

## FINANCIAL FRAUD DETECTION USING VALUE AT RISK WITH MACHINE LEARNING IN SKEWED DATA

K. KUMARASWAMY<sup>1</sup>, CH. RASHMITHA<sup>2</sup>, B. SHARVANI SHARMA<sup>3</sup>, E. MANISHA<sup>4</sup>

ASSISTANT PROFESSOR <sup>1</sup>, UG STUDENT <sup>2,3,4</sup>

DEPARTMENT OF IT, MALLA REDDY ENGINEERING COLLEGE FOR WOMEN (UGC-AUTONOMOUS),  
MAISAMMGUDA, HYDERABAD, TELANGANA-500100

### ABSTRACT

The significant losses that banks and other financial organizations suffered due to new bank account (NBA) fraud are alarming as the number of online banking service users increases. The inherent skewness and rarity of NBA fraud instances have been a major challenge to the machine learning (ML) models and happen when non-fraud instances outweigh the fraud instances, which leads the ML models to overlook and erroneously consider fraud as non-fraud instances. Such errors can erode the confidence and trust of customers. Existing studies consider fraud patterns instead of potential losses of NBA fraud risk features while addressing the skewness of fraud datasets. The detection of NBA fraud is proposed in this research within the context of value-at-risk as a risk measure that considers fraud instances as a worst-case scenario. Value-at-risk uses historical simulation to estimate potential losses of risk features and model them as a skewed tail distribution. The risk-return features obtained from value-at-risk were classified using ML on the bank account fraud (BAF) Dataset. The value-at-risk handles the fraud skewness using an adjustable threshold probability range to attach weight to the skewed NBA fraud instances. A novel detection rate (DT) metric that considers risk fraud features was used to measure the performance of the fraud detection model. An improved fraud detection model is achieved using a K-nearest neighbor with a true positive (TP) rate of 0.95 and a DT rate of 0.9406. Under an acceptable loss tolerance in the banking sector, value-at-risk presents an intelligent approach for establishing data-driven criteria for fraud risk management.

### INTRODUCTION

The Association of Certified Fraud Examiners (ACFE) 2022 released a financial fraud report stating that 2,110 fraud cases involving industries in financial sectors in 133 countries resulted in losses of around \$3.6 billion [1]. Financial fraud can be termed as the deliberate employment of unlawful procedures or tactics to obtain financial gain [2]. The consequences of financial fraud can potentially disrupt economies, raise living expenses, and undermine consumer confidence [3]. Forms of financial fraud

include insurance fraud, money laundering, new bank account fraud, credit and debit card fraud, mortgage fraud, and many more [4,5]. The act of opening an account to commit fraud at banks or other financial organizations is known as “new bank account(NBA) fraud” [6]. Fraud not only results in immediate financial losses and erodes public confidence in institutions, but has broader consequences, affecting customers and financial systems through market instability and contributing to larger macroeconomic downturns [7]. Fraud datasets typically exhibit some properties including skewness, evolving patterns, highly dimensional, and restricted access to relevant information. Specifically, fraud skewness which represents the majority fraud class over the non-fraud class has been a major concern to studies, as it affects the performance of fraud detection model. The Skewed fraud instances can have a bad influence on machine learning algorithms such as distance-based algorithms [8]. Previous efforts in tackling fraud involve developing rule-based expert systems, statistical methods, machine learning, and risk-based methods [9], [10]. Due to the cost of maintenance and the inefficiency of rule-based methods [10], decision-makers decide to utilize statistical methods such as autoregressive models to handle financial fraud [11], [12], [13]. The complex patterns and high dimensional nature of frauds make the statistical methods less effective, as such machine learning models were deployed [10], [14]. However, some of the studies that utilize machine learning techniques were found to have a high False Positive (FP) rate [15], [16], [17]. Machine learning models can potentially handle high-dimensional data and complex patterns of fraud instances. To evaluate the effectiveness of machine learning model, Jesus et al. [18] presented the first domain-specific and real world bank account fraud (BAF) dataset. The datasets were generated using generative adversarial networks (GANs) and evaluated using light gradient boosting method (LGBM). The study [18], [19] utilizes 25 sets of hyper parameter configurations to optimize the LGBM model, utility aware reweighing was used to handle the class skewness of BAF dataset. The study [15] utilizes stacking in ensembled learning with majority voting to evaluate the BAF dataset and address the changing fraud patterns.



[CC BY 4.0 Deed Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/)

This article is distributed under the terms of the Creative Commons CC BY 4.0 Deed Attribution 4.0 International attribution which permits copy, redistribute, remix, transform, and build upon the material in any medium or format for any purpose, even commercially without further permission provided the original work is attributed as specified on the Ninety Nine Publication and Open Access pages <https://turcomat.org>

The study [20] uses federated learning in addressing data privacy issues of BAF dataset and deep neural networks to classify fraud instances. These studies achieve good performance in addressing BAF challenges; However, the studies do not consider the potential losses of fraud risk features. To our knowledge, little research exists that employs machine learning techniques in NBA fraud detection. The detection of NBA fraud is proposed in this paper within the context of risk management that uses value-at-risk to considers skewed fraud instances as a worst-case scenario. To adequately estimate the losses of fraud risks, value-at-risk was augmented with expected loss and expected shortfall of frauds which further quantifies the mean and extreme loss effects respectively. These risk measures combination will allow the quantification of risks across mean, worst-case, and extreme scenarios. Value-at risk employs historical simulation to estimate potential losses of risk features. The risk-return features obtained from value at-risk are based on assessing their risk exposure to fraud risk. The risk-return features are sent as input to the NBA fraud detection model. Different machine learning models were trained; However, the K-nearest neighbor out performed other models. The contributions of this paper are:

- This paper used an extreme value theorem to model the tails (potential losses) instead of the fraud pattern.
- This paper used value-at-risk to model the skewness of fraud instances more efficiently.
- This paper utilized historical simulation to estimate value-at-risk as it makes no assumptions on any distribution.
- This paper used novel detection rate performance metrics to capture the overall performance in detection of NBA fraud instances that incorporate risk fraud factors.

## LITERATURE REVIEW

### Aggregate earnings informativeness and economic shocks: international evidence

- [Yuto Yoshinaga, M. Nakano](#)
- Published in [Asia-Pacific Journal of...](#) 17 November 2019

Our study proposes the usage of aggregate earnings to forecast future GDP growth. Using empirical analyses with global quarterly data, we investigate whether aggregate-level profitability drivers, which are components of aggregate earnings, are relevant for forecasting GDP growth. After confirming that aggregate-level profitability drivers are useful for forecasting future GDP growth worldwide, we show that considering the effects of crises improves the forecast model of GDP growth. In addition, we suggest that predicting GDP growth using aggregate-level profitability drivers is relevant for stock valuation in developed countries, but not in emerging countries

### Incorporating Financial Statement Information to Improve Forecasts of Corporate Taxable Income

- [Daniel Green, E. Henry](#), +1 author [G. Plesko](#)

- Published in [Social Science Research...](#) 5 February 2020

We examine whether public financial statement information is incrementally useful in forecasting confidential taxable income. More precise firm-level taxable income forecasts can improve policymakers' modeling of the tax system and the analysis of proposed changes in corporate tax law, while more accurate macro-level forecasts of corporate taxable income can improve estimates of corporate tax revenues, a significant component of the federal budget. We find the addition of financial statement information improves firm- and industry-level estimates of future taxable income by primarily providing more timely information, but also through accruals. Our results suggest that macroeconomic forecasts of taxable income may be further improved by the aggregation of firm-level forecasts that are generated using financial statement information. Importantly, our results are driven primarily by tax information in financial statements. We also contribute to the research on the information content of financial statement information for forecasting economic activity.

### Using Economic Links between Firms to Detect Accounting Fraud

- [Chenchen Li, Ningzhong Li, F. Zhang](#)
- Published in [Social Science Research...](#) 3 February 2021

We explore whether accounting fraud can be detected using the information of firms economically linked to a focal firm. Specifically, we examine whether customer information disclosed by a supplier firm, combined with customers' accounting information, helps to detect the supplier's revenue fraud. We first confirm the economic link between the supplier and customers by showing a strong positive correlation between the supplier's sales growth and the growth rate of total customer purchases. We then introduce two variables based on customer accounting information—the discrepancy between supplier sales growth and customer purchase growth and customer excess purchases—and show that they are predictive of supplier revenue fraud. We conduct a battery of cross-sectional tests to further examine the two fraud predictors and generally find results to vary cross-sectionally in a predictable way. Finally, the out-of-sample tests indicate adding the two variables to Dechow et al.'s (2011) model increases fraud prediction accuracy.

## EXISTING SYSTEM

Many studies in the literature utilize statistical methods in evaluating financial fraud. Specifically, significant studies were found to utilize ordinary least squares (OLS) regression and autoregressive (AR) models for financial fraud evaluation. Using the Tehran Stock Exchange dataset, the study [21] uses a regression model to investigate the association between auditor characteristics and fraud detection in emerging economies. The authors provide useful information for improving the reliability of the findings. Using pooled OLS and panel regressions, the study [22] investigates the effect of political alignment on corporate fraud convictions, offering insights into the connection between politics and fraud. An existing system presents the financial fraud assessment from the perspective of risk mitigation. The existing studies utilize different risk measures such as value-at-risk (VaR), expected loss, and expected shortfall

to assess the level of risk of fraud. The study [29] offers strategies for breaking down the risk of fraud, identifying potential fraudsters, and enabling more targeted anti-fraud measures by tying the motivation of the fraud triangle to human tendencies that lead to specific actions as well as the meta-model of fraud together. Regression analysis is utilized in the study [30] to look at how enterprises manage risk to determine how control environments, risk assessments, control activities, information and communication, and monitoring contributed to fraud prevention and detection efforts in Indonesian firms. The study [35] identified a positive correlation between fraud risk assessment and management and the efficient use of forensic accounting using chi-square, fisher test, and correlation, however, there is no relationship between fraud risk assessment and management in terms of techniques causing fraud. The study [9] examines fraud using ensemble learners for anomaly detection and also handles data skewness, a triage model that receives input from the ensemble model, and a risk model that estimates the financial losses. The authors successfully provide an effective fraud risk-based detection, from machine learning techniques to risk assessment, but do not to evaluate fraud detection by first considering the risk component before subjecting it to machine learning detection. An existing system presents studies that utilize machine learning techniques for the classification of fraud applications. The majority of the presented studies consider the detection while addressing the skewed nature of fraud instances. Sampling methods, hybrid methods, and other novel methods are majorly used to overcome the skewed nature of fraud datasets. The study [36] addresses class skewness in credit card fraud using quantum machine learning (QML) and support vector machines (SVM). The results show that classic machine learning techniques are still useful for non-time series data, whereas QML applications can be used for time-series-based and highly skewed data. Quantum neural network (QNN) achieves good performance in fraud detection by the study [37]. The study [38] trained different machine learning models, all of which were using default implementations and parameters, XGBoost performed more accurately than any other models. The effectiveness of telecom fraud is assessed in the study [39] using a dynamic graph neural network (DGNN), the authors effectively present a suggested method for resolving the issue of telecom fraud detection in extensive phone social network.

#### Disadvantages:

- Most existing studies do not consider potential losses of fraud risk features, but fraud instances happen rarely and cause big losses when they occur.
- Fraud instances are inherently skewed compared to non-fraud instances, producing a highly skewed distribution.
- Fraud patterns tend to have more irregular and extreme values, while models like logistics regression or regression assume normality and predictions may produce an inaccurate result.

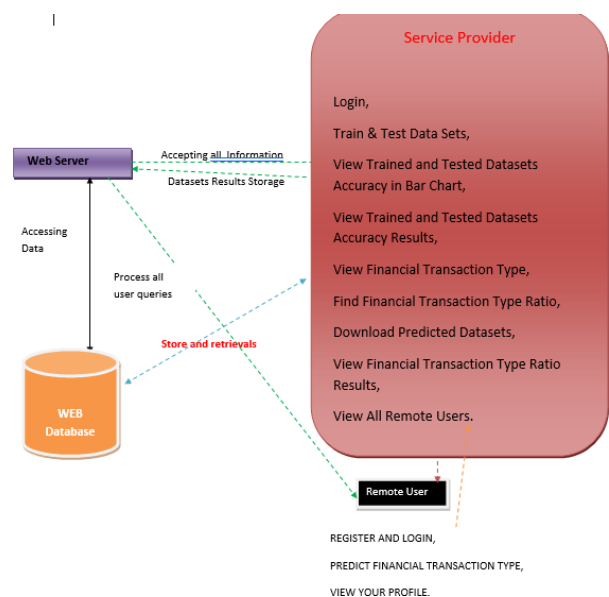
**PROPOSED SYSTEM** The proposed design implemented in which it describes the steps and process involved in NBA fraud detection. Value-at-risk being an important part of this research

is designed to model the severe and extreme fraud risk features, it also focuses on rare fraud instances that are detrimental and very costly when occurred. However, the rare cases that are mostly skewed can distort machine learning algorithms especially distance based like KNN. The value-at-risk can handle the fraud skewness through the utilization of adjustable threshold probability ranges (confidence level) unlike the conventional methods that employ constant fraud probability weight that's attached to the skewed fraud instances. The preprocessed, extracted and engineered features were sent as input to value-at-risk for simulation. Meanwhile, a distance based KNN is designed for adjustability to detect fraudulent features through identifying rare clusters with nearest neighbor distance  $k$ . The confidence level chosen considers the rare fraud cases as higher risk features that would result in fewer training sets, particularly for the KNN model with hyperparameter  $k$ . The fraud detection model requires the optimization of  $k$  to a lower setting to sufficiently model the fraudulent features in the rare cluster. The distance weight of KNN is imperative in inhibiting fraud skewness by assigning a higher weight to near instances which in turn facilitates efficient detection of skewed instances.

#### Advantages:

- This paper used an extreme value theorem to model the tails (potential losses) instead of the fraud pattern.
- This paper used value-at-risk to model the skewness of fraud instances more efficiently.
- This paper utilized historical simulation to estimate value-at-risk as it makes no assumptions on any distribution.
- This paper used novel detection rate performance metrics to capture the overall performance in detection of NBA fraud instances that incorporate risk fraud factors.

#### IMPLEMENTATION SYSTEM ARCHITECTURE



#### MODULES

- **Service Provider**

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Train & Test Data Sets, View Trained and Tested Datasets Accuracy in Bar Chart, View Trained and Tested Datasets Accuracy Results, View Financial Transaction Type, Find Financial Transaction Type Ratio, Download Predicted Datasets, View Financial Transaction Type Ratio Results, View All Remote Users.

- **View and Authorize Users**

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

- **Remote User**

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGINPREDICT FINANCIAL TRANSACTION TYPE, VIEW YOUR PROFILE.

## RESULT



## CONCLUSION

The value-at-risk-based fraud detection model presented in this paper enables the quantification and mitigation of fraud risk features and at the same time overcome the influence of skewed fraud instances which is very crucial in solving financial fraud challenges. The value-at-risk attach confidence probability weight to the rare fraud cases with nearest neighbor distance k. The distance weight of KNN is imperative in inhibiting class skewness by assigning a higher weight to near instances which in turn facilitates efficient detection of skewed instances. The deployment of expected short fall and expected loss by value at risk allows quantification of risk across mean, worst-case, and extreme scenarios enabling aggregation of their strengths. Therefore, an accurate fraud detection system assists organizations in making effective choices and reducing the overall expense of fraud detection and prevention. This paper does not consider the time windows in the experiment. However, the major challenge is the lack of data availability in NBA fraud detection.

## REFERENCES

- [1] ACFE. Association of Certified Fraud Examiners (ACFE) 2022 Report to the Nations. Accessed: 2023. [Online]. Available: <https://legacy.acfe.com/report-to-the-nations/2022/>
- [2] T. Ashfaq, R. Khalid, A. S. Yahaya, S. Aslam, A. T. Azar, S. Alsafari, and I. A. Hameed, "A machine learning and blockchain based efficient fraud detection mechanism," Sensors, vol. 22, no. 19, p. 7162, Sep. 2022.
- [3] N. S. Alfaiz and S. M. Fati, "Enhanced credit card fraud detection model using machine learning," Electronics, vol. 11, no. 4, p. 662, 2022.
- [4] A. Alfaadhel, I. Almomani, and M. Ahmed, "Risk-based cybersecurity compliance assessment system (RC2AS)," Appl. Sci., vol. 13, no. 10, p. 6145, May 2023.
- [5] D. Sarma, W. Alam, I. Saha, M. N. Alam, M. J. Alam, and S. Hossain, "Bank fraud detection using community detection

- algorithm,” in Proc. 2<sup>nd</sup> Int. Conf. Inventive Res. Comput. Appl. (ICIRCA), Jul. 2020, pp. 642–646.
- [6] A. Pagano, “Digital account opening fraud on demand deposit accounts: An assessment of available technology,” Ph.D. thesis, Utica College, Utica, NY, USA, 2020.
- [7] Shuftipro. New Account Fraud—A New Breed of Scams. Accessed: 2023. [Online]. Available: <https://shuftipro.com/reports-whitepapers/newaccount-fraud.pdf>
- [8] R. Sasirekha, B. Kanisha, and S. Kaliraj, “Study on class imbalance problem with modified KNN for classification,” in *Intelligent Data Communication Technologies and Internet of Things*, vol. 101. Singapore: Springer, 2022, pp. 207–217, doi: [https://doi.org/10.1007/978-981-16-7610-9\\_15](https://doi.org/10.1007/978-981-16-7610-9_15).
- [9] P. Vanini, S. Rossi, E. Zvizdic, and T. Domenig, “Online payment fraud: From anomaly detection to risk management,” *Financial Innov.*, vol. 9, no. 1, p. 66, Mar. 2023, doi: 10.1186/s40854-023-00470-w.
- [10] X. Zhu, X. Ao, Z. Qin, Y. Chang, Y. Liu, Q. He, and J. Li, “Intelligent financial fraud detection practices in post-pandemic era,” *Innovation*, vol. 2, no. 4, Nov. 2021, Art. no. 100176, doi: 10.1016/j.xinn.2021.100176.
- [11] M. Monge, C. Poza, and S. Borgia, “A proposal of a suspicion of tax fraud indicator based on Google Trends to foresee Spanish tax revenues,” *Int. Econ.*, vol. 169, pp. 1–12, May 2022, doi: 10.1016/j.inteco.2021.11.002.
- [12] S. Kannan and K. Somasundaram, “Autoregressive-based outlier algorithm to detect money laundering activities,” *J. Money Laundering Control*, vol. 20, no. 2, pp. 190–202, May 2017, doi: 10.1108/jmlc-07-2016-0031.
- [13] B. Xiao, B. Lei, W. Lan, and B. Guo, “A blockwise network autoregressive model with application for fraud detection,” *Ann. Inst. Stat. Math.*, vol. 74, no. 6, pp. 1043–1065, Dec. 2022, doi: 10.1007/s10463-022-00822-w.
- [14] G. Moschini, R. Houssou, J. Bovay, and S. Robert-Nicoud, “Anomaly and fraud detection in credit card transactions using the ARIMA model,” in Proc. 7th Int. Conf. Time Forecasting, Jul. 2021, p. 56, doi: 10.3390/engproc2021005056.
- [15] A. A. Alhashmi, A. M. Alashjaee, A. A. Darem, A. F. Alanazi, and R. Effghi, “An ensemble-based fraud detection model for financial transaction cyber threat classification and countermeasures,” *Eng., Technol. Appl. Sci. Res.*, vol. 13, no. 6, pp. 12433–12439, Dec. 2023, doi: 10.48084/etasr.6401.