The Power of Ensemble Methods: A Comparative Study of Machine Learning, Deep Learning, and LLMs for Financial Fraud Detection

Zahra Rezaei
School of Electrical and Computer Engineering, University of Oklahoma, Norman, Oklahoma, USA 73019

Sara Safi Samghabadi
School of Electrical and Computer Engineering, Islamic Azad University, Karaj Branch, Iran

Mohammad Amin Amini
Department of Computer Engineering, Islamic Azad Islamic University of Jasb, Markazi, Iran

Yaser Mike Banad
School of Electrical and Computer Engineering, University of Oklahoma, Norman, Oklahoma, USA 73019

***Abstract*— This study explores the efficacy of machine learning (ML), deep learning (DL), and large language models (LLMs) for detecting financial fraud in company reports. Comparing models like a Voting Classifier ensemble, Seq2Seq, and FinBERT (an LLM tuned for financial data), we found that traditional ML ensembles outperform complex DL and LLM models in this context. The Voting Classifier achieved the highest accuracy (91.2%), followed by Seq2Seq (83%) and FinBERT (74%), underlining the strength of ensemble methods for fraud detection.**

Keywords— Financial Fraud Detection, Machine Learning Models, Deep Learning Models, and Large Language Models.

I. INTRODUCTION

The fraud detection and prevention (FDP) market, valued at $19.5 billion in 2017, is projected to reach over $63 billion by 2023. In 2018, 23% of internet users experienced online identity theft. [1] AI-enabled fraud detection spending is expected to exceed $10 billion by 2027, a 57% growth since 2022.

The study [2] compared machine learning and large language models (LLMs) for fraud detection. Traditional models like Random Forest and SVM excelled, with FinBERT performing well among LLMs.The paper [3] applied logistic regression and SVM to payment fraud, achieving high accuracy despite imbalanced data.This paper [4] proposes a two-layer machine learning approach for detecting and classifying DNS over HTTPS (DoH) traffic. The study demonstrates that LGBM and XGBoost algorithms achieve near-perfect accuracy in distinguishing benign from malicious DoH traffic, with SourceIP and DestinationIP as key features. This paper [5] analyzed financial ratios to detect fraud in statements, using logistic regression on a dataset from Lithuania. The paper [6] examined logistic regression for credit card fraud, showing that novel preprocessing techniques improved accuracy. This paper [7] explored ML techniques, noting that models like decision trees and neural networks adapt to complex fraud. The article [8] focused on the Iranian stock market, finding that boosted regression tree models improved fraud detection accuracy. This research paper [9] reviewed supervised and unsupervised learning methods for transaction fraud detection, emphasizing hybrid approaches for adaptability. Lastly, paper [10]evaluates four machine learning algorithms for network intrusion detection in IoT infrastructure, using the CICIDS2017 dataset. Boosted machine learning techniques outperform others, achieving over 99% accuracy, and providing valuable insights for effective NIDS selection.

II. PROPOSED METHODS

The dataset used in this study is a collection of financial filings from various companies submitted to the U.S. Securities and Exchange Commission (SEC) [11]. This study addresses financial fraud detection using a multi-phase approach with traditional machine learning, deep learning, and a specialized language model. Phase one uses models like Logistic Regression and Decision Trees, combining them with voting to improve accuracy. Phase two employs a Seq2Seq deep learning model for sequential data processing, while the final phase applies Prosusai/Finbert, a financial language model, to capture domain-specific nuances, enhancing overall fraud detection. Logistic Regression is a linear binary classification model that estimates class probability using a linear combination of features, applying L2 regularization (penalty='l2') with C=1.0 to prevent overfitting. The Decision Tree Classifier is a non-linear model that predicts through learned decision rules, limited to a maximum depth of 1 to control complexity. K-Nearest Neighbors (KNN) is non-parametric, classifying instances by the majority among k=5 nearest neighbors. Multinomial Naive Bayes, assuming feature independence, performs well for text classification. Voting Mechanism combines "Soft and Hard Voting," with Soft Voting enhancing accuracy through weighted averaging. The Voting Classifier is trained on TF-IDF-transformed text data to prioritize informative features.

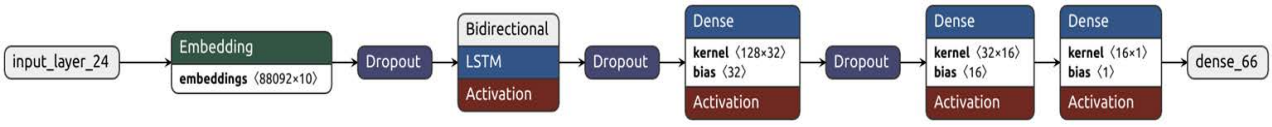

Figure 1: Seq2Seq Model Architecture for Fraud Detection

DOI 10.1109/AIxDKE63520.2024.00031

The Seq2Seq model in **Figure 1** addresses financial fraud detection by sequentially processing text data. An embedding layer converts words to numerical representations, with a bidirectional LSTM capturing context, followed by dense layers and a sigmoid layer predicting fraud probability. FinBERT, a pre-trained BERT model adapted for financial language, uses Reuters and financial sentiment datasets for nuanced domain understanding.

Training Strategies: FinBERT employs Slanted Triangular Learning Rates, Discriminative Fine-Tuning, and Gradual Unfreezing to avoid "catastrophic forgetting" during fine-tuning, achieving 97% accuracy on the Financial PhraseBank and 86% on the full dataset, surpassing previous models [12].

III. Results

Table 1 evaluates four fraud detection models—hard voting, soft voting, Seq2Seq, and FinBERT—using accuracy, precision, recall, and F1-score on both training and test data. "Voting Soft" performs best, with a training accuracy of 0.91 and test accuracy of 0.912, achieving high precision and recall, while FinBERT has the lowest scores, with a test accuracy of 0.74, indicating challenges in detecting fraud. Confusion matrices reveal that both voting classifiers have fewer misclassifications and higher accuracy, whereas FinBERT shows more errors with increased false positives and negatives. The Seq2Seq model has balanced predictions but slightly more false negatives than the voting classifiers.

IV. Conclusion

This study analyzes machine learning, deep learning, and LLMs for financial fraud detection, highlighting that the ensemble-based Voting Classifier achieves the highest accuracy (91.2%) by leveraging complementary model strengths. While LLMs (e.g., FinBERT) and Seq2Seq models show solid performance (74% and 83% accuracy), they fall short of the ensemble's accuracy. The findings emphasize the importance of model selection and the impact of data quality and model adaptability. Future research should enhance data quality, explore hybrid approaches, and incorporate domain-specific knowledge to advance fraud detection accuracy and reliability, contributing to financial system integrity.

Table 1. Performance Metrics for Model Classifiers

Models	Train-accu	Train-Prec	Train-Recall	Train-F1	Test-accu	Test-Prec	Test-Recall	Test-F1
Hard Voting	0.89	0.85	0.93	0.88	0.88	0.83	0.93	0.88
Soft Voting	0.91	0.89	0.96	0.91	**0.912**	**0.84**	**1.0**	**0.90**
Seq2Seq Model	0.83	0.78	0.91	0.84	0.83	0.77	0.94	0.85
LLM-Finbert	0.75	0.686	0.91	0.79	0.74	0.68	0.94	0.79

V. References

[1] E. H. Dyvik, "Size of the fraud detection and prevention (FDP) market worldwide from 2016 to 2023," Statista, 12 8 2024. [Online]. Available: https://www.statista.com/statistics/786778/worldwide-fraud-detection-and-prevention-market-size/#statisticContainer.

[2] A. S. Kedia, "Enhancing Financial Fraud Detection: A Comparative Analysis of Large Language Models and Traditional Machine Learning and Deep Learning Approaches," *A report the degree of Master of Science Brunel University,* 2022-2023.

[3] A. Oza, "Fraud Detection using Machine Learning," *Stanford CS229 Machine Learning,* 2018.

[4] S. R. Yaser Banad, "Detecting Malicious DNS over HTTPS Traffic in Domain Name System using Machine Learning Classifiers," *Journal of Computer Sciences and Applications,* August 2020.

[5] Z. G. Rasa Kanapickiene, "The Model of Fraud Detection in Financial Statements by Means of Financial Ratios," *20th International Scientific Conference Economics and Management - 2015 (ICEM-2015),* 2015.

[6] N. O. A. Hala Z Alenzi, "Fraud Detection in Credit Cards using Logistic Regression," *(IJACSA) International Journal of Advanced Computer Science and Applications,* 2020.

[7] Oluwabusayo Adijat Bello, *et al.* "Machine Learning Approaches for Enhancing Fraud Prevention in Financial Transactions," *International Journal of Management Technology,* pp. 85-108, 2023.

[8] Jafar Nahri Aghdam Ghalejoogha, *et al.* "Detecting financial fraud using machine learning techniques," *Int. J. Nonlinear Anal. Appl,* p. 199–214, 2024.

[9] O. Kazeem, "Fraud Detection using Machine Learning," *DOI: 10.13140/RG.2.2.12616.29441,* September 2023.

[10] Yaser Banadaki, "Design of intrusion detection systems on the internet of things infrastructure using machine learning algorithms," *Proceedings Volume 11594, NDE 4.0 and Smart Structures for Industry, Smart Cities, Communication, and Energy; 115940J (2021) https://doi.org/10.1117/12.2584499,* 27 April 2021.

[11] A. Shushil, "Financial Statement Fraud Data Kaggle," Kaggle, 1 12 2023. [Online].

[12] Z. Genc, "medium," FinBERT: Financial Sentiment Analysis with BERT, 7 2020. [Online]. Available: https://medium.com/prosus-ai-tech-blog/finbert-financial-sentiment-analysis-with-bert-b277a3607101. [Accessed 2024].