Recent trends in Management and Commerce

Vol: 7(1), 2026

REST Publisher; ISBN: 978-81-936097-6-7

Website: https://restpublisher.com/book-series/rmc/

Real Time AI Fraud Detection & Financial Risk Management

B. Uma Mahesware

St. Joseph's College of Arts and Science for Women (Autonomous), Hosur, Tamil Nadu, India.

Abstract: *In the modern financial landscape, fraud detection systems must evaluate transactions within a critical window of less than 200 milliseconds. Traditional machine learning (ML) models, while efficient at processing structured data, often fail to detect novel fraud tactics or interpret unstructured information. This paper proposes a multi-model ensemble architecture that combines the speed of predictive ML with the contextual reasoning of Encoder Large Language Models (LLMs). By utilizing a two-stage workflow where only ambiguous cases are escalated to an LLM, the system achieves a balance between computational efficiency and high-precision accuracy, significantly reducing false positives and manual human review.*

Keywords: *Artificial Intelligence, Predictive Machine Learning, spoofing, or fishy phrasing, Escalation, Deep analysis, decision trees, random forests, and gradient boosting machines.*

1. INTRODUCTION

The primary challenge in fraud detection is answering a binary question—"Is this fraud?"—before money moves. Financial institutions have historically leaned on AI models to recognize patterns and make decisions at scale. However, as fraudulent tactics become more sophisticated, traditional models often struggle with "borderline" cases or data that does not fit neatly into structured columns. Artificial Intelligence (AI) is fundamentally transforming the financial industry by enabling institutions to predict risks and detect fraud as they occur. This paper examines the evolution of risk management— specifically addressing credit, market, operational, and liquidity risks—through the lens of AI-driven solutions. By synthesizing traditional machine learning algorithms with the contextual reasoning of Encoder Large Language Models (LLMs), financial institutions can achieve a more proactive and accurate defense against rising online threats.

2. THE TWO-MODEL STRATEGY

The proposed architecture uses an ensemble of two distinct types of AI:
Predictive Machine Learning (ML): Utilizing algorithms like Random Forest and Gradient Boosting, these models are optimized for structured data (e.g., transaction amount, time, and location). They are fast and computationally "cheap" but can be "pattern-bound," missing new types of scams that use clever wording.

Encoder Large Language Models (LLMs): These are non-generative models like **BERT** or **RoBERTa** that focus on natural language understanding. They are "language savvy" and can analyze **unstructured data**—such as wire memos, transaction descriptions, or merchant names—to detect urgency, spoofing, or fishy phrasing.

The Workflow: Efficiency through Escalation

To maintain speed and reduce costs, the system does not run every transaction through the intensive LLM. Instead, it follows a tiered workflow:
Initial Screen: All transactions are first processed by the Predictive ML model.

Immediate Decision: If the model has high confidence (clearly legitimate or clearly fraud), a final action is taken immediately.
Escalation: Only borderline or ambiguous cases are escalated to the Encoder LLM.
Deep Analysis: The LLM evaluates the structured data alongside unstructured context (like customer notes) to provide a more nuanced assessment, significantly reducing false positives.

3. INFRASTRUCTURE REQUIREMENTS

Because Encoder LLMs have millions or billions of parameters, they require significant processing power, often involving GPU acceleration. The video emphasizes that for these models to work in real- time at the point of transaction, specialized AI accelerator chips and on-chip acceleration are necessary to handle low-latency inference.
BERT is an encoder Large Language Model (LLM), which is a non-generative AI model specifically designed for natural language understanding tasks like text classification and sentiment analysis. Unlike traditional machine learning that relies on structured data in spreadsheets, BERT identifies urgency and spoofing by analyzing unstructured data through a context-aware lens. To identify urgency, the model reads between the lines of unstructured texts such as wire memos or fund transfer descriptions—to detect nuanced language patterns. For example, phrases like "Please rush" or "urgent investment guaranteed 200% ROI" trigger higher risk scores because the model recognizes them as linguistic patterns common in scam scenarios. To detect spoofing, BERT analyzes merchant names and free-form addresses for subtle signs of manipulation or associations with known fraud cases that traditional models often ignore. It processes this information by comparing the input against millions of fraud patterns, allowing it to connect dots and understand why a specific phrasing or context looks "fishy". Ultimately, BERT's strength lies in being "language savvy," which enables it to extract key information and recognize fraudulent intent within text that does not fit into neat data columns

Traditional Predictive Machine Learning

Most current fraud detection platforms utilize predictive ML algorithms such as logistic regression, decision trees, random forests, and gradient boosting machines. These models are trained on large, labeled datasets to recognize patterns in structured data, including transaction amounts, timestamps, locations, and merchant categories.
Advantages: These models offer microsecond latency, low compute requirements, and an easy- to-follow audit trail.
Limitations: They are "pattern-bound," meaning new scams using clever wording or novel tactics can evade detection. Furthermore, they generally ignore unstructured data like free-form text descriptions and images.

The Role of Encoder Large Language Models (LLMs)

To address the limitations of predictive ML, the proposed architecture introduces a second AI model: a non-generative Encoder LLM (e.g., BERT or RoBERTa). Unlike decoder LLMs used for chatbots, encoder models focus on natural language understanding.
Capabilities: Encoder LLMs can extract key information from unstructured data, such as reading a wire memo or an online funds transfer description to detect urgency or phrasing common in scams (e.g., "Refund for overpayment. Please rush").
Contextual Reasoning: These models are "language savvy" and can connect dots that humans or traditional models might miss, which helps in reducing false positives.

Proposed Multi-Model Ensemble Workflow

The core of this research is a two-stage workflow designed to maximize efficiency:
Initial Assessment: All incoming transactions first pass through the predictive ML model. If the model returns a high-confidence score (either clearly legitimate or clearly fraud), a final decision is made immediately.
Escalation: If the predictive model returns a low-confidence or ambiguous score, the transaction is escalated to the Encoder LLM.
Deep Analysis: The LLM processes both the original structured features and any available unstructured data (e.g., customer profile notes or transaction descriptions) to provide a context- aware assessment.
Final Decision: A decision engine combines findings from both models to approve or flag the transaction, avoiding the need for immediate human intervention.

Computational Considerations and Infrastructure

Encoder LLMs are computationally intensive, often requiring millions or billions of parameters and GPU acceleration. To maintain the required millisecond-level response time, the architecture requires specialized hardware. On-chip AI acceleration and specialized AI accelerator chips are essential to handle low-latency inference at the point of the transaction.

In an increasingly digital world, the volume of online transactions has necessitated the adoption of sophisticated AI technologies to safeguard assets and maintain customer trust. Financial security now relies on the ability to identify, assess, and manage threats to institutional stability. Risk management aims to mitigate various dangers, including credit risk (borrower default), market risk (stock or currency fluctuations), operational risk (system failures or cyber threats), and liquidity risk (cash shortages).

Foundational AI Technologies in Finance

AI enhances financial stability through its immense predictive power, utilizing machine learning algorithms such as decision trees and support vector machines (SVMs).

Pattern Recognition: These algorithms analyze massive datasets to identify patterns indicative of potential risk.

Adaptability: Unlike static systems, AI models learn from and adapt to new data, allowing them to improve their forecasting capabilities over time.

Predictive Analytics: Beyond analyzing historical records, AI can forecast future events, such as predicting market downturns or identifying high-risk loan applications prior to approval.

The infrastructure requirements include the following components:

Real-Time Data Pipelines

The system requires a constant stream of real-time data, which is generated by transactions the moment they occur. This enables the infrastructure to assess current risk levels instantly. To provide a comprehensive view, this pipeline must also integrate historical data (benchmarks) and external data, such as economic indicators or news events.

Specialized AI Hardware

Processing these transactions within the industry-standard 200-millisecond window requires significant computational power. As discussed previously, this often necessitates:

GPU Acceleration: Essential for running complex models, especially if using a multi-model ensemble that includes Encoder LLMs.

AI Accelerator Chips: Specialized hardware designed for low-latency inference at the point of the transaction.

On-Chip Acceleration: Integrating AI processing directly into the transaction hardware to minimize delays.

Scalable Machine Learning Models

The infrastructure must host predictive machine learning algorithms, such as decision trees and support vector machines (SVMs). These algorithms are critical because they:

Analyze massive datasets to find patterns indicating risk.

Adapt to new data, allowing the system to constantly improve its forecasting.

Provide predictive analytics to spot high-risk applications or market downturns before

Data Sources and Real-Time Analysis

For AI to be effective in risk management, it must leverage three primary data streams:

Historical Data: Past transaction records serve as benchmarks for normal behavior.

Real-Time Data: Transactions generated in the moment allow for instantaneous risk assessment and fraud detection.

External Data: Economic indicators, news events, and social media sentiment provide a comprehensive view of the global financial environment.

4. CONCLUSION

By adopting a multi-model ensemble, businesses can process straightforward cases with minimal overhead while applying deep, contextual analysis to trickier scenarios. This hybrid approach ensures that the "costly" LLM is only utilized when necessary, maintaining system efficiency while significantly improving overall accuracy in detecting

sophisticated fraud. The integration of AI into financial services is no longer optional but a requirement for stability in a digital economy. By combining the pattern-matching speed of traditional machine learning with the linguistic intelligence of Encoder LLMs, institutions can create a layered defense system. This approach not only identifies risks proactively but also significantly reduces false positives, ensuring a secure and efficient financial landscape

REFERENCES

[1]. Below is a list of sources that were mentioned or referenced in the study:
[2]. Fraud Detection with AI: Ensemble of AI Models improve precision & speed-IBM Technology
[3]. AI in financial Services-codetech academy
[4]. AI in risk management: Mitigating uncertainity with machine learning