# 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

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.**

**‌**