Schneider, Moritz; Brühl, Rolf

Disentangling the black box around CEO and financial information-based accounting fraud detection: machine learning-based evidence from publicly listed U.S. firms

Journal of Business Economics

Provided in Cooperation with:
Springer Nature

Suggested Citation: Schneider, Moritz; Brühl, Rolf (2023) : Disentangling the black box around CEO and financial information-based accounting fraud detection: machine learning-based evidence from publicly listed U.S. firms, Journal of Business Economics, ISSN 1861-8928, Springer, Berlin, Heidelberg, Vol. 93, Iss. 9, pp. 1591-1628, https://doi.org/10.1007/s11573-023-01136-w

This Version is available at:
https://hdl.handle.net/10419/311070

Journal of Business Economics (2023) 93:1591–1628
https://doi.org/10.1007/s11573-023-01136-w

ORIGINAL PAPER

Disentangling the black box around CEO and financial information-based accounting fraud detection: machine learning-based evidence from publicly listed U.S. firms

Moritz Schneider[1] · **Rolf Brühl**[1]

Accepted: 28 December 2022 / Published online: 4 February 2023

Abstract

This study investigates the predictive power of CEO characteristics on accounting fraud utilizing a machine learning approach. Grounded in upper echelons theory, we show the predictive value of widely neglected CEO characteristics for machine learning-based accounting fraud detection in isolation and as part of a novel combination with raw financial data items. We employ five machine learning models well-established in the accounting fraud literature. Diverging from prior studies, we introduce novel model-agnostic techniques to the accounting fraud literature, opening further the black box around the predictive power of individual accounting fraud predictors. Specifically, we assess CEO predictors concerning their feature importance, functional association, marginal predictive power, and feature interactions. We find the isolated CEO and combined CEO and financial data models to outperform a no-skill benchmark and isolated approaches by large margins. Nonlinear models such as Random Forest and Extreme Gradient Boosting predominantly outperform linear ones, suggesting a more complex relationship between CEO characteristics, financial data, and accounting fraud. Further, we find CEO Network Size and CEO Age to contribute second and third strongest towards the best model's predictive power, closely followed by CEO Duality. Our results indicate U-shaped, L-shaped, and weak L-shaped associations for CEO Age, CEO Network Size, CEO Tenure, and accounting fraud, consistent with our superior nonlinear models. Lastly, our empirical evidence suggests that older CEOs who are not simultaneously serving as chairman and CEOs with an extensive network and high inventory are more likely to be associated with accounting fraud.

Keywords Accounting fraud · CEO characteristics · Corporate governance · Machine learning

JEL Classification C63 · K22 · K42 · M41 · M48

✉ Moritz Schneider
mschneider@escp.eu

Extended author information available on the last page of the article

1 Introduction

Accounting fraud cases are frequently and globally occurring events causing extensive financial and non-financial damage to employees, businesses, investors, and society.[1] The Association of Certified Fraud Examiners (2020) estimates that fraud damages for organizations amount to 5% of total revenue or $4.5 trillion yearly, with financial statement fraud being the rarest but most costly fraud.[2] Timely accounting fraud detection to mitigate the associated costs would be valuable to investors, regulators, and auditors (Bao et al. 2020).[3] However, detecting accounting fraud is challenging (Bao et al. 2020). First, due to the severe class imbalance between detected fraud and non-fraud cases (Beneish 1999) and *"partial observability of fraud"* (Wang et al. 2010, p. 2256), there is a need for substantial sample sizes. Second, aggregated financial information does not fully reflect information asymmetry related to organizational behavior (Campbell and Shang 2022).

While the vast majority of accounting fraud detection studies have mainly considered predictors based on financial information, scant empirical studies address non-financial information. Although the associations between non-financial characteristics and financial misconduct have been theoretically and empirically established (e.g., Beasley 1996; Dechow et al. 1996; Johnson et al. 2009; Troy et al. 2011; Wahid 2019; Zahra et al. 2005), only a few studies have considered such characteristics for machine learning-based accounting fraud detection (Bertomeu et al. 2021; Fanning and Cogger 1998; Kim et al. 2016; Wang et al. 2018, 2020). However, none of them primarily focus on CEO characteristics, whereas we aim at improving predictions by studying the joint predictive power of these predictors and raw financial data.[4]

We specifically investigate CEO-related data for three reasons. First, we look for alternative predictors, as aggregated financial information can be altered to disguise underlying firm-related manipulations (Lewis 2013). Second, while the CEO and CFO represent top managers likely to be associated with a firm's accounting outcomes (e.g., Gupta et al. 2020; Troy et al. 2011), the literature typically considers the CEO as the most powerful character of a firm. Moreover, studies suggest that CEO power influences the interaction with CFOs, as they find a CEO to hold power over the CFO's accounting behavior (e.g., Feng et al. 2011; Friedman 2014) and that pressure constitutes a primary cause for earnings management by CFOs (Dichev et al. 2013). Lastly, various CEO characteristics have been established concerning outcomes related to accounting fraud. Because upper echelons theory suggests that the top management team's characteristics have predictive power on organizational

[1] In this study, we use the terms "accounting fraud " and "misstatement" synonymously, similar to previous literature (e.g., Bao et al. 2020).

[2] This loss has been obtained by projecting an estimate of a 5% loss in revenue against 2019's gross world product (GWP) of $90.52 trillion (Association of Certified Fraud Examiners 2020).

[3] We follow prior literature (e.g., Bao et al. 2020) and use the terms "prediction" and "detection" interchangeably throughout our work.

[4] We follow Bao et al.'s (2020) terminology and refer to "raw financial data items" as financial information that can be directly obtained from the financial statements rather than computed as human expert-identified financial ratios.

decisions (Hambrick and Mason 1984; Hambrick 2007), we consider CEO-related information. The latter is theoretically motivated and empirically associated with financial misconduct by previous literature (e.g., Ali and Zhang 2015; Bhandari et al. 2018; Dechow et al. 1996; Ho et al. 2015; Huang et al. 2012; Troy et al. 2011; Zahra et al. 2005).

Moreover, prior machine learning-based detection studies focused on prediction improvements but lacked interpretability. While prediction and explanation are two distinct research goals (Shmueli 2010), understanding the drivers behind a model's predictive power is paramount for financial applications (Sigrist and Hirnschall 2019). Consequently, we introduce state-of-the-art model-agnostic techniques to the accounting fraud detection literature, opening the black box surrounding model prediction. In particular, we rely on the permutation-based feature importance (Breiman 2001) and SHapley Additive exPlanation (SHAP) dependence plots (Lundberg and Lee 2017).

We use five established machine learning algorithms to assess and disentangle the CEO characteristics' out-of-sample predictive power for accounting fraud in isolation and a novel combination with raw financial data items. We expect that the inclusion of literature and theory-derived CEO characteristics captures and adds additional non-financial, and latent firm-related insights, resulting in models that outperform a no-skill benchmark and solely data-based financial models, respectively. Guided by Schnatterly et al. (2018) and the well-established fraud triangle framework (Cressey 1950), we derive CEO characteristics that reflect pressure, opportunity, and rationalization to commit fraud. Our reasoning is similar to recent literature that constitutes the additional value of complementing accounting with other firm-related information for predicting firm outcomes (e.g., Bertomeu et al. 2021; Cheynel and Levine 2020). Further, we specifically investigate the contribution of the variables toward the prediction, their marginal effects, functional form, and interactions to understand the drivers behind the predictive performance.

Following prior literature (e.g., Bao et al. 2020; Beasley 1996; Brown et al. 2020; Cecchini et al. 2010; Dechow et al. 2011; Perols et al. 2017; Purda and Skillicorn 2015), we use material accounting misstatements published in the SEC's Accounting and Auditing Enforcement Releases (AAERs) and provided by Dechow et al. (2011) as binary fraud measure. While established accounting fraud detection literature uses human expert-identified financial ratios to predict accounting fraud (e.g., Beneish 1999; Cecchini et al. 2010; Dechow et al. 2011), we instead incorporate their underlying raw financial data items directly. Thus, we follow recent empirical evidence of the predictive superiority of raw financial data items over financial ratios (Bao et al. 2020). Our sample covers publicly listed U.S. firms for 2000–2018 and contains matched financial and CEO data of 30,178 firm-years, including 198 fraudulent firm-years.

Consistent with our expectations, we find empirical evidence that suggests a robust predictive performance of machine learning models based on CEO characteristics over a no-skill benchmark for accounting fraud detection. Further, we show the additional predictive value of CEO characteristics combined with raw financial predictors compared to isolated models across all classifiers. Interestingly, nonlinear models such as random forest (RF) and extreme gradient boosting (XGB)

predominantly outperform their linear counterparts, suggesting more complex associations between CEO and financial data and accounting fraud. With *CEO Network Size*, *CEO Age*, and *CEO Duality*, we find half of the considered CEO characteristics included in the top 10 essential features. Consistent with our superior nonlinear models, our results indicate a U-shaped association between *CEO Age*, an L-shaped association for *CEO Network Size,* and a weak L-shaped relationship between *CEO Tenure* and accounting fraud. Lastly, we extend the literature's knowledge by visualizing interactions between essential features. The results suggest that older CEOs not simultaneously serving as chairman and CEOs with a network of up to 2500 connections and high inventory are more likely to be associated with accounting fraud.

This study complements previous accounting fraud detection literature in various aspects. First, to our knowledge, we are the first to focus on both CEO characteristics and assess the joint predictive power of CEO characteristics combined with raw financial data items in machine learning models for out-of-sample accounting fraud detection. Second, we address the shortage of empirical research investigating nonlinear relationships between corporate governance and financial misconduct (Velte 2021). Third, we follow the call by Doornenbal et al. (2021) that invoked research to incorporate more (interpretable) machine learning techniques to uncover currently hidden and more complex associations and allow for future theory advancements. Our results suggest future research's potential to address more complex relationships between *CEO Age*, *CEO Network Size*, *CEO Tenure,* and accounting fraud.

2 Theoretical background and research questions development

Our study draws on two literature streams. Firstly, our research relates to the substream of accounting fraud literature that develops out-of-sample accounting fraud detection models. Secondly, we draw on literature about fraud antecedents on the individual level of corporate governance—the CEO (Velte 2021). The first literature stream can be separated into explanatory and predictive approaches (Shmueli 2010). While many studies focused on investigating causal relationships between financial and non-financial firm-specific characteristics as antecedents of accounting fraud (e.g., Beasley 1996; Brazel et al. 2009; Dechow et al. 1996; Schrand and Zechman 2012), the rise of technological development initiated a transition towards more predictive approaches. Thus, having established associations of relevant antecedents for accounting fraud, many studies started to employ regression- or machine learning-based approaches to predict accounting fraud. Regression-based models appear to be the most frequently applied technique for accounting fraud detection within the accounting and information systems literature (Albizri et al. 2019). Among others, Beneish (1999) constructed a probit model to predict the likelihood of accounting fraud based on accounting variables. They showed that the model could identify about 50% of firm-years with manipulated earnings before public disclosure (Beneish 1999). Another well-known example is the study of a logit model developed by Dechow et al. (2011), which investigated financial information of about 16,000 firm-years including 2190 AAERs, and predicted misstating firms with an overall accuracy of 63.7% and a recall of 68.6%. This study is still considered a

competitive model and is used as a benchmark for more recent algorithms (Bao et al. 2020).

While many studies have included regression-based models for accounting fraud detection (e.g., Bao et al. 2020; Bertomeu et al. 2021; Craja et al. 2020; Larcker and Zakolyukina 2012), a shift toward more advanced machine learning approaches has occurred along with technical developments. Thus, the majority of studies investigated the predictive power of random forests (RF) (e.g., Bertomeu et al. 2021; Craja et al. 2020; Wang et al. 2020; Whiting et al. 2012), support vector machines (SVM) (e.g., Bertomeu et al. 2021; Cecchini et al. 2010; Craja et al. 2020; Perols et al. 2017; Purda and Skillicorn 2015), and neural networks (NN) (e.g., Craja et al. 2020; Fanning and Cogger 1998; Green and Choi 1997; Ravisankar et al. 2011). Recently, ensemble methods like RUSBoost (Bao et al. 2020), extreme gradient boosting (XGB) (Craja et al. 2020), or gradient boosted regression tree (Bertomeu et al. 2021) have been introduced.

Prior studies deployed financial ratios identified by experts and empirically shown to be associated with accounting fraud or material misstatements (e.g., Beneish 1997, 1999; Cecchini et al. 2010; Dechow et al. 2011). For instance, Beneish (1999, pp. 26–28) considered *"Days' sales in receivables index"*, *"Gross margin index"*, *"Asset quality index"*, *"Sales growth index"*, *"Depreciation index"*, *"Sales, general, and administrative expenses index"*, *"Leverage index"*, and *"Total accruals to total assets"*. Similar categorizations have been established by Dechow et al. (2011). They categorized financial characteristics into *"Accrual quality"*, *"Performance"*, *"Nonfinancial measures"*, *"Off-balance-sheet activities"*, and *"Market-related incentives"* (Dechow et al. 2011, pp. 34–41). However, Bao et al. (2020) recently deviated from including the predominantly used financial ratios and empirically showed the superior performance of the underlying raw financial data items. In particular, they derived a combined set of 28 underlying raw financials from well-established financial ratios utilized by Cecchini et al. (2010) and Dechow et al. (2011). Among others, the variables included stock information, such as *"Common Shares Outstanding"* or *"Price Close—Annual"*, balance-sheet information like *"Current Assets—Total"*, *"Account Payable—Trade"* or *"Cash and Short-Term Equivalents"*, as well as income information, such as *"Depreciation and Amortization"* or *"Net Income (Loss)"* (Bao et al. 2020, p. 229).[5] Robustness checks with different raw financials validated the predictive power of the identified 28 financials (Bao et al. 2020).

Only a few machine learning-based studies employed prediction models using non-financial predictors. An exception is the sub-stream of literature that investigates text-based predictors either in isolation or in combination with financials (e.g., Brown et al. 2020; Craja et al. 2020; Hobson et al. 2012; Larcker and Zakolyukina 2012; Purda and Skillicorn 2015). However, besides this literature stream, hardly any study either combines financial and non-financial predictors, such as board characteristics (Fanning and Cogger 1998; Wang et al. 2020), executive compensation (Kim et al. 2016), governance, audit or business data (Bertomeu et al. 2021) or builds their model solely on non-financials, such as board data (Wang et al. 2018).

[5] For a complete list of variable names see Table 7 in Bao et al. (2020).

Interestingly, several studies exist that investigate non-financial antecedents related to accounting fraud but are widely neglected by prior machine learning-based prediction approaches (e.g., Ali and Zhang 2015; Dechow et al. 1996; Huang et al. 2012; Troy et al. 2011; Schrand and Zechman 2012).

The second literature stream we draw on focuses on accounting fraud-related antecedents at the CEO level, which has been widely neglected within machine learning-based accounting fraud studies. Inspired by Schnatterly et al. (2018), we consider six CEO characteristics derived from prior literature and categorize them within the well-established fraud triangle framework (Cressey 1950; Dorminey et al. 2012; Trompeter et al. 2013), common in the audit literature (Dorminey et al. 2012): pressure (*CEO Tenure*, *CEO Network Size*), opportunity (*CEO Duality*), and rationalization (*CEO Age*, *CEO Gender*, *CEO MBA*).[6]

According to the fraud triangle, these three interacting antecedent groups precede fraud, where higher interaction and manifestation of these elements result in higher fraud risk (Dorminey et al. 2012).

Pressure (or Incentives). Pressure (or incentives) represents the perceived motivation that forces (incentivizes) the actor to behave fraudulently (Dorminey et al. 2012). CEOs' perceived pressure (or incentives) to misbehave can originate from various sources, such as career concerns. Thus, Ali and Zhang (2015) argue that CEO and firm performance are linked (Fama 1980) and that the CEO has an interest in optimizing her future career perspectives, such as *"compensation, reappointments or managerial autonomy"* (Ali and Zhang 2015, p. 61). Therefore, the external labor market needs to perceive the CEO as a well-performer (e.g., Fama 1980). However, especially at the beginning of a CEO's tenure at a firm, the market is likely uncertain about a CEO's ability (Gibbons and Murphy 1992) due to a lack of historical CEO performance information (Ali and Zhang 2015). As unfavorable market valuations of the CEO's ability could hamper her career perspectives, the CEO is incentivized to improve the market's perception of her abilities in the early years (Ali and Zhang 2015). Consistent with this argumentation, Ali and Zhang (2015) find earnings overstatements to be greater in CEOs' early years than in the later years of their tenure. However, the horizon problem[7] suggests that departing CEOs have incentives to behave opportunistically in their final years of tenure to boost their short-term compensation (Dechow and Sloan 1991). Ali and Zhang (2015) consistently find that earnings are overstated in the CEO's final years when controlling for early year overstatements.

Similarly, social psychology suggests that human behavior is driven by social norms and the expected judgment of one's behavior by others (Cialdini et al. 1991).

[6] We also considered established compensation-related variables which were suggested to provide incentives for fraudulent action (e.g., Johnson et al. 2009; Zhang et al. 2008), such as CEO ownership and CEO options to compensation (Koch-Bayram and Wernicke 2018). However, following Cecchini et al. (2010), we removed them from the final sample because of missing values > 25%.

[7] Following Smith and Watts (1982) we refer to the "horizon problem" as executives considering to leave the firm and focusing on short-term performance due to earnings-based compensation.

Corporate misconduct could result in reputational damages that mitigate future career prospects (Karpoff 2011). Thus, anticipated social capital losses may result in a disincentive to misbehave (Atanasov et al. 2012). Based on this reasoning, Bhandari et al. (2018) find the number of CEO connections negatively associated with earnings management and financial restatements.

Opportunity. Opportunity describes the perceived possibility of the actor committing fraud without fearing detection or punishment (Dorminey et al. 2012). CEO power is typically considered the main characteristic of fraud opportunity on the CEO level (Schnatterly et al. 2018), commonly measured by CEO duality (Velte 2021). Particularly, Jensen (1993) theorizes that if a CEO simultaneously serves as chairman of the board, the oversight of the management is reduced. Based on this argument, Dechow et al. (1996) find firms with CEO duality to show a higher likelihood of receiving SEC AAERs, consistent with the prior expectations.

Rationalization. Lastly, rationalization embodies the actor's integrity to internally justify the fraudulent act as morally reasonable (Dorminey et al. 2012). According to Schnatterly et al. (2018), individuals try to resolve a moral trade-off between their fraudulent behavior and societal ethics (e.g., Cressey 1950; Trompeter et al. 2013). Thus, a CEO's ethical socialization is likely important for rationalizing fraudulent behavior. CEO age, gender, and business education (MBA) may be associated with ethical processes and, in turn, fraudulent acts (Schnatterly et al. 2018).

First, Troy et al. (2011) argue that prior literature suggests negative associations between age and unethical (e.g., Hunt and Chonko 1984; Kelley et al. 1990) or risk-taking behavior (e.g., Brouthers et al. 2000; Hambrick and Mason 1984; Markóczy 1997). The authors further suggest that older individuals behave less risky, as they are more likely to adhere to organizational rules (Child 1974), are more morally developed (Kelley et al. 1990), construe a code of conduct more strictly (Serwinek 1992), and are less likely to succumb to external pressures (Daboub et al. 1995; Price and Norris 2009). Based on these findings, Troy et al. (2011) posit that older CEOs are less likely to rationalize the costs of the risk and the respective consequences of being discovered than their younger counterparts. Consistently, the literature finds CEO age negatively related to accounting fraud (Huang et al. 2012; Troy et al. 2011).

Second, different socializations between genders contribute towards diverging individual ethical values (e.g., Mason and Mudrack 1996; Weeks et al. 1999), which can result in more ethical decision-making at work (Dawson 1995). In particular, female practitioners are more ethically sensitive and risk-averse than their male counterparts (e.g., Weeks et al. 1999). Following these studies, Ho et al. (2015) find female CEOs positively associated with accounting conservatism. Building on this, Schnatterly et al. (2018) propose an association between CEO gender and financial misconduct.

Lastly, Troy et al. (2011) draw on prior literature that indicates positive associations between education and moral development (Freeman and Gilbert 1988; Rest and Thoma 1985), information processing (Wiersema and Bantel 1992), and better decision-making (e.g., Fiske and Taylor 1991). The authors follow the

"preponderance of literature" (Troy et al. 2011, p. 265) and Barker and Mueller (2002), who suggest that business education establishes analytical skills that avoid negative business consequences. Based on this argumentation, Troy et al. (2011) argue that the findings on education levels can be extended to business education, which establishes fundamental knowledge of accounting and the potential negative consequences of misbehavior. This is, in turn, expected to result in CEOs being less likely to rationalize accounting fraud (Troy et al. 2011). Consistent with this reasoning, Troy et al. (2011) find significant negative associations between firms with CEOs holding business degrees and accounting fraud.

Grounded in upper echelons theory (Hambrick and Mason 1984) and the proposed individual predictive value of the CEO characteristics on accounting fraud, we argue that the combination of CEO characteristics could reveal a large share of a CEO's underlying values and cognitive processes linked to accounting fraud behavior. Therefore, we explore if machine learning models based on these characteristics perform better in separating fraudulent from non-fraudulent firm-years than a random guessing threshold. Hence, the first research question can be stated as follows:

RQ1a Do machine learning models for accounting fraud detection based on CEO characteristics achieve a predictive performance superior to random guessing?

Bao et al. (2020) suggest that although most accounting fraud detection literature relies on human expert-identified financial ratios, raw financial data items capture a superior predictive value compared to financial ratios. However, following the argumentation of Campbell and Shang (2022), aggregated financial information only partially reflects the company's inner workings, including behavioral patterns and values, which are likely associated with misconduct. Similarly, Bertomeu et al. (2021) propose that accounting variables add predictive value to their accounting fraud detection model, primarily through their complementary effect with other information sources. Following this line of reasoning, we fit in the scant but growing accounting fraud prediction literature, which suggests that complementing accounting information with non-financial data (e.g., business, text, corporate governance data) leads to superior predictive models (e.g., Bertomeu et al. 2021; Craja et al. 2020; Wang et al. 2020).[8] We argue to incorporate additional latent firm-specific information associated with accounting fraud and neglected by financial predictors. This information increase likely results in higher predictive performance for accounting fraud detection, which is essential to detect and reduce costs on time. Following this reasoning, in combination with raw financial data items, we posit that CEO characteristics deliver superior predictive performance than financial and CEO models in isolation. Hence, the second research question can be stated as follows[9]:

[8] Many studies investigate non-financial information and their associations with accounting fraud (e.g., Dechow et al. 1996). While we follow this stream and believe the disclosed information to capture valuable latent information, we acknowledge its potentially decreased reliability due to a lack of external auditing of reports that include this information. We thank an anonymous reviewer for raising this point.

[9] See Online Appendix A for a visual presentation of the main research framework, including RQ1a and b.

RQ1b Do machine learning models for accounting fraud detection based on a combination of raw financial and CEO information (CEO + FIN) outperform isolated approaches (CEO, FIN)?

Predictive performance and causal inferences are two distinct objectives within machine learning (Shmueli 2010). While the vast body of the accounting fraud detection literature focuses on incorporating machine learning models to improve predictive performance, hardly any study also sheds light on variable-specific insights (Bao et al. 2020; Bertomeu et al. 2021; Wang et al. 2020). This, however, is of particular interest to researchers and practitioners alike, as understanding the drivers for predictions is considered paramount for financial applications (Sigrist and Hirnschall 2019). Thus, we follow these studies and a recent call for interpretable machine learning techniques by Doornenbal et al. (2021) and investigate the model's feature importance.

However, we deviate from prior studies that performed impurity-based feature importance (e.g., Bao et al. 2020) and rely on permutation-based feature importance developed by Breiman (2001). This practice delivers more robust results (Strobl et al. 2007) and was recently proposed by Doornenbal et al. (2021). Following the previously established argumentation of the importance of CEO characteristics for the predictive power of accounting fraud detection, we investigate the feature importance of our best-performing model, with a particular focus on the newly incorporated CEO characteristics. Thus, the third research question can be framed as follows:

RQ2a How influential are individual CEO characteristics for accounting fraud detection within the best CEO + FIN model?

Furthermore, our study aims to open the black box of accounting fraud detection beyond feature importance, providing insights into functional forms, directions, and main effects of CEO-related predictors. We introduce the novel SHAP dependence plot by Lundberg and Lee (2017) to the accounting fraud literature to disentangle these associations. We are interested in how CEO characteristics contribute to the model predictions and whether the associations comply with previous literature or diverge in structural complexity. Thus, we state the following research question:

RQ2b How do the individual CEO characteristics contribute toward accounting fraud detection within the best CEO + FIN model?

Interaction effects complement our analysis to disentangle feature importance beyond main effects. Consequently, the interdependencies of CEO characteristics and financial information are particularly interesting for machine learning-based accounting fraud detection. As before, we rely on SHAP dependence plots and visualize the interaction between two of the most crucial CEO characteristics and raw financials. This extends current knowledge towards the associations around CEO and financial information within accounting fraud detection models. The related research question can be formulated as follows:

RQ2c How do the essential CEO characteristics interact with each other and raw financials within the best CEO + FIN model?

3 Modeling approach

3.1 Algorithms

This study uses five prediction models. Besides the traditional logistic regression (LR), more advanced machine learning techniques, including SVM, RF, XGB, and NNs, are employed. LR is often considered a benchmark model for more advanced machine learning approaches (e.g., Bao et al. 2020; Bertomeu et al. 2021; Craja et al. 2020). Besides its application for predictions, the LR allows for inferences of the partial effects of X on Y. More advanced machine learning models often face limitations in interpretation (Zhao and Hastie 2021). Therefore, interpretability is commonly understood as an advantage of LR.

The SVM developed by Cortes and Vapnik (1995) is a more advanced prediction model. The underlying idea of SVM is rooted in a maximum margin hyperplane that perfectly separates training data into two classes by constructing a hyperplane within a p-dimensional feature space (James et al. 2021). The algorithm chooses the function coefficients to maximize the margin to the closest observed training data (James et al. 2021). A support vector classifier allows some observations to be incorrectly linearly classified, regulated by the C's tuning parameter (James et al. 2021).[10] The function varies in linear, polynomial, or radial form. A support vector classifier with a nonlinear kernel is known as SVM (James et al. 2021).

RF, introduced by Breiman (2001), is a popular ensemble learning technique that delivers a competitive predictive performance (Hastie et al. 2009). They address decreased generalizability caused by overfitting, which is the main challenge of a predictive model (Shmueli 2010), by various randomization elements. Contrasting the previously described models, ensemble learning models represent a more complex machine learning technique that combines various base estimators. Specifically, the RF incorporates multiple classification and regression trees (CART) (Breiman et al. 2017). For classification problems, base learner classification trees typically conduct each split to minimize the impurity of the resulting nodes, as measured by the Gini index (Breiman et al. 2017).[11] To compute the RF's predictions, it trains

[10] While $C=0$, the training data is perfectly classified following the maximum margin hyperplane and leading to narrow margins, $C>0$ allows for C misclassifications and, hence, is characterized by large margins (James et al. 2021).

[11] While other splitting criteria exist for classification trees, such as the entropy, we focus on the Gini index, as this approach has been set as default for random forests within the scikit-learn package in Python (Pedregosa et al. 2011) and has been implemented within our study. For a description of RF's default parameters see, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

these multiple decision trees individually and averages over each tree's results (Hastie et al. 2009).

Moreover, the RF adds randomization by training the base models on bootstrapped training data samples (James et al. 2021). The RF incorporates its multiple decision trees with additional random feature selection of a subset m of features p for splitting the nodes as base models (Breiman 2001).[12] RFs are relatively robust against outliers and noise while being faster than similar algorithms (Breiman 2001). Another advantage is partial explainability which allows for variable importance estimations based on the algorithm's inherent splitting procedure (Breiman 2001).

Another tree-based algorithm that has been considered in a wide range of prediction tasks is the recently developed XGB algorithm by Chen and Guestrin (2016). Boosting algorithms, including XGB, are among the best-performing learning methods (Hastie et al. 2009). Like RF, they draw on an ensemble of multiple decision trees (James et al. 2021). However, boosting algorithms do not produce bootstrapped samples but sequentially add newly grown small decision trees to the currently fitted model to update the model's residuals and slowly progress toward a more accurate prediction (James et al. 2021). Thus, each new tree heavily depends on the previous ones (James et al. 2021). The algorithm draws on Friedman's (2001) established gradient boosting algorithm and extends it in scalability (Chen and Guestrin 2016). Thus, the algorithm is computationally more efficient than Friedman's (2001) original gradient boosting algorithm (Climent et al. 2019). For boosting algorithms, the number of trees, also called iterations, the learning rate, and the decision trees' complexity are typically considered for model tuning (James et al. 2021).

Lastly, NNs represent another class of established algorithms for accounting fraud detection. NNs are rooted in the seminal work about perceptrons by Rosenblatt (1958), who developed a probabilistic model of information processing within the brain. The basic idea of NNs can be explained with a single-layer neural network that takes p inputs as vector $X = \{X_1, X_2, \ldots, X_p\}$, called the input layer (James et al. 2021). A hidden layer follows the input layer. It consists of K hidden nodes that each receive all inputs from the input layer, sum the weighted inputs, and add a bias term (James et al. 2021). Next, it transforms these linear functions into probability values between 0 and 1 using nonlinear activation functions $g(z)$ (James et al. 2021). Three main function types exist: the sigmoid function, the hyperbolic tangent, and the rectified linear unit (ReLu). Afterward, activations A_k are passed on to the output layer (James et al. 2021). In a single-layer NN, the function's parameters and weights are typically estimated by minimizing the log loss for a qualitative response or the squared-error loss for a quantitative response (James et al. 2021). More recent NNs typically extend this structure by consisting of multiple hidden layers through which

[12] Typically, $m = \sqrt{p}$ to decorrelate the trees, with smaller m in situations with correlated features (James et al. 2021).

a chain of transformations is performed, similar to the outlined approach (James et al. 2021).[13]

3.2 Evaluation metrics

Following the vast majority of related research (e.g., Bao et al. 2020; Bertomeu et al. 2021; Cecchini et al. 2010; Craja et al. 2020; Larcker and Zakolyukina 2012; Perols et al. 2017), we assess the models' predictive out-of-sample performances mainly using the AUC. Larcker and Zakolyukina (2012) highlight that alternative, cutoff-dependent measures rely on determining a cutoff value to classify probabilities into binary classes. When chosen ambiguously, this could result in the misclassification of observations (Larcker and Zakolyukina 2012), a challenge within accounting fraud due to the uncertainty about actual misclassification costs (Bao et al. 2020). Additionally, these measures are sensitive to class distributions (Larcker and Zakolyukina 2012). Relying on the AUC mitigates such limitations while simultaneously establishing some comparability with recent literature.[14] The AUC numerically describes the integrated area under a curve depicting the relationship between the true positive rate (TPR) (y-axis) and the false positive rate (FPR) (x-axis) in a two-dimensional feature space ranging from 0 to 1 on each axis (Fawcett 2006). Following Bradley (1997, p. 1146), the AUC can be computed using the trapezoidal integration as follows: $\sum_i \{\left(1-\beta_i * \Delta\alpha\right) + \frac{1}{2}[\Delta(1-\beta) * \Delta\alpha]\}$, with $\alpha = \text{P(FP)} = \text{FPR}$ and $1-\beta = \text{P(TP)} = \text{Sensitivity} = \text{TPR}$.[15] The benchmark AUC for random guesses is 0.5, where any model exceeding this threshold outperforms random models, and a perfect model yields an AUC of 1 (Fawcett 2006).

However, we follow prior studies and complement this metric to provide additional insights into the models' overall predictive performance. We consider sensitivity, specificity, and accuracy as additional metrics (e.g., Bao et al. 2020; Cecchini et al. 2010; Craja et al. 2020). Sensitivity measures the correctly classified minority firm-years (here fraud; TP) of all investigated firm-years (TP + FP) as $\frac{\text{TP}}{\text{TP+FP}}$. Specificity, measured as $\frac{\text{TN}}{\text{TN+FP}}$, displays the correctly identified non-fraudulent firm-years (TN) as a ratio out of all negative firm-years (TN + FP). Additionally, we report the models' accuracy, defined as $\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$, to display the overall correct classifications. While we acknowledge the limited validity of this measure within imbalanced data settings, we report it to present a complete overview of the overall performance.

[13] For a detailed description of NNs and parameter minimization using backpropagation see James et al. (2021).

[14] It must be noted that while establishing some comparability in metrics by using the AUC, additional heterogeneity in research designs between most studies persists.

[15] Also, $\Delta(1-\beta) = \left(1-\beta_i\right) - \left(1-\beta_{i-1}\right)$ and $\Delta\alpha = \alpha_i - \alpha_{i-1}$ (Bradley 1997). For details see Bradley (1997).

4 Sample and research design

4.1 Sample

We follow previous literature in constructing our sample from three distinct primary sources. First, we collect the 28 raw financial data items Bao et al. (2020) utilized for all publicly listed U.S. firms through COMPUSTAT from 2000 to 2018 and find 202,529 firm-year observations. We limit these observations beginning in 2000, as some CEO characteristics from BoardEx are only available from that year onwards. Second, consistent with a vast body of accounting fraud literature (e.g., Bao et al. 2020; Brown et al. 2020; Cecchini et al. 2010; Craja et al. 2020; Dechow et al. 2011; Perols et al. 2017), we use material accounting misstatements published in the SEC's Accounting and Auditing Enforcement Releases (AAERs) as binary fraud measure. We obtain the AAERs from 1982 to 2018 from Dechow et al. (2011) and match them to the firm-year observations from 2000 to 2018 using CIKs. AAERs can span over multiple firm-years, representing several consecutive materially misstated years. We follow most prior accounting detection literature and consider all misstated years within our sample. Following Wang et al. (2020), we argue that the CEO and the firm could have intervened at any time.[16] Lastly, we gather CEO data from BoardEx from 2000 to 2018. Inspired by Gupta et al. (2020), we identify CEOs by only considering observations with role titles, including "CEO" or "Chief Executive Officer". We matched the CEO and financial data items. Following previous literature (e.g., Brown et al. 2020; Dechow et al. 2011, 2012; Perols et al. 2017; Purda and Skillicorn 2015), we exclude financial services firms due to structurally missing financial items.

Due to mismatches in company and director identifiers, duplicates, and non-CEO data, the final unbalanced sample consists of 30,178 firm-years, including 198 firm-years for which AAERs have been issued. Table 1 depicts an overview of the sample selection process. Typical for accounting fraud research, the number of SEC identified misstatement years within our sample is small. This could lead to limitations through algorithmic over-specification, which should be considered when interpreting our results. However, our sample's absolute number of fraud cases is comparable to that of previous studies (e.g., Cecchini et al. 2010; Craja et al. 2020; Wang et al. 2020) and corresponds to 0.66% of all included firm-years, which is consistent with prior literature (e.g., Bao et al. 2020; Beneish 1999; Bertomeu et al. 2021). The data constitute 4281 unique, publicly-traded U.S.-located firms and 6,581 distinct CEOs.

Variables. The combined final data set includes 35 variables. The dependent dummy variable *misstate* is coded as 1 if an AAER has been issued for a firm-year and 0 otherwise. The considered 28 raw financial data items identified by Bao

[16] As hardly any prior accounting fraud detection studies addressed *"serial fraud"* (Bao et al. 2020, p. 203) as a potential limitation (Bao et al. 2020; Perols 2011; Perols et al. 2017), we follow the majority of predictive studies that did not specifically address this. However, we note that this procedure could result in a performance assessment that is too optimistic. Although the results are quite comparable to most prior studies, they must be interpreted with caution.

et al. (2020, p. 229) are *"Common Shares Outstanding"*, *"Current Assets—Total"*, *"Sale of Common and Preferred Stock"*, *"Property, Plant and Equipment—Total"*, *"Account Payable—Trade"*, *"Cash and Short-Term Investments"*, *"Price Close—Annual—Fiscal"*, *"Retained Earnings"*, *"Inventories—Total"*, *"Common/Ordinary Equity—Total"*, *"Debt in Current Liabilities—Total"*, *"Depreciation and Amortization"*, *"Receivables—Total"*, *"Cost of Goods Sold"*, *"Assets—Total"*, *"Long-Term Debt Issuance"*, *"Income Before Extraordinary Items"*, *"Long-Term Debt—Total"*, *"Interest and Related Expense—Total"*, *"Income Taxes—Total"*, *"Current Liabilities—Total"*, *"Sales/Turnover (Net)"*, *"Income Taxes Payable"*, *"Investment and Advances—Other"*, *"Liabilities—Total"*, *"Short-Term Investments—Total"*, *"Net Income (Loss)"*, and *"Preferred/Preference Stock (Capital)—Total"*. Additionally, this study includes six CEO-related variables established as antecedents of accounting fraud. *CEO Gender* represents an indicator variable, coded as 1 for male and 0 for female CEOs. *CEO MBA* has been operationalized as a dummy variable representing 1 if a CEO obtained an MBA and 0 otherwise. *CEO Duality* is operationalized through a binary dummy variable with 1 representing a CEO who also holds the chairperson position and 0 otherwise. *CEO Age* describes the age of the CEO at the beginning of a given firm-year. *CEO Tenure* describes the number of years the CEO has been in the position as CEO at the beginning of a given firm-year. Following prior literature, we consider these two variables at the beginning of the respective firm-year (e.g., Karpoff et al. 2008).[17] *CEO Network Size* represents the number of overlaps from the respective CEO's education, work, and further activities.[18]

Table 2 represents selected summary statistics for the variables of particular interest—CEO characteristics.[19] Precisely, Panel A of Table 2 displays the descriptive statistics for the entire sample and shows the average firm-year to have a CEO of about 54 years of age, a network of 1311 social contacts from work, education, or other activities, serving the company in his current position for about 4 years, be male (96%), without MBA (36%) and not serving as a chairman of the board (46%). Overall, these results appear in line with prior literature.[20]

Panel B of Table 2 represents the univariate differences in CEO characteristics between the fraudulent and non-fraudulent firm-year groups. We apply students *t* tests and the more robust Welch's tests for differences in unequal samples and variances. Accordingly, we find significantly smaller *CEO Age*, *CEO Network Size,* and *CEO Tenure* in fraudulent firm-years, while the ratio of firm-years with *CEO Duality* is significantly higher. This descriptive evidence is consistent with prior literature (e.g., Ali and Zhang 2015; Bhandari et al. 2018; Dechow et al. 1996; Huang et al. 2012; Troy et al. 2011). No statistically significant differences can be found for *CEO MBA* and *CEO Gender*. However, the mean value of *CEO MBA* is similar to the

[17] We thank an anonymous reviewer for raising this point.

[18] Complete variable definitions are provided in Online Appendix B.

[19] See Online Appendices C and D for a correlation matrix and descriptive statistics of all features.

[20] Only *CEO Network Size* appears to be diverging from prior literature which is due to a different measuring approach conducted by prior studies (e.g., Bhandari et al. 2018). While Bhandari et al. (2018) further disentangle and specifically focus on different components of network size, we argue that it is more useful to rely on more readily available data when constructing a prediction model.

Table 1 Sample selection overview

	Compustat firm-years	BoardEx firm-years	AAER firm-years
Initial Sample from Compustat (2000–2018)	**202,529**		
Initial Sample from AAER (2000–2018)			**1055**
Less: missing CIK	– 22,865		– 93
Less: duplicates	– 21,878		– 4
Sample Merged Compustat-AAER (2000–2018)	**157,786**		**838**
Initial CEO Sample from BoardEx (2000–2018)		**120,261**	
Less: non-matching identifiers (ISIN)		– 25,496	
Less: non-US headquarters		– 12,577	
Less: duplicates		– 14,070	
Sample BoardEx		**68,118**	
Less: non-matching identifiers & years	– 103,143		– 471
Sample Merged Compustat—BoardEx	**54,643**		**367**
Less: financial industry (SIC 6000–6999)	– 12,873		– 61
Less: missing values	– 11,592		– 108
Final Sample (firm-years 2000–2018)	**30,178**		**198**

Bold values indicates the important values

descriptive result of accounting fraud firms in Koch-Bayram and Wernicke (2018). The higher rate of male CEOs in fraudulent firm-years is in line with prior literature suggesting more conservative accounting behavior by firms led by female CEOs (Ho et al. 2015).

4.2 Research design

This study performs a research design similar to Craja et al. (2020) and Doornenbal et al. (2021). Thus, consistent with Craja et al. (2020), we apply five distinct classification models to detect accounting fraud out-of-sample. We perform various models to establish robustness across multiple prediction approaches and allow for potential insights into the complexity of the association. This argumentation is similar to Doornenbal et al. (2021), who employ a linear and more complex RF model to assess the degree of nonlinearity within the leadership-trait paradigm. Specifically, the models considered within this study are LR, SVM, RF, XGB, and a NN. Following Craja et al. (2020), we train and test the selected models incorporating different groups of variables.[21] While we only consider CEO-related variables within the

[21] Consequently, we follow Craja et al.'s (2020) depiction of the variable groups and similarly refer to them as CEO, FIN, and CEO + FIN.

Table 2 Descriptive statistics for CEO-characteristics

Panel A full sample

	Mean	SD	Median	Min	Max
CEO Age	53.89	7.66	54.00	27.00	86.00
CEO Network Size	1310.67	1463.59	829.00	5.00	17,168.00
CEO Tenure	3.94	5.28	2.10	0.00	54.90
CEO Duality	0.46	0.50	0.00	0.00	1.00
CEO MBA	0.36	0.48	0.00	0.00	1.00
CEO Gender	0.96	0.19	1.00	0.00	1.00
$n = 30{,}178$					

Panel B By fraud status

	Fraudulent firm-years		Non-fraudulent firm-years			
	Mean	SD	Mean	SD	*t* test *p*-value	Welch's test *p* value
CEO Age	50.83	8.45	53.91	7.65	**< 0.001*****	**< 0.001*****
CEO Network Size	977.14	1036.86	1312.88	1465.76	**0.001*****	**< 0.001*****
CEO Tenure	3.24	3.89	3.94	5.29	**0.062***	**0.012****
CEO Duality	0.58	0.50	0.46	0.50	**0.002*****	**0.002*****
CEO MBA	0.36	0.48	0.36	0.48	0.869	0.869
CEO Gender	0.97	0.16	0.96	0.19	0.359	0.269
n	198		29,980			

Panel A represents key summary statistics of the CEO characteristics for the full sample. Panel B depicts the means, standard deviations (SD), and *p* values for these variables between the fraudulent and non-fraudulent firm-year groups. All results are rounded to two decimals except for the *p* values, which are rounded to three decimals. The superscripts *, **, and *** represent the standard statistical significance levels of 10, 5, and 1% of a two-tailed students *t* test or Welch's test

Boldvalues indicates the significance values

selected models to answer RQ1a, we perform the detection models on the isolated CEO (CEO) and financial variables (FIN) as well as on the combined set of features (CEO + FIN) to answer RQ1b. Referring to prior studies, we perform permutation-based feature importance (e.g., Doornenbal et al. 2021) to rank the features by importance for model building and introduce the novel SHAP dependence plots by Lundberg and Lee (2017) on the best performing model. We assess the feature ranking, the direction of the association, functional form, and main and interaction effects to investigate RQ2a to RQ2c. Following previous literature, we mainly use the AUC to evaluate the models' capability to separate fraudulent and non-fraudulent firm-years correctly (e.g., Bao et al. 2020; Bertomeu et al. 2021; Craja et al. 2020).

Model Development. This study conducts an extensive model development process, including data pre-processing, hyperparameter tuning, and resampling. To address

the potential spurious effects of outliers which could significantly influence outlier-sensitive models (LR, NN), we follow previous studies and winsorize financial variables at the 1% and 99% levels (Beneish 1999; Bertomeu et al. 2021; Dechow et al. 2011; Green and Choi 1997).[22] Consistent with Craja et al. (2020), we perform a random and stratified sample split to keep a constant fraud-to-nonfraud ratio across training and test samples.[23] Following common machine learning practice, we split the data to reach a training set of 70% (21,124, incl. 139 fraud-years) and a hold-out test set of 30% (9054, incl. 59 fraud-years) of the total firm-year observations. We normalize the data sets to address potential scaling biases across LR, SVM, and NN features. As tree-based methods are insensitive to different scales, we follow Sigrist and Hirnschall (2019) and do not scale the RF and XGB models. To prevent our models from potentially misidentifying fraud years caused by an imbalanced data bias, we follow recent literature (e.g., Bao et al. 2020; Craja et al. 2020) and address the severely imbalanced data using random under sampling (RUS). RUS randomly removes majority class observations (i.e., nonfraud) to reach a targeted nonfraud-to-fraud ratio. While this approach could result in a biased majority class sample,[24] van Hulse et al. (2007) demonstrate the improved predictive performance of resampled models, with RUS performing best in most experimental settings, including settings relying on the AUC. As our data constitute a similar absolute fraud number to Craja et al. (2020), we follow them and target a nonfraud-to-fraud ratio of 4:1.[25] This reduces the severe imbalance and keeps a reasonable training size that allows for proper model training while still acknowledging non-fraudulent observations as the majority class (Craja et al. 2020).[26]

We apply all models to the isolated (CEO, FIN) and combined feature sets (CEO + FIN). Because different models are based on various meta-parameters that can be tuned to optimize their predictive power on out-of-sample testing, we apply the well-established grid-search algorithm to find each model's optimal values. The algorithm utilizes fivefold cross-validation to establish robustness in predictive performance over five varying training and validation set combinations to prevent the trained model from overfitting. Overall, the algorithm repeatedly runs through a grid of specified parameter values to train and validate the model on all potential combinations of parameters and identifies the parameter combination that results in the best AUC averaged over the five left-out validation sets.

Following this procedure for the CEO + FIN data, we find the optimal parameters for the SVM to incorporate a complexity parameter $C = 10$ and polynomial degree

[22] Analyses without winsorization show qualitatively similar results, which are available upon request.

[23] Following previous literature (e.g., Bao et al. 2020) and common machine learning practice, we set a random seed equaling zero to allow for random number generations that enable replications of our results for any random element.

[24] We thank an anonymous reviewer for raising this point.

[25] Acknowledging that the precise determination of a nonfraud-to-fraud ratio can hardly be supported by theoretical arguments or empirical evidence, we also compute our out-of-sample performance comparison with the adjacent ratios of 3:1 and 5:1. The untabulated analyses show qualitatively similar results, suggesting our findings to be robust.

[26] Specifically, we apply the outlined normalization and RUS approaches within grid-search and cross-validation to adapt to repeatedly changing training and validation sets.

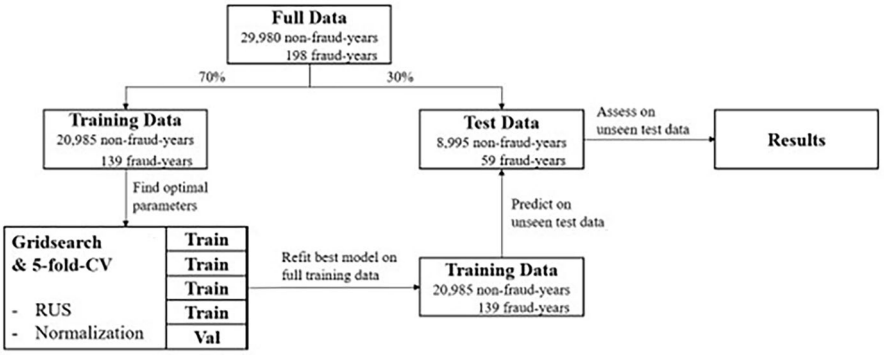

Fig. 1 Overview of model development and testing process

$d=2$. For the RF model, we find the number of trees $M=100$, the maximum number of features $m=2$, and tree depth $T=\infty$.[27] We find the final XGB model to include $M=500$ boosting iterations, a maximum depth $T=10$, and a learning rate $v=0.1$, and the NN consists of $h=100$, $a=$"relu", $s=$"adam", and $l2=0.05$. Following prior literature (e.g., Bao et al. 2020; Perols 2011), the LR has not been tuned. We follow an equivalent approach for the isolated (CEO, FIN) models. See Appendix A for details on the selection of the tuning parameters.[28] An overview of the model development and testing process is visualized in Fig. 1.

5 Results

5.1 Performance evaluation

RQ1a Do machine learning models for accounting fraud detection based on CEO characteristics achieve a predictive performance superior to random guessing?

Having found optimal hyperparameters, we assess the predictive power of these models on unseen test data. We follow prior literature and use the AUC to evaluate the models' predictive power. Model comparison primarily relies on the AUC differences in magnitude rather than statistical differences (e.g., Bao et al. 2020; Bertomeu et al. 2021; Cecchini et al. 2010; Craja et al. 2020).

Table 3 Panel A displays the prediction results for the final models based solely on CEO data (CEO). Our results indicate that all models exceed the benchmark

[27] We draw on Sigrist and Hirnschall (2019) in denoting the parameters. Thus, ∞ represents a tree's indefinite depth.

[28] Our results are based on computations in Python. Specifically, we draw on the scikit-learn (Pedregosa et al. 2011), SHAP (Lundberg and Lee 2017), Alibi (Klaise et al. 2021), and imblearn (Lemaître et al. 2017) packages.

AUC for no-skill models of 0.5 by a large margin. Even the weakest model, the LR, outperforms the random benchmark by about 11%-points. The strongest predictive model, the XGB, even outperforms the benchmark by 31%-points, closely followed by the RF with 29%-points. These results suggest robust empirical evidence of CEO characteristics' joint predictive power toward firm-year accounting fraud detection across various prediction models.

Interestingly, we also find superior tree-based models, suggesting a nonlinear relationship between CEO characteristics and accounting fraud. We find the XGB as the most sensitive model when considering the other metrics. It correctly identifies TPs out of all investigated firms (TP, FP) in about 85% of firm-years. However, compared to the RF model, the XGB performs weaker in correctly identifying non-fraudulent firm-years (TN) out of all non-fraudulent firm-years (0.6819), also resulting in lower accuracy (0.6830). The LR and especially the SVM models appear to have difficulties achieving high sensitivity.

RQ1b Do machine learning models for accounting fraud detection based on a combination of raw financial and CEO information (CEO + FIN) outperform isolated approaches (CEO, FIN)?

In addition to assessing CEO-based models, we compare the models for CEO data combined with raw financial data items drawn from Bao et al. (2020) and expect superior predictive performance. Table 3 Panel C shows the results for the combined model (CEO + FIN). We find that the RF outperforms the other models with a test AUC (sensitivity) of 0.9285 (0.8644), followed by the XGB with a score of 0.9018 (0.7458). The weakest prediction model, the LR, achieves an AUC score (sensitivity) of 0.7590 (0.6610). Thus, the strongest model outperforms the weakest by a large AUC margin of about 17%-points. Related to specificity (0.8987) and accuracy (0.8977), the XGB outperforms all other models. Again, the tree-based models strongly exceed the performance of the other models when considering the AUC. This is consistent with Craja et al. (2020), who found that the RF and XGB models outperformed the LR, SVM, and NN models for finance data. Again, this provides empirical evidence for a potential nonlinear relationship between these predictors and accounting fraud. Further, Panel B of Table 3 presents the results of the isolated financial (FIN) models for comparison.

When comparing the results of the combined models (CEO + FIN) with equivalent models on the isolated data (CEO, FIN), we find the combined data models to outperform these separated approaches across all models. While the AUC deltas are higher for the isolated CEO models than the isolated FIN models, we still find relevant increases in prediction results for the combined models over the FIN models. Thus, especially among the weaker performing models (LR, NN, SVM), the added predictive power of the combined models on out-of-sample data reaches an increase of 7%-points for the LR, 6%-points for the NN, and about 5%-points for the

Table 3 Out-of-sample performance comparison of different feature models

Panel A CEO data (CEO)					
Model	AUC	Sensitivity	Specificity	Accuracy	Delta AUC (No skill)
LR	0.6148	0.5593	0.6804	0.6796	0.1148
SVM	0.6252	0.3729	0.8316	0.8286	0.1252
RF	0.7919	0.6610	**0.8744**	**0.8744**	0.2919
XGB	**0.8089**	**0.8475**	0.6819	0.6830	0.3089
NN	0.7073	0.6949	0.6784	0.6785	0.2073
Panel B finance data (FIN)					
Model	AUC	Sensitivity	Specificity	Accuracy	Delta AUC (CEO)
LR	0.6884	0.4576	**0.8822**	**0.8794**	0.0736
SVM	0.7209	0.5593	0.8328	0.8310	0.0957
RF	**0.9133**	0.8983	0.8327	0.8331	0.1214
XGB	0.8849	**0.9322**	0.6854	0.6870	0.0760
NN	0.7936	0.7458	0.7675	0.7674	0.0863
Panel C CEO and finance data (CEO + FIN)					
Model	AUC	Sensitivity	Specificity	Accuracy	Delta AUC (FIN)	Delta AUC (CEO)
LR	0.7590	0.6610	0.7676	0.7670	0.0706	0.1442
SVM	0.7748	0.5763	0.8888	0.8868	0.0539	0.1496
RF	**0.9285**	**0.8644**	0.8743	0.8742	0.0152	0.1366
XGB	0.9018	0.7458	**0.8987**	**0.8977**	0.0169	0.0929
NN	0.8581	0.7458	0.8410	0.8404	0.0645	0.1508

This table's presentation is inspired by Craja et al. (2020). The scores of all metrics are rounded to 4 decimals. The Delta AUC scores rely on the rounded 4-decimal AUC scores

Bold values indicates the important values

SVM. While the added predictive value appears smaller for the RF and XGB models (about 2%-points each), this still represents an economically significant improvement, as prior studies find similar improvement ranges. The most similar study by Craja et al. (2020) finds the combination of financial and textual models to outperform models based on a combination of financial and linguistic data by 2%-points of AUC for RF and 3% for XGB.[29] Despite more diverging data pre-processing, models, and testing approaches, Bao et al. (2020) show the combination of the novel RUSBoost model and raw data to outperform the financial ratio-based LR model by Dechow et al. (2011) by 5.3%-points as measured by an average AUC score on unseen test data. They also show a predictive improvement of the LR and RUSBoost

[29] We note that Craja et al. (2020) diverge from our study in their use of different financial data and other aspects, such as imputing missing values. Consequently, these studies' results can only be compared with caution.

models by 2%-points and 6%-points for changing from 14 financial ratios to 28 raw financials (Bao et al. 2020).

Overall, the results suggest the added predictive power by combining CEO and raw financial data items compared to equivalent models based on their isolated predictors (CEO, FIN). The results appear to be robust across all tested prediction models.[30]

5.2 Model interpretability

RQ2a How influential are individual CEO characteristics for accounting fraud detection within the best CEO + FIN model?

Interpretability is relevant to identifying the main drivers behind an algorithm classifying firm-years into fraudulent and non-fraudulent. Consequently, many feature-importance techniques have been proposed (see Molnar (2022) for an introduction). We follow recent literature (e.g., Doornenbal et al. 2021) and apply the permutation-based feature importance to our best performing CEO + FIN model, the RF.[31] We do so rather than following prior accounting fraud literature (Bao et al. 2020; Bertomeu et al. 2021) since Strobl et al. (2007) showed empirical evidence for the superiority of the permutation-based approach. According to Strobl et al. (2007), impurity-based feature importance can lead to more unreliable feature importance estimates than permutation-based ones. Specifically, as the importance of impurity-based features systematically discriminates against lower-cardinality features (Strobl et al. 2007), we would expect the impurity-based feature importance to strongly bias the importance against the low-cardinality variables *CEO Duality*, *CEO MBA*, and *CEO Gender*. While this would limit the accurate interpretation of the variables' importance,[32] applying the permutation-based approach mitigates this potential bias.

Permutation-based feature importance was developed by Breiman (2001) to provide a relative ranking of the contributions of a model's features toward its prediction. The technique shows the total feature importance, including main and second-order effects (Molnar 2022). The idea is to introduce noise by permuting a feature's values, keeping the other features constant, and measuring the increase in the model's prediction error (Molnar 2022). The higher the difference in prediction error after permutation, the more the model depends on this (permuted) feature and vice versa (Molnar 2022). However, we would like to point out that the ranking is not linked to the variables' statistical significance (Shmueli 2010). Features that are not

[30] Robustness tests comparing equal variable models support the models' increased predictive performance by complementing FIN with CEO characteristics. See Online Appendix E for detailed results.

[31] See https://scikit-learn.org/stable/modules/permutation_importance.html for a more detailed explanation of the function and a computational example.

[32] To validate this expectation, we also estimated the impurity-based feature importance ranking on the best performing CEO + FIN RF model and found strong evidence of lower cardinality variables being discriminated, whereas *CEO Age* (1) and *CEO Network Size* (7) retained a top 10 rank. The untabulated results are available upon request.

significantly associated with the target might still be necessary for the out-of-sample predictive performance of a model (Gow et al. 2016).

We follow Doornenbal et al. (2021) and employ the permutation-based feature importance with 200 feature permutations on the test set. However, while Doornenbal et al. (2021) based their feature importance method on the regression-based root mean squared error (RMSE), we consider the decrease in AUC for our imbalanced classification setting. This seems reasonable, as the AUC is used to find the best model. Panel A of Table 4 visualizes the permutation feature importance and the respective mean AUC decrease.

Columns 1 and 3 of Panel B of Table 4 compare our RF model's top 10 most essential features on CEO + FIN and Bao et al. (2020). Following Bao et al. (2020), we investigate the 10 most essential features. For the best performing RF model on the test sample, our results suggest *Inventories—Total, CEO Network Size*, *CEO Age*, *Receivables—Total*, *Investment and Advances—Other, Liabilities—Total, Property, Plant, and Equipment—Total (Gross)*, *Cash and Short-Term Investments*, *CEO Duality* and *Interest and Related Expense—Total* as the essential features, ranked from highest to lowest.[33] Interestingly, this implies two of the three most important features to be CEO characteristics. Mainly, *CEO Network Size* contributes second most towards the predictions, followed by *CEO Age*. When permuting the most critical variables, *CEO Network Size* and *CEO Age* decrease the mean AUC by about 1.8 and 1.2%-points.

Additionally, with *CEO Duality,* a third CEO variable ranks within the top 10 most essential features. Thus, half of the CEO variables enter the top 10 despite their strong numerical inferiority compared to the incorporated raw financials. This suggests the strong predictive power of some CEO characteristics in the novel combination with raw financials for accounting fraud detection. However, the remaining CEO characteristics only rank 19 (*CEO Tenure*), 21 (*CEO MBA*), and 30 (*CEO Gender*). This result suggests the rather mediocre importance of *CEO Tenure* and *CEO MBA* and the weak contribution of *CEO Gender* towards the model's predictive power. Especially when introducing noise to *CEO Gender* while keeping the other variables' values constant, it has nearly no effect on the AUC. Consistent with similar literature (e.g., Bertomeu et al. 2021), our relatively low mean AUC decreases indicate that the final model heavily draws on the combinations of variables rather than a few strongly predictive ones.[34]

Inspired by Bao et al. (2020), we relate those findings to prior literature. For *CEO Network Size*, Bhandari et al. (2018) suggest a negative association between CEO connections and accounting fraud. Concerning the importance of *CEO Age* for accounting fraud, Troy et al. (2011) argue and empirically show that younger

[33] As stated by Bertomeu et al. (2021), the results of this feature importance ranking should be interpreted with caution. Thus, correlations between variables can bias the results, as is typical for multivariate descriptive analyses (Bertomeu et al. 2021). However, when testing our results by comparing the RF and XGB feature rankings, we find a large overlap (7/10, incl. *CEO Network Size, CEO Age* and *CEO Duality*) in the top 10 features, suggesting that our results are robust. The results are presented in Sect. 5.2, Table 4 Panel B, and verbally in Sect. 5.3.

[34] This explanation is consistent with our untabulated impurity-based feature importance analysis. Thus, most features show a relative importance of about 3%, indicating the essentiality of variable combination.

Table 4 Feature importance of the best performing RF Model on CEO + FIN test data

Panel A Permutation feature importance

Rank	Predictor	Mean AUC decrease
1	Inventories—Total	0,018402
2	CEO Network Size	0,017624
3	CEO Age	0,011711
4	Receivables—Total	0,011329
5	Investment and Advances—Other	0,009656
6	Liabilities—Total	0,009263
7	Property, Plant, and Equipment—Total (Gross)	0,009229
8	Cash and Short-Term Investments	0,008977
9	CEO Duality	0,008436
10	Interest and Related Expense—Total	0,008409
11	Sale of Common and Preferred Stock	0,006927
12	Long-Term Debt—Total	0,006790
13	Income Taxes Payable	0,006324
14	Common Shares Outstanding	0,004984
15	Debt in Current Liabilities—Total	0,004929
16	Current Assets—Total	0,004559
17	Long-Term Debt—Issuance	0,004486
18	Common/Ordinary Equity—Total	0,004325
19	CEO Tenure	0,003247
20	Retained Earnings	0,003063
21	CEO MBA	0,002446
22	Income Before Extraordinary Items	0,001946
23	Price Close—Annual—Fiscal	0,001849
24	Depreciation and Amortization	0,001669
25	Income Taxes—Total	0,001589
26	Preferred/Preference Stock (Capital)—Total	0,001420
27	Current Liabilities—Total	0,001022
28	Net Income (Loss)	0,000587
29	Assets—Total	0,000231
30	CEO Gender	0,000042
31	Accounts Payable—Trade	− 0,000436
32	Cost of Goods Sold	− 0,000554
33	Short-Term Investments—Total	− 0,002645
34	Sales/Tturnover (Net)	− 0,004064

Panel B Top 10 feature importance—test comparisons

	CEO + FIN		Bao et al. (2020)
Rank	RF	XGB	RUSBoost
1	**Inventories—Total**	*Common/Ordinary Equity—Total*	*Common Shares Outstanding*
2	**CEO Network Size**	**CEO Network Size**	Current Assets, Total

Table 4 (continued)

Panel B Top 10 feature importance—test comparisons

	CEO + FIN		Bao et al. (2020)
Rank	RF	XGB	RUSBoost
3	**CEO Age**	**CEO Age**	*Sale of Common and Preferred Stock*
4	**Receivables—Total**	**Receivables—Total**	*Property, Plant, and Equipment, Total*
5	**Investment and Advances—Other**	**Investment and Advances—Other**	Account Payable, Trade
6	Liabilities—Total	***Inventories—Total***	*Cash and Short-Term Investments*
7	**Property, Plant, and Equipment—Total (Gross)**	CEO Duality	*Price Close, Annual, Fiscal*
8	*Cash and Short-Term Investments*	property, Plant, and Equipment—Total (Gross)	Retained Earnings
9	**CEO Duality**	*Sale of Common and Preferred Stock*	*Inventories, Total*
10	Interest and Related Expense—Total	*Common Shares Outstanding*	*Common/Ordinary Equity—Total*

Panel A represents the permutation-based feature importance ranking, and the respective mean AUC decreases after permutation for the best CEO + FIN model, RF. Panel B displays our top 10 most important features and compares them with the top 10 features by Bao et al. (2020) as a benchmark. We note, however, that the comparison with Bao et al. (2020) requires caution, as we use permutation-feature importance instead of impurity-based feature importance. Variables with asterisks signify CEO characteristics that were not included in Bao et al. (2020). Variables within the top 10 feature importance rankings of RF and XGB are bold. Variables in the RF/XGB and Bao et al.'s (2020) RUSBoost are written in italics

CEOs rationalize accounting fraud significantly stronger than older CEOs. Similarly, Huang et al. (2012) investigate financial reporting quality and *CEO Age* and found a positive association. Considering the association between *CEO Duality* and accounting fraud, we assume this variable to be important as the power of the CEO strongly increases when also serving as chairman of the board, which in turn results in higher earnings management and financial misconduct (e.g., Dechow et al. 1996; Yang et al. 2017).

Consequently, this literature suggests a positive relationship between *CEO Duality* and accounting fraud. However, further empirical evidence in this study's setting would be required to validate this expectation. While these CEO characteristics are suggested to contribute significantly to the model's overall prediction, we also find that some characteristics appear relatively irrelevant. Interestingly, *CEO Gender* shows a neglectable contribution to the prediction despite previously found empirical evidence for a significant association between gender and accounting fraud (e.g., Gupta et al. 2020; Ho et al. 2015; Wahid 2019). We assume that the small absolute and relative number of female CEOs in our data set partially explains this finding. Only 1135 (3.8%) of the included firm-years represent observations with female CEOs.

Additionally, we find no statistically significant difference in means between fraudulent and non-fraudulent firm-years for *CEO Gender* in Sect. 4.1. This suggests that firm-years do not significantly differ in female CEO ratios within our sample. Noteworthy, within the top 10, half of the remaining raw financials have been identified among the 10 most essential features by Bao et al. (2020). In particular, *Inventories – Total*, *Property, Plant and Equipment – Total (Gross),* and *Cash and Short-Term Investments* overlap with Bao et al. (2020).[35] This suggests that our best-performing RF model identifies a similar pattern for raw financials as Bao et al. (2020).

RQ2b How do the individual CEO characteristics contribute toward accounting fraud detection within the best CEO + FIN model?

While feature importance rankings allow for interpretations of the overall influences of variables towards the prediction outcome, they do not further open the features' total importance. Thus, they lack information on the structural form, direction of the association, and main and second-order effects between X_i and $f(X_i)$. The SHAP dependence plot is a state-of-the-art interpretable technique that disentangles the association of a prediction model's features (Lundberg and Lee 2017). SHAP dependence plots are a novel visualization alternative to the primarily used Partial Dependence Plots (PDPs) by Friedman (2001) and the more recent Accumulated Local Effect (ALE) plots by Apley and Zhu (2020) (Molnar 2022). Similar to PDP and ALE plots, SHAP dependence plots estimate and visualize the structural relationship between a feature and the prediction without assuming a functional form a priori. Like PDPs, they can identify more complex associations. However, they deviate from PDPs and ALE plots in additionally displaying prediction variance (Molnar 2022) and their interpretation of the main effect. SHAP aims to explain a prediction by calculating every feature's contribution to the prediction (Molnar 2022) based on the game-theoretic Shapley values (Shapley 1953). Molnar (2022) describes the approximation of Shapley values by Štrumbelj and Kononenko (2014) as follows:

$$\hat{\varnothing}_j = \frac{1}{M}\sum_{m=1}^{M}(\hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m)),$$

where $\hat{f}(x_{+j}^m)$ is x's prediction with a random amount of predictor values. Lundberg and Lee (2017, p. 4) use the *"Shapley values of a conditional expectation function of the original model"* as SHAP values. They estimate the features' marginal contributions to a prediction $f(x)$ as the differences between the conditional expectations using the respective feature and the observation's unconditional expectation E[$f(z)$] (Lundberg and Lee 2017).[36] SHAP dependence plots could be considered the most

[35] However, note that Bao et al. (2020) diverged in some instances from our study and, for instance, followed the impurity-based feature importance rather than the permutation feature importance. Nevertheless, Bao et al. (2020) is most comparable to this study in raw financial predictors and our findings suggest some similarity in feature importance of the identified model patterns despite those differences.

[36] For a detailed description of SHAP see Lundberg and Lee (2017). For an introduction see Molnar (2022).

intuitive global interpretation plot as it plots the feature instances' values on the x-axis and their respective Shapley values on the y-axis (Molnar 2022). We employ SHAP dependence plots and visualize the main effects of our CEO characteristics using SHAP values in Fig. 2.[37]

For *CEO Age*, our results suggest a tendency of a U-shaped association with accounting fraud. While the accounting fraud likelihood is the highest with low *CEO Age*, the predicted accounting fraud probability decreases with CEOs becoming older, up to about 60 years. Beyond this age, the likelihood of accounting fraud increases to around 70 and stabilizes after that.[38] This result is partly consistent with previous literature (Huang et al. 2012; Troy et al. 2011). Troy et al. (2011) find a significant difference in *CEO Age* between fraudulent and non-fraudulent firms, with younger CEOs being related to a higher accounting fraud likelihood. These results are supported by Huang et al. (2012) who suggest a positive (negative) relationship between *CEO Age* and accounting quality (financial restatements).

However, while they also tested for a potentially curvilinear relationship between *CEO Age* close to retirement and reporting quality, they found no statistically significant coefficient for distinguishing between *CEO Age* under and beyond 62 (Huang et al. 2012). In contrast to this finding and in line with their reasoning and prior literature that suggests a potential increase in the likelihood of earnings management shortly before retirement (e.g., Davidson et al. 2007; Dechow and Sloan 1991), we find empirical evidence of a U-shaped tendency.

Concerning *CEO Network Size*, our results suggest a negative L-relationship with a substantial decline in the accounting fraud likelihood until about 1000 to 2000 connections and stabilization at higher network levels. However, the limited data in the latter region requires caution when analyzing this movement. Accordingly, a smaller network is associated with a high likelihood of the firm-year being identified as fraudulent. This finding partly aligns with previous research by Bhandari et al. (2018) that suggests a negative association between *CEO Network Size* and accounting fraud. Accordingly, an increase in network size could be associated with a lower accounting fraud likelihood due to a CEO's more assertive pursuit of keeping a good reputation with increasing network size (Bhandari et al. 2018). However, this does not explain why accounting fraud likelihood somewhat stabilizes at high levels of social connections and why the decrease does not seem to be strictly linear.

CEO Tenure shows the highest contribution toward a higher accounting fraud likelihood in concise years of service, which decreases toward about 3–4 years and stabilizes beyond with high variance.[39] This indicates a nonlinear, somewhat

[37] We also estimate the main effects of nonbinary CEO characteristics using PDP and ALE plots, finding similar results. We only tested the robustness of these features as current Python packages (e.g., Alibi) do not allow for an estimation of categorical features for ALE. The untabulated results are available from the authors upon request.

[38] Untabulated robustness tests using PDP and ALE plots suggest a similar but even stronger U-shaped association.

[39] Robustness tests using PDP and ALE plots show a similar association. However, they suggest slight tendencies to decrease further for CEO tenures between 10 and 15 years and stabilization for longer tenures.

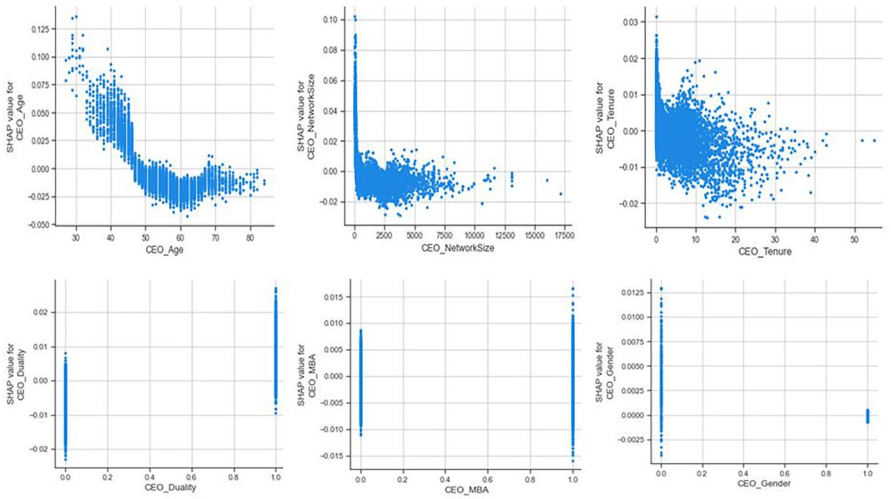

Fig. 2 SHAP dependence plots visualizing the main effects of CEO characteristics

L-shaped association. Consistent with previous literature, one potential explanation for our results could be the uncertainty surrounding the CEO's ability in the early years, which might lead the CEO to manipulate financial statements in the CEO's favor (Ali and Zhang 2015). Thus, Ali and Zhang (2015) found that earnings overstatement is higher in CEO's early years, up to 3 years. Although this study investigates the magnitude rather than the likelihood of earnings management, it might partly provide evidence for the observed association. This could explain the relatively high accounting fraud likelihood in the early years and the stabilization close to 10 years of service.

Additionally, the horizon problem suggests that CEOs engage in extraordinary earnings overstatement in their final years when controlling for early year overstatement (Ali and Zhang 2015). While the high variance requires caution when interpreting our results, this could partly explain the minor increase and positive outliers at about 5–10 years but does not explain the weak negative and stabilization tendency in accounting fraud likelihood beyond this tenure.

Concerning binary features, our results indicate that the essential *CEO Duality* variable exhibits a positive association, consistent with prior literature that suggests firm-years with CEOs serving as chairman have a higher accounting fraud likelihood (e.g., Dechow et al. 1996; Yang et al. 2017). Thus, a CEO's high power over the board received through the combination of two important management positions could lead to ineffective monitoring (Jensen 1993) and, hence, a higher opportunity which could explain the higher accounting fraud likelihood (Dechow et al. 1996).

Consistent with Troy et al. (2011), who find business education to reduce the likelihood of accounting fraud, our results suggest no clear but if at all a weak negative association between *CEO MBA* and accounting fraud.

Interestingly, Fig. 2 suggests a tendency of a weak positive association between female CEOs and the likelihood of accounting fraud. This contrasts with prior literature that provided empirical evidence of firms with female CEOs conducting more conservative accounting than male-led firms (Ho et al. 2015). However, this result can only be interpreted with caution, as there is only a small amount of data and a high variance for the female class (here, # 795 or 3.8%).

Overall, our results indicate nonlinear relationships for the nonbinary features, suggesting a more complex relationship between these CEO characteristics and the likelihood of accounting fraud. This is consistent with our findings, which showed that the nonlinear, tree-based models outperform the linear ones.

RQ2c How do the essential CEO characteristics interact with each other and raw financials within the best CEO + FIN model?

We also analyze interaction effects, introducing SHAP dependence plots by Lundberg and Lee (2017) to the accounting fraud literature for visualizing feature interactions. As a state-of-the-art alternative to PDP and ALE, SHAP also allows separating interaction from main effects (Molnar 2022). Inspired by Sigrist and Hirnschall (2019), we analyze feature interactions for two exemplary essential feature combinations. The results for two interaction effects of *CEO Duality* with *CEO Age* and *CEO Network Size* with *Inventories—Total* are visualized in Fig. 3. The graphs illustrate the SHAP values for accounting fraud likelihood on the combination of two variables simultaneously (Molnar 2022).[40]

Concerning the interaction effect of *CEO Duality* and *CEO Age*, our findings suggest a counterintuitive interaction. We find older CEOs who are not the chairman of a firm to show a higher likelihood of accounting fraud as compared to younger CEOs. For CEOs who are also the chairman of a company, however, younger CEOs are more likely to be associated with accounting fraud. While the overall finding is consistent with prior literature that suggests *CEO Duality* be positively associated with accounting fraud likelihood (e.g., Dechow et al. 1996), *CEO Age* does not seem to mitigate this association equally for both scenarios as suggested by the literature (Huang et al. 2012).

Investigating the interactions of *CEO Network Size* and *Inventories—Total*, we find a relatively small *CEO Network Size* associated with minor *Inventories*—Total. Notably, for CEOs with connections around 1000–2000, the accounting fraud

[40] However, we would like to note that SHAP dependence plots represent no causal model and interaction effects could simply be driven by confounders, requiring cautious interpretation (Molnar 2022).

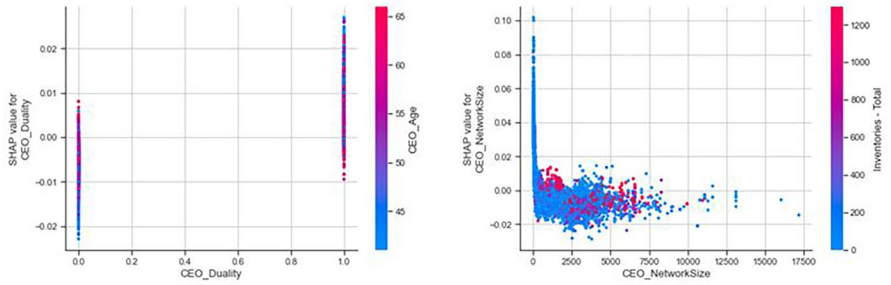

Fig. 3 SHAP dependence plots visualizing interaction effects

likelihood is highest for higher inventory firm years. Given a specific network size, higher inventories appear to be associated with medium to higher accounting fraud likelihood.

5.3 Alternative feature importance evaluation

The previous results of permutation feature importance are based on the best performing CEO + FIN RF only. However, as we noted, correlated features could lead to a different ranking for diverging prediction models. Thus, we follow Bertomeu et al. (2021) and test the robustness of our feature importance ranking by comparing it with other tree-based models, here the second-best performing model, XGB.

Table 4 Panel B compares the permutation feature importance of the top 10 most important features across the RF and XGB models. Further, the RUSBoost model of Bao et al. (2020) is provided as a benchmark. Comparing the most essential RF and XGB models' features, we find substantial overlap in feature inclusions while the order changes slightly in many instances. Thus, both models overlap in 7 of 10 variables. This includes all essential CEO characteristics, providing further evidence of the strong importance of these characteristics toward the models' predictive performance.

Additionally, we find a substantial overlap with raw financials, also identified as necessary by Bao et al. (2020). This especially holds for the XGB model, which shows even more substantial overlap with 5 of 7 raw financials identified by Bao et al.'s (2020) RUSBoost model. While this result could be rooted in the more similar nature of algorithms drawing on boosting, it further suggests a similar detection pattern to our models and validates these raw financials' importance for accounting fraud detection.

6 Limitations and research recommendations

Our work is not without limitations. Although not unique to our research, this study utilizes published SEC AAERs as accounting fraud proxy. This has two main disadvantages. First, it assumes a perpetrator's intention that can hardly be verified.

However, the assumption of intention seems reasonable, as the SEC is mainly expected to investigate and publish cases they believe in proving intention for in a court of law (Dechow et al. 1996; Feroz et al. 1991). Second, AAERs only reflect detected and enforced cases by the SEC, likely resulting in a significant fraction of hidden fraud cases (Dechow et al. 2011; Karpoff et al. 2017). While predominantly used within the prediction literature, future research might consider complementing their findings with other financial misconduct metrics to increase validity. Karpoff et al. (2017) suggest alternative proxies from the Governmental Accounting Office, Audit Analytics, and Stanford Securities Class Action Clearinghouse to complement SEC AAERs. However, while they might mitigate some limitations at the expense of introducing others, limitations inherent to the research field of accounting fraud, such as the small number of fraud observations, prevail.

Our sample covers publicly listed U.S. firms. While generalizability to other contexts is limited, recent European financial scandals (e.g., Carillion, Wirecard) and the subsequent audit and corporate governance reform initiatives by the EU Commission suggest the potential interest of our research to stakeholders from other regions, as well. The recently adopted European Single Access Point (ESAP), which aims at providing centralized financial information on EU companies, might improve research on European companies. Additionally, few studies consider non-financial predictors, such as CEO characteristics, control-, risk management-, and internal audit systems, for machine learning-based accounting fraud detection (e.g., our study and Bertomeu et al. 2021). Interestingly, the Sarbanes Oxley Act of 2002 highlighted the importance of internal governance systems. Similarly, related antecedents of sustainability and compliance information might also represent an interesting future research avenue.[41]

We particularly acknowledge limitations in the algorithms' black-box character regarding machine learning-based research. While we address this caveat by incorporating model-agnostic methods, interpretations should be considered cautiously. In particular, causality can hardly be established. Consistent with our results suggesting multiple research opportunities of complex relationships between CEO characteristics and accounting fraud, we strengthen recent calls for advanced interpretable machine learning (Doornenbal et al. 2021) and complexity-driven research (Velte 2021). To summarize, we encourage research in the European context that focuses on non-financial factors, interpretable and non-linear relationships.

7 Conclusion

This study investigates five predictive models based on well-established raw financial data items and CEO characteristics in isolation (CEO, FIN) and a novel combination of these predictors (CEO + FIN). We consider the CEO instead of other top executives, such as the CFO, as the literature typically considers the CEO to be a firm's most powerful character and to hold additional power over the CFO's

[41] We thank an anonymous reviewer for raising these potential future research areas.

accounting behavior (e.g., Feng et al. 2011; Friedman 2014). Various CEO characteristics have been established concerning outcomes related to accounting fraud. Consistent with this line of argumentation, we find all isolated CEO models to outperform random guesses by a large AUC margin. For all combined data models (CEO + FIN), we find them to outperform the isolated models by large margins, with RF performing the best. These results suggest the complementary predictive value of CEO characteristics within machine learning-based accounting fraud detection. We confirm prior empirical evidence by Craja et al. (2020) that showed the superiority of tree-based models, suggesting nonlinear relationships between financial predictors and accounting fraud, and extend it to CEO characteristics.

While we rely on research design elements of similar studies (Bao et al. 2020; Craja et al. 2020), we diverge from these studies in various ways. Thus, to our knowledge, we are the first to focus on the predictive power of CEO characteristics in isolation and combination with raw financials for machine learning-based accounting fraud detection. Additionally, diverging from most prior literature, we address the typical issue of black-box models and introduce model-agnostic techniques to gain feature-related insights. Thus, we disentangle the novel combination models' predictive performance drivers utilizing permutation-feature importance and introduce SHAP dependence plots to the accounting fraud detection literature. Our results suggest that *CEO Network Size*, *CEO Age*, and *CEO Duality* are among the top 10 most substantial contributors to the model's predictions. Robustness checks confirm these results. We are also the first to extend these findings and show the L-shaped, U-shaped, and L-shaped main effects of *CEO Network Size, CEO Age*, and *CEO Tenure* within machine learning models, respectively. We suggest strong, weak, and neglectable main effects for *CEO Duality*, *CEO MBA*, and *CEO Gender*. Moreover, we indicate complex interactions between *CEO Duality* and *CEO Age* and *CEO Network Size* and *Inventories—Total*. Thus, older CEOs not serving as the chairman and CEOs with a network of up to 2500 and high inventory are more likely to be associated with accounting fraud.

Our study extends current knowledge of CEO characteristics and accounting fraud detection in multiple ways. First, we indicate that CEO characteristics effectively detect accounting fraud within machine learning models in isolation and complement raw financial data items. Second, opening the black box, we find empirical evidence for nonlinear relationships between nonbinary CEO characteristics and accounting fraud and complex interactions, suggesting future research potential to advance theories and develop novel hypotheses.

Our research also has practical implications. Stakeholders interested in detecting accounting fraud, such as auditors and authorities, might incorporate or improve predictive models, including CEO characteristics, to reduce detection time. CEO-based models could also be used for widely excluded firms (e.g., the financial industry) with diverging financial statement requirements. Incorporating SHAP also allows for explaining predictions, which is essential for applying such models in practice. Novel insights into CEO characteristics' functional form and interactions on accounting fraud likelihood might provide ground for new policy directions or research theories.

Appendix A

Tuning parameter selection

Following prior literature (e.g., Bao et al. 2020; Perols 2011; Wang et al. 2020), we do not tune the LR model. For the SVM, we follow previous studies (e.g., Perols 2011; Shin et al. 2005) and tune the complexity parameter $C \in \{1, 10, 50, 75, 100\}$ and Perols (2011) in tuning the polynomial kernel-based SVM concerning its degree $d \in \{0.5, 1, 2, 5, 10\}$. After searching the parameter combinations to optimize the cross-validation AUC, we yield a final model with $C = 10$ and $d = 2$ for the combined feature model. For tuning the RF, we follow Sigrist and Hirnschall (2019) and consider the number of decision trees M, the maximum depth of each base tree T, and the maximum number of randomly chosen features for decision tree splits m to be tuned. We consider $M \in \{100, 500, 1000, 1500, 2000, 2500, 3000, 3500\}$ and $T \in \{3, 5, 10, \infty\}$.[42] Following standard machine learning practices (e.g., James et al. 2021), we investigate values around $m \approx \sqrt{p}$.[43] Specifically, we consider lower values of m leading to higher decorrelation of the individually grown decision trees, primarily used data sets with potentially correlated predictors (James et al. 2021). Thus, we consider $m \in \{1, 2, 3, 4, 5, 6\}$ for the combined CEO + FIN model. We find optimal parameters for the number of trees $M = 100$, the maximum number of features $m = 2$, and tree depth $T = \infty$. Following Sigrist and Hirnschall (2019) for tuning boosted trees, the XGB is tuned concerning the number of boosting iterations M, maximum depth of each decision tree T, and the learning rate v.[44] We consider parameter combinations of $M \in \{100, 500, 1000, 1500, 2000, 2500, 3000, 3500\}$, $T \in \{3, 5, 10, \infty\}$ and $v \in \{0.01, 0.1, 0.001\}$. Further, we set early stopping rounds to 10 to speed up the computing process. The final XGB model includes $M = 500$ boosting iterations, a maximum depth $T = 10$, and a learning rate $v = 0.1$. For the hyperparameter tuning of the NN, we find the best model by searching through the parameter combinations of hidden layer sizes $h \in \{10, 50, 100, 200, 300, 400, 500\}$, the activation functions $a \in$ {hyperbolic tan function, rectified linear unit function}, solver functions $s \in$ {"sgd", "adam"} and the l2 penalty $l2 \in \{0.0001, 0.05\}$. Validating these parameter combinations, we find the optimal model to consist of $h = 100$, $a =$ "relu", $s =$ "adam," and $l2 = 0.05$. The tuning parameter selection for FIN and CEO are visualized in Table 5.

[42] Following Sigrist and Hirnschall (2019), we denote an indefinite maximum tree depth as ∞.

[43] As p represents to total number of predictors, m represents a subsample as its square root, here $\sqrt{34} = 6$.

[44] James et al. (2021) also highlight the importance of tuning these meta-parameters for boosting algorithms.

Table 5 Hyperparameter tuning finance (FIN) and CEO (CEO) models

Method	Hyperparameter	Grid range	Final value CEO	Final value FIN	Final value CEO + FIN
LR	–	–	–	–	–
SVM	C	1, 10, 50, 75, 100	10	50	10
	Polynomial Degree	0.5, 1, 2, 5, 10	5	2	2
RF	# Estimators	100, 500, 1000, 1500, 2000, 2500, 3000, 3500	500	500	100
	Max. Depth	3, 5, 10, ∞	∞	∞	∞
	Max. Features	1, 2, 3, 4, 5, 6	1	1	2
XGB	Learning Rate	0.001, 0.01, 0.1	0.01	0.1	0.1
	Max. Depth	3, 5, 10, ∞	∞	10	10
	# Estimators	100, 500, 1000, 1500, 2000, 2500, 3000, 3500	2500	100	500
	Early Stopping	10	10	10	10
NN	Hidden Layer Size	10, 50, 100, 200, 300, 400, 500	500	100	100
	Activation	“tanh”, “relu”	“relu”	“relu”	“relu”
	Learning Rate	0.0001, 0.05	0.0001	0.0001	0.05
	Solver	“sgd”, “adam”	“adam”	“adam”	“adam”

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11573-023-01136-w.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The datasets analyzed for the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Albizri A, Appelbaum D, Rizzotto N (2019) Evaluation of financial statements fraud detection research: a multi-disciplinary analysis. Int J Discl Gov 16:206–241. https://doi.org/10.1057/s41310-019-00067-9

Ali A, Zhang W (2015) CEO tenure and earnings management. J Account Econ 59:60–79. https://doi.org/10.1016/j.jacceco.2014.11.004

Apley D, Zhu J (2020) Visualizing the effects of predictor variables in black box supervised learning models. J R Stat Soc 82:1059–1086. https://doi.org/10.1111/rssb.12377

Association of Certified Fraud Examiners (2020) Report to the nations: 2020 global study on occupational fraud and abuse. https://acfepublic.s3-us-west-2.amazonaws.com/2020-Report-to-the-Nations.pdf. Accessed 5 Apr 2022

Atanasov V, Ivanov V, Litvak K (2012) Does reputation limit opportunistic behavior in the VC industry? Evidence from litigation against VCs. J Financ 67:2215–2246

Bao Y, Ke B, Li B, Yu J, Zhang J (2020) Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. J Account Res 58:199–235. https://doi.org/10.1111/1475-679X.12292

Barker VL, Mueller GC (2002) CEO characteristics and firm R&D spending. Manage Sci 48:782–801. https://doi.org/10.1287/mnsc.48.6.782.187

Beasley MS (1996) An empirical analysis of the relation between the board of director composition and financial statement fraud. Account Rev 71:443–465

Beneish MD (1997) Detecting GAAP violation: implications for assessing earnings management among firms with extreme financial performance. J Account Public Policy 16:271–309. https://doi.org/10.1016/S0278-4254(97)00023-9

Beneish MD (1999) The detection of earnings manipulation. Financ Anal J 55:24–36. https://doi.org/10.2469/faj.v55.n5.2296

Bertomeu J, Cheynel E, Floyd E, Pan W (2021) Using machine learning to detect misstatements. Rev Acc Stud 26:468–519. https://doi.org/10.1007/s11142-020-09563-8

Bhandari A, Mammadov B, Shelton A, Thevenot M (2018) It is not only what you know, it is also who you know: CEO network connections and financial reporting quality. Auditing 37:27–50. https://doi.org/10.2308/ajpt-51821

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn 30:1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Brazel JF, Jones KL, Zimbelman MF (2009) Using nonfinancial measures to assess fraud risk. J Account Res 47:1135–1166
Breiman L (2001) Random forests. Mach Learn 45:5–32
Breiman L, Friedman J, Stone CJ, Olshen RA (2017) Classification and regression trees. Routledge
Brouthers KD, Brouthers LE, Werner S (2000) Influences on strategic decision-making in the Dutch financial services industry. J Manag 26:863–883
Brown NC, Crowley RM, Elliott WB (2020) What are you saying? Using topic to detect financial misreporting. J Account Res 58:237–291. https://doi.org/10.1111/1475-679X.12294
Campbell DW, Shang R (2022) Tone at the bottom: Measuring corporate misconduct risk from the text of employee reviews. Manage Sci 68:7034–7053. https://doi.org/10.1287/mnsc.2021.4211
Cecchini M, Aytug H, Koehler GJ, Pathak P (2010) Detecting management fraud in public companies. Manage Sci 56:1146–1160. https://doi.org/10.1287/mnsc.1100.1174
Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. https://doi.org/10.1145/2939672.2939785
Cheynel E, Levine CB (2020) Public disclosures and information asymmetry: a theory of the mosaic. Account Rev 95:79–99. https://doi.org/10.2308/accr-52447
Child J (1974) Managerial and organizational factors associated with company performance part I. J Manage Stud 11:175–189. https://doi.org/10.1111/j.1467-6486.1974.tb00693.x
Cialdini RB, Kallgren CA, Reno RR (1991) A focus theory of normative conduct: a theoretical refinement and reevaluation of the role of norms in human behavior. Adv Exp Soc Psychol 24:201–234. https://doi.org/10.1016/S0065-2601(08)60330-5
Climent F, Momparler A, Carmona P (2019) Anticipating bank distress in the eurozone: an extreme gradient boosting approach. J Bus Res 101:885–896. https://doi.org/10.1016/j.jbusres.2018.11.015
Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297
Craja P, Kim A, Lessmann S (2020) Deep learning for detecting financial statement fraud. Decision Support Syst 139:113421. https://doi.org/10.1016/j.dss.2020.113421
Cressey DR (1950) The criminal violation of financial trust. Am Sociol Rev 15:738–743
Daboub AJ, Rasheed AA, Priem RL, Gray DA (1995) Top management team characteristics and corporate illegal activity. Acad Manag Rev 20:138–170
Davidson WN, Xie B, Xu W, Ning Y (2007) The influence of executive age, career horizon and incentives on pre-turnover earnings management. J Manage Governance 11:45–60. https://doi.org/10.1007/s10997-007-9015-8
Dawson LM (1995) Women and men, morality and ethics. Bus Horiz 38:61–68
Dechow PM, Sloan RG (1991) Executive incentives and the horizon problem: an empirical investigation. J Account Econ 14:51–89
Dechow PM, Sloan RG, Sweeney AP (1996) Causes and consequences of earnings manipulation: an analysis of firms subject to enforcement actions by the SEC. Contemp Account Res 13:1–36
Dechow PM, Ge W, Larson CR, Sloan RG (2011) Predicting material accounting misstatements. Contemp Account Res 28:17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x
Dechow PM, Hutton AP, Kim JH, Sloan RG (2012) Detecting earnings management: a new approach. J Account Res 50:275–334. https://doi.org/10.1111/j.1475-679X.2012.00449.x
Dichev ID, Graham JR, Harvey CR, Rajgopal S (2013) Earnings quality: evidence from the field. J Account Econ 56:1–33. https://doi.org/10.1016/j.jacceco.2013.05.004
Doornenbal BM, Spisak BR, van der Laken PA (2021) Opening the black box: uncovering the leader trait paradigm through machine learning. Leadersh Q. https://doi.org/10.1016/j.leaqua.2021.101515
Dorminey J, Fleming AS, Kranacher M-J, Riley RA (2012) The evolution of fraud theory. Issues Account Educ 27:555–579. https://doi.org/10.2308/iace-50131
Fama E (1980) Agency problems and the theory of the firm. J Polit Econ 88:288–307. https://doi.org/10.1017/CBO9780511817410.022
Fanning KM, Cogger KO (1998) Neural network detection of management fraud using published financial data. Int J Intell Syst Account Finance Manag 7:21–41
Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27:861–874. https://doi.org/10.1016/j.patrec.2005.10.010
Feng M, Ge W, Luo S, Shevlin T (2011) Why do CFOs become involved in material accounting manipulations? J Account Econ 51:21–36. https://doi.org/10.1016/j.jacceco.2010.09.005
Feroz EH, Park K, Pastena VS (1991) The financial and market effects of the SEC's accounting and auditing enforcement releases. J Account Res 29:107–142

Fiske ST, Taylor SE (1991) Social cognition, 2nd edn. McGraw-Hill, New York
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232
Friedman HL (2014) Implications of power: when the CEO can pressure the CFO to bias reports. J Account Econ 58:117–141. https://doi.org/10.1016/j.jacceco.2014.06.004
Freeman RE, Gilbert D (1988) Corporate strategy and the search for ethics. Prentice Hall, Englewood Cliffs, NJ
Gibbons R, Murphy KJ (1992) Optimal incentive contracts in the presence of career concerns: theory and evidence. J Polit Econ 100:468–505
Gow ID, Larcker DF, Reiss PC (2016) Causal inference in accounting research. J Account Res 54:477–523. https://doi.org/10.1111/1475-679X.12116
Green BP, Choi JH (1997) Assessing the risk of management fraud through neural network technology. Auditing 16:14–28
Gupta VK, Mortal S, Chakrabarty B, Guo X, Turban DB (2020) CFO gender and financial statement irregularities. Acad Manag J 63:802–831. https://doi.org/10.5465/amj.2017.0713
Hambrick DC (2007) Upper echelons theory: an update. Acad Manag Rev 32:334–343. https://doi.org/10.5465/amr.2007.24345254
Hambrick DC, Mason PA (1984) Upper echelons: the organization as a reflection of its top managers. Acad Manag Rev 9:193–206. https://doi.org/10.5465/amr.1984.4277628
Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer
Ho SSM, Li AY, Tam K, Zhang F (2015) CEO gender, ethical leadership, and accounting conservatism. J Bus Ethics 127:351–370. https://doi.org/10.1007/s10551-013-2044-0
Hobson JL, Mayew WJ, Venkatachalam M (2012) Analyzing speech to detect financial misreporting. J Account Res 50:349–392. https://doi.org/10.1111/j.1475-679X.2011.00433.x
Huang H-W, Rose-Green E, Lee C-C (2012) CEO age and financial reporting quality. Account Horiz 26:725–740. https://doi.org/10.2308/acch-50268
Hunt SD, Chonko LB (1984) Marketing and machiavellianism. J Mark 48:30–42
James G, Witten D, Hastie T, Tibshirani R (2021) An introduction to statistical learning: with applications in R, 2nd edn. Springer, NY
Jensen MC (1993) Modern industrial revolution, exit, and the failure of internal control systems. J Financ 48:831–880
Johnson SA, Ryan HE, Tian YS (2009) Managerial incentives and corporate fraud: the sources of incentives matter. Rev Finance 13:115–145. https://doi.org/10.1093/rof/rfn014
Karpoff J (2011) Does reputation work to discipline corporate misconduct? In: The Oxford University Handbook. Oxford University Press, Oxford, U.K.
Karpoff JM, Scott Lee D, Martin GS (2008) The consequences to managers for financial misrepresentation. J Financ Econ 88:193–215. https://doi.org/10.1016/j.jfineco.2007.06.003
Karpoff JM, Koester A, Lee DS, Martin GS (2017) Proxies and databases in financial misconduct research. Account Rev 92:129–163. https://doi.org/10.2308/accr-51766
Kelley SW, Ferrell OC, Skinner SJ (1990) Ethical behavior among marketing researchers: an assessment of selected demographic characteristics. J Bus Ethics 9:681–688. https://doi.org/10.1007/BF00383395
Kim YJ, Baik B, Cho S (2016) Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. Expert Syst Appl 62:32–43. https://doi.org/10.1016/j.eswa.2016.06.016
Klaise J, van Looveren A, Vacanti G, Coca A (2021) Alibi explain: algorithms for explaining machine learning models. J Mach Learn Res 22:1–7
Koch-Bayram IF, Wernicke G (2018) Drilled to obey? Ex-military CEOs and financial misconduct. Strateg Manag J 39:2943–2964. https://doi.org/10.1002/smj.2946
Larcker DF, Zakolyukina AA (2012) Detecting deceptive discussions in conference calls. J Account Res 50:495–540. https://doi.org/10.1111/j.1475-679X.2012.00450.x
Lemaître G, Nogueira F, Aridas CK (2017) Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 18:1–5
Lewis CM (2013) "Keynote address". The 26th XBRL International Conference
Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
Markóczy L (1997) Measuring beliefs: accept no substitutes. Acad Manag J 40:1228–1242

Mason ES, Mudrack PE (1996) Gender and ethical orientation: a test of gender and occupational socialization theories. J Bus Ethics 15:599–604. https://doi.org/10.1007/BF00411793
Molnar C (2022) Interpretable machine learning: a guide for making black box models explainable, 2nd edn. https://christophm.github.io/interpretable-ml-book/
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirlon B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
Perols JL (2011) Financial statement fraud detection: an analysis of statistical and machine learning algorithms. Auditing 30:19–50. https://doi.org/10.2308/ajpt-50009
Perols JL, Bowen RM, Zimmermann C, Samba B (2017) Finding needles in a haystack: using data analytics to improve fraud prediction. Account Rev 92:221–245. https://doi.org/10.2308/accr-51562
Price M, Norris DM (2009) White-collar crime: corporate and securities and commodities fraud. J Am Acad Psychiatry Law 37:538–544
Purda L, Skillicorn D (2015) Accounting variables, deception, and a bag of words: assessing the tools of fraud detection. Contemp Account Res 32:1193–1223. https://doi.org/10.1111/1911-3846.12089
Ravisankar P, Ravi V, Rao GR, Bose I (2011) Detection of financial statement fraud and feature selection using data mining techniques. Decis Support Syst 50:491–500
Rest JR, Thoma SJ (1985) Relation of moral judgment development to formal education. Dev Psychol 21:709–714
Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. Psychol Rev 65:386–408
Schnatterly K, Gangloff KA, Tuschke A (2018) CEO wrongdoing: a review of pressure, opportunity, and rationalization. J Manag 44:2405–2432. https://doi.org/10.1177/0149206318771177
Schrand CM, Zechman SL (2012) Executive overconfidence and the slippery slope to financial misreporting. J Account Econ 53:311–329. https://doi.org/10.1016/j.jacceco.2011.09.001
Serwinek PJ (1992) Demographic & related differences in ethical views among small businesses. J Bus Ethics 11:555–566. https://doi.org/10.1007/BF00881448
Shapley LS (1953) A value for n-person games. Contrib Theory of Games 2:307–317
Shin K-S, Lee TS, Kim H (2005) An application of support vector machines in bankruptcy prediction model. Expert Syst Appl 28:127–135. https://doi.org/10.1016/j.eswa.2004.08.009
Shmueli G (2010) To explain or to predict? Stat Sci. https://doi.org/10.1214/10-STS330
Sigrist F, Hirnschall C (2019) Grabit: Gradient tree-boosted Tobit models for default prediction. J Bank Finance 102:177–192. https://doi.org/10.1016/j.jbankfin.2019.03.004
Smith CW, Watts RL (1982) Incentive and tax effects of executive compensation plans. Aust J Manag 7:139–157. https://doi.org/10.1177/031289628200700204
Strobl C, Boulesteix A-L, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform 8:25. https://doi.org/10.1186/1471-2105-8-25
Štrumbelj E, Kononenko I (2014) Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 41:647–665. https://doi.org/10.1007/s10115-013-0679-x
Trompeter GM, Carpenter TD, Desai N, Jones KL, Riley RA (2013) A synthesis of fraud-related research. Auditing 32:287–321. https://doi.org/10.2308/ajpt-50360
Troy C, Smith KG, Domino MA (2011) CEO demographics and accounting fraud: who is more likely to rationalize illegal acts? Strateg Organ 9:259–282. https://doi.org/10.1177/1476127011421534
van Hulse J, Khoshgoftaar TM, Napolitano A (2007) Experimental perspectives on learning from imbalanced data. Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007. 935–942. https://doi.org/10.1145/1273496.1273614
Velte P (2021) The link between corporate governance and corporate financial misconduct. A review of archival studies and implications for future research. Manag Rev Q. https://doi.org/10.1007/s11301-021-00244-7
Wahid AS (2019) The effects and the mechanisms of board gender diversity: evidence from financial manipulation. J Bus Ethics 159:705–725. https://doi.org/10.1007/s10551-018-3785-6
Wang TY, Winton A, Yu X (2010) Corporate fraud and business conditions: evidence from IPOs. J Financ 65:2255–2292. https://doi.org/10.1111/j.1540-6261.2010.01615.x
Wang R, Lee C-J, Hsu S-C, Lee C-Y (2018) Corporate misconduct prediction with support vector machine in the construction industry. J Manag Eng 34:04018021

Wang R, Asghari V, Hsu S-C, Lee C-J, Chen J-H (2020) Detecting corporate misconduct through random forest in China's construction industry. J Clean Prod 268:122266. https://doi.org/10.1016/j.jclepro.2020.122266

Weeks WA, Moore, Carlos, W., McKinney JA, Longenecker JG (1999) The effects of gender and career stage on ethical judgment. J Business Ethics 20:301–313

Whiting DG, Hansen JV, McDonald JB, Albrecht C, Albrecht WS (2012) Machine learning methods for detecting patterns of management fraud. Comput Intell 28:505–527. https://doi.org/10.1111/j.1467-8640.2012.00425.x

Wiersema MF, Bantel KA (1992) Top management team demography and corporate strategic change. Acad Manag J 35:91–121

Yang D, Jiao H, Buckland R (2017) The determinants of financial fraud in Chinese firms: does corporate governance as an institutional innovation matter? Technol Forecast Soc Chang 125:309–320. https://doi.org/10.1016/j.techfore.2017.06.035

Zahra SA, Priem RL, Rasheed AA (2005) The antecedents and consequences of top management fraud. J Manag 31:803–828. https://doi.org/10.1177/0149206305279598

Zhang X, Bartol KM, Smith KG, Pfarrer MD, Khanin DM (2008) Ceos on the edge: earnings manipulation and stock-based incentive misalignment. Acad Manag J 51:241–258. https://doi.org/10.5465/amj.2008.31767230

Zhao Q, Hastie T (2021) Causal interpretations of black-box models. J Business & Econ Stat 39:272–281. https://doi.org/10.1080/07350015.2019.1624293

Authors and Affiliations

Moritz Schneider[1] · **Rolf Brühl**[1]

Rolf Brühl
rbruehl@escp.eu

[1] ESCP Business School, Heubnerweg 8-10, 14059 Berlin, Germany