# INTRODUCTION

## Background

**Understanding Financial Fraud and it’s Impact**

In today’s fast-paced world, financial transactions are happening continuously, every millisecond. Most of the financial institutions are moving into digitalization in order to supply the emerging demand arising from consumers globally. With the significant growth of the industry and improved digitalization, there has been a corresponding increase in scammers and fraudsters. Consequently, fraudulent activities have become a major concern for businesses, banks and consumers from around the world as such activities are becoming both more sophisticated and more frequent.

Financial fraud can happen in many forms. Credit card frauds, insurance frauds, identity theft and money laundering are some of the most prominent ways of fraud in the present society. The implications of such transactions are affecting not only individuals and businesses but also the stability of entire financial systems. The global cost of financial fraud is staggering, with estimates running into trillions of dollars annually, which in turn impacts consumer trust, raises operational costs for businesses, and undermines economic growth (ACFE, 2022)

**Traditional Approaches to Financial Fraud Detection**

Given the significant impact of financial fraud, detecting and preventing such activities has become a critical area of focus. Over the years, various machine learning models have been deployed to detect fraudulent activities by analyzing patterns within large and complex datasets. Some of the most commonly used models include:

**Random Forest:** This ensemble learning method constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is known for its robustness and high accuracy, particularly in handling large datasets with many features (Breiman, 2001).

**Neural Networks:** These are inspired by the human brain's structure and function. Neural networks, particularly deep learning models, have shown remarkable performance in identifying complex patterns within data. However, they often require large amounts of labeled data and can be computationally expensive (LeCun, Bengio, & Hinton, 2015).

**XGBoost**: An advanced implementation of gradient-boosted decision trees, XGBoost is designed for speed and performance. It has been widely adopted for its ability to handle imbalanced datasets, which are common in fraud detection scenarios. However, like other tree-based methods, it may struggle with generalization if the fraud patterns are highly dynamic (Chen & Guestrin, 2016).

**Limitations of Traditional Models**

Despite the popularity of these traditional fraud detection models, there are concerns about their effectiveness in identifying new and evolving patterns of fraudulent transactions. These models highly rely on the historical data, rather than identifying evolving patterns of fraudulent transactions. This highlights a raising question about the abilities of these traditional models in detecting emerging fraud trends.

**Introduction to Retrieval-Augmented Generation (RAG)**

In recent years, advancements in the artificial intelligence, particularly in the field of natural language processing (NLP), have introduced few possibilities having the potential to overcome thus limitations. One such advancement is Retrieval Augmented Generation (RAG), a novel approach which combines retrieval based models with generative models. RAG has shown promising results in different domains such like question answering and text generation, by leveraging the large corpus of relevant information during inference. In RAG, a retrieval mechanism is employed to gather relevant documents or data segments from an external corpus, which are then fed into a generative model (like GPT) to produce a final output (Lewis et al., 2020). This hybrid approach allows the model to generate informed responses based on both the input data and the retrieved information, leading to more accurate and contextually rich predictions.

This research explores the potential of integrating Retrieval-Augmented Generation into traditional fraud detection models to address the limitations of current approaches. The study aims to evaluate the effectiveness of RAG in improving the detection of fraudulent transactions, particularly in scenarios where existing models may fall short, such as in cases of novel or subtle fraud patterns. By enhancing the ability of fraud detection systems to dynamically retrieve and utilize relevant information, this research seeks to contribute to the development of more robust and adaptable solutions for combating financial fraud.

## Problem Statement

Financial fraud continues to be a significant threat to the global economy. While fraudulent transactions are becoming increasingly sophisticated and more frequent, the ability to detect and prevent of thus transactions is becoming further challenging. Traditional models like Random Forest Classification, Neural Network models and even XGBoost are widely equipped around the world in detection of fraudulent activities. However, these models fall short in identifying new or evolving fraud patterns due to their reliance of historical data and incapability of retrieving external, contextually relevant information (Breiman, 2001; LeCun, Bengio & Hinton, 2015; Chen & Guestrin, 2016).

As fraudsters continuously refine their tactics, there is a critical demand for a more advanced model that can dynamically integrate external knowledge and provide context aware analysis. Retrieval Augmented Generation (RAG) provides a promising solution by combining both retrieval based models with generative models. This enhance the detection of complex and emerging fraud patterns. However, the application of RAG in the context of financial fraud detection remains underexplored.

This project aims to address the gap by investigating the potential of RAG in improving the accuracy and effectiveness of financial fraud detection systems. By leveraging the capabilities of RAG, this research seeks to develop a more robust and adaptive approach to combat financial fraud.

## Justification of the Study

Due to prevailing economic issues around the world, financial fraud has increased significantly throughout the world. According to the UK Finance annual fraud report 2024, losses due to fraud in the UK reached £1.17 billion in 2023 (UK Finance, 2024). This implies the current systems in place are not successful enough, demanding a more advanced model which can detect novel and complex patterns of fraud.

Retrieval Augmented Generation (RAG) has shown promising results in the realm of Natural Language Processing and widely used in creating question answering systems. In this research I tried the potential of leveraging the retrieval and generation capabilities of a RAG model and incorporate It into identifying fraudulent transactions.

In the realm of data science and machine learning, usage of embeddings, vector stores and querying is underexplored in the context of non-textual datasets. Hence, there was a urge to utilize RAG model along with embeddings, vector stores, indexing and querying in a structured transactional dataset, in this case the financial fraud transaction dataset. I took this opportunity to investigate this un-explored area.

In this project, I have utilized traditional models like Random Forest, Neural Network and XGBoost to set up a benchmark to compare the performance of RAG integrated model. Also I have tried multiple measuring metrics when it comes to finding the nearest matching transaction based on the similarity score. Additionally, creating a protype model of generating responses based on the user query prompt is also considered, as the need for such model is helpful in detecting fraudulent transactions more effectively.

## Research Questions

This research aims to answer the following questions.

1. What are the capabilities of Retrieval-Augmented Generation (RAG) in identifying fraudulent transactions?
2. How does the effectiveness of Retrieval-Augmented Generation in detecting fraudulent transactions compare to traditional machine learning methods?
3. How can a prototype model be developed using generative AI to retrieve related transactions based on user queries?

## Aims and Objectives

The primary aim of this research is to explore and evaluate the effectiveness of Retrieval Augmented Generation (RAG) in detection of fraudulent transactions within financial dataset.

**Objectives**

**Assess the Capabilities of RAG:** Investigate how Retrieval-Augmented Generation (RAG) can be utilized to identify and analyze fraudulent transactions

**Compare RAG with Traditional Methods:** Evaluate the performance of RAG in detecting fraudulent transactions relative to established machine learning techniques, such as Random Forest, Neural Networks, and XGBoost, in terms of accuracy, adaptability, and effectiveness.

**Develop a Prototype Model:** Create and implement a prototype model that leverages generative AI to retrieve and analyze related transactions based on user queries, demonstrating the practical application of RAG in fraud detection.

**Analyze Model Performance:** Assess the prototype model’s performance in real-world scenarios, including its ability to identify novel fraud patterns and provide actionable insights compared to traditional detection methods.

# REVIEW OF LITERATURE

**Financial Fraud Detection**

Financial fraud is a persistent issue vastly impacting not only the individuals and financial institutions but also the global economy itself. Early detection of fraudulent activities is crucial due to the growing fraud schemes. Various machine learning models have been developed and deployed to address this challenge.

**Traditional models for Fraud Detection**

Traditional models for fraud detection primarily rely on rule based system and statistical methods. These systems utilize pre-defined set of rules identified by the subject experts to flag suspicious activities. These rules are more often derived from historical data and domain knowledge (Phua, Lee, Smith & Gayler, 2010). Statistical methods, including logistic regression, is also widely used to model the probability of fraud occurrences (Bolton & Hand, 2002). While this models provides a foundation for fraud detection, these models mostly struggle with scalability and adaptability to new and evolving patterns.

**Random Forest**

Random Forest is a robust ensemble learning method widely known for high accuracy in fraud detection models. It operates by constructing multiple decision trees during training and outputting the mode of the classes for classification tasks (Breiman, 2001). Research by Abdallah, Maarof, and Zainal (2016) demonstrated that Random Forest outperforms other algorithms, such as Decision Trees and Logistic Regression, in detecting financial fraud due to its ability to handle large datasets and reduce overfitting.

**Neural Network**

Neural network, especially deep learning models have shown promising results in the context of fraud detection. These models are inspired by the human brain's structure. Neural network model consist of multiple layers which process input data through weighted connections (Goodfellow, Bengio & Courville, 2016). Wei, Hu, and Zhang (2019) applied deep learning techniques to detect fraudulent credit card transactions. It has achieved higher detection rates compared to traditional methods. These models can capture complex patterns and relationship of data. Hence, neural network models are capable of identifying subtle and sophisticated fraud schemes.

**XGBoost**

XGBoost is an implementation of gradient-boosted decision trees. This has emerged as a powerful tool for fraud detection due to its efficiency and performance (Chen & Guestrin, 2016). These models particularly capable of handling large scale datasets and provide better generalization. A study by Pourhabibi et al. (2020) highlighted XGBoost's superior accuracy in fraud detection tasks. This ensures the model's ability to focus on hard-to-classify cases by adjusting weights iteratively.

**Class Imbalance and SMOTE**

One of the significant challenges when dealing with financial fraud analysis dataset is the class imbalance problem. The ratio of majority vs minority class is highly different. These imbalance can lead to biased models that favors the major class resulting in poor detection of fraudulent transactions (Jing et al., 2018).

The Synthetic Minority Over-sampling Technique (SMOTE) is a widely-used method to address this issue. SMOTE generates synthetic data for the minority class by interpolating between existing minority instances, thus balancing the dataset (Chawla et al., 2002). In the context of fraud detection SMOTE has shown a significant improvement in the performance of models by providing a more balanced representation of classes. Balancing the weight among the classes is critical for effective model training (Dal Pozzolo et al., 2015).

**Embedding, Vector Stores and Similarity Search**

Embeddings convert categorical variables into continuous vector representation, capturing the semantic relationship among them. This approach is undoubtedly useful in financial data where categorical features like ‘transaction type’ , ‘transaction behavior patterns’ are prevalent (Mikolov et al., 2013).

Vector stores allow for efficient storage and retrieval of these embeddings. These vector stores also allows similarity search on these embeddings. Similarity search is a technique used to find instances in the data according to the query instance provided. In the context of financial fraud this basically mean detecting transactions which are similar to known fraudulent transacitons (Johnson, Douze & Jégou, 2019).

**Retrieval Augmented Generation (RAG)**

RAG is an emerging approach that combines both retrieval mechanisms and generative models. RAG has proven increase in performance of tasks like question answering and information retrieval. Also RAG holds the potential for fraud detection as well. RAG works by retrieving relevant documents or data points from a large corpus and using this information to guide the generation process in models like transformers (Lewis et al., 2020). This method can be beneficial in fraud detection systems to pull in past related cases like past incidents to provide more context for identifying new fraudulent activities. For instance, RAG could enhance the performance of fraud detection model by simply integrating insights from past fraud cases and then using them to make decisions. By this, the model will lead to more accurate and informed predictions (Khattak et al., 2022).  
  
**Sentence Transformers**

Sentence Transformer models are designed to generate semantically meaningful sentence embeddings that can be used for tasks like similarity search, clustering and classification (Reimers & Gurevych, 2019). A study by Wang et al. (2021) demonstrated the effectiveness of Sentence Transformers in financial text classification, which can be modified and applied to fraud detection models.

## Related work in the area

The use of Retrieval-Augmented Generation (RAG) in fraud detection is a new and emerging area of research. RAG is widely used in question answering models. Rag retrieves contextual information from a large dataset and uses it to guide predictions, improving detection accuracy.

Lewis et al. (2020) introduced RAG and showcased its effectiveness in tasks requiring extensive knowledge. They found that RAG outperforms traditional transformer models by incorporating a retrieval step which brings external knowledge. While they primary has focused on Natural Language Processing tasks, the principals of RAG too can be applied to fraud detection. It can retrieve past transaction fraud patterns or similar cases, informing the detection model.

RAG can identify complex fraud schemes that traditional models might miss. For example, Khattak et al. (2022) used RAG in a legal question-answering system. They have shown how RAG could generate precise answers by retrieving and integrating relevant cases. Similarly, this can be applied in the domain of fraud detection as well. Integrating RAG into this domain could lead to more nuanced and informed decision making, especially in complex and more sophisticated cases where traditional models might fail.

While the direct application of RAG in the domain of fraud detection is still in early stages, there are lot of potential benefits.

# RESEARCH METHOD

## Data Overview

The dataset used in this project is taken from Kaggle.com repository. According to the repository, the dataset is sourced from BankSim, comprising an aggregated sample of transactional data from a Spanish bank. Dataset consists of 594,643 transactions from unique 4112 users. The target variable(isFraud) identifies fraudulent transactions, while the remaining seven columns include a time step identifier, personal information about the payer such as gender, age group, zip code and transaction details including merchant name, category, and transactional amount. Out of all the transactions in the dataset only 7200 records are identified as fraud, making the dataset highly imbalanced.

A screenshot of a computer

Description automatically generated

A screenshot of a phone

Description automatically generated

## Review of Columns

Original dataset consists of below columns.

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Description | Keep ? | Explanation |
| Step | Time-step identifier | Yes | Values ranging from 0-180. Assuming a time-step represents a one transactional day. |
| customer | Customer identifier | Yes | Categorical feature. Random unique number to identify customer, ensuring the anonymity |
| age | Age group of customer | Yes | Values ranging from 0-6. |
| gender | Gender of the customer | Yes | Unique values are ‘M’, ‘F’ and ‘U’ |
| zipCodeOri | Zip code of the originated location | No | Removing as the value is static throughout the dataset |
| Merchant | Identification of the merchant | Yes | Categorical feature. Random unique number to identify customer, ensuring the anonymity |
| zipMerchant | Zip code of the merchant location | No | Removing as the value is static throughout the dataset |
| Category | Transaction type | Yes | Categorical feature. Specify for which category the transaction is originated |
| Amount | Transaction amount | Yes | Numerical Feature. |
| Fraud | is Fraud yes or no | Yes | Target variable. If fraud=0 means legitimate transaction. Fraud =1 means transaction is identified as Fraud. |

## Data Pre-processing

References

 Association of Certified Fraud Examiners (ACFE). (2022). *2022 Report to the Nations: Global Study on Occupational Fraud and Abuse*. ACFE.

 Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

 Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

 LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, 521(7553), 436-444.

 Lewis, P., Perez, E., Piktus, A., Karpukhin, V., Goyal, N., Kuang, Z., ... & Riedel, S. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

** Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.**

** Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.**

** LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, 521(7553), 436-444.**

**UK Finance. (n.d.). *Annual Fraud Report 2024*. [online] Available at:** [**https://www.ukfinance.org.uk/policy-and-guidance/reports-and-publications/annual-fraud-report-2024**](https://www.ukfinance.org.uk/policy-and-guidance/reports-and-publications/annual-fraud-report-2024)**.**

** Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68, 90-113.**

** Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 17(3), 235-255.**

** Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.**

** Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.**

** Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.**

** Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. *IEEE Symposium Series on Computational Intelligence*.**

** Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.**

** Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*.**

** Jing, H., Ngai, E. W. T., Lee, S. Y., & Xing, X. (2018). Fraud detection for online businesses: A perspective from blockchain technology. *Internet Research*, 28(5), 1068-1087.**

** Khattak, M. A., Hashmi, M., Ahmad, J., & Kang, B. H. (2022). Using Retrieval-Augmented Generation (RAG) to generate correct answers for legal case question answering. *Applied Sciences*, 12(6), 3198.**

** Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.**

** Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.**

** Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.**

** Pourhabibi, T., Ong, K. L., Saeed, A., & Yahyapour, R. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 133, 113303.**

** Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.**

** Wang, X., Zhang, Y., & Liu, W. (2021). Financial text classification using pre-trained language models. *Information*, 12(9), 364.**

** Wei, W., Hu, L., & Zhang, Y. (2019). A novel ensemble method for credit card fraud detection using optimized deep belief networks. *Soft Computing*, 23(17), 7939-7950.**