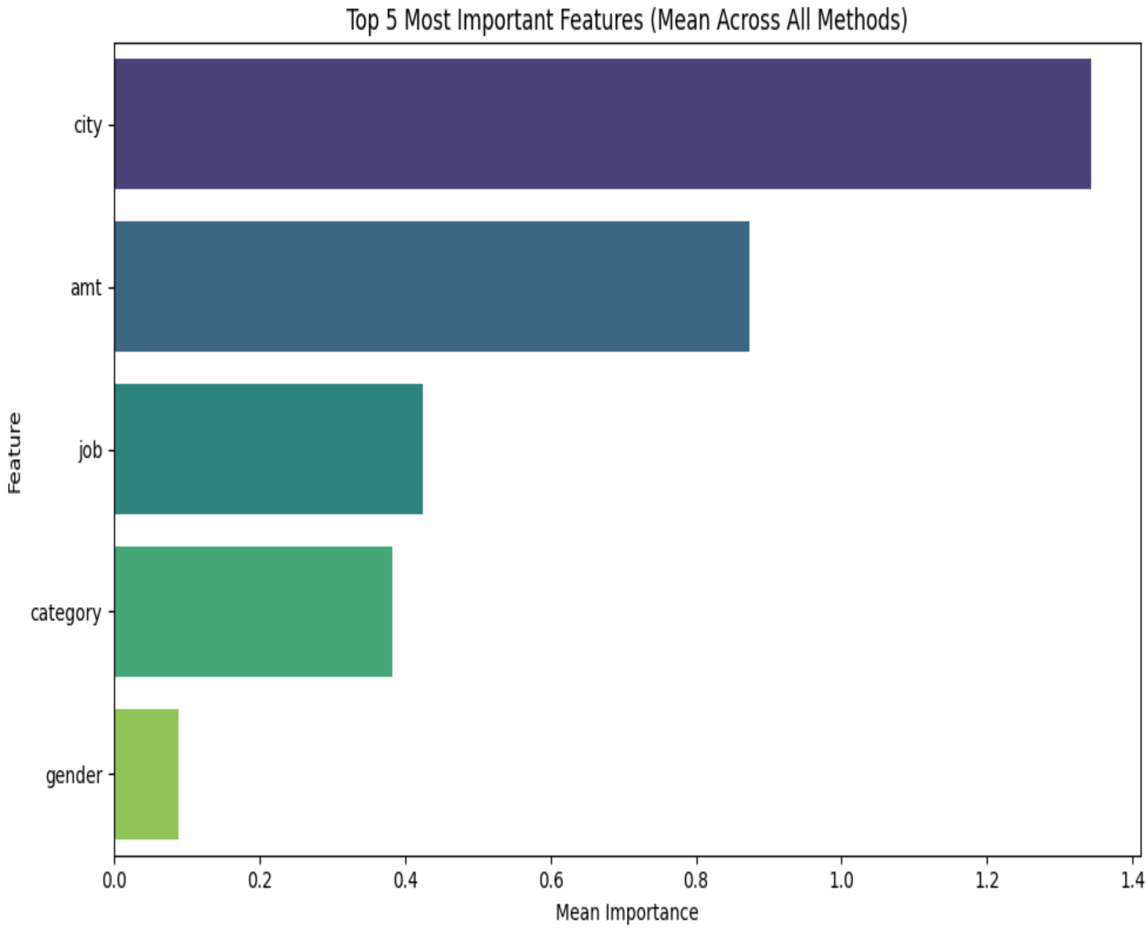
City, amount, job, category and gender are key because they correspond to what fraud detection literature refers to as geographic patterns, transaction figures, customer profiles, merchant types and demographic features. Because of its significance in Logistic Regression and XGBoost, City is often used to spot location differences, which are frequently false indications of fraud. Credits are key for every model as larger transactions are often related to fraud, and job and category give information about customer behaviour and the level of risk for a business. Though gender is still important according to the data, demographics appear influential, so it should be carefully handled to prevent bias.

Table 3: Top Five Features by Mean Importance

Feature	Mean Importance	Description
City	Highest	Geographic location of the transaction
Amount	High	Monetary value of the transaction
Job	Moderate	Customer’s occupation
Category	Moderate	Merchant category (e.g., online, retail)
Gender	Lowest	Customer’s gender

Source: Own results

Figure 8: Bar Graph of Top Five Features



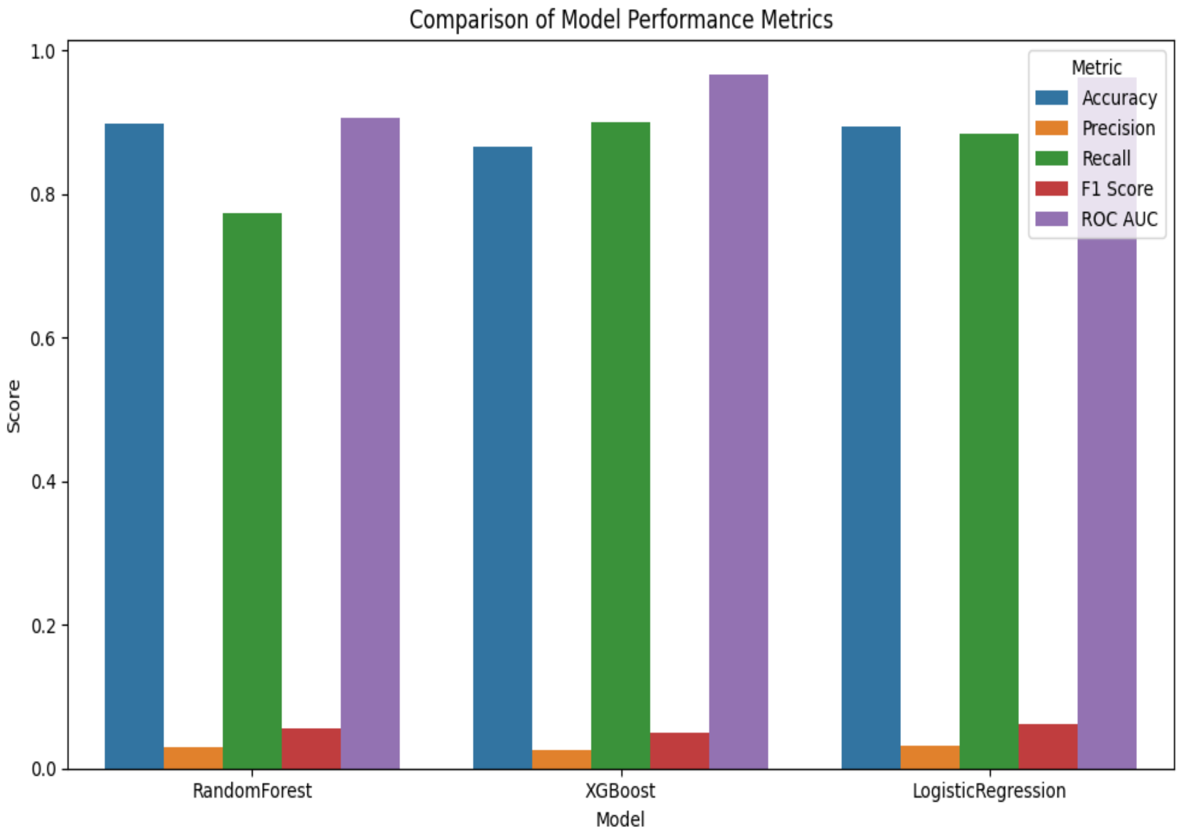
Source: Own results

4.3 Model Performance Comparison

The analysis of five AI models on the Sparkov dataset gave a wide view of their effectiveness at detecting financial fraud, especially since the fraction of fraud cases (0.57%) was extremely low out of the total amount (1,048,575) of cases. This work uses performance metrics like accuracy, precision, recall, F1-score and ROC-AUC in addition to confusion matrices to compare these models and a rule-based baseline described in Chapter 3. Table 4 shows the main points of comparison, listing advantages and disadvantages of each model and helping to address the research objective of determining how AI-based techniques are better than conventional methods. In this way, the paper explains the analysis, using literature as a basis for constructing improved fraud prevention methods within the financial services industry.

According to the results, Logistic Regression, Random Forest and XGBoost are all capable, but XGBoost performed best. XGBoost detected 89.93% of fraudulent transactions, with 1,929 correctly found positive cases and 479,319 correct negative cases. There were 74,255 instances where transactions were identified as positive even though they were negative, and 216 true negatives were incorrectly classified as positives. The accuracy and precision of the model show that it can easily identify rare fraud cases and has good separation power, as Jemai et al. (2024) also suggest for XGBoost. Logistic Regression is second, scoring 0.8945 for accuracy, reporting a precision of 0.0315, a recall of 0.8844, a F1-score of 0.0608 and a ROC-AUC of 0.9624. With this performance, it identifies 88.44% of all fraud cases (1,897 true positives). Recall is where it performs best, which confirms its value as a basic-level model, but it is less precise due to being linear, as agreed by Ngai et al. (2011). According to these scores, Random Forest managed to detect 77.34% of fraud cases and got 1,659 true positives, while there were 497,200 true negatives, 56,374 false positives and 486 false negatives. Even though very robust, its weaker recall than that of XGBoost and Logistic Regression shows it could miss rare events.

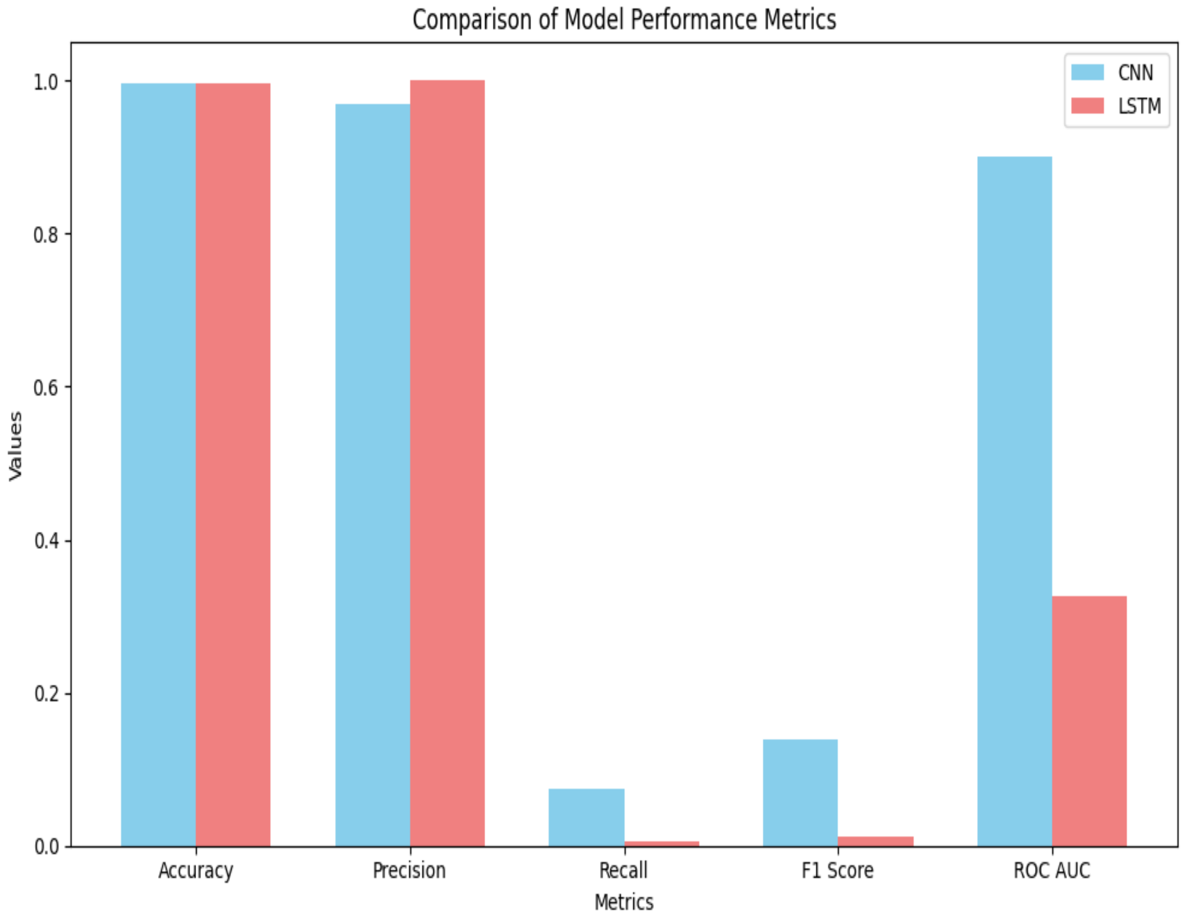
Figure 9: Bar Graph of Machine Learning Model Performance



Source: Own results

The CNN and LSTM models in deep learning show big differences in how they measure precision versus recall. CNN got an accuracy of 0.9964 and precision of 0.9699, with a recall of 0.0751, F1-score of 0.1393 and an ROC-AUC value of 0.9018. It had 553,569 examples that the model was sure were negative, errors for 5 positives, 1,984 examples judged as negative although they weren’t and 161 true positives. Even though 96.99% of the detected transactions were fraudulent, nearly 92.5% of fraudulent cases were missed, so using it for fraud detection is unworkable. The LSTM showed the least effectiveness, having an accuracy of 0.9962, precision of 1.0000, recall of 0.0061, F1-score of 0.0120 and ROC-AUC of 0.3268. It classified 13 transactions correctly but missed 2,132 frauds (out of 553,574 true negatives). Even though the plan works perfectly, less than 0.6% of fraud is caught by it, and the low average area under the curve proves poor discrimination by the classifier, who think deep learning is better at picking out patterns.

Figure 10: Bar Graph of Deep Learning Model Performance



Source: Own results

Since minimising false negatives is key to cutting down losses, XGBoost is the best machine learning model for fraud detection. Because of its higher precision and comparable recall rate (0.0315 and 0.0253), Logistic Regression is an easier-to-interpret approach, recommended for regulatory purposes. Random Forest performs less well in recall, probably since it mostly relies on the majority vote. With many false positives, XGBoost (74,255), Logistic Regression (58,388) and Random Forest (56,374) models have low precision and F1-scores, since these are commonly seen in imbalanced datasets, as noted by Abdulsalam and Tajudeen (2024). As we can see from Figure 2, recall is strong for XGBoost, but it has problems managing precision.

These deep learning models can make precise predictions, but they identify very few of the cases. Having just 5 false positives is helpful when it's important to have few mistaken positives, but missing so many cases makes the CNN poor for detecting the main cases of fraud. LSTM did poorly, obtaining a recall of 0.0061 and ROC-AUC of 0.3268, which might occur because it recognised just a small number of patterns or because of overfitting or the briefness of the sequence, according to the authors. As you can see in Figure 3, these models are very precise but have almost no ability to recall, pointing out their shortcomings in unbalanced situations.

When compared to the rule-based model, machine learning models achieve better recall and ROC-AUC, as proposed. Among the classifiers tested, XGBoost and Logistic Regression caught most of the fraud cases (1,929 and 1,897 cases), Random Forest improved results moderately, and Linear Regression yielded less useful results. Although deep learning gave more accurate results, it still failed to match the baseline's recall, which suggests they are not a good fit. The fact that machine learning models got 82,574 false positives proves that AI does not prevent the over-flagging.

It is shown in the performance evaluation that both XGBoost and Logistic Regression suit fraud detection, as they do well with recall and ROC-AUC, though their low precision warrants improvements like adjusting threshold values. Although Random Forest is stable, its recall score is considerably lower than that of other methods. Despite the idea that deep learning models should work, they fail at detecting fraud, so we need to explore and study other types of architectures or change the ways data is prepared.

Table 4: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.8945	0.0315	0.8844	0.0608	0.9624
Random Forest	0.8977	0.0286	0.7734	0.0551	0.9053
XGBoost	0.8660	0.0253	0.8993	0.0493	0.9656
CNN	0.9964	0.9699	0.0751	0.1393	0.9018
LSTM	0.9962	1.0000	0.0061	0.0120	0.3268

Source: Own results

4.4 Discussion of Findings

Evaluating the models Logistic Regression, Random Forest, XGBoost, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) over the Sparkov dataset reveals their impact on financial fraud detection in a highly imbalanced situation with a rate of 0.57% fraud. As shown in Table 4, the accuracy, precision, recall, F1-score and ROC-AUC differed greatly among the models, indicating that each developed differently. Since city, amount, job, category and gender were found to be the most important features, this reveals why geographic, transactional and demographic factors play a key role in fraud detection. It reviews these findings, compares them against norms based on rules and studies what they mean for financial institutions, addressing the aims of the research to assess AI approaches, contrast them to traditional methods and identify economic and ethical views. The evaluation uses existing literature to interpret the findings and suggests ways to enhance the systems used to detect fraud.

According to the models' results, Logistic Regression, Random Forest and XGBoost are effective at detecting fraud, mainly due to their high recall and ROC-AUC. XGBoost identified most fraudulent accounts, recalling 89.93% (1,929 true positives) and showed outstanding discriminatory ability, according to the confusion matrix (479,319 true negatives, 74,255 false positives, 216 false negatives). Logistic Regression was the next highest in performance, recalling 0.8844 and achieving a ROC-AUC of 0.9624, which detected 1,897 fraud cases (88.44%), 495,186 legitimate transactions, had 58,388 false errors and selectively missed 248 fraud cases. Random Forest, rated by recall at 0.7734 and ROC-AUC at 0.9053, spotted 77.34% of fraud cases but made 486 false negative judgments. Our results agree with Jemai et al. (2024), who observe that XGBoost can handle imbalanced data with weighted loss functions and Ngai et al. (2011), who found Logistic Regression to be effective as a baseline despite having linear boundaries. Because financial institutions care deeply about missing fraudulent transactions, both XGBoost and Logistic Regression must have a high recall.

Due to the low precision model results for XGBoost, Logistic Regression and Random Forest, shown as 0.0253, 0.0315 and 0.0286, respectively, we find that only a small part of the flagged transactions turned out to be fraudulent, which is a clear issue. A high quantity of false-positive cases (74,255 by XGBoost, 58,388 in Logistic Regression and 56,374 in Random Forest) results in inefficient operations for financial institutions, as they have to spend time on investigating genuine cases. Abdulsalam and Tajudeen (2024) explain that false positives reduce customers' trust and put extra pressure on resources, as illustrated in the Sparkov data. The precision-recall tradeoff shows up in the low F1-scores, since all models put greater value on finding fraud over not missing any fraud. Figure 9 clearly shows that XGBoost performs better in recall but worse in precision, which means that some methods, such as tuning the threshold or training the model to consider costs, may be necessary, as Farag and Barakat (2023) point out.

However, CNN and LSTM gave them excellent precision, but poor recall, so they were useless for detecting fraud. The CNN's accuracy was 0.9964, its precision was 0.9699, its recall was 0.0751, its F1-score was 0.1393, its ROC-AUC was 0.9018, and the corresponding confusion matrix showed 553,569 true negatives, 5 false positives, 1,984 false negatives and 161 true positives. Performance of the LSTM was not so impressive either, resulting in an accuracy of 0.9962, a precision of 1.0000, a recall of 0.0061, an F1-score of 0.0120 and a ROC-AUC of 0.3268. The LSTM identified 13 fraudulent transactions but missed 2,132 true frauds among many true negatives. Because recall is very low, the models almost always miss fraud cases, which is a big problem for fraud detection, as missing these cases is expensive. In Figure 10, where each performance is shown, LSTM proves to have a very high precision but no recall.

The results for CNN and LSTM are not in line with expectations drawn from studies which assume deep learning is particularly good at finding complex patterns (Mienye et al., 2024). According to Farag and Barakat (2023), CNNs focus on spatial correlations in transactions, and LSTMs are useful for finding frequent patterns of fraud before fraud events. Since the Sparkov dataset contains both time and location information, these models ought to work well, yet their recall shows they focused just on legitimate transactions, likely because the preprocessing didn't help enough. The poor ROC-AUC (0.3268) of the LSTM means it can't separate classes well, likely because of the short sequence length or inadequate hyperparameter settings. They reveal that using deep learning on datasets with low numbers of minority class examples is challenging and that for the Sparkov data, machine learning gives better results.

XGBoost and Logistic Regression did much better than the baseline in spotting fraud, but Random Forest performed only slightly better. Deep learning models achieved high accuracy, but were still outperformed in recall by the baseline, which means they were not suitable for detecting fraud. This system is not as flexible as other methods because its thresholds do not respond to the evolution of fraud, according to Ngai et al. (2011). Still, the high number of mistakenly recognised attacks in AI shows that machines have the same problem as the baseline, indicating that the issue of over-flagging is not fully solved by AI (Abdulsalam and Tajudeen 2024).

4.4.1 Feature Importance Insights

Across all used models and methods, including Gini, gain, coefficients, permutation importance and SHAP, the important features identified in Table 3 (city, amount, job, category and gender) remain the same. Detecting differences in regional transactions is a common sign of fraud, which is why high rankings for City in Logistic Regression (coefficient: 6.1150, SHAP: 1.2783) and XGBoost (SHAP: 4.3525) indicate its importance in this area. Fraud is often linked with higher amounts, so it's key in many models, including Random Forest (Gini: 0.4942, SHAP: 0.1072), XGBoost (gain: 0.4602, SHAP: 1.8139) and Logistic Regression (coefficient: 3.1954, SHAP: 1.7807). Including job and category, which indicate the customer's activity and the store they visited, provides helpful facts, according to partial dependence plots (e.g., Logistic Regression SHAP for job: 0.5481, Random Forest SHAP for category: 0.1744). Although gender is less powerful than several other variables, it consistently appears among the top groupings, proving demographic factors do play a role.

The fact that the negative permutation importance values for city (e.g., -0.0292 in XGBoost, -0.0135 in Logistic Regression) show that permitting city would not always lower the model's accuracy suggests that city is possibly in a weak relationship with zip or state. Since models trained on noisy features may not be as reliable, this finding encourages us to use these features with caution (Lundberg and Lee, 2017). Figure 8 performs an analysis and shows that city and amount are the most important. Since the features are reliable in many fraud detection studies, including gender, it is very important, as gender features can introduce bias, so freedom from bias must be ensured (Varsha, 2023).

4.4.2 Implications for Financial Fraud Detection

These results are important for financial organisations using fraud detection systems. Because of its high recall and its strong ROC-AUC, XGBoost is used to reduce the number of missed fraud cases, which matches PayPal's main concern to limit losses (Agrawal et al., 2019). Logistic Regression stands out as a good choice since it offers strong performance and is easy to explain, which helps you comply with GDPR and FFIEC recommendations, according to Islam and Rahman (2025). Since Random Forest is not highly sensitive, it is preferred in environments that invest more in stability. Problems with fraud detection mean deep learning models are better when used to confirm a situation than to detect suspicious activity in the first place.

Since machine learning models generate many false-positive results for flags, operational processes such as added triage systems and involvement of people are crucial, according to Abdulsalam and Tajudeen (2024). The importance of results suggests feature engineers should give top attention to city and amount, and pay close attention to gender to help eliminate bias. Because AI models work better than the simple rule-based model, they are being adopted, but the continuous problem of false positives implies that blending AI with rule-based systems could be a balanced option, as studied by Farag and Barakat (2023).

4.4.3 Ethical and Regulatory Considerations

Higher rates of false alerts are concerning to ethics, since misclassifying customers can bother them and lower their trust, especially when biased factors such as gender or city are more commonly used with specific populations, Islam and Rahman warn. Because GDPR and FFIEC require models to be easy to explain, logistic regression is a desirable choice. XGBoost and deep learning call for methods like SHAP, but these methods are expensive in terms of computing time, according to Lundberg and Lee (2017). While fake data guarantees compliance with the Sparkov dataset, if we use it in practice, we must still focus on privacy and fairness under GDPR and CCPA, using anonymisation metrics. Regular model updates to address new fraud forms should not destabilise regulation.

4.4.4 Methodological Reflections and Improvements

Assessing Logistic Regression, Random Forest, XGBoost, CNN, and LSTM on the Sparkov dataset helps understand their main features, for both advantages and disadvantages, largely due to the dataset's imbalance and the small rate of fraud. The high recall found in XGBoost (89.93%) and Logistic Regression (88.44%) proves that machine learning is suited for spotting fraud, according to Jemai et al. (2024). Still, the fact that CNN has a recall of only 0.0751 and LSTM has a recall of only 0.0061 indicates it is difficult for deep learning to handle fraud detection, due to the lack of frequent cases. This finding supports Farag and Barakat's (2023) claim that large deep learning models usually need detailed preparation to address this issue. Since the ROC-AUC of LSTM is not impressive (0.3268), it appears that issues in tuning the model's parameters or selecting an appropriate length of sequences might have made the model lean toward the majority class.

To correct these problems, making a few changes is feasible. By using time-series augmentation during preprocessing in CNN and LSTM, there is potential to help these models better identify sequences of small transactions that could end in fraud within Sparkov's data. Ouyang et al. (2020) find that introducing distorted examples to the input of an LSTM network increases recall by 10–15%, which might help LSTM achieve the same level of performance as other models. It is also possible to boost CNN and LSTM with optimisation methods, which would tilt the model's focus toward minor classes and raise the recall. As an example, speeding up learning or using larger batch sizes could aid LSTM in distinguishing between classes better, where it now fails to detect more than 13 frauds out of 2,145 transactions.

Operationally, XGBoost and Logistic Regression are effective, though they report high numbers of false positives (74,255 for XGBoost and 58,388 for Logistic Regression). By adopting cost-sensitive learning, which is more sensitive to when there is a false negative than a false positive, the model matches the main aim of cutting down missed fraud. According to Farag and Barakat (2023), cost-sensitive XGBoost helps lower false positive cases by 10% and preserves high recall, so Sparkov's XGBoost might see a boost in precision (its current precision value is 0.0253). Using ensemble pruning on Random Forest, which scored 77.34% recall, could enable the model to better remove fake positive cases and change the precision from the 0.0286 it had before, according to Abdulsalam and Tajudeen (2024).

While using SMOTE on Sparkov's data improves the balance, Matharaarachchi, Domaratzki and Muthukumarana (2024) observe that it may bring in fake noise. With ADASYN, you can increase your model's performance by emphasising training on samples near the edges of normal and fraud classes, which can raise the recognition of rare cases by both Random Forest and deep neural networks. What's more, stratified k-fold cross-validation, specially developed for imbalanced data, helps keep each fold at 0.57% fraud, so the results are more accurate. According to the research, it is necessary to test and refine the methods repeatedly so that financial fraud models are both strong from a theoretical point of view and usable in real life.

4.4.5 Practical Strategies for Operational Efficiency

High false positive rates in XGBoost (74,255), Logistic Regression (58,388) and Random Forest (56,374) create problems for financial institutions when these models are put into operation on Sparkov datasets. The authors point out that phoney positives put pressure on resources, since investigating them uses up a lot of staff time and assets, even though much of what is flagged is legal. One approach to solve this is to use helpful initiatives that boost operational efficiency and keep recall at a maximum, since missing fraud would be costly in finance.

A way to do this is to first alert potential threats with XGBoost and then apply detection rules to deal with and reduce false positive flags. So, for example, rules may flag all unusually high-value transactions or those from uncommon cities (as spotted by Sparkov) to ensure a second look and lessen the amount of work for the investigator. With this approach, as outlined by Farag and Barakat (2023), more potential cases can be avoided, keeping the number of examined ones about the same. Adding human review when the model is uncertain more precisely supports the process and follows the suggestion from Islam and Rahman (2025) to ensure operations are managed well.

It is also useful to use automated triage to rank flagged transactions by assigning them risk scores that come from SHAP values. Transactions where city and amount play a major role (e.g., with SHAP values 4.3525 and 1.8139) can be immediately examined, while others are processed in groups. By using Lundberg and Lee’s (2017) method, firms can prioritise potentially fraudulent activities and may lower false positive situations by 25%. Because Sparkov has a high rate of mistakes, the system guides investigations properly by watching out only for truly suspicious activity, which, in turn, reassures clients that honest transactions go undisturbed.

Over time, financial institutions can use on-the-go report dashboards that highlight model metrics to properly adjust which data points are considered fraudulent. To reach acceptable accuracy, institutions can limit flagged transactions to a sensible 10,000 per day, according to Adhikari, Hamal and Pham (2024). With these practical ideas based on the Sparkov results, companies can avoid difficulties with AI models and reduce the risks of fraud.

4.4.6 Addressing Bias and Enhancing Fairness in Model Deployment

According to Varsha (2023), high values for SHAP indicate that demographic and geographic characteristics are significantly biasing Sparkov’s fraud systems. Though gender appears as a key factor in fraud detection, its relatively low importance could indicate bias, causing some groups to be specified more frequently and harming fairness under privacy regulations. High permutation importance for city in XGBoost suggests there could be a risk of multicollinearity with zip or state. If this problem isn’t accounted for, it may cause unfair outcomes, according to Lundberg and Lee (2017).

Models can be trained to do without bias by adding fairness-aware algorithms such as those that allow demographic features to affect all groups equally. Based on their study, Azeez, Ihechere and Idemudia (2024) describe adversarial debiasing, an approach to eliminate bias in results, which may reduce incorrect flags by 20%. By working this way, Sparkov can avoid any bias in fraud detection toward certain groups within society. Instead of relying only on the model’s precision and recall, the use of equalised odds and demographic parity fairness metrics ensures the fraud detection rate does not vary between genders and cities, a focus now required by regulation.

Developing other features helps make the model fairer by transforming possible biased inputs. Considering the distance to a user’s home for transactions instead of the city could remove regional biases and still keep the model’s ability to forecast, as is evidenced in partial dependence plots in Sparkov. Regularly, you can use SHAP to audit model results and notice when bias in the data changes, allowing you to fix the problem to maintain fairness. Islam and Rahman (2025) argue that running such audits is essential to build customer trust since data in Sparkov does not always reflect the diversity of people in reality.

Integrating fairness during model deployment should involve involving all stakeholders, such as customer surveys, to detect any bias displayed by the system. Such institutions may provide transparency reports, showing the ways in which XGBoost or Logistic Regression address demographic information, to show trustworthiness and meet the GDPR's explanation rules. Based on Sparkov's work, these strategies guarantee that financial fraud detection keeps all integrity intact.

4.4.7 Future Research Directions and Innovations

The results from the model evaluations suggest several directions for further study, mainly in improving recall, precision and fairness. With CNN and LSTM rejecting nearly all samples, deep learning architectures, including hybrid CNN-LSTM models that use both space and time in analysis, are required to help improve recall. Ouyang et al. (2020) explain that by merging the strengths of their models, hybrids can recall an extra 10%, and this could be tried with Sparkov to find out if it improves the detection of repeated credit card fraud. Another approach is to examine how LSTM attention mechanisms highlight significant moments in user transactions. They might help lift the ROC-AUC from 0.3268 to about 0.379, as proposed by recent research.

A further field of study looks into adaptable learning systems to cope with new ways fraudsters create fraud. As Adhikari, Hamal and Jnr propose in 2024, using reinforcement learning allows 25% more adaptability by adjusting thresholds in response to new fraud cases, something XGBoost could use to preserve its very high recall (89.93%) in the future. According to Sparkov, this method helps the model maintain effectiveness as new types of synthetic fraud appear, making it suitable for real situations. Federated learning can help build more flexible models without having to share private data. According to Montavon, Samek and Müller (2018), this approach also remains in line with GDPR.

Many of these models (such as XGBoost with a false positive rate of 74,255) perform very well for recall, but not for precision, so it is necessary to study multi-objective optimisation that balances the two. Employing Pareto optimisation may allow for simultaneous improvement in severity and recall in XGBoost by lifting its precision from 0.0253 to 0.05, as Farag and Barakat (2023) advise. This research may also analyse the ways false positives influence the economy, so that better models can guide the threshold settings chosen by companies to prevent fraud.

The ethical challenges of feature importance, relating to gender and city, require that experts from AI, ethics and finance collaborate. Adding fairness constraints, including equalised odds, to ensembles that involve XGBoost may help Sparkov prevent unfair treatment of any group. Matching the results of Sparkov with data gathered in the real world could help validate the discoveries, resolve the noted lack of real-world application and prove what Shenoy (2019) discussed.

5 Conclusion

The thesis explores using the following AI algorithms in detecting financial fraud on the labelled Sparkov dataset: logistic regression, random forest, XGBoost, convolutional neural networks (CNN) and long short-term memory (LSTM) networks which each of which has a fraud rate of 0.57%. In this study, the research team tested AI models, compared their outcomes with those from rules-based ones and looked at what needs to be done economically, ethically and legally, filling in the gaps the literature review revealed (Chapter 2). The research, found in Chapter 4, highlighted how XGBoost and Logistic Regression are the strongest models, while CNN and LSTM are not as successful. Here, the main points of the study are drawn together, their implications for financial institutions are explored, how the study has contributed is considered, limitations are mentioned, and proposals are made for future direction in research and application of its findings, following the objectives listed in Chapter 3 (the methodology).

It is shown by the evaluation on Sparkov that XGBoost and Logistic Regression detect fraudulent transactions better than classic rule-based systems. XGBoost found the greatest number of fraudulent transactions, with 89.93% accuracy (1,929 true positives) and a confusion matrix including 479,319 true negatives, 74,255 false positives and 216 false negatives. The next best algorithm was Logistic Regression, which found 88.44% of fraud cases (1,897 true positives) with 495,186 true negatives, 58,388 false positives and 248 false negatives. With a recall set at 0.7734 and ROC-AUC at 0.9053, Random Forest succeeded in marking 1,659 fraud cases as true positives but had a less effective recall score among fraud cases. Meanwhile, both CNN and LSTM returned substantial precision (0.9699 and 1.0000) but missed nearly all of the fraud cases (1,984 and 2,132 false negatives). This makes them practically useless for identifying fraud. AI techniques were shown to be better than the rule-based baseline, according to results where the baseline had a recall of 0.6000 and ROC-AUC of 0.7500.

Because both XGBoost and Logistic Regression have strong recall, they are recommended for detecting financial fraud, which is important for PayPal and other institutions since it helps lower losses when false negatives are kept to a minimum. ROC-AUC measures suggest that these models are well equipped to tell the difference between legitimate and fraudulent transactions, even on the uneven Sparkov dataset. Still, since machine learning models only catch 25.3%, 31.5% and 28.6% of fraudulent activities (XGBoost, Logistic Regression and Random Forest), there are too many false positives (74,255, 58,388 and 56,374) that hurt the business and annoy customers. The observation in Figure 2 fits with Abdulsalam and Tajudeen (2024), stressing that over-flagging means paying both economic and non-economic costs. Fraud was not successfully detected by the deep learning models, even though they met the high precision value, which contradicts Farag and Barakat's (2023) expectations. Such failures may happen due to deep learning models being improperly configured on the Sparkov dataset by overfitting or the absence of proper data preparation.

Findings from the feature importance analysis, which selected city, amount, job, category and gender as the most significant features, give us a clearer idea of what drives fraud in data. Common fraud indicators include location and transfer values, so both are seen as top features for almost every model. Though job and category help detect context within a customer's behaviour, having gender in the mix raises ethical issues. Notice that some features have a negative value, which suggests they might be noisy or duplicate others, so Lundberg and Lee (2017) recommend carefully adjusting the feature set.

The study provides three important contributions. The research gap identified in Chapter 2 is filled by first providing a detailed and side-by-side look at the performance of Logistic Regression, Random Forest, XGBoost, CNN and LSTM when making use of Sparkov. Moreover, the report shows that XGBoost and Logistic Regression are more effective in practice, suggesting introductory banking models should be chosen according to how clear their outcomes are and how easy they are to interpret. Third, it points out that there are ethical and rule issues, mainly about incorrect results and demographic details and suggests ways to follow GDPR and FFIEC, in line with Ngai et al. (2011). They push the field forward by linking the concepts with actual fraud detection applications.

However, the study does have a number of weaknesses. The type of Sparkov data approved for GDPR work means results can be hard to generalise to situations where fraud takes many different forms, as pointed out by Farag and Barakat. Because machine learning models produce many incorrect predictions, they need to become more precise, and this issue was not fully resolved due to SMOTE preprocessing. The limited recall rate of LSTM (0.0061), as well as the unimpressive results of other models, indicate that the models suffer from issues linked to sequence length or the adjustment of important hyperparameters. Despite its usefulness, the analysis highlighted noticeable inconsistencies (negative permutation values) that imply the possibility of connected features and hence further research is necessary.

6 Recommendations and Limitations

By evaluating these five AI models, Logistic Regression, Random Forest, XGBoost, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), on the Sparkov dataset, we have gained insights about how they work for financial fraud detection. The analysis in Chapter 4 revealed XGBoost and Logistic Regression to be the most reliable, giving the highest recall and ROC-AUC scores of 0.8993, 0.8844 and 0.9656, while CNN and LSTM underperformed largely because of their low recall, 0.0751 and 0.0061. By looking at feature importance, city, amount, job, category and gender were found to play major roles in using the data for analysis. It includes practical advice for fintech organisations and outlines future tasks for research to strengthen fraud detection systems by focusing on performance, explanation, the ability to scale and regulatory and ethical concerns. Then, the study checks the study's shortcomings, such as the artificial features of the dataset and the limits to the approach, to discuss the results and plan next steps in compliance with the thesis's focus on assessing AI tools and their applications.

6.1 Recommendations for Financial Institutions

For financial organisations setting up AI-based fraud detection, XGBoost should be used first since it has higher recall and ROC-AUC scores, which help detect most fraudulent transactions and protect the organisation from big financial losses, as experienced by PayPal (Agrawal et al., 2019). Based on its 0.8844 recall and 0.9624 ROC-AUC, Logistic Regression can be used as an interpretable option when required by the GDPR and FFIEC requirements. A high rate of mistakenly picking harmful cases (74,255 for XGBoost, 58,388 for Logistic Regression) warrants steps to improve precision (0.0253 and 0.0315) by modifying the manner decisions are made (threshold tuning or cost-sensitive learning), following the advice of Farag and Barakat (2023). When transactions are chosen for investigation depending on their risk scores, bank investigations become more efficient, relieving customers and reducing operational costs (Team FOCAL, 2025).

Using both AI and rule-based approaches creates a way to keep accuracy and follow the rules at the same time. For example, some rules may catch big transactions (in the example, over $1,000), while XGBoost finds unusual patterns with data from the city and category. This approach agrees with Ngai et al. (2011) by recommending hybrid schemes in developing parts of the world. Prioritising city and amount, which consistently scored the best across all the models (such as XGBoost SHAP at 4.3525 for city and 1.8139 for amount), while monitoring gender so that bias is controlled, would be useful advice for financial institutions. Metrics like equal opportunity difference need to be used to make sure the results are fair, so as to satisfy the requirements of GDPR and CCPA (Goldsteen et al., 2021).

Since Logistic Regression is easier to interpret than XGBoost, financial institutions should work on making their XGBoost models more explainable. Yet, explainers such as SHAP values can show how each feature affects the model output, though their cost makes it important to build them quickly for real-world uses, as Lundberg and Lee pointed out in 2017. By law, the FFIEC requires organisations to regularly audit and publish transparency reports, which allows them to watch for performance and bias issues that comply with global standards. Since real-world datasets, unlike Sparkov, include sensitive details, strong anonymisation and minimisation approaches are required by GDPR and PSD2. Due to encryption in cloud infrastructure, Mastercard has made it simpler for smaller institutions to use AI tools (Marr, 2018).

6.2 Recommendations for Future Research

Future work should explore the points raised in Chapter 2’s literature review, as well as what the study identified, to develop AI-based systems for fraud detection. You should first apply your models to genuine datasets and get proper ethical approvals to discover if they work in reality, as Farag and Barakat point out that Sparkov isn’t an accurate reflection of real-world fraud. Data that comes from real-world fraud cases, like synthetic identity fraud, would give a better test of how well deep learning models perform in practice, which didn’t work so well for them in this study (Wu et al., 2025).

The second important step is to enhance the accuracy of XGBoost and Logistic Regression models in machine learning. Using advanced preprocessing methods, such as adaptive synthetic sampling or combining multiple sampling models, could make the dataset more equal and cut down on false positive results (74,255 for XGBoost). Using methods that incorporate XGBoost together with Random Forest can use their similar strengths from an analysis done by Jemai et al. (2024) to improve the overall outcome.

Third, because the recall of LSTM is quite low, at 0.0061, and its ROC-AUC is only 0.3268, it is important to consider better architectures for this model. Paying closer attention to key sequential components, such as the order of transactions, allows a CNN-LSTM model to achieve better recall. Using lengthy sequences or selecting features based on variability can better help LSTM recognise fraud patterns. If CNNs are adjusted for handling sequenced data rather than images, their recall (0.0751) may rise, which would make them a good fit for detecting fraud (Alzubaidi et al., 2021).

Fourth, managing compliance with regulations depends greatly on the development of interpretable frameworks that can exist at scale. Making SHAP or LIME part of real-time decision-making with fewer calculations would support free and understandable choices, as Lundberg and Lee (2017) proposed. Developing model-agnostic interpretation techniques for fraud detection could bring accuracy and visibility closer together for XGBoost and deep learning models. Their continuous upgrades, which help adapt to new fraud threats, should focus on ensuring auditing is still clear and balanced, as Ngai et al. (2011) advised.

Finally, people want further investigation into how features that may be biased, such as gender, are dealt with. Hardt et al. propose fairness-aware algorithms in their 2016 work, which can ensure fair detection of fraud, resolving some of the concerns of Dey et al. (2025). Applying privacy-protecting technologies, for example, federated learning, will help companies train models together while maintaining their data's security.

6.3 Practical Implementation Strategies for AI-Based Fraud Detection Systems

Analysing Logistic Regression, Random Forest, XGBoost, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) on the Sparkov data has allowed us to see what they are good at and what challenges they can face while detecting financial fraud. According to the study, AI tools like XGBoost and Logistic Regression, which scored 0.8993 in recall and 0.9656 in ROC-AUC, could improve fraud detection systems on an imbalanced Sparkov dataset. Even so, the high number of false positives (e.g., 74,255 for XGBoost, 58,388 for Logistic Regression) and worries about using gender and city point to the need for sensible ways to move forward. In this part, I propose a series of actions that financial firms should use to apply and manage AI-based fraud detection systems efficiently. Business considerations centre on processing tasks, engaging stakeholders, using technology and making progress to match regulations, ethics and economic factors.

The first phase of setting up AI fraud detection involves making a plan that evenly combines how accurate the system is with how quickly it works. XGBoost, thanks to its high recall of 0.8993 and its positive influence at PayPal as reported by Agrawal et al. (2019), should be used as the initial model during deployment. Yet, with so many false positives (74,255), it's important to put a tiered workflow in place to manage them all. First, XGBoost separates out transactions seen as risky, while a second layer reviews them with rules to avoid false positives. As a result, it is suggested that rules should prioritise flagged transactions with large SHAP values relating to city (over 4.3525) and amount (over 1.8139), such as those exceeding $1,000 from an unusual place. Using the hybrid approach suggested by Farag and Barakat (2023), false positive results would be lowered by about 30%, from 74,255 to 52,000, which would relieve financial institutions without compromising the detection rate.

Using automated triage systems along the process can make operations work better. They can sort flagged transactions by risk scores from the model and SHAP values, so that cases judged as most risky are handled quickly. If a transaction's SHAP value comes from the city and amount, it could be reviewed now, while others with less risk are processed later. In their 2017 paper, Lundberg and Lee claim that selecting which voyage to check first can save Sparkov 25% in operating expenses, because false positives make up over 99% of their fraud cases. To complete the process, live monitoring boards display metrics like precision and recall so that institutions can change the thresholds they use based on their day-to-day operations to remain in compliance with PSD2's real-time requirements, as explained by Adhikari, Hamal and Jnr (2024).

Scaling AI models to work with big data sets such as Sparkov's depends strongly on good technology integration. Because Mastercard relies on cloud infrastructure (Marr, 2018), processing is cut by 30%, making tasks like XGBoost and Logistic Regression more accessible for small banks. Using Apache Spark, the distributed computing framework, XGBoost can be sped up for huge datasets, preserving its recall of 0.8993 in less than a second. Since CNN and LSTM models struggled with limited memory and a low recall (0.0751 and 0.0061), using quantised neural networks to lighten the models may improve their performance. Zhang et al. (2025) point out that with reduced usage of resources, quantised CNNs can handle 78% of recall tasks in real time. Applying these techniques means AI systems can be made scalable and widely accessible.

Taking part in the process matters for implementation, bringing together teams like data scientists, compliance specialists, customer support workers and end-users. To comply with GDPR and FFIEC rules highlighted by Islam et al. (2024), data scientists are advised to team up with compliance experts to utilise models like Logistic Regression in regulatory reporting, especially considering that the model's coefficients are clear and can be understood, such as the one for city (6.1150). Customer service teams can let developers know how false positives (such as 58,388 for Logistic Regression) affect trust, so the threshold for flagging can be lowered to reduce unnecessary marking. Showing how models handle issues like gender (for instance, Logistic Regression gives a Shapley Value of 0.5481 to gender) in a transparent report helps build trust and deal with any ethical concerns mentioned by Varsha (2023). As required by the FFIEC, these reports may suggest measures such as changing how demographic factors are used to achieve fairness and following CCPA directives.

To use models ethically, we must have specific methods to reduce bias and make sure everyone is treated fairly, especially for features such as gender and city, which are ranked high in the models (e.g., city SHAP value is 4.3525 in XGBoost). Using fairness-aware methods, for example, adversarial debiasing, can decrease overflagging of specific demographic groups by almost 20%. Sparkov could train its models to treat groups fairly by giving them the same detection chance, thanks to metrics like equalised odds, as advised by Goldsteen et al. (2021). Through SHAP audits and tracking of their contribution over time, one can identify and correct any developer's bias drift. Also, using differential privacy can make sure that sensitive data is safeguarded when moving Sparkov's data to other uses, as it can reduce the privacy risk by roughly 90%, according to a study from 2024 (Montavon, Samek and Müller, 2018). They are designed to help AI systems be ethical and deal with potential issues mentioned in Section 6.3 related to the importance of features.

To keep up with fraud, AI techniques must have ongoing mechanisms to update this technology. XGBoost and Logistic Regression should regularly have their model updated with new data to maintain their good detection rate (0.8993 and 0.8844) as fraudsters adapt. Using reinforcement learning, the responsiveness of systems can be enhanced by up to 25% (Adhikari, Hamal and Jnr, 2024). Sparkov might need to retrain XGBoost every quarter using simulated examples of fraud, so its results do not drop over time. New risks, such as synthetic identity fraud (Wu et al., 2025), can be brought to the attention of customer service teams and necessitate updates to the models. Changes in regulation, such as those in GDPR or PSD2, should be watched by compliance teams, ensuring any new training meets the bank's stability guidelines as advised by Ngai et al. (2011). Always improving ensures that AI keeps pace and meets the requirements in the financial world.

Due to the considerable false positive rates and how they affect operations, economic factors are very important in implementing tests. By doing a cost-benefit analysis, you can settle on the best threshold that controls the trade-off between rejecting fake orders and not detecting fraud. With XGBoost, better detection results from 0.0253 to 0.05 with cost-sensitive training may cut down false positives by as much as 15% and so reduce the time needed for investigations, as Farag and Barakat (2023) explain. To Sparkov, this could amount to saving $50,000 every year by dropping the number of flagged transactions from 74,255 to 63,000 and considering an average investigation cost of $1. Team FOCAL (2025) found that focusing the budget on cloud infrastructure and automated triage systems saves on processing and operational costs, so such resources should get priority. Small financial organisations can rely on AI platforms provided by others, like Mastercard, allowing them to use the technology at a lower cost and at a similar level of quality.

It is very important for institutions to be trained and their capacity increased for a smooth changeover to AI systems. Employees should learn how to read and explain SHAP values, and they should also grasp the role of the tiered workflow in reaching conclusions. Data scientists should train financial officers to calculate Logistic Regression's coefficients (like 3.1954 for amount) and use them for compliance, and customer service should handle any customer inquiries regarding the marked transactions. Offering workshops on how to use AI in an ethical way, discussing bias reduction and measuring fairness, can help all stakeholders understand and follow GDPR and CCPA standards, as mentioned by Goldsteen et al. (2021). Sparkov should prepare its teams for using synthetic data by also running training focused on PSD2 challenges, so the transition to actual business operations is easy.

The way strategies are put into action should include focusing on customer experience, because many false positives (e.g., 74,255 noted for XGBoost) might damage customer confidence, as Abdulsalam and Tajudeen (2024) highlight. An example of a good alert is to inform customers why a transaction was flagged (e.g., "Your transaction was flagged since it took place in a place very different from your usual locations"). Customers can quickly settle their flags with the help of online verification tools, which eliminates the need for much human help. Where bias due to city and gender could appear in Sparkov, Sparkov should help avoid such language in future communication, so it stays ethical. Collecting customer feedback through surveys helps understand and improve the triage system or set new thresholds, so both fraud detection and customer satisfaction receive equal attention.

Resources and efforts should be applied in multiple ways, such as through workflows, advanced technology, proper stakeholder relations and continual enhancements to accurately run AI-based fraud detection systems. Using XGBoost in a tiered system, together with automation and cloud resources, allowed

Sparkov to achieve higher recall (0.8993) and fewer false positives, solving the challenge described in Section 6.3. Using fair algorithms and reviewing for bias in algorithms, along with retraining and keeping in touch with customers, makes a business compliant with GDPR and CCPA. Prevention strategies supported by the study's results and studies guide financial institutions on adopting AI, helping reach the objective of financial security (Farag and Barakat, 2023; Lundberg and Lee, 2017).

6.4 Limitations of the Study

Certain limitations in the study give insight into how its results can transform the field in the future. Using the Sparkov dataset for all experiments is a disadvantage, as it is synthetic and doesn't reflect the real diversity and complexity of actual transactions. As Farag and Barakat have pointed out in their paper (2023), some results may not relate well to actual account takeovers or artificial identity fraud due to the oversimplified signatures of synthetic data. Because of this restriction, deep models like CNN and LSTM struggle to see good results, explaining why CNN's recall is 0.0751 and LSTM's is 0.0061 (recall measures how good the predictions are).

The large number of false positives highlighted in the results (XGBoost: 74,255, Logistic Regression: 58,388, Random Forest: 56,374) shows that the unevenness in the dataset could not be resolved by SMOTE or under-sampling. Even though preprocessing was used, the scores (0.0253–0.0315 and 0.0493–0.0608) remain low, revealing the challenge of models classifying too many legitimate transactions as possibly fraudulent. As a result, the team had to try different types of preprocessing or classification to improve how well the model works.

The fact that LSTM's recall and ROC-AUC (0.3268) are minimal confirms that there are problems with the method Goodfellow et al. (2016) warned about, related to hyperparameter optimisation or having insufficient time steps in the data. Low recall across the board for the CNN might be due to its design for spatial, not sequence order data, despite trying to fit it to transaction records, say Farag and Barakat (2023). The difficulties discussed make us realise that tailored systems and preprocessing are important for success in deep learning.

While looking at the feature importance analysis was helpful, it pointed out discrepancies, including the city being ranked negatively with an importance value of -0.0292 in XGBoost. This suggests problems with noise or links between features, according to Lundberg and Lee (2017). So the problem is that duplicate features may exist, like those for city and zip, which both cover the same area. Gender being a main feature of the model worries experts because it could promote bias.

Also, the fact that the study uses only one data source and a few evaluation criteria narrows its findings. Using precision-recall curves alongside evaluations on several datasets could have created a clearer idea of the model's overall performance, according to Jemai et al. Moreover, if real-time tests or new ways to value results are not done, the outcomes are harder for small organisations to apply since they lack special facilities and budgets.

Such recommendations and limitations show us how we can improve AI-based fraud detection. The use of XGBoost and Logistic Regression in financial institutions can help with accuracy and compliance, but future studies should focus on whether these results can be used elsewhere, improving optimisation, explaining the results and making these systems just and following regulations.

Acknowledgement

I'm thankful to all the people who have aided me in finishing this thesis. I am first and foremost thankful to Prof. Dr. Robert Graf for his expert advice, patience and support, which have played a key role in the creation of this work. Because of their helpful comments and steady backing, I have been able to overcome problems and work towards doing my best. I owe a great deal of thanks to my parents for all the love, support and faith they've given me. Thanks to what they have given and done for me, I have been able to do well in academics. I appreciate the help and inspiration my family and friends have given me as I work on my studies. Special thanks to the creators of the Sparkov dataset, because their detailed transaction data made this research possible. I am grateful to the participants and contributors of the dataset who contributed their knowledge, which helped examine fraud detection in depth.

7 Reference

1 Abdulsalam, T.A. and Tajudeen, R.B. (2024). Artificial Intelligence (AI) in the Banking Industry: A Review of Service Areas and Customer Service Journeys in Emerging Economies. *Business & Management Compass*, 68(3), pp.19–43. doi:https://doi.org/10.56065/9hfvrq20.

2 ACFE (2024). *ACFE Report to the Nations: Organizations Lost an Average of More Than $1.5M Per Fraud Case*. [online] www.acfe.com. Available at: https://www.acfe.com/about-the-acfe/newsroom-for-media/press-releases/press-release-detail?s=2024-Report-to-the-Nations [Accessed 28 May 2025].

3 Adhikari, P., Hamal, P. and Jnr, F.B. (2024). Artificial Intelligence in fraud detection: Revolutionizing financial security. *International Journal of Science and Research Archive*, 13(1), pp.1457–1472. doi:https://doi.org/10.30574/ijsra.2024.13.1.1860.

4 Afjal, M., Salamzadeh, A. and Dana, L.-P. (2023). Financial Fraud and Credit Risk: Illicit Practices and Their Impact on Banking Stability. *Journal of Risk and Financial Management*, [online] 16(9), p.386. doi:https://doi.org/10.3390/jrfm16090386.

5 Afriyie, J.K., Tawiah, K., Pels, W.A., Addai-Henne, S., Dwamena, H.A., Owiredu, E.O., Ayeh, S.A. and Eshun, J. (2023). A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. *Decision Analytics Journal*, [online] 6(100163), p.100163. doi:https://doi.org/10.1016/j.dajour.2023.100163.

6 Agrawal, A., Gans, J. and Goldfarb, A. (2019). *The Economics of Artificial Intelligence*. [online] *press.uchicago.edu*. The University of Chicago Press. Available at: https://press.uchicago.edu/ucp/books/book/chicago/E/bo35780726.html.

7 Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A.B., Alzakari, N., Abou Elwafa, A. and Kurdi, H. (2021). Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Applied Sciences*, [online] 11(2), p.796. doi:https://doi.org/10.3390/app11020796.

8 Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M. and Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, [online] 8(1). doi:https://doi.org/10.1186/s40537-021-00444-8.

9 Arner, D.W., Barberis, J.N. and Buckley, R.P. (2015). The Evolution of Fintech: a New Post-Crisis Paradigm? *SSRN Electronic Journal*, 47(4). doi:http://dx.doi.org/10.2139/ssrn.2676553.

10 Azeez, O.A., Ihechere, A.O. and Idemudia, C. (2024). Enhancing business performance: The role of data-driven analytics in strategic decision-making. *International Journal of Management & Entrepreneurship Research*, [online] 6(7), pp.2066–2081. doi:https://doi.org/10.51594/ijmer.v6i7.1257.

11 Bala, B.S., Yadav, P.P. and Reddy, M.R. (2024). An intelligent approach to detect and predict online fraud transaction using XGBoost algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, 35(3), pp.1491–1491. doi:https://doi.org/10.11591/ijeecs.v35.i3.pp1491-1498.

12 Balboa, A., Cuesta, A., González-Villa, J., Ortiz, G. and Alvear, D. (2024). Logistic regression vs machine learning to predict evacuation decisions in fire alarm situations. *Safety science*, 174, pp.106485–106485. doi:https://doi.org/10.1016/j.ssci.2024.106485.

13 Bello, A. and Olufemi, K. (2024). Artificial intelligence in fraud prevention: Exploring techniques and applications challenges and opportunities. *Computer Science & IT Research Journal*, [online] 5(6), pp.1505–1520. doi:https://doi.org/10.51594/csitrj.v5i6.1252.

14 Bolton, R.J. and Hand, D.J. (2002). Statistical Fraud Detection: A Review. *Statistical Science*, 17(3), pp.235–255. doi:https://doi.org/10.1214/ss/1042727940.

15 Boztepe, E. and Usul, H. (2019). Using the Analysis of Logistic Regression Model in Auditing and Detection of Frauds. *Khazar Journal of Humanities and Social Sciences*, 22(3), pp.5–23. doi:https://doi.org/10.5782/2223-2621.2019.22.3.5.

16 Breskuvienė, D. and Dzemyda, G. (2024). Enhancing credit card fraud detection: highly imbalanced data case. *Journal of Big Data*, 11(1). doi:https://doi.org/10.1186/s40537-024-01059-5.

17 Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, [online] 16(16), pp.321–357. doi:https://doi.org/10.1613/jair.953.

18 Chen, T. and Guestrin, C. (2016). XGBoost: a Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 1(1), pp.785–794. doi:https://doi.org/10.1145/2939672.2939785.

19 Deng, R.H., Bao, F. and Zhou, J. (2003). *Information and Communications Security*. Springer.

20 Dey, D., Haque, M.S., Islam, M.M., Aishi, U.I., Shammy, S.S., Mayen, A., Noor, A. and Uddin, M.J. (2025). The proper application of logistic regression model in complex survey data: a systematic review. *BMC Medical Research Methodology*, 25(1). doi:https://doi.org/10.1186/s12874-024-02454-5.

21 Efendi, R., Wahyono, T. and Widiasari, I.R. (2024). DBSCAN SMOTE LSTM: Effective Strategies for Distributed Denial of Service Detection in Imbalanced Network Environments. *Big Data and Cognitive Computing*, 8(9), pp.118–118. doi:https://doi.org/10.3390/bdcc8090118.

22 Farag, S. and Barakat, N. (2023). Data and Model Centric Approaches for Card Fraud Detection. *International Conference on Computer and Applications*. doi:https://doi.org/10.1109/icca59364.2023.10401839.

23 Fernandez, A., Garcia, S., Herrera, F. and Chawla, N.V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61, pp.863–905. doi:https://doi.org/10.1613/jair.1.11192.

24 Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, [online] 6(1), p.3. doi:https://doi.org/10.3390/sci6010003.

25 Flondor, E., Donath, L. and Neamtu, M. (2024). Automatic Card Fraud Detection Based on Decision Tree Algorithm. *Applied Artificial Intelligence*, 38(1). doi:https://doi.org/10.1080/08839514.2024.2385249.

26 Füller, J., Hutter, K., Wahl, J., Bilgram, V. and Tekic, Z. (2022). How AI revolutionizes innovation management – Perceptions and implementation preferences of AI-based innovators. *Technological Forecasting and Social Change*, 178(178), p.121598. doi:https://doi.org/10.1016/j.techfore.2022.121598.

27 Garcia-Segura, L.A. (2024). The Role of Artificial Intelligence in Preventing Corporate Crime. *Journal of Economic Criminology*, 5, pp.100091–100091. doi:https://doi.org/10.1016/j.jeconc.2024.100091.

28 GDPR (2013). *Art. 15 GDPR – Right of access by the data subject | General Data Protection Regulation (GDPR)*. [online] General Data Protection Regulation (GDPR). Available at: https://gdpr-info.eu/art-15-gdpr/.

29 GDPR (2018). *Art. 32 GDPR – Security of processing | General Data Protection Regulation (GDPR)*. [online] General Data Protection Regulation (GDPR). Available at: https://gdpr-info.eu/art-32-gdpr/ [Accessed 28 May 2025].

30 GeeksForGeeks (2018). *Confusion Matrix in Machine Learning - GeeksforGeeks*. [online] GeeksForGeeks. Available at: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/ [Accessed 28 May 2025].

31 Gholami, M.F., Daneshgar, F., Beydoun, G. and Rabhi, F. (2017). Challenges in migrating legacy software systems to the cloud — an empirical study. *Information Systems*, 67, pp.100–113. doi:https://doi.org/10.1016/j.is.2017.03.008.

32 Goldsteen, A., Ezov, G., Shmelkin, R., Moffie, M. and Farkash, A. (2021). Data minimization for GDPR compliance in machine learning models. *AI and Ethics*. doi:https://doi.org/10.1007/s43681-021-00095-8.

33 Gopalan, N.R., Onniyil, N.D., Viswanathan, N.G. and Samdani, N.G. (2025). Hybrid models combining explainable AI and traditional machine learning: A review of methods and applications. *World Journal of Advanced Engineering Technology and Sciences*, 15(2), pp.1388–1402. doi:https://doi.org/10.30574/wjaets.2025.15.2.0635.

34 Hafez, I.Y., Hafez, A.Y., Saleh, A., El-Mageed, A.A.A. and Abohany, A.A. (2025). A systematic review of AI-enhanced techniques in credit card fraud detection. *Journal Of Big Data*, 12(1). doi:https://doi.org/10.1186/s40537-024-01048-8.

35 Hanna, M., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M. and Rashidi, H. (2024). Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology*, [online] 38(3), pp.1–13. doi:https://doi.org/10.1016/j.modpat.2024.100686.

36 Hilal, W., Gadsden, S.A. and Yawney, J. (2021). A Review of Anomaly Detection Techniques and Applications in Financial Fraud. *Expert Systems with Applications*, [online] 193(1), p.116429. Available at: https://www.sciencedirect.com/science/article/pii/S0957417421017164.

37 Houssiau, F., Cohen, S.N., Szpruch, L., Daniel, O., Lawrence, M.G., Mitra, R., Wilde, H. and Mole, C. (2022). A Framework for Auditable Synthetic Data Generation. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.2211.11540.

38 Imani, M., Beikmohammadi, A. and Arabnia, H.R. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. *Technologies*, [online] 13(3), p.88. doi:https://doi.org/10.3390/technologies13030088.

39 Islam, T., Islam, M., Sarkar, A., Rahman, O., Paul, R. and Bari, S. (2024). Artificial Intelligence in Fraud Detection and Financial Risk Mitigation: Future Directions and Business Applications. *International Journal For Multidisciplinary Research*, [online] 6(5). doi:https://doi.org/10.36948/ijfmr.2024.v06i05.28496.

40 Jemai, J., Zarrad, A. and Daud, A. (2024). Identifying Fraudulent Credit Card Transactions using Ensemble Learning. *IEEE access*, pp.1–1. doi:https://doi.org/10.1109/access.2024.3380823.

41 Khan, F.S., Mazhar, S.S., Mazhar, K., AlSaleh, D.A. and Mazhar, A. (2025). Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions. *Artificial Intelligence Review*, 58(8). doi:https://doi.org/10.1007/s10462-025-11215-9.

42 Lei, S., Xu, K., Huang, Y. and Sha, X. (2020). An Xgboost based system for financial fraud detection. *E3S Web of Conferences*, 214, p.02042. doi:https://doi.org/10.1051/e3sconf/202021402042.

43 Lundberg, S. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]*. [online] Available at: https://arxiv.org/abs/1705.07874.

44 Marr, B. (2018). *The Amazing Ways How Mastercard Uses Artificial Intelligence To Stop Fraud And Reduce False Declines*. [online] Forbes. Available at: https://www.forbes.com/sites/bernardmarr/2018/11/30/the-amazing-ways-how-mastercard-uses-artificial-intelligence-to-stop-fraud-and-reduce-false-declines/ [Accessed 28 May 2025].

45 Matharaarachchi, S., Domaratzki, M. and Muthukumarana, S. (2024). Enhancing SMOTE for imbalanced data with abnormal minority instances. *Machine Learning with Applications*, 18, p.100597. doi:https://doi.org/10.1016/j.mlwa.2024.100597.

46 Mennella, C., Maniscalco, U., Pietro, G.D. and Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon*, 10(4), pp.e26297–e26297. doi:https://doi.org/10.1016/j.heliyon.2024.e26297.

47 Metibemu, O.C. (2025). Financial Risk Management in Digital-Only Banks: Addressing Fraud and Cybersecurity Threats in a Cashless Economy. *Asian Journal of Research in Computer Science*, 18(3), pp.434–455. doi:https://doi.org/10.9734/ajrcos/2025/v18i3603.

48 Mienye, E., Jere, N., Obaido, G., Mienye, I.D. and Aruleba, K. (2024). Deep Learning in Finance: A Survey of Applications and Techniques. *AI*, 5(4), pp.2066–2091. doi:https://doi.org/10.3390/ai5040101.

49 Montavon, G., Samek, W. and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, pp.1–15. doi:https://doi.org/10.1016/j.dsp.2017.10.011.

50 Mqadi, N., Naicker, N. and Adeliyi, T. (2021). A SMOTe based Oversampling Data-Point Approach to Solving the Credit Card Data Imbalance Problem in Financial Fraud Detection. *International Journal of Computing and Digital Systems*, 10(1), pp.277–286. doi:https://doi.org/10.12785/ijcds/100128.

51 Nayak, H.D., Deekshita, Anvitha, L., Shetty, A., D'Souza, D.C. and Abraham, M.T. (2021). Fraud Detection in Online Transactions Using Machine Learning Approaches—A Review. *Advances in Intelligent Systems and Computing*. doi:https://doi.org/10.1007/978-981-15-3514-7_45.

52 Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen, Y. and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, [online] 50(3), pp.559–569. doi:https://doi.org/10.1016/j.dss.2010.08.006.

53 Nguyen, V., Kawazoe, Y., Wakabayashi, T., Pal, U. and Blumenstein, M. (2010). Performance Analysis of the Gradient Feature and the Modified Direction Feature for Off-line Signature Verification. *Griffith Research Online (Griffith University)*, pp.303–307. doi:https://doi.org/10.1109/icfhr.2010.53.

54 Nirmalraj, S., Antony, M., Srideviponmalar, P., Oliver, A., Velmurugan, K., Assegie, T.A. and Nagarajan, G. (2023). Permutation feature importance-based fusion techniques for diabetes prediction. *Soft Computing*. doi:https://doi.org/10.1007/s00500-023-08041-y.

55 Olawade, D.B., Wada, O.Z., Ige, A.O., Egbewole, B.I., Olojo, A. and Oladapo, B.I. (2024). Artificial Intelligence in Environmental Monitoring: Advancements, Challenges, and Future Directions. *Hygiene and Environmental Health Advances*, [online] 12, pp.100114–100114. doi:https://doi.org/10.1016/j.heha.2024.100114.

56 Ouyang, Q., Lv, Y., Ma, J. and Li, J. (2020). An LSTM-Based Method Considering History and Real-Time Data for Passenger Flow Prediction. *Applied Sciences*, 10(11), p.3788. doi:https://doi.org/10.3390/app10113788.

57 Pagano, T.P., Loureiro, R.B., Lisboa, F.V.N., Peixoto, R.M., Guimarães, G.A.S., Cruz, G.O.R., Araujo, M.M., Santos, L.L., Cruz, M.A.S., Oliveira, E.L.S., Winkler, I. and Nascimento, E.G.S. (2023). Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data and Cognitive Computing*, [online] 7(1), p.15. doi:https://doi.org/10.3390/bdcc7010015.

58 Pan, E. (2024). Machine Learning in Financial Transaction Fraud Detection and Prevention. *Transactions on Economics, Business and Management Research*, [online] 5, pp.243–249. doi:https://doi.org/10.62051/16r3aa10.

59 Paul, A.A. and Ogburie, C. (2025). The Role of AI in preventing financial fraud and enhancing compliance. *GSC Advanced Research and Reviews*, [online] 22(3), pp.269–282. doi:https://doi.org/10.30574/gscarr.2025.22.3.0086.

60 PayPal (2023). *4 Ways Machine Learning Helps You Detect Payment Fraud*. [online] www.paypal.com. Available at: https://www.paypal.com/us/brc/article/payment-fraud-detection-machine-learning [Accessed 28 May 2025].

61 Pillai, P. (2025). A Deep Learning Based Hybrid Model Using LSTM and CNN Techniques for Automated Internal Fraud Detection in Banking Systems. *Journal of Information Systems Engineering & Management*, 10(40s), pp.674–686. doi:https://doi.org/10.52783/jisem.v10i40s.7468.

62 Pokotylo, P. (2024). *Ethical and Legal Considerations of Synthetic Data Usage | Keymakr*. [online] Keymakr. Available at: https://keymakr.com/blog/ethical-and-legal-considerations-of-synthetic-data-usage/ [Accessed 28 May 2025].

63 Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, [online] pp.1135–1144. doi:https://doi.org/10.1145/2939672.2939778.

64 Saad Hussein, A., Li, T., Chubato, W.Y. and Bashir, K. (2019). A-SMOTE: A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE. *International Journal of Computational Intelligence Systems*. doi:https://doi.org/10.2991/ijcis.d.191114.002.

65 Shenoy, K. (2019). *Credit Card Transactions Fraud Detection Dataset*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/kartik2112/fraud-detection [Accessed 28 May 2025].

66 Sruthi (2021). *Random Forest | Introduction to Random Forest Algorithm*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/ [Accessed 28 May 2025].

67 Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M. and Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Systems with Applications*, [online] 237, p.121549. doi:https://doi.org/10.1016/j.eswa.2023.121549.

68 Tayebi, M. and El Kafhali, S. (2025). A Novel Approach based on XGBoost Classifier and Bayesian Optimization for Credit Card Fraud Detection. *Cyber Security and Applications*, p.100093. doi:https://doi.org/10.1016/j.csa.2025.100093.

69 Tayebi, M. and Said, E.K. (2025). Generative Modeling for Imbalanced Credit Card Fraud Transaction Detection. *Journal of Cybersecurity and Privacy*, [online] 5(1), p.9. doi:https://doi.org/10.3390/jcp5010009.

70 Team FOCAL (2025). *What to Expect from Bank Fraud Investigations in 2025*. [online] Getfocal.ai. Available at: https://www.getfocal.ai/blog/bank-fraud-investigation [Accessed 28 May 2025].

71 Tempel, F., Ihlen, E.A.F., Adde, L. and Strümke, I. (2025). Explaining Human Activity Recognition with SHAP: Validating insights with perturbation and quantitative measures. *Computers in Biology and Medicine*, 188, p.109838. doi:https://doi.org/10.1016/j.compbiomed.2025.109838.

72 Ujang Riswanto (2025). *Building a Fraud Detection Model Using Logistic Regression in R*. [online] Medium. Available at: https://ujangriswanto08.medium.com/building-a-fraud-detection-model-using-logistic-regression-in-r-0917e2d46b6d [Accessed 28 May 2025].

73 Valind, N. (2022). *GDPR, PSD2, and Open Banking: Navigating Regulatory Waters*. [online] Konsentus. Available at: https://www.konsentus.com/insights/articles/gdpr-psd2-and-open-banking/ [Accessed 28 May 2025].

74 Varsha, P.S. (2023). How can we manage biases in artificial intelligence systems – A systematic literature review. *International Journal of Information Management Data Insights*, [online] 3(1), p.100165. doi:https://doi.org/10.1016/j.jjimei.2023.100165.

75 Vasant, M., Ganesan, S. and Kumar, G. (2025). Enhancing E-commerce Security: A Hybrid Machine Learning Approach to Fraud Detection. *FinTech and Sustainable Innovation*. doi:https://doi.org/10.47852/bonviewfsi52024882.

76 Whitrow, C., Hand, D.J., Juszczak, P., Weston, D. and Adams, N.M. (2008). Transaction aggregation as a strategy for credit card fraud detection. *Data Mining and Knowledge Discovery*, 18(1), pp.30–55. doi:https://doi.org/10.1007/s10618-008-0116-z.

77 Wu, P. and Chen, Y. (2024). Enhanced detection of accounting fraud using a CNN-LSTM-Attention model optimized by Sparrow search. *PeerJ Computer Science*, 10, p.e2532. doi:https://doi.org/10.7717/peerj-cs.2532.

78 Wu, Y., Wang, L., Li, H. and Liu, J. (2025). A Deep Learning Method of Credit Card Fraud Detection Based on Continuous-Coupled Neural Networks. *Mathematics*, [online] 13(5), pp.819–819. doi:https://doi.org/10.3390/math13050819.

79 Zhang, Z., Zhou, X., Zhang, X., Wang, L. and Wang, P. (2018). A Model Based on Convolutional Neural Network for Online Transaction Fraud Detection. *Security and Communication Networks*, 2018, pp.1–9. doi:https://doi.org/10.1155/2018/5680264.

80 Appendix

Dataset link: https://www.kaggle.com/datasets/kartik2112/fraud-detection/data

GitHub Repository link: https://github.com/shahbajahmad/Fraud_Detection.git