ISSN 1611-1699 / eISSN 2029-4433

VILNIUS TECH
Vilnius Gediminas Technical University

JOURNAL of BUSINESS ECONOMICS & MANAGEMENT

2025
Volume 26
Issue 1
Pages 210–230

https://doi.org/10.3846/jbem.2025.23443

IS FRAUD DETECTION FEASIBLE WITHOUT TRAINING DATA? TESTING AN EXPERT-BASED APPROACH

Serdar BENLİGİRAY [1], Ahmet ONAY [2]✉, Fatma YAŞLI ŞEN [3]

[1]*Faculty of Business Administration, Anadolu University, Eskisehir, Turkey*
[2, 3]*Transportation Vocational School, Eskisehir Technical University, Eskisehir, Turkey*

Article History:
- received 4 May 2024
- accepted 6 January 2025

Abstract. We aim to derive a fraud detection approach applicable to conditions where historical fraud data is absent, inadequate, or outdated for making predictions. To this end, we propose a new approach to fraud detection based on expert opinion, enabling tailored tools for various conditions of economic/institutional environments. For this, we determined the relative importance of common fraud indicators based on a widely used model in the literature. We then used this information to formulate a scoring alternative to conventional versions, which uses either the original coefficients or the coefficients obtained from training the model. Finally, these scoring alternatives were compared by their detection performances. The design of this research demanded a multifaceted dataset consisting of expert opinions, financial statement data of non-financial companies in the Istanbul Stock Exchange, and local regulatory authority's notifications on fraudulent companies. The analysis of the detection performances indicates that the proposed alternative scoring method poses a feasible alternative with competitive performance and fewer data requirements. This research's approach sidesteps the training data requirement and provides financial analysts, auditors, and regulatory bodies a versatile classifier for various use cases regarding financial data, such as detecting fraudulent financial activity, as demonstrated in this study.

Keywords: financial statement fraud, fraud detection, beneish model, probit regression, expert knowledge, best-worst method, ROC-AUC analysis.

JEL Classification: M41, M42, M48.

✉Corresponding author. E-mail: ***ahmet_onay@eskisehir.edu.tr***

1. Introduction

Financial statement manipulation, fraudulent financial reporting, or financial statement fraud (hereafter, fraud, in brief) is the malpractice of distorting financial performance by intentionally altering accounting statements to deceive and mislead the public (Rezaee, 2005, p. 279). According to Association of Certified Fraud Examiners [ACFE], financial statement fraud, while the least common type of occupational fraud, is by far the most costly (ACFE, 2024, p.10). This strand of fraud encompasses a wide spectrum of actions, ranging from understating expenses and inflating assets to recording fictitious revenues, all of which significantly distort a company's true financial position. Fraud poses a serious threat to the efficient functioning of capital markets. Such manipulations can mislead regulatory authorities, investors, and

other stakeholders, leading them to form an inaccurate impression of the financial condition and performance of the company. These malpractices may serve a specific purpose, such as inflating the share price, concealing debt repayment capacity, or presenting a misleading financial image to investors. Fraud is a major concern for participants in financial markets. In this respect, fraud detection is critical to preserving integrity, reliability, and transparency (Dechow et al., 2011, p. 17).

Fraud detection is an important task for financial analysts, auditors, and regulators to ensure the efficient functioning of financial markets. However, detecting fraud is challenging as it requires knowledge about its nature and how it is practiced (Kassem & Higson, 2012). In this context, various models developed for fraud detection have become important tools. These can enable regulators to impose deterrent sanctions on manipulators and promulgate effective standards and regulations to prevent fraud. With such tools, financial analysts can protect investors by providing more accurate advice. Auditors can utilize these at every stage of the audit process, from client acceptance to audit risk identification (Albrecht et al., 2018). Models can provide useful financial information to the stakeholders in decision-making processes regarding financial reporting (International Accounting Standards Board [IASB], 2018), hence reducing the information risks of companies' shareholders, employees, or creditors and improving market efficiency (Perols & Lougee, 2011).

For more than two decades, fraud detection models have been developed to produce accurate and reliable estimations by computations on the samples representing actuality (Beneish, 1999; Dechow et al., 2011; Bao et al., 2020). To attain the best performance possible, these models require rigorous effort to include as much historical fraud data as possible by examining regulators' notifications or publicly disclosed fraud lawsuits and cases (Beneish, 1999; Spathis, 2002; Dechow et al., 2011; Perols, 2011; Repousis, 2016; Craja et al., 2020). In this process, the government's accountability mechanisms such as the Government Accountability Office Financial Statement Restatement Database, databases of lawsuits against companies such as the Stanford Law Database on Shareholder Lawsuits, reports that regulators require publicly traded companies to publish, such as the MD&A section and other occasional disclosures, should be examined in detail. However, accessing such data often poses a significant obstacle in constructing fraud detection models. In economies where regulatory agencies do not disclose fraudulent companies to the public, this becomes more challenging due to the unavailability of official fraud data.

Conventional detection models producing a probabilistic scoring output capture the yearly change in particular financial ratios obtained from items in financial statements (Beneish, 1999, p. 26; Perols, 2011, p. 28; Dechow et al., 2011). These types of models basically exploit the anomalies in the financials caused by fraudulent activities. Thus, to label a fraudulent company's financials as one unit of firm-year fraud data in the ground truth dataset, the company must not be announced for such activities for the previous year. This requirement eliminates the use of sequenced firm-years of fraud data for a company announced to have committed prolonged fraudulent activity over the past years, which results in a significant portion of the fraud data being dropped out. For instance, Perols (2011) merely utilized around one-tenth of all fraud observations in his analysis. One way to enhance the size of the fraud dataset is to expand the scope with less severe issues, such as misstatements in general (Achakzai & Peng, 2022; Bertomeu et al., 2021); however, a tool developed from this scope would be unspecific to typical frauds and the detection performance varies for sub-categories of them (Beneish & Vorst, 2022). An alternative source with abundance is the restatements. However, although restatement announcements can inform about future financial statement

fraud disclosures, not all restatements can be directly associated with fraud (Qiu et al., 2019). Parallel to this, Papík and Papíková (2020, p. 75) utilized a fraud detection tool to predict restatements, resulting in lower performance. In sum, the inadequacy of historical fraud data may pose another problem, mainly when the data is insufficient due to inappropriateness or the lack of data.

One way to overcome the difficulties mentioned above is to skip the modeling stage and use the presets of an existing detection model to produce company scores and classify those with off-limit scores as fraudulent companies (Repousis, 2016; Tahmina & Naima, 2016; Halilbegovic et al., 2020; Maniatis, 2022; Khatun et al., 2022). However, using the original coefficients and the cut-off point performs worse than the alternative of adapting these parameters to economic conditions with distinctive institutional governance characteristics, accounting practices, and rules. Lastly, classification approaches based on historical data are susceptible to underperform in sudden shifts due to remarkable changes in rules and practices, all of which affect the significance of the indicators comprising the model. As a noteworthy example, Bao et al. (2020) curtailed the most current data in their research and ended their dataset in 2008 because "the regulators reduced the enforcement of accounting fraud starting from around 2009". They also stated that there was a significant shift in U.S. firms' fraudulent behavior historically. Transformational change in the economy is another common reason for curtailing potential data. Duan et al. (2024, p. 3) discarded around one-third of the fraud incidents from their analysis data for such reasons. Our research is motivated by two main research questions (RQ) in the context of the limitations of existing fraud detection approaches presented thus far:

RQ.1: Can expert opinions be a basis for an alternative fraud detection tool to overcome the implementation challenges of current approaches?

Expert opinion-based methodologies can help overcome the challenges we have identified for the fraud detection models in the literature. Experts' experience and insights can play a critical role in model development and fraud detection, especially when there is a lack of data or limited access. Furthermore, expert opinions on the relative importance of fraud indicators reveal prominent types of fraudulent activity in their native environment. Lokanan (2017, p. 903) distinguishes macro factors from individuals' drivers of fraudulent behavior. These factors are related to the socio-cultural environment, which is also reflected in the rules and practices of an economic environment. Domain experts native to a particular environment are useful for interpreting such reflections of fraud on financial reports. The main objective of this study is to develop a fraud detection tool based on expert opinions instead of utilizing verified ground truth data on fraudulent activities. The experts are expected to be native to the concepts of accounting fraud and financial manipulation, and the Beneish Model (Beneish et al., 2013) is a seminal work in this area. Hence, this model is selected to be the basis for the experts' evaluations on the relative weights of importance for the comprising fraud indicators in a particular economic environment. The coefficients are weighted by Best-Worst Method (BWM), a Multi-Criteria Decision Making (MCDM) methodology proposed by Rezaei (2015, 2016). These coefficients are then used to calculate expert opinion-based fraud scoring labeled as Exp-score. To the best of our knowledge, this study is the first to apply such an MCDM methodology to weight the coefficients of a specific fraud detection model.

RQ.2: Does an expert opinion-based alternative compete with the conventional approaches for the same fraud detection model?

A consequential objective of our study is to compare the detection performance of the proposed alternative with the basis model's conventional utilization modes, which are either to use the presets or to re-run the model with the specific data. For this purpose, probit regression analysis is used to obtain coefficients adapted to a specific economic environment, which is Turkey, for this particular study. The scores calculated with these coefficients are named the Z-scores, whereas the original name of the scoring is preserved for the coefficients of the original study, the M-scores. The ease of application advantage is expected to be traded off by the lower detection performance of these scoring alternatives. Among these, the most practical is the M-score. Conversely, the Z-score requires computation and data gathering for the ground truth. Here, we posit that the EXP-score resides in the midst of this spectrum as an intermediate alternative in both performance and the prerequisites for the calculations. To compare these alternatives, we conducted a Receiver Operating Characteristic (ROC) analysis, a robust technique in evaluation detection performance, and provided additional statistical tests for the area under the ROC curve, which we believe to be fruitful for both the researchers and the practitioners of fraud detection.

The remainder of the paper is structured as follows: The subsequent section introduces the related work in the field. The research methodology, analysis, and findings are presented in sequence. The study is concluded in the final section.

2. Related work

Over the last quarter century, recurring accounting scandals have motivated researchers to develop tools for detecting fraud, which often work with financial statement data. The items of financial statements have underlying patterns of interrelations caused by appropriate recordings of normal business activities. Unusual changes in these patterns may signal irregular business activities, implying cover-ups of a fraud act or motivational conditions for potential fraud. Many fraud detection models focus on capturing such irregularities through financial ratios. Research adopting various methodologies found financial ratios to be functional in fraud detection.

Financial reports are the primary data source for the studies, and there are common inputs borrowed from previous research (Ali et al., 2023; Rahman & Zhu, 2023; Zainudin & Hashim, 2016; Perols, 2011; Dikmen & Küçükkocaoğlu, 2010). Here, a seminal one is Beneish's model. Some research (Dechow et al., 2011; Perols, 2011; Cecchini et al., 2010) achieved higher detection success by using models with more inputs computed on relatively larger fraud datasets. The most frequently used classification tool is logistic regression (Shahana et al., 2023). Recent studies have also utilized methods such as Support Vector Machine (SVM), Artificial Neural Networks, Bayesian Classifiers, Decision Trees, or a combination of these by utilizing ensemble learning methods, which are some of the numerous classifiers adapted from the prolific research field of machine learning (Ramzan & Lokanan, 2024; Shahana et al., 2023; Ashtiani & Raahemi, 2022).

This section on related work aims to present a framework for determining the basis for a comparable tool derived from expert evaluations rather than ascertaining the state-of-the-art among popular classifiers. However, it is worth noting that conventional methods, i.e. Beneish's M-score and Dechow F-score, remain prominent in various benefit-cost settings of prediction and provide a net benefit over many other alternatives, including financial kernel SVM presented in Cecchini et al. (2010) and ensemble learning alternative as proposed by Bao et al. (2020) (Beneish & Vorst, 2022). Achakzai and Peng (2022) reported a close detection

performance for the stand-alone version of logistic regression and the RUSBoost classifier. In recent studies, logistic regression has been used in combination with RUSBoost (Achakzai & Peng, 2022) and XGBoost (Zhao & Bai, 2022) to enhance classification performance.

The general construct of the datasets comprises fewer fraudulent companies in proportion to the non-fraudulent control companies. In our review of similar research, the medians of the ratio and the size of fraud data are 0.07 and 76.5, respectively. Some studies have adopted an approach where fraudulent and non-fraudulent companies are matched in the sample, a fifty-fifty fraud ratio (Spathis, 2002; Zainudin & Hashim, 2016; Gepp et al., 2021). However, the most common approach is observed to construct more realistic datasets with lower ratios of fraud in the datasets. The limited amount of fraud data in previous studies (Ashtiani & Raahemi, 2022, p .72513) may also be attributed to the difficulties in collecting ground truth data in this particular research field. This is a strong motivation for us to develop a detection tool based on expert opinions rather than depending on the verification of past incidents.

MCDM techniques are effective when gathering structured insights from domain experts. These techniques have been applied to many finance-related settings, including capital budgeting and financial planning, investment appraisal, auditing, portfolio management, and bankruptcy prediction/credit scoring, where the last two had greater attention lately (Marqués et al., 2020). In many MCDM settings, machine-learning algorithms attain higher prediction performance but lack explainability and have lower operating transparency than MCDM techniques (Černevičienė & Kabašinskas, 2022). MCDM methods such as Analytic Hierarchy Process (AHP) require input for criteria weights, and the domain experts are the fundamental source for this; however, other methods such as SAW and TOPSIS can directly assign the weights equal importance, or databases and artificial intelligence-based estimation can be used alternatively (Martinkutė-Kaulienė et al., 2021, p. 65). Here, we focus on the related work utilizing expert knowledge in finance and economics, which exhibits a similar motivation and approach to our study.

The expert knowledge-based approach has been adopted for various finance and economics problems in the literature. Zhao et al. (2021) designed an MCDM model with an integration of expert knowledge using AHP for the portfolio selection. An application scheme similar to fraud prediction is bankruptcy prediction, as its consequences, prevention motivations, and the techniques used are alike. Domain knowledge and expert opinion are fruitful for the feature selection stage in the bankruptcy prediction or credit scoring processes (Mokrišová & Horváthová, 2023; Lappas & Yannacopoulos, 2021). In this line of research, Mokrišová and Horváthová (2023) measured ROC-AUC (Area Under the Curve) the detection performances of the selected feature sets, which are formed by logistic regression and expert opinion-based approach. Their research question and design are akin to those explored in our paper. Expert insight also provides illuminating guidance for internal audit and control processes. Panigrahi (2011) proposed a knowledge-driven framework for internal fraud detection, combining forensic auditor judgment and data analytics. This integration takes advantage of the efficiency and practicality of utilizing domain-specific expertise weighted by an MCDM method to develop adaptive internal systems for detecting fraudulent activities.

Fraud detection is a research area that has not been extensively explored by utilizing expert opinions and/or MCDM methods. Hooda et al. (2018) compared the contemporaneous tools based on multiple criteria modeled as an MCDM. Their paradigm does not directly contribute to the classification of fraudulent firms. Huang et al. (2017) addressed the application problems of the methods requiring large feature sets. Beginning with an extensive feature set, they narrowed it down to a limited list of features obtained in an MCDM scheme involving

domain experts. This research presents the relative weights of the measures categorized by the fraud triangle. Lin et al. (2015) utilized classifiers, including logistic regression, to rank the potential features of financial statement fraud and compared them with the importance ranking of the features derived from auditing experts' evaluations by using the AHP method. The result shows inconsistencies between the expert evaluations and mechanical classifiers. The paradigm employed by Lin et al. (2015) bears some resemblance to the approach delineated here in this study. However, the significance of their expert evaluations was indecisive because those evaluations were not transformed into a comparable classifier.

Hamal and Senvar (2022) addressed the determination of which financial ratios are more valuable for fraud detection as an MCDM problem. Their study aims to determine which financial ratios are more important among the many financial ratios included in the fraud models in the literature. Gepp et al. (2021), instead of expert opinions, utilized data from previous studies but served a similar purpose as Hamal and Senvar (2022). However, the general idea of mixing variables may not be effective in enhancing the detection performance. Accordingly, Bao et al. (2020) confirmed that more predictors do not necessarily improve prediction performance. A mixed selection of variables derived from common sense misses the complementary aspect of the model constituents. This is because the variables of some models can only indicate fraud collectively, not individually (Beneish et al., 2013, p. 76). A modification to this approach is to divert the experts' evaluations to a specific set of variables defined by a theoretically sound model. This approach can also be defined as fine-tuning an existing detection tool by expert judgment. Our study aims to provide the literature with a feasible fraud detection tool by adopting this idea. The next section gives a detailed explanation of the research design and data.

3. Research design and data

The research presented in this study is based on the basic idea that expert opinion-based scoring can effectively compete with conventional alternatives of financial fraud detection. Accordingly, the research is designed in two stages: the first stage reproduces the reference scoring alternatives and constructs a novel scoring mode based on expert judgments, and the second stage compares the classification performances of these. In the financial manipulation concept, the original idea for such scoring is to utilize the coefficients of a probit regression model, i.e., the Beneish Model. This model forms a suitable foundation for creating a scoring based on expert opinion. The arguments for choosing this model to be the basis can be summarized as follows:

- The components are financial statement-based ratios with which the experts are familiar to evaluate the relative importance.
- Constructed to be tested on the ground truth data gathered from the filings of regulatory bodies.
- Provides a well-appreciated financial manipulation score for various use cases from scholars to practitioners in the financial market (Morris, 2009; Vincent, 2012; Zumbrun, 2023).

The original coefficients of this model have been used widely as presets of a financial manipulation scoring formula; however, a formula derived by specifically run probit regression coefficients is expected to be superior in classification performance. These two modes constitute the reference scoring alternatives of this research. On the other hand, the proposed alternative is constructed by referring to the experts to obtain the relative weights of importance

for the components of the scoring formula identical to the reference alternatives. BWM is utilized in this process, an MCDM method that provides consistency with fewer comparison requirements than its alternatives. In the second stage, the performances of these scoring methods are analyzed through the ROC curve plots and AUC. Each process explained thus far requires a particular combination of data sources, which are explained below. The flow of the whole process is visualized in Figure 1.

The variables of the Beneish Model require financial data for both fraudulent and non-fraudulent companies. However, fraudulent company activities require verification by the regulatory authority announcements. This study's research environment is the Turkish economy, and the primary regulatory body responsible for financial malpractices in Turkey is Capital Markets Board (CMB). CMB issues weekly bulletins, which include financial reporting misconduct and related fraudulent activities of the publicly listed companies on the Borsa Istanbul. These are accessed through CMB's database, which dates back to 2013. This database is similar to the Accounting and Auditing Enforcement Releases by the U.S. Securities Exchange Commission. In the data collection process of this research, the bulletins are scrutinized to build the dataset for the fraudulent companies verified by the market regulators. Besides the significant financial fraud issues, these releases encompass various issues that can be considered in minor relation or indirectly related to financial reporting misconduct, e.g., late reporting and missing auditor's report. Thus, the companies are determined to be fraudulent if they are disclosed for acts of financial manipulation and fraud resulting in CMB's enforcement actions, i.e., corrective action directives, injunctions, monetary penalties, and prohibition/suspension orders.

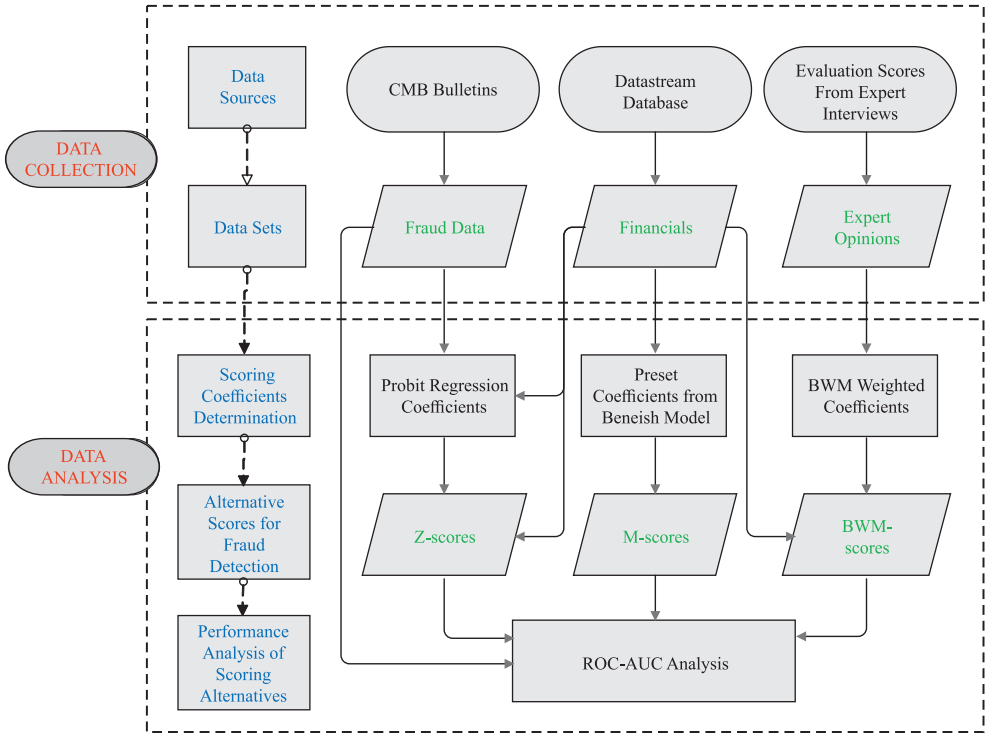

Figure 1. Research design diagram

A comparison of the performance of alternative modes for determining the scoring parameters requires non-fraud data as for the training dataset for the probit model. In similar studies, the compositions of the complete dataset differ by the presumptions and research design, and this affects the percentage of the fraud data, varying from high up to fifty percent (one-to-one match) and down to three percent (reflecting reality) (Cecchini et al. 2010, p. 1152). The financials of the publicly traded companies that have never been on the CMB's releases are presumed to be non-fraudulent. However, it is possible that the authorities failed to detect and disclose all the incidents. This problem is possibly more aggravated in emerging economies with weaker governance mechanisms. Being cautious and conservative, a target ratio of fraudulent firm-year data to form a realistic dataset for an emerging economy can be around one-tenth. According to these arguments, the analysis data is completed to the size of 300 firm-year with the data of the publicly traded non-financial companies non-disclosed in the CMB's releases for the given period. The financial data is obtained from the Refitiniv Eikon-Datastream Database.

The design of this research also requires expert opinion to derive an alternative metric based on the weights of importance of the financial fraud indicators given in the Beneish Model. The expert opinion data was derived by consulting scholars and forming their responses according to the methodological requirements. Participants were formerly informed about the research objective and supplied with a brief note about the basics of the basis model. The expertise of the total of fifty-one participants was accounting and finance. Twenty-three of them specialized in auditing, fourteen of them were financial accounting specialists and the remainder were experts in cost accounting, economics and finance. The following sections present the analyses and discuss their findings.

4. Fraud detection by financial manipulation scoring

Companies can be scored for their susceptibility to financial fraud by processing their reported financials. Such scores can be derived from the coefficients of a probit regression output. As the probit model given below, we modeled the probability of a firm's financials being manipulated, that is $Y_i = 1$.

$$\mathrm{P}\left[Y_i = 1 \mid X_{1i},\ldots,X_{Ki};\beta_0,\ldots,\beta_K\right] = \Phi\left(\beta_0 + \sum_{k=1}^{K}\beta_k X_{ki}\right). \tag{1}$$

In Eq. (1), $\Phi(\cdot)$ is the Cumulative Distribution Function (CDF) of the standard normal distribution. The explanatory variables are determined in line with the Beneish Model. The coefficients to be used for scoring are obtained by probit regression output, which is presented in comparison with the original paper's output in Table 1. In general terms, the probit regression model transforms the response variable Y from binary condition (i.e., one for manipulated financials, and zero otherwise) to continuous data using the cumulative normal distribution function.

$$Y = \Phi\left(X\beta + \in\right); \tag{2}$$

$$\Phi^{-1}Y = X\beta + \in; \tag{3}$$

$$Y' = X\beta + \in. \tag{4}$$

In the closed form of the model given in Eq. (4), $X\beta$ is utilized for generating Z-scores for firm-year data. These coefficients do not convey information on the magnitudes for a

constant effect on the probability output by the nature of the model. Yet, positive coefficients translate to the higher the values of *X*, the more likely the expected condition to occur. The coefficients of a probit regression can practically be used for scoring an entity as an indicator of a certain condition. In this regard, subsequent research using the Beneish Model as a classifier conventionally utilized its original set of coefficients and often referred to these scores as M-scores. These coefficients are given in Table 1 section B. Beneish's analysis is based on the financials of companies in the United States in the late 90's. Nevertheless, its coefficients have been applied in various economic conditions ever since, which requires a perpetual examination of the generalization for these parameters. In this stage of our research, probit regression is computed with the reproduced dataset, which is quite similar to the original study. The outcome of this analysis is presented in Table 1, section A. The regression outcome presented in this table supports the model's generalization with the consistent results obtained from a considerably different economic environment and time. However, the coefficients differ, to a degree, reflecting the particular importance of the indicators affecting the estimated probability of a company committing manipulation in that specific economic environment. In this essence, specialized Z-scores are expected to be superior to general M-scores in terms of classification performance.

Table 1. Probit regression output in comparison to the original model output

A. Probit Regression Output					χ^2 (p-value)	Pseudo R^2	Data Set (Manipulator/Control)	
					0.001	0.300	31 / 269	
Cons	DSRI	GMI	AQI	SGI	DEPI	SGAI	LVGI	TATA
–3.332	0.950	0.045	0.201	0.424	0.247	–0.025	–0.317	–2.514
(–4.99)	(4.95)	(0.36)	(1.68)	(3.28)	(2.02)	(–0.07)	(–0.79)	(–3.32)
B. Beneish Model Original Output					χ^2 (p-value)	Pseudo R^2	Data Set (Manipulator/Control)	
					0.001	0.371	24 / 648	
Cons	DSRI	GMI	AQI	SGI	DEPI	SGAI	LVGI	TATA
–4.840	0.920	0.528	0.404	0.892	0.115	–0.172	–0.327	4.679
(–11.01)	(6.02)	(2.20)	(3.20)	(5.39)	(0.70)	(–0.71)	(–1.22)	(3.73)
(t-statistics in parentheses)								

Financial analysts in practice and scholars in theory widely utilize M-scores as an indicator: firms with off-limit scores are considered potential manipulators. However, the efficiency of this method is dependent on its prediction performance. M-score has thus far proven its effectiveness in various applications on distinctive datasets gathered from different economies and economic conditions. However, specifically run probit regression coefficients are normally the most efficient predictors for the data of concern. Hereafter, the predictions from these specialized coefficients are called Z-scores. A classification by the probit Z-scores with a threshold level (or cut-off point) is demonstrated in Figure 2. This figure exhibits the classifier's performance and visualizes the confusion matrix, which is a summary of all correct and false predictions generated for the specific dataset. In a binary classification as such, the predictions are classified into two classes. In Figure 2, the class of concern is the manipulators, and thus, they are conveniently called positives, while the others are called negatives.

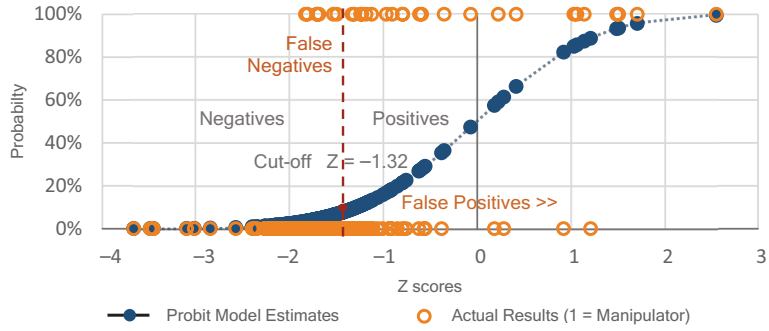

Figure 2. Probit Z-score prediction performance on CDF graph

In the original classification application of the model, the threshold level is determined in favor of better sensitivity (or, True Positive Rate (TPR)), arguing that undetected financial manipulations (false negatives) would jeopardize traders' portfolios (Beneish, 1999, p. 34). Given the trade-off between sensitivity and precision, this approach determines the cut-off point sub-optimally for classification precision, which is the true prediction rate for the positives (Fawcett, 2006, p. 862). However, precision is an important aspect of a detection method if the model is used as an early warning system for a detailed (hence, costly) investigation with limited sources (Cecchini et al., 2010, p. 1156). As this is the case for the other major parties of the financial system, such as regulatory bodies and independent auditors, precision is of importance as well as sensitivity.

Lowering the threshold increases the number of positive predictions and, therefore increases TPR with a degree of precision loss. Put differently, lowering precision inflates positive predictions, increases the false positive rate (FPR), and decreases specificity (or, True Negative Rate (TNR)). To meet the needs of the aforementioned parties in the financial market, our study utilizes an all-around measurement of classification performance as a cost-agnostic metric for each type of classification error. This approach is explicitly divergent from the objective of the original cut-off point (–1.78) for the M-score. Indeed, the cut-off point in Figure 2 (–1.32) is determined to maximize correct classification probability by using the Youden Index (Martínez-Camblor & Pardo-Fernández, 2019, p. 2). This index, the maximum difference between TPR and FPR, is error cost neutral and does not evaluate the utility differences of the classifications (Baker & Kramer, 2007, p. 346).

5. Methodology for the analysis of classification performance

The performance of a classifier can be measured at a particular threshold with one measurement (TPR, FPR, TNR, etc.). Such single-metric can be reasonable for a particular application. Nevertheless, how to choose the correct threshold value remains unclear. ROC plots TPR against FPR changing in the same direction at the range of zero to one for each axis. Thus, the curve is a two-dimensional reflection of the performance across various threshold levels. Despite significant research in accounting, the use of ROC analysis in inferential studies has recently gained prominence (Zhou et al., 2024; Bačo et al., 2023; Durana et al., 2022; Brown et al., 2020). However, this method is predominant in research treating fraud detection as a generic classification problem (Rahman & Zhu, 2024; Khan et al., 2022; Zhao & Bai, 2022). An

elaborate AUC comparison of fraud detection modes is presented by Cecchini et al. (2010), and Abbasi et al. (2012) provide monetary performance measurements of varying error costs for both investors and regulators in addition to AUC. Benesih and Vorst (2022) iterate this alternative strand of performance comparison for these parties, adding the auditor's side of view.

ROC analysis is alternatively used in model validation procedures, which require the research data to be split into training and testing subsets (Saito & Rehmsmeier, 2015, p. 2). However, this may impair the model development stage as a significant amount of interpretable information is given up for the validation process unless the data is large enough. ROC-AUC can also be used as a post-estimation tool for performance comparison of distinct models trained with the complete dataset (Cleves, 2002, p. 306). The fraud data gathered for this study is a condensed extraction from the regulatory authority's reports released from the beginning, adequate to train a probit regression similar to relevant research. This makes the data less expendable for other procedures than training, and hence, ROC analysis for the Z-scores in our study encompasses all the data and serves as a goodness of fit indicator of the probit model in broad terms. Besides, Z-score classification is constructed as a benchmark in the research design, representing the best-performing classifier, i.e., the gold standard with the theoretical best ROC-AUC performance for the given model.

The alternatives to a gold standard are often disregarded as they typically underperform. Nevertheless, an alternative classification measure can be feasible with competitive performance and fewer input requirements. In this regard, applications using financial manipulation scoring as a tool depend widely on M-scores with the preset coefficients rather than reproducing them on the ground truth data constructed by gathering the verified true manipulators' data. This one-size-fits-all formula seems appealing, although using the presets for different economies adopting distinct legislations/regulations lowers the classification performance. In this context, a better and still practical classifier can be constructed using tailored coefficients derived from expert opinion. Expert opinion is expected to adapt the coefficients to various economies' distinctive regulations and conjunctures affecting the manipulation indicators. The coefficients are obtained by the BWM, and used for calculating the expert opinion-based scores abbreviated as EXP-scores. The next sub-section introduces BWM and presents the outcome to be used as EXP-score coefficients. ROC-AUC analysis of Z-scores, M-scores, and EXP-scores are to be presented right after the following section.

6. MCDM coefficients for expert opinion-based scoring

In the expert opinion-based scoring alternative, the coefficients of the scoring components are posited as relative weights of importance for the eight constituents of the fraud detection model. Since the cognitive ability of an expert as a decision-maker is limited for many concurrent evaluations of information, MCDM approaches are useful in evaluating multiple factors and determining their relative importance. In this study, BWM is used to determine the importance weights of the factors. Using BWM based on pairwise comparisons, it is possible to effectively and easily infer expert judgments for determining the importance of the fraud indicators constituting the model.

BWM is an MCDM using mathematical modeling (Rezaei, 2015). It is considered as an improved version of the AHP method based on pairwise comparisons. In AHP, $n(n-1)/2$ pairwise comparisons are required for n factors, meaning that the number of comparisons swiftly inflates for an exceeding number of factors, which is the case in this study. Potential concentration losses and/or the limitations of cognitive capacity to compare multiple factors

concurrently may lead to repetitive inconsistencies in pairwise comparison matrices in such cases. In BWM method, there are only best-to-other and other-to-worst comparisons with a total of 2n–3 comparisons and the rest is solved by mathematical optimization. Thus, BWM requires fewer pairwise comparisons and provides more consistent results than AHP (Rezaei, 2015, 2016). This method efficiently determines the importance weights of the evaluated elements by means of a mathematical model that uses as input the experts' determination of the best and worst criteria and only the pairwise comparison data associated with them. Our study is the first to apply BWM to determine the importance of the indicators in a financial fraud detection model. The steps of BWM are as follows (Rezai, 2015, 2016):

Step 1: Determination of the related criteria set $\{c_1, c_2, c_3 \ldots . c_n\}$.

Step 2: Identifying the best criterion (e.g. most desirable or important) and the worst criterion (e.g. least desirable or important).

Step 3: Pairwise comparisons for the Best-to-Others. By comparing the preference of the best criterion *B* over the other criteria *j*, the following vector A_B consisting of a_{Bj} is determined (Eq. (5)). A rating scale between one to nine is used for comparisons, where one is equally important and nine is extremely more important.

$$A_B = \left(a_{B1}, a_{B2}, a_{B3} \ldots . a_{Bn}\right), \text{ where } a_{BB} = 1. \tag{5}$$

Step 4: Pairwise comparisons for the Others-to-Worst. By comparing the preference of the all criteria j over the worst criteria *W*, the following vector A_W consisting of a_{jW} is determined (Eq. (6)). A rating scale between one to nine is used for comparisons, where one is equally important and nine is extremely more important.

$$A_W = \left(a_{1W}, a_{2W}, a_{3W} \ldots . a_{nW}\right)^T, \text{ where } a_{WW} = 1. \tag{6}$$

Step 5: Developing the optimization model to identify the optimal weights of the factors

$$\left(w_1^*, w_2^*, w_3^* \ldots w_n^*\right), \text{ where } a_{Bj} = \frac{w_B}{w_j} \text{ and } a_{jW} = \frac{w_j}{w_W}.$$

To satisfy the equations conditions, BWM minimizes the maximum absolute differences

$$\left|\frac{w_B}{w_j} - a_{Bj}\right| \text{ and } \left|\frac{w_j}{w_W} - a_{jW}\right| \text{ for all } j \text{ values.}$$

Also, provided that the criteria weights are non-negative and the sum of their weights is one, the model is constructed as presented in Eq. (7):

$$\min_j \max \left\{ \left|\frac{w_B}{w_j} - a_{Bj}\right|, \left|\frac{w_j}{w_W} - a_{jW}\right| \right\} \quad s.t. \quad \sum_j w_j = 1 \quad w_j \geq 0 \text{ for all } j. \tag{7}$$

Hence, the model can be transferred to the Eq. (8):

$$\begin{aligned} &\min \varepsilon s.t. \left|\frac{w_B}{w_j} - a_{Bj}\right| \leq \varepsilon \text{ for all } j \quad \left|\frac{w_j}{w_W} - a_{jW}\right| \leq \varepsilon \text{ for all } j \\ &\sum_j w_j = 1 \; w_j \geq 0 \text{ for all } j. \end{aligned} \tag{8}$$

For BWM using a nonlinear min-max model, Rezai (2016) proposed the following linear model that provides a single solution as presented in Eq. (9):

$$\min \varepsilon^L \; s.t. \left| w_B - a_{Bj} w_j \right| \le \varepsilon^L \text{ for all } j \left| w_j - a_{jW} w_W \right| \le \varepsilon^L \text{ for all } j$$
$$\sum_j w_j = 1 w_j 0 \text{ for all } j. \tag{9}$$

Solving Eq. (9) yields the decision variables $(w_1^*, w_2^*, w_3^* \ldots w_n^*)$ and ε^{L*} as the optimal weights and a direct consistency indicator, respectively. ε^{L*} is desired to be close to zero as the outcome for the comparisons be more reliable.

Experts' data with high levels of consistency ratio are discarded in the process, and conventionally, values lower than 0.1 are recommended for reliable evaluations. Filtering the data with such restriction provides more consistency while decreasing the number of eligible evaluations. Table 2 presents three sets of importance weights determined by Eq. (9) from the expert datasets filtered with altered restriction conditions, and the average consistency ratios are given at the bottom line of the table.

Table 2. Importance weights of the indicators by consistency limits

Indicator	No Restriction	Consistency < 0.11	Consistency < 0.10
DSRI-Days Sales in Receivables Index (F_1)	0.189	0.186	0.171
GMI-Gross Margin Index (F_2)	0.145	0.163	0.173
AQI-Asset Quality Index (F_3)	0.150	0.146	0.149
SGI-Sales Growth Index (F_4)	0.140	0.152	0.130
DEPI-Depreciation Index (F_5)	0.090	0.092	0.082
SGAI-Sales, General, and Admin. Exp. (F_6)	0.071	0.079	0.075
LVGI-Leverage Index (F_7)	0.110	0.089	0.098
TATA-Total Accruals to Total Assets (F_8)	0.105	0.093	0.121
Average Consistency Ratio	0.119	0.093	0.086

The remainder of the analysis is conducted on the most restricted expert base, whose output is given in the rightmost column in Table 2. EXP-scores are calculated using the coefficients given in this column. Detailed data from the BWM evaluations are available upon request from the authors. The following section compares the detection performance of the proposed expert opinion-based scoring, that is EXP-score, with the formerly introduced alternatives, namely M-score and Z-score.

7. ROC-AUC analysis for the scoring methods of fraud detection

Three distinct approaches for determining the parameters of the scoring formula have been discussed thus far. The ease of application advantage is expected to be traded off by the lower detection performance of the methods associated with these approaches. Among the given alternatives, the most practical is M-score, which uses the presets and only requires the financials of testing companies. Z-score requires the most rigorous effort: obtaining historical data of true manipulators' financials besides the others, and processing these. Here, EXP-

score is posited as an intermediate alternative to those requiring less input than Z-score while providing some degree of adjustment to improve the detection performance. In this regard, given approaches of financial fraud detection are expected to have ordinal performance: processing the ground-truth data is the best, and deriving from expert opinion is better than using the universal preset. Testing this assumption can reveal the potential of EXP-score for many situations, such as the absence of ground-truth data or swiftly changing conditions that urge a need for forward-looking adjustment on the scoring parameters.

In this section, the performances of the scoring methods are analyzed through ROC-AUC Analysis. To initiate, empirical estimations of the cut-off points for each scoring alternative are presented in Table 3. Combined with the coefficients given in the former sections, cut-off points complement ready-to-use and up-to-date scoring parameters for further research in classification applications with similar data. As discussed earlier, the optimization objective of the cutpoints is defined to maximize the Youden Index, which equally weights true positive and true negative rates. The remaining measurements are calculated for the optimal cutpoint. Based on these, Z-score performs best, M-score underperforms, and EXP-score performs between these two.

Table 3. Empirical cut-off point estimations for scoring alternatives

	Z-Score	M-Score	EXP-Score (All Experts)	EXP-Score (Auditing Specialists)
Optimal Cutpoint	–1.320	–1.982	1.075	1.058
Youden Index (J)	0.625	0.357	0.506	0.544
Sensitivity (TPR) at Cutpt.	0.774	0.581	0.581	0.645
Specificity (TNR) at Cutpt.	0.851	0.776	0.925	0.899
AUC by Optimal Cutpoint	0.812	0.678	0.753	0.772

EXP-score coefficients are obtained from experts with a variety of specializations in the field. These are grouped into auditing, cost accounting, financial accounting, and finance. The importance weights of the indicators for each area of expertise are also calculated and available upon request from the authors. Some types of expertise can be considered more informative on the weights of the fraud indicators. For instance, auditing specialists are prominent in knowledge and experience about fraudulent activities. In this essence, an alternative score is derived from the auditing specialists, arguing that refining the expert base may increase the classification performance. Despite a slight decrease in TNR, this version of EXP-score is better in other aspects and, therefore, adopted for further analysis. A final note on the outcome in Table 3 is that EXP-scores better detect non-manipulators than the conventional alternatives at the optimal cutpoint. For every potential cutpoint, detection performances of the alternatives are visualized in ROC curves given in Figure 3.

Figure 3 supports the initial findings of the superiority of EXP-score over M-score by exhibiting the EXP-score curve being above M-score's curve throughout the line. Here, Z-score resides at the top, representing the gold standard. Visual comparison of the sizes of ROC areas is easy as the curves do not cross each other; however, exact AUC measurements are also given in the legend in Figure 3. In this regard, EXP-score poses a feasible alternative to Z-score by offering better performance than the presets with marginally less effort. The AUCs are also compared statistically; the results are presented in Table 4.

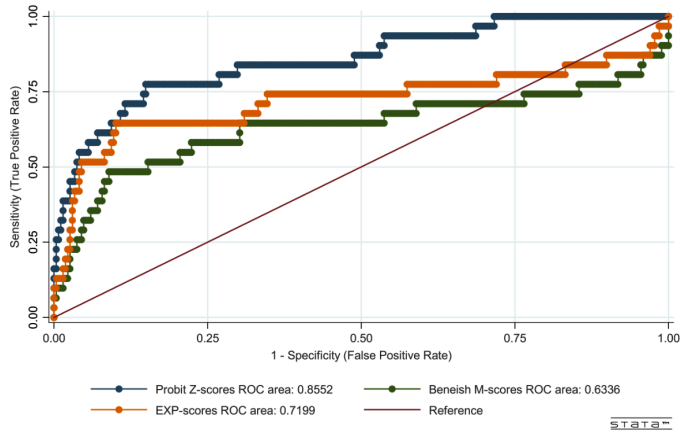

Figure 3. ROC plot for Z-scores, M-scores and EXP-scores

Table 4. Descriptive statistics and the tests of equality of AUCs for scoring alternatives

A. Descriptive Statistics of ROC Areas					
	Obs.	ROC Area	Standard Error	Confidence Interval (%95)	
				Lower Limit	Upper Limit
Z-score	300	0.855	0.040	0.777	0.933
M-score	300	0.634	0.072	0.492	0.775
EXP-score	300	0.720	0.067	0.589	0.852
B. Chi-squared Tests of Equality for AUCs					
Testing Hypothesis			df	Chi2	Prob>Chi2
H_{01}: AUC(Z-score) = AUC(M-score) = AUC(EXP-score)			2	8.330	0.016
H_{02}: AUC(Z-score) = AUC(M-score)			1	8.052	0.005
H_{03}: AUC(Z-score) = AUC(EXP-score)			1	3.811	0.051
H_{04}: AUC(M-score) = AUC(EXP-score)			1	3.308	0.069

The basic method of statistical comparison is to check for overlapping confidence intervals to determine if the differences are insignificant. The descriptive statistics and the confidence intervals for the ROC areas are given in Table 4 section A. Discrete confidence intervals of Z-score and M-score indicate statistical difference at a ninety-five percent confidence level. On the other hand, the overlapping intervals for EXP-score do not indicate any significant difference between both. The differences are explored more thoroughly by Chi-squared Tests of Equality for the AUCs suggested by DeLong et al. (1998). The outcomes for hypothesis testing of the comparison combinations are presented in Table 4 section B. All the hypotheses are rejected with statistical significance in this section. Statistically proven differences between AUCs indicate the aforementioned order for the detection performances of the alternatives. Recall that Z-scores predict its training data, resulting in overfitting that inflates AUCs. Considering this, EXP-score stands out with its relative predictive ability in the final evaluation. The next sections consolidate the key insights and their implications from the analyses.

8. Discussion

This section presents a series of implications for the analysis outcome based on our review of the related work. First, following Lin et al. (2015), we comparatively evaluate the prominence of the indicators in machine creation (Z-score) and expert opinion (EXP-score) by ranking each method's outputs, which are the magnitude of the probabilistic effects and the importance of weights, respectively. Three of the first four indicators remain in both rankings of two classifiers. Six of the indicators remain in the range of a maximum of two ranks away from each other. This can be interpreted as the experts mostly agree with mechanical inference. On the other hand, Total Accruals to Total Assets (TATA) leaps forward by three, while Depreciation Index (DEPI) falls back by four in the rankings. This translates into the expert being more cautious about accruals and being negligent in depreciation practices to a degree. Days Sales in Receivables Index (DSRI) is the most influential indicator of fraud for each scoring formula, so the experts confirm the susceptibility of an indication of fraud for inflated receivables. These findings accentuate the difference in subjective perceptions of the experts and an objective classifier.

The adoption of the proposed approach is not limited to the creation of a scoring formula as exhibited in this study. Alternatively, expert opinions as developed here can be evaluated as complementary to a more complex machine learning classifier, e.g., an additional feature in ensemble learning (Duan et al., 2024; Bao et al., 2020), in combination with a gradient boosting machine (Zhao & Bai, 2022), or a diverse predictor to be embedded in a meta-classifier including RUSBoost and logistic regression (Achakzai & Peng, 2022). An alternative use case of the MCDM framework is in the feature selection process (Lappas & Yannacopoulos, 2021; Huang et al., 2017) in feature engineering, which is a specialty of applying domain knowledge and theoretical foundations to data science (Duan et al., 2024, p. 182). This research will encourage such applications by providing consistent results of expert evaluations and machine-generated output on feature evaluation. The approach and the paradigm provided in this study also enrich the literature that addresses deficiencies of historical data-driven detection methods through, e.g., prediction point in time horizon (Duan et al., 2024); supervision aspect of the training phase (Carcillo et al., 2021); look-ahead bias (Rahman & Zhu, 2024; Bao et al., 2020); infeasibilities of gathering big data (Huang et al., 2017; Lin et al., 2015); and the complexity hindering the potential real-life applications (Papík & Papíková, 2020, p. 66).

Expert knowledge is an out-of-the-box aid for many hard data-driven methods in solving various problems in finance and economics domains. Extant studies applied this rewarding approach to many finance-related settings, including capital budgeting and financial planning, investment appraisal (Černevičienė & Kabašinskas, 2022; Marqués et al., 2020), portfolio management (Zhao et al., 2022) and bankruptcy prediction/credit scoring (Mokrišová & Horváthová, 2023; Lappas & Yannacopoulos, 2021). To the best of our multidisciplinary literature survey, the present study is the first to use the MDCM application for fraud detection utilizing expert opinions. Financial fraud risk is somewhat similar to bankruptcy risk, as the realization of both results in a decrease in the stock price – a negative condition when managing portfolios. Regarding these, fraud detection is internally related to those two basic lines of research in the finance area. In this regard, the application provided in this paper contributes to the utilization of the acquisition of expert knowledge on some of finance's core domains.

Lenard and Alam (2008) emphasize the value of expert knowledge as a crucial element in knowledge management, arguing that human insights contribute significantly to decision support systems and adaptive models in fraud detection. This aligns with the EXP-score

model's design to be robust to changes in data and adaptable to various financial environments. Moreover, Eining et al. (1997) highlight that decision aids relying on expert judgments play a pivotal role at many stages of the audit process, from client acceptance to analytical procedures for risk assessment and fraud detection by auditors. In line with these, the EXP-score model demonstrates the practical importance of structured expert input as a decision aid for auditing businesses, complementing established fraud detection tools and enhancing the robustness of auditing practices.

This study contributes to the field of financial fraud detection in several ways. First, it introduces a feasible application of expert-based scoring (EXP-score) in fraud detection, offering a viable alternative when traditional data-driven approaches are challenging to implement. Second, it demonstrates how expert knowledge can be systematically incorporated into quantitative fraud detection techniques, potentially improving the adaptability and robustness of fraud detection models. Finally, the proposed approach provides a framework that can be replicated and tested in various financial markets, contributing to the development of more flexible and universally applicable fraud detection methodologies.

9. Conclusions

This research stemmed from the question of whether expert opinions can be a basis for a fraud detection tool that overcomes the challenges of existing methods. A well-grounded detection tool derived from expert evaluations is arguably less restrictive on the data requirement. This flexibility widens its applicability to various conditions, making it more convenient compared to rival methods. However, the feasibility of a proposed alternative is ultimately determined by its detection performance. Based on these, the objective of this study is to construct a financial fraud detection tool using an expert opinion-based approach and to gauge its comparative detection performance with its basic alternatives.

The research objective here requires comparable alternatives that only differ in their parameter determination approach. For this, we determined a fraud detection model as a basis and then produced three distinct fraud scores from three distinct parameter sets of this model. Each parameter set was obtained by adopting a different approach: the first set is directly taken by the original model; the second set is the coefficients of specifically run probit regression; and the third set is the relative importance of the model indicators obtained by seeking expert opinions by an efficient MCDM method, BWM. We labelled these scores as M-scores, Z-scores, and EXP-scores, respectively.

In the reproduction stage of the Z-score parameters, regression analysis output exhibited a similar construct to the original. A slight deviation in the coefficients can be interpreted as a variation in adapting the model to the research data. The congruity of these outputs also implies that the analysis data was appropriately gathered to test the fraud detection performance of an alternative technique. In the performance comparison stage, the ROC curve plot indicates that the expert opinion-based EXP score is a feasible alternative, offering better performance than the M-score and still requiring less effort than the Z-score. The statistical analysis on ROC-AUC confirms this inference.

The approach endorsed in this study has the significant advantage of being independent of verified past incidents as ground truth. A fraud detection tool that relies on historical data may become obsolete during periods of sudden changes in rules and practices (e.g., enactment of regulations such as the Sarbanes-Oxley Act, International Financial Reporting Standards, and unprecedented conditions such as the COVID-19 pandemic, inflation waves or

even financial crises). Expert opinions are fruitful in developing a forward-looking detection tool that considers the specific impacts of such sudden changes. However, this argument awaits to be provided by evidence from specifically designed analysis in further research.

In this study, we aimed to demonstrate the feasibility of an expert opinion-based approach for fraud detection in an emerging economy, Turkey. However, the research environment may pose a limitation on the generalizability of the results to other economic contexts. Nevertheless, we believe the potential of such an approach is broader for the most fortified financial systems with strong control mechanisms because of the adaptive nature of the fraudulent activity. Evolving policies force potential frauds to be ingenious and inventive. At this point, human evaluation is valuable to swiftly describe these before they are reflected in the financial data. The use of this approach is not limited to the method demonstrated in this study. Alternatively, expert opinions can be evaluated as complementary to the inputs of more complex machine learning techniques. To this end, the proposed approach contributes to the integration of human wisdom into historical data-driven classifiers and encourages broader applications of varying financial circumstances. We invite researchers to contribute to this strand of research by testing the generalization of such an approach for various conditions and integrating it into their classifiers.

Funding

This work was supported by Eskisehir Technical University Scientific Research Projects Agency (No. 24ADP148).

Author contributions

Serdar Benligiray designed the research model, contributed to the literature review and data gathering, conducted the analyses and made the revisions. Ahmet Onay reviewed the literature, gathered the data, designed the presentation of the research model, and contributed to the revisions. Fatma Yaşlı-Şen conducted the MCDM method.

References

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). MetaFraud: A metalearning framework for detecting financial fraud. *MIS Quarterly*, *36*(4), 1239–1327. https://doi.org/10.2307/41703508

Achakzai, M. A. K., & Peng, J. (2023). Detecting financial statement fraud using dynamic ensemble machine learning. *International Review of Financial Analysis*, *89*, Article 102827. https://doi.org/10.1016/j.irfa.2023.102827

Albrecht, W. S., Albrecht, C. C., Albrecht C. O., & Zimbelman, M. F. (2018). *Fraud examination*. Cengage Learning.

Ashtiani, M. N., & Raahemi, B. (2022). Intelligent fraud detection in financial statements using machine learning and data mining: A systematic literature review. *IEEE Access*, *10*, 72504–72525. https://doi.org/10.1109/ACCESS.2021.3096799

Association of Certified Fraud Examiners. (2024). *Occupational Fraud 2024: A Report to the Nations*. https://www.acfe.com/rttn

Bačo, T., Baumöhl, E., Horváth, M., & Výrost, T. (2023). Beneish Model for the detection of tax manipulation: Evidence from Slovakia. *Ekonomicky Casopis/Journal of Economics*, *71*(3), 185–201. https://doi.org/10.31577/ekoncas.2023.03.01

Baker, S. G., & Kramer, B. S. (2007). Peirce, Youden, and receiver operating characteristic curves. *The American Statistician*, *61*(4), 343–346. https://doi.org/10.1198/000313007X247643

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded US firms using a machine learning approach. *Journal of Accounting Research*, *58*(1), 199–235. https://doi.org/10.1111/1475-679X.12292

Beneish, M. D. (1999). The detection of earnings manipulation. *Financial Analysts Journal*, *55*(5), 24–36. https://doi.org/10.2469/faj.v55.n5.2296

Beneish, M. D., & Vorst, P. (2022). The cost of fraud prediction errors. *Accounting Review*, *97*(6), 91–121. https://doi.org/10.2308/TAR-2020-0068

Beneish, M. D., Lee, C. M. & Nichols, D. C. (2013). Earnings manipulation and expected returns. *Financial Analysts Journal*, *69*(2), 57–82. https://doi.org/10.2469/faj.v69.n2.1

Bertomeu, J., Edwige C., Eric F., & Wenqiang P. (2021). Using machine learning to detect misstatements. *Review of Accounting Studies*, *26*, 468–519. https://doi.org/10.1007/s11142-020-09563-8

Brown, N. C., Crowley, R. M., & Elliott, W. B. (2020). What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research*, *58*, 237–291. https://doi.org/10.1111/1475-679X.12294

Carcillo, F., Le Borgne, Y. A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, *557*, 317–331. https://doi.org/10.1016/j.ins.2019.05.042

Černevičienė, J., & Kabašinskas, A. (2022). Review of multi-criteria decision-making methods in finance using explainable artificial intelligence. *Frontiers in Artificial Intelligence*, *5*, Article 827584. https://doi.org/10.3389/frai.2022.827584

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Detecting management fraud in public companies. *Management Science*, *56*(7), 1146–1160. https://doi.org/10.1287/mnsc.1100.1174

Cleves, M. A. (2002). From the help desk: Comparing areas under receiver operating characteristic curves from two or more probit or logit models. *The Stata Journal*, *2*(3), 301–313. https://doi.org/10.1177/1536867X0200200307

Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. *Decision Support Systems*, *139*, Article 113421. https://doi.org/10.1016/j.dss.2020.113421

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary Accounting Research*, *28*, 17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845. https://doi.org/10.2307/2531595

Dikmen, B., & Küçükkocaoğlu, G. (2010). The detection of earnings manipulation: the three-phase cutting plane algorithm using mathematical programming. *Journal of Forecasting*, *29*(5), 442–466. https://doi.org/10.1002/for.1138

Duan, W., Hu, N., & Xue, F. (2024). The information content of financial statement fraud risk: An ensemble learning approach. *Decision Support Systems*, *182*, Article 114231. https://doi.org/10.1016/j.dss.2024.114231

Durana, P., Blazek, R., Machova, V., & Krasnan, M. (2022). The use of Beneish M-scores to reveal creative accounting: Evidence from Slovakia. Equilibrium. *Quarterly Journal of Economics and Economic Policy*, *17*(2), 481–510. https://doi.org/10.24136/eq.2022.017

Eining, M. M., Jones, D. R., & Loebbecke, J. K. (1997). Reliance on decision aids: An examination of auditors' assessment of management fraud. *Auditing: A Journal of Practice & Theory*, *16*(2), 1–19.

Fawcett, T. (2006). Introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Gepp, A., Kumar, K., & Bhattacharya, S. (2021). Lifting the numbers game: Identifying key input variables and a best-performing model to detect financial statement fraud. *Accounting & Finance*, *61*(3), 4601–4638. https://doi.org/10.1111/acfi.12742

Halilbegovic, S., Celebic, N., Cero, E., Buljubasic, E., & Mekic, A. (2020). Application of Beneish M-score model on small and medium enterprises in Federation of Bosnia and Herzegovina. *Eastern Journal of European Studies*, *11*(1), 146–163.

Hamal, S., & Senvar, O. (2022). A novel integrated AHP and MULTIMOORA method with interval-valued spherical fuzzy sets and single-valued spherical fuzzy sets to prioritize financial ratios for financial accounting fraud detection. *Journal of Intelligent & Fuzzy Systems*, *42*(1), 337–364. https://doi.org/10.3233/JIFS-219195

Hooda, N., Bawa, S., & Rana, P. S. (2018). Fraudulent firm classification: A case study of an external audit. *Applied Artificial Intelligence*, *32*(1), 48–64. https://doi.org/10.1080/08839514.2018.1451032

Huang, S. Y., Lin, C. C., Chiu, A. A., & Yen, D. C. (2017). Fraud detection using fraud triangle risk factors. *Information Systems Frontiers*, *19*, 1343–1356. https://doi.org/10.1007/s10796-016-9647-9

International Accounting Standards Board. (2018). *The conceptual framework for financial reporting*. London.

Kassem, R., & Higson, A. (2012). The new fraud triangle model. *Journal of Emerging Trends in Economics and Management Sciences*, *3*(3), 191–195.

Khan, A. T., Cao, X., Li, S., Katsikis, V. N., Brajevic, I., & Stanimirovic, P. S. (2022). Fraud detection in publicly traded US firms using Beetle Antennae Search: A machine learning approach. *Expert Systems with Applications*, *191*, Article 116148. https://doi.org/10.1016/j.eswa.2021.116148

Khatun, A., Ghosh, R., & Kabir, S. (2022). Earnings manipulation behavior in the banking industry of Bangladesh: The strategical implication of Beneish M-score model. *Arab Gulf Journal of Scientific Research*, *40*(3), 302–328. https://doi.org/10.1108/AGJSR-03-2022-0001

Lappas, P. Z., & Yannacopoulos, A. N. (2021). A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. *Applied Soft Computing*, *107*, Article 107391. https://doi.org/10.1016/j.asoc.2021.107391

Lenard, M. J., & Alam, P. (2008). The use of fuzzy logic and expert reasoning for knowledge management and discovery of financial reporting fraud. In M. E. Jennex (Ed.), *Knowledge management: Concepts, methodologies, tools, and applications* (pp. 1013–1028). IGI Global. https://doi.org/10.4018/978-1-59904-933-5

Lin, C. C., Chiu, A. A., Huang, S. Y., & Yen, D. C. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems*, *89*, 459–470. https://doi.org/10.1016/j.knosys.2015.08.011

Lokanan, M. (2017). Theorizing financial crimes as moral actions. *European Accounting Review*, *27*(5), 901–938. https://doi.org/10.1080/09638180.2017.1417144

Maniatis, A. (2022). Detecting the probability of financial fraud due to earnings manipulation in companies listed in Athens Stock Exchange Market. *Journal of Financial Crime*, *29*(2), 603–619. https://doi.org/10.1108/JFC-04-2021-0083

Martínez-Camblor, P., & Pardo-Fernández, J. C. (2019). The Youden Index in the generalized receiver operating characteristic curve context. *The International Journal of Biostatistics*, *15*(1), 1–20. https://doi.org/10.1515/ijb-2018-0060

Martinkutė-Kaulienė, R., Skobaitė, R., Stasytytė, V., & Maknickienė, N. (2021). Comparison of multicriteria decision-making methods in portfolio formation. *Finance, Markets and Valuation*, *7*(2), 60–72. https://doi.org/10.46503/qwpu4486

Marqués, A. I., García, V., & Sánchez, J. S. (2020). Ranking-based MCDM models in financial management applications: Analysis and emerging challenges. *Progress in Artificial Intelligence*, *9*(3), 171–193. https://doi.org/10.1007/s13748-020-00207

Mokrišová, M., & Horváthová, J. (2023). Domain knowledge features versus LASSO features in predicting risk of corporate bankruptcy – DEA approach. *Risks*, *11*(11), Article 199. https://doi.org/10.3390/risks11110199

Morris, G. D. L. (2009). How a group of business students sold Enron a year before the collapse. *Financial History*, (Spring/Summer), 12–15.

Papík, M., & Papíková, L. (2020). Detection models for unintentional financial restatements. *Journal of Business Economics and Management*, *21*(1), 64–86. https://doi.org/10.3846/jbem.2019.10179

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory*, *30*(2), 19–50. https://doi.org/10.2308/ajpt-50009

Perols, J. L., & Lougee, B. A. (2011). The relation between earnings management and financial statement fraud. *Advances in Accounting*, *27*(1), 39–53. https://doi.org/10.1016/j.adiac.2010.10.004

Panigrahi, P. K. (2011, June). A framework for discovering internal financial fraud using analytics. In *2011 International Conference on Communication Systems and Network Technologies* (pp. 323–327). IEEE. https://doi.org/10.1109/CSNT.2011.74

Qiu, S., He, H.-Q., & Luo, Y.-s. (2019). The value of restatement to fraud prediction. *Journal of Business Economics and Management*, *20*(6), 1210–1237. https://doi.org/10.3846/jbem.2019.10489

Rahman, M. J., & Zhu, H. (2023). Predicting accounting fraud using imbalanced ensemble learning classifiers – evidence from China. *Accounting & Finance*, *63*(3), 3455–3486. https://doi.org/10.1111/acfi.13044

Rahman, M. J., & Zhu, H. (2024). Detecting accounting fraud in family firms: Evidence from machine learning approaches. *Advances in Accounting*, *64*, Article 100722. https://doi.org/10.1016/j.adiac.2023.100722

Ramzan, S., & Lokanan, M. (2024). The application of machine learning to study fraud in the accounting literature. *Journal of Accounting Literature*. https://doi.org/10.1108/JAL-11-2022-0112

Repousis, S. (2016). Using Beneish model to detect corporate financial statement fraud in Greece. *Journal of Financial Crime*, *23*(4), 1063–1073. https://doi.org/10.1108/JFC-11-2014-0055

Rezaee, Z. (2005). Causes, consequences, and deterrence of financial statement fraud. *Critical Perspectives on Accounting*, *16*(3), 277–298. https://doi.org/10.1016/S1045-2354(03)00072-8

Rezaei, J. (2015). Best-worst multi-criteria decision-making method. *Omega*, *53*, 49–57. https://doi.org/10.1016/j.omega.2014.11.009

Rezaei, J. (2016). Best-worst multi-criteria decision-making method: Some properties and a linear model. *Omega*, *64*, 126–130. https://doi.org/10.1016/j.omega.2015.12.001

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall plot is more informative than the ROC Plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, *10*(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432

Shahana, T., Vilvanathan, L., & Bhat, A.R. (2023). State of the art in financial statement fraud detection: A systematic review. *Technological Forecasting and Social Change*, *192*, Article 122527. https://doi.org/10.1016/j.techfore.2023.122527

Spathis, C. T. (2002). Detecting false financial statements using published data: Some evidence from Greece. *Managerial Auditing Journal*, *17*(4), 179–191. https://doi.org/10.1108/02686900210424321

Tahmina, A., & Naima, J. (2016). Detection and analysis of probable earnings manipulation by firms in a developing country. *Asian Journal of Business and Accounting*, *9*(1), 59–82.

Vincent, M. (2012, June 15). M-Score' flags dubious company earnings. *Financial Times* [Electronic Version]. Retrieved January 14, 2024, from http://www.ft.com

Zainudin, E. F., & Hashim, H. A. (2016). Detecting fraudulent financial reporting using financial ratio. *Journal of Financial Reporting and Accounting*, *14*(2), 266–278. https://doi.org/10.1108/JFRA-05-2015-0053

Zhao, K., Dai, Y., Ji, Y., & Jia, Z. (2021). Decision-Making model to portfolio selection using Analytic Hierarchy Process (AHP) with expert knowledge. *IEEE Access*, *9*, 76875–76893. https://doi.org/10.1109/ACCESS.2021.3082529

Zhao, Z., & Bai, T. (2022). Financial fraud detection and prediction in listed companies using SMOTE and machine learning algorithms. *Entropy*, *24*(8), Article 1157. https://doi.org/10.3390/e24081157

Zhou, Y., Xiao, Z., Gao, R., & Wang, C. (2024). Using data-driven methods to detect financial statement fraud in the real scenario. *International Journal of Accounting Information Systems*, *54*, Article 100693. https://doi.org/10.1016/j.accinf.2024.100693

Zumbrun, J. (2023, March 24). Accounting-Fraud indicator signals coming economic trouble. *Wall Street Journal* [Electronic Version]. Retrieved February 4, 2024, from http://www.wsj.com