A Semantic Approach in Detecting Accounting Fraud: Document Embeddings

Ying Zhang Zhaoque (Chosen) Zhou

Current Version: September 2025

Ying (Zoey) Zhang (zoeyying.zhang@umanitoba.ca) is with Asper School of Business, University of Manitoba, and Zhaoque (Chosen) Zhou (zzhou22@albany.edu) is with Massry School of Business at University at Albany, SUNY.

A Semantic Approach in Detecting Accounting Fraud: Document Embeddings

Abstract

Accounting fraud poses ongoing challenges for regulators and market participants, yet detection remains imperfect and the true incidence of misconduct is likely higher than official statistics suggest. This study introduces the use of document embeddings in fraud detection. Document embeddings are unsupervised, computationally efficient, and static vectors that capture the semantic content of entire 10-K filings. They do not require fine-tuning and labeling, which makes them replicable, scalable, and free from the forward-looking problem.

We integrate these embeddings with XGBoost, a classifier well-suited to imbalanced data, and combine them with financial and linguistic variables to predict fraudulent firm-years, which represent less than one percent of observations. Our results show that embeddings substantially improve predictive performance. Incorporating embeddings raises PR-AUC from 0.84% (financial variables only) to 1.87%, and with contrastive learning to 4.48%. SHAP analysis indicates that embedding dimensions dominate as key predictors and contribute more strongly than grouped financial variables. This is the first accounting study to apply document embeddings to disclosures, demonstrating their value for fraud detection and opening the door to broader applications in textual research.

Keywords: Document Emebddings, XGBoost, Misstatement, Fraud

Acknowledgements

The authors gratefully acknowledge the financial support provided by Asper School of Business (University of Manitoba), Olin Business School (Washington University in Saint Louis), and Massry School of Business (University at Albany, SUNY). We also thank Lorien Stice-Lawrence for valuable comments and suggestions. The authors declare that they have no financial or personal relationships that could be viewed as potential conflicts of interest related to this work.

1 Introduction

Accounting fraud remains a persistent concern for regulators, practitioners, and academics, as undetected cases undermine market integrity and investor trust. To improve detection methods, we employ document embeddings, unsupervised and efficient vectors that encode the semantic content of entire 10-K filings, and combine them with XGBoost to enhance fraud-detection[1] performance in imbalanced data.

The importance of improving detection has been stressed repeatedly by regulators. As Mary Jo White, former SEC Chair, noted, "Investors depend on comprehensive and accurate financial reporting, and so our fundamental objective is to raise the bar of compliance by issuers and their auditors, and we will use all our tools to do so." U.S. Securities and Exchange Commission (2015).[2] Surveys further suggest that the ture incidence of fraud is likely higher than official statistic indicate. EY's 14th Global Fraud Survey[3] found that many finance professionals would justify unethical conduct to meet targets, and the SEC itself acknowledges that existing tools remain imperfect, particularly for subtle or persistent manipulations. U.S. Securities and Exchange Commission (2022).[4]

Academic research has long examined fraud detection through accounting ratios and financial statement items, finding that earnings, liabilities, and related figures are especially prone to manipulationBeneish (1999); Summers and Sweeney (1998); Green and Choi (1998); Dechow et al. (2011); Cecchini et al. (2010). Other work highlights the role of governance and oversight. For example, regulatory distance reduces scrutiny Kedia and Rajgopal (2011), stronger whistleblower incentives increase detection likelihood Dyck et al. (2010), and compensation

[1] We use the terms *detection* and *prediction* interchangeably in this study, as detection performance is evaluated through out-of-sample predictions.

[2] https://www.sec.gov/newsroom/speeches-statements/keynote-2015-aicpa-white

[3] EY (2016). *EY's 14th Global Fraud Survey: Corporate misconduct – individual consequences*. Available at: https://eyfinancialservicesthoughtgallery.ie/wp-content/uploads/2016/08/EY_Global_Fraud_Survey.pdf

[4] https://www.sec.gov/newsroom/speeches-statements/munter-statement-fraud-detection-101122

structures can encourage earnings manipulationDechow et al. (1996); Beasley (1996); Farber (2005); Dechow et al. (2010); Summers and Sweeney (1998); Beneish (1997, 1999). Even seamingly minor misstatemnts matter. Cheynel et al. (2024) document that managers adjust earnings just below materiality thresholds to evade detection.

In addition to financial variables, textual analysis of disclosures has also been used in fraud research. Early studies applied bag-of-words (BOW) and topic modelingLi (2008); Loughran and McDonald (2011); Brown et al. (2020); Lo et al. (2017); Hoberg and Lewis (2017), which capture tone, themes, and complexity but reduce text to surface-level patterns. BOW ignores syntax and semantics, and topic models rely on co-occurrence statistics; such methods cannot distinguish semantic equivalence (”queen” vs ”female king”) or contextual meaning (”bank” as a financial institution vs riverbank).

Recent developments in natural language processing provide alternative tools. BERT-based models capture local semantics but require heavy fine-tuning, struggle with long filings such as 10-Ks, and face the ”looking-forward problem”, since fine-tuning relies on post hoc outcomes. Document embeddings address these issues by generating fixed-length dense vectors that represent entire filings in a continuous semantic space Mikolov et al. (2013); Le and Mikolov (2014). Generated by an unsupervised algorithm, embeddings capture semantics without forward-looking labels or intensive fine-tuning, while remaining efficient, replicable, and scalable, making them well suited for fraud detection research.

We apply this approach by preprocessing 10-K filings, generating document embeddings with a pre-trained sentence-transformer, and running XGBoost for models that combine embeddings with financial and linguistic variables. Prior work has relied on machine learning approaches such as Support Vector Machines (SVM) Cecchini et al. (2010), neural network classifier Green and Choi (1998), ensemble methods such as RUSBoost Bao et al. (2020), random forest Dong et al. (2018); Purda and Skillicorn (2015), and topic modelings Brown et al. (2020); Hoberg and Lewis (2017). A persistent challenge across these studeis is class imbalance, as fraudulent firm-years

account for less than one percent of observationsWalauskis and Khoshgoftaar (2025); Bao et al. (2022). XGBoost is particularly effective in this setting, addressing imbalance without heavy under-sampling and using regularization to reduce overfitting.

Our findings show that embedding-based features add explanatory and predictive value in addition to financial and linguistic variables. [5] Approximately half of the embedding dimensions differ significantly between fraudulent and nonfraudulent firms, capturing distinctive semantic patterns linked to misreporting. Incorporating embeddings into XGBoost increases PR-AUC from 0.84% (financial variables only) 1.87%, and adding contrastive learning further raises performance to 4.48%. SHAP analysis confirms that document embeddings dominate as key predictors. We also compare embedding importance with grouped financial variables. The results show that embeddings contribute more strongly than any single financial category.

This study contributes to the literature in several ways. First, it introduces document embeddings as a novel, static, and efficient representation of disclosure content in accounting research. Because document embeddings are unsupervised, generated independently of observed outcomes, they are free from both intensive fine-tuning and the forward-looking problem. The static vectors ensure scalability and replicability of results. Second, it shows that embeddings, when integrated with XGBoost, substantially improve fraud prediction in highly imbalanced datasets, a long-standing challenge in the literature. Third, it provides evidence that textual semantics contain predictive signals of misreporting outside traditional financial features, with embeddings dominating dictionary-based textual features. More broadly, this study is the first in accounting research to apply document embeddings to the analysis of corporate disclosures, opening the door to further applications of this method across a wide range of disclosure and reporting questions.

[5] We use the terms *variables* and *features* interchangeably, and *linguistic* and *dictionary-based* interchangeably throughout this study.

2 Literature Review

2.1 Accounting Fraud, Causes, and Impacts

Misstatements in financial reporting can arise from both unintentional causes, such as clerical errors or changes in accounting principles, and intentional manipulations, which constitute accounting fraud. The term "AAER" (Accounting and Auditing Enforcement Release) specifically refers to accounting fraud cases investigated by the SEC. These cases involve intentional misstatements or fraudulent activities intended to deceive investors or misrepresent financial performance. SEC Rule 10b-5 prohibits fraudulent conduct related to the purchase or sale of securities and serves as the primary legal foundation for SEC enforcement against securities fraud. A strong inference of intent to deceive, often evidenced through evidence of personal gain or insider trading, distinguishes fraudulent misstatements from unintentional errors Schrand and Zechman (2012). Prior literature has documented significant associations between accounting irregularities and management incentives, such as equity incentives Efendi et al. (2007); Erickson et al. (2006).

The Enron scandal brought widespread public attention to the severe economic and ethical consequences of accounting fraud, affecting both the fraudulent firms themselves and society at large. Kedia and Philippon (2007) find that firms tend to hire excessively and invest disproportionately to mimic truly productive peers and that managers exercise stock options during periods of misreporting. Once the misreporting is uncovered, these firms often respond by laying off employees and reducing overinvestment, which paradoxically results in improvements in measured productivity. Firms caught engaging in fraudulent behavior also suffer from credit rating downgrades, increased cost of capital, loss of investor confidence, diminished investment efficiency, weakened operating performance, and substantial legal and reputational penalties Karpoff et al. (2008); Dechow et al. (1996); Burns and Kedia (2006); Hribar and Jenkins (2004).

Prior literature has extensively documented the significant roles that corporate governance and inventive compensation play in promoting earnings manipulation Dechow et al. (1996); Beasley (1996); Farber (2005); Dechow et al. (2010); Summers and Sweeney (1998); Beneish (1997, 1999); Efendi et al. (2007); Erickson et al. (2006). Agrawal and Cooper (2015) find a strong association between insider trading and misstatements, demonstrating that top executives of firms issuing restatements tend to sell substantially more stock during the misstated period than beforehand. Beyond management incentives, board characteristics also affect the likelihood of misstatements. For example, a higher proportion of independent directors reduces fraud risk Beasley (1996), while factors such as management-dominated boards, CEOs who serve simultaneously as board chairpersons, CEOs who are firm founders, and the absence of effective audit committees or outside blockholders increase the risk of fraudulent reporting Dechow et al. (1996). These findings suggest that both financial incentives and non-financial governance factors play important roles in the likelihood of accounting fraud.

2.2 Challenges in Detecting Accounting Fraud

Many fraud cases go undetected due to factors such as management's strategic manipulation, limited resources and staffing within the SEC task force, and varying levels of whistleblower incentives. U.S. Securities and Exchange Commission, Office of Inspector General (2023) highlights that leadership must curtail investigative work because of unfilled positions, limiting the agency's capacity to examine and investigage cases as thoroughly as a fully staffed workforce would allow. EY Global Integrity Report (2024) reveals that 20% of employees report experiencing or witnessing retaliation against whistleblowers, while 37% express reluctance to report issues due to concerns about future career progression. Cheynel et al. (2024) find that managers strategically adjust fraudulent activities to remain just beneath materiality thresholds to avoid detection, and that managers who evade detection repeatedly tend to gradually increase fraudulent amount over time. Kedia and Rajgopal (2011) find that firms located closer to SEC headquarters and in regions

with greater past SEC enforcement activity are less likely to restate their financial statements. Dyck et al. (2010) investigate the incentives motivating various types of whistleblowers, noting that employees and short sellers are often driven by monetary rewards, journalists and financial analysts by career and reputational concerns, while auditors face weaker incentives to expose fraud due to potential conflicts of interest. Call et al. (2018) further demonstrate that whistleblower involvement significantly improves the outcomes of enforcement actions, underscoring whistleblowers as valuable sources of information for regulators in fraud investigations and prosecutions.

2.3 Accounting Fraud Detection Models

Traditional accounting fraud detection models primarily rely on financial statement items, as accounting fraud is often evidenced through abnormal financial figures, such as overstated earnings or misstated liabilities Green and Choi (1998); Summers and Sweeney (1998); Beneish (1999). Dechow et al. (2011) propose a comprehensive fraud prediction model that incorporates a wide range of financial statement variables, including accruals quality measures, performance metrics, nonfinancial indicators, off-balance-sheet items, and market-related incentives, to predict misstatements. The authors document the effectiveness of financial-based ratios in predicting the existence of misstatements.

Beyond the traditional approach, machine learning techniques have been found effective in improving fraud detection. Cecchini et al. (2010) apply Support Vector Machines (SVM) to pick financial ratios in misstatement prediction model and document superior predictive performance compared to human-selected financial ratios. In addition to financial ratios, studies also highlight the informative value of linguistic features in textual disclosures for detecting accounting fraud. These features capture language characteristics of textual disclosures, including emotion words, management's self-references, tentative versus certain words, general knowledge phrases, linguistic complexity, tone, and activation language Larcker and Zakolyukina (2012); Lo et al. (2017); Purda and Skillicorn (2015); Humpherys et al. (2011). However, traditional

Bag-of-words (BOW) approaches have limitations. Dictionaries used in BOW approaches may not fully align with business contexts, and they do not account for the semantic meaning or contextual implications of words. Moreover, individual words or language tones alone may be difficult to interpret. For example, Purda and Skillicorn (2015) find that reports written during fraudulent periods tend to use more positive words to overstate optimistic expectations, while reports after fraud periods contain more negative words. To gain deeper insights into disclosure content, Hoberg and Lewis (2017) use LDA to show that fraudulent firms systematically produce abnormal disclosure, and that this signal predicts fraud out-of-sample. Brown et al. (2020) use Bayesian topic modeling to analyze the thematic content of financial statement disclosures and find that certain topics are associated with a higher likelihood of misreporting.

A growing number of machine learning techniques have been found effective in fraud detection, and they have been applied to both financial-based models and textual analysis. For example, Cecchini et al. (2010) employ SVM to develop fraud prediction model; Bao et al. (2020) use an ensemble learning approach, RUSBoost, based on raw financial statement data and find that it outperforms traditional ratio-based logistic regression models. Dong et al. (2018) highlight the informativeness of social media features in fraud predictions and show that random forest models outperform SVM in predicting fraud. In the textual domain, Purda and Skillicorn (2015) use SVM and random forests decision trees to construct a BOW dictionary and analyze intra-firm changes in language use over time, finding that data-driver language tool outperforms predefined word lists in classifying fraud events.

3 Methodology and Data

3.1 Document Embeddings

A document embedding is a numerical representation of text that summarizes its meaning in a compact form. In this study, we generate document embeddings for 10-K filings through a multi-

step process. We begin by preprocessing the filings, obtained from Bill McDonald's repository,[6] to ensure consistency across firms and years. This involves removing non-informative content such as headers, tables, and boilerplate disclaimers, converting text to lowercase, and standardizing characters. Because 10-Ks are lengthy, we divide them into smaller units, such as sentences or paragraphs, so that the model can process the content efficiently while retaining the context of each section.

Each segment is then encoded into a 384-dimensional vector using a pre-trained language model known as all-MiniLM-L6-v2. This model, developed by Microsoft Research and widely adopted in the research community, is designed to capture the meaning of sentences in a way that balances accuracy and computational efficiency. Once segment-level vectors are generated, we combine them into a single document-level embedding using weighted averaging. This step ensures that both local meaning (e.g., how a sentence conveys risk or optimism) and overall thematic patterns (e.g., tone across the filing) are represented in a single vector. The end result is a fixed-length, 384-dimensional vector that serves as a "semantic fingerprint" of the entire filing. Importantly, this representation is consistent across documents of different lengths, making it suitable for large-scale empirical analysis.

This approach offers important advantages over traditional textual measures. Bag-of-words counts reduce documents to word frequencies and ignore context, while dictionary-based sentiment scores (such as the Loughran–McDonald word lists) treat all appearances of a word as equal regardless of meaning. Topic modeling can uncover broad themes but is unable to distinguish subtle shifts in how words are used. Document embeddings, by contrast, capture relationships between words and their context, allowing them to identify more nuanced patterns of disclosure. Compared with earlier embedding methods such as Doc2Vec, which often depend on training settings and random variation, modern pre-trained models provide stable and replicable representations suitable for financial text.

[6] https://sraf.nd.edu/sec-edgar-data/cleaned-10x-files/

It is also important to distinguish between document embedding methods and BERT-based approaches. While BERT and its variants are widely used in natural language processing research, applying them directly to 10-K filings is challenging because their limited input size requires filings to be heavily split, and task-specific fine-tuning is both resource-intensive and difficult to replicate. Document embeddings avoid these issues by producing a single, ready-to-use representation for each filing. They are generated through unsupervised algorithms, making them free from labeled outcomes and independent of forward-looking information. They are also static vectors, ensuring consistent representations across documents and time. These characteristics make document embeddings computationally efficient, scalable, and replicable, while still capturing the semantic content of disclosures in a form that is easy to incorporate into empirical models.

Prior work has shown that document embeddings improve performance in classification and clustering tasks compared with bag-of-words and topic models Le and Mikolov (2014); Lau and Baldwin (2016). In our setting, embeddings provide a scalable and context-aware way of representing 10-K filings, allowing us to capture subtle linguistic cues in disclosure narratives that help distinguish between fraudulent and nonfraudulent firms. This makes them a powerful complement to the financial and linguistic variables commonly used in accounting research.

3.2 Data Description

We follow Dechow et al. (2011) and use the Accounting and Auditing Enforcement Releases (AAERs) dataset to identify accounting misstatement cases[7][8]. The dataset has been updated through December 31, 2021 . The AAER dataset records 1,816 unique firm-level misstatement events issued between 1982 and 2021. These events include both quarterly and annual financial misstatements. If a quarterly misstatements is corrected in a later period and the corresponding 10-K for the same year shows no misstatement, the firm-year is not classified as fraudulent. To

[7] https://sites.google.com/usc.edu/aaerdataset/home/

[8] Dechow, P.M., Ge, W., Larson, C.R., and Sloan, R.G. (2011) Predicting Material Accounting Misstatements. Contemporary Accounting Research, 28: 17-82.

align with our model predictors, we focus exclusively on annual misstatements. Because a single misstatement event can affect multiple years, the AAER annual dataset contains 2, 196 firm-year observations, higher than 1,816 unique events.

To construct the financial statement-based factors, we follow Bao et al. (2020), combining raw financial data items from Dechow et al. (2011) and Cecchini et al. (2010), resulting in 28 financial statement variables for our accounting fraud detection model. We also incorporate Loughran-McDonald (LM) linguistic features to capture the predictive power of linguistic information in detecting accounting fraud. These features measure language complexity (Complexity), sentiment (Negative, Positive, Uncertainty, Litigious, Strong modal, Weak modal, Constraining), negation word counts (Negation), and file size. We further merge the sample with 384 document embedding vector factors, as described in Section 3.1, to form our final dataset. Because the LM 10X summaries dataset begins in 1994, and no misstatement events were identified in 2020 or 2021 in AAER datasets[9], our final sample spans 1994 to 2019, comprising 640 misstated firm-years and 84,499 firm-years without misstatement. The misstated firm-years account for 0.75% of the total observations.

Table 1 presents the descriptive statistics for the 28 financial predictors and linguistic features for the nonfraudulent and fraudulent firm-years. Among the 28 financial predictors, only 9 (32.14%) differ significantly between the two groups at the 1% levels. Specifically, fraudulent firm-years have higher current assets (act), accounts payable (ap), inventory (invt), current liabilities (lct), receivables (rect), proceeds from the sale of common and preferred stock (sstk), and stock price (prcc). Two income statement accounts, cost of goods sold (cogs) and sales (sale), are significantly different across the two groups at 5% level. Notably, except for stock price, these accounts are either core working capital accounts or primary drivers of earnings. They also constitute the key components of ratios measuring liquidity (current ratio), profitability, leverage, and market valuation, the financial metrics most closely monitored by investors, creditors, and

[9]Tracing prior filings can take several years; to mitigate undetected fraud cases due to investigation time lags, we exclude 2020 and 2021 from our sample.

regulators, making them prime manipulation targets for manipulation. Fraudulent firm-years also exhibit lower income (ib and ni), reflecting underperformance relative to the nonfraudulent firm-years. Regarding the LM linguistic factors, 4 out of the 11 factors (36.36%) are significant at the 1% level. Fraudulent firm-year 10-Ks are generally more complex to read (Complexity), contain more legal words, indicating heightened legal risk and regulatory scrutiny, and use more weak modal words, suggesting reduced confidence in future performance and subtle linguistic cues of managerial hedging.

To compare with financial and linguistic factors, we also conduct two-sample t-tests on each element of the 384×1 document embedding vector. Figure 1 presents the distribution of t-statistics for these features. Among the 384 tests, 187 elements (48.70%) exhibit significant differences between fraudulent and nonfraudulent firm-years at the 1% level ($|t\text{-}stat| > 2.576$). In other words, nearly half of the document embedding factors differ significantly across the two groups, a higher proportion than the 32.14% and 36.36% observed for financial and linguistic factors, respectively. These results indicate that document embedding factors capture distinct patterns between fraudulent and nonfraudulent firms, suggesting they may provide additional predictive power in fraud detection models.

To examine the predictive power of document embedding factors, we follow Dechow et al. (2011) and Bao et al. (2020) and use the 28 financial items as predictors first to examine a baseline logistic regression model. However, one persistent challenge in fraud detection is the rarity of successful fraud identifications, which lead to highly imbalanced datasets Walauskis and Khoshgoftaar (2025); Bao et al. (2022). In such datasets, models tend to be biased toward the majority class, in this case, nonfraudulent firms, resulting in high overall accuracy but poor detection of rare fraudulent events. Additionally, over-tuning machine learning models on limited samples can produce unstable performance across subsamples. To address these issues, we employ XGBoost to train the imbalanced datasets and predict accounting fraud.

3.3 Contrastive Learning and XGBoost

In the fraud detection setting, where only a small fraction of firms (0.75%) engage in financial misstatements, traditional supervised models often struggle in such highly imbalanced contexts, as the limited number of fraudulent cases provides insufficient signal and leads to overfitting. Motivated by the evidence in Table 1, which shows that 48.70% of the elements in the document embedding vectors differ significantly at the 1% level between fraudulent and nonfraudulent firms, we introduce contrastive learning to exploit these distinctions in the embedding space.

Contrastive learning provides a representation learning framework that differentiates between similar and dissimilar data pairs by constructing an embedding space in which "similar" observations are drawn closer together while "dissimilar" ones are pushed apart. The differences observed from document embedding vectors between nonfraudulent and fraudulent firms suggest that fraud leaves detectable linguistic or structural traces in financial disclosures, even when such traces are subtle or non-obvious to humans. Contrastive learning leverages these patterns by structuring the learning task around similarity rather than direct classification.

To operationalize this approach, we define similar pairs as two firms from the same class, either two nonfraudulent firms or two fraudulent firms, and dissimilar pairs as a combination of a fraudulent and a nonfraudulent firm. During training, the model minimizes the distance between embeddings of similar pairs while maximizing the distance between disimilar pairs, refining the embedding space so that firms with similar disclosure characteristics cluster together while fraudulent and non-fraudulent firms are clearly separated. This pairwise design also naturally generates a more balanced training signal. Each fraudulent firm can be paired with multiple non-fraudulent firms to create diverse dissimilar pairs, while the numerous non-fraudulent firms can form abundant similar pairs among themselves. This strategy expands the effective size of the training set while preserving balance between similar and dissimilar examples, which is essential for learning a discriminative embedding space.

Once trained, the model produces an embedding space where firms with similar disclosure characteristics—whether fraudulent or non-fraudulent—cluster together. This enriched and balanced representation enables downstream classifiers, such as XGBoost, to deliver stronger and more stable performance, even under severe class imbalance. In this way, contrastive learning not only improves fraud detection accuracy but also provides a scalable and robust framework for addressing imbalanced datasets frequently encountered in financial applications.

Extreme Gradient Boosting (XGBoost) is a scalable machine learning algorithm developed by Chen and Guestrin (2016) that extends the gradient boosting framework, which integrates multiple weak learners to form a strong predictive model. A central challenge in applying machine learning to imbalanced datasets is the risk of overfitting. Overfitting arises when a model is overly complex relative to the size or variability of the data, causing it to capture noise and idiosyncratic patterns in the training set. This greatly reduces the model's ability to generalize to testing set. The issue is especially pronounced in highly imbalanced settings; in our study, fraudulent firm-years constitute less than one percent of the sample. Under such conditions, a model may appear to achieve high overall accuracy simply by predicting nonfraudulent outcomes but fail to detect actual fraud. In accounting fraud detection, identifying true fraudulent cases is far more important than correctly classifying nonfraudulent firms.

Prior research demonstrates that gradient boosting methods achieve strong performance in detecting minority classes Chen and Guestrin (2016); Buda et al. (2018); Sun et al. (2009). Within this family of algorithms, XGBoost stands out as a highly optimized and robust implementation that incorporates several mechanisms to mitigate overfitting in imbalanced datasets. Specifically, it applies both L1 (Lasso) and L2 (Ridge) regularization to decision tree weights, thereby penalizing overly complex trees and reducing the likelihood of memorizing noise within the minority class. In addition, XGBoost addresses class imbalance by assigning higher weights to minority observations, ensuring that rare cases are given greater consideration during model training. The algorithm further enhances generalization through depth control and sequential boosting, which incrementally corrects errors from earlier iterations, while early stopping prevents excessive

training once additional trees fail to improve performance.

To evaluate prediction model performance, we rely on the Precision-Recall Area Under the Curve (PR-AUC) metric. Unlike traditional metrics such as overall accuracy or ROC-AUC, which can be misleading when the positive class is rare, PR-AUC specifically focuses on the model's ability to identify the minority class correctly. In particular, Precision measures the proportion of predicted fraud cases that are actually fraudulent, while Recall measures the proportion of actual fraud cases successfully detected by the model. PR-AUC summarizes this trade-off across all classification thresholds ranging between 0 and 1, providing a single, threshold-independent measure of predictive power. While ROC-AUC also considers performance across thresholds, it equally weights the nonfraudulent and fraudulent classes, making it less informative when positive cases are rare. In contrast, PR-AUC emphasizes performance on the fraud class, capturing how well a model balances detecting actual frauds against avoiding false alarms. This focus is particularly meaningful in accounting fraud detection, where identifying true fraudulent cases is far more important than correctly classifying nonfraudulent firms.

4 Empirical Results

4.1 Comparison of Various Models Performance

Figure 2 highlights three key findings. First, the fusion of XGBoost with contrastive learning delivers substantial performance gains over both the logistic and standalone XGBoost models across all variable combinations. The document-embedding-only specification achieves a PR-AUC of 4.42%, while the full specification incorporating all variables attains the highest PR-AUC of 4.48%. These improvements underscore the value of leveraging pairwise relationships in the embedding space, which enhance the model's ability to distinguish fraudulent from non-fraudulent firms. Notably, the PR-AUC of the fusion model is consistently and meaningfully above the 0.36% baseline, indicating a significant lift in predictive accuracy.

Second, the inclusion of document embeddings (Emb)—either alone or combined with financial metrics (Fin) and word frequency measures (Text)—produces notable improvements in detection performance. For the logistic and XGBoost models, the addition of embeddings raises PR-AUC by 0.02% and 0.92% on average, respectively, relative to models using only financial or textual variables. This pattern suggests that the semantic and structural signals embedded in the textual representations capture dimensions of fraud risk that are not fully reflected in traditional financial or surface-level textual measures.

Finally, the results highlight the complementarity of multiple variable sets. Specifications that integrate embeddings with financial and textual measures generally achieve higher PR-AUC than those relying on a single data source. This finding indicates that fraud detection benefits from a more comprehensive representation of firms, combining quantitative financial indicators with surface-level textual cues and deeper semantic information. Overall, these results demonstrate the effectiveness of advanced embedding techniques and contrastive learning in building robust, scalable models for fraud detection.

4.2 Which Features are Important?

To assess the marginal importance of individual predictors, we compute Shapley Additive exPlanations (SHAP) values for all out-of-sample observations. SHAP assigns each feature a value representing its contribution to the model's prediction for a given observation. The method builds on Shapley values from cooperative game theory, in which a feature's contribution is defined as its average marginal effect across all possible subsets of predictors.

This approach is particularly valuable in fraud detection, where model interpretability is as important as predictive accuracy. SHAP values can be either positive or negative, indicating the direction in which a feature shifts the prediction relative to the baseline level. A positive value implies that the feature increases the predicted probability of fraud for that observation, while a negative value reduces it. The magnitude of the SHAP value reflects the strength of the

feature's impact on the prediction for a given observation. Unlike traditional measures of feature importance, such as standardized regression coefficients, SHAP captures both the magnitude and direction of feature contributions at the observation level, making it particularly well-suited for fraud detection. Because our purpose is to evaluate each feature's contribution to the model rather than interpret the directions at this stage, we assess each feature's overall importance by averaging absolute SHAP values across all observations. This measure indicates how much, on average, each feature influences the model's predictions.

Formally, the prediction for each out-of-sample observation k can be expressed in log-odds form as

$$logit(p(k)) = \beta_0 + \sum_{i=1}^{M} \text{SHAP}_i^k, \tag{1}$$

where $p(k)$ is the predicted probability, β_0 is the baseline log-odds, and SHAP_i^k denotes the SHAP value of variable j for observation k, and M indicates the number of features included in the model. The overall importance (global importance) for feature i across all observations can be expressed as

$$G_i = \frac{1}{N} \sum_{k=1}^{N} \left| \text{SHAP}_i^{(k)} \right|, \tag{2}$$

where N is the number of observations in the testing sample.

Figure 3 presents the top 20 features ranked by global importance (mean absolute SHAP) across all three groups of features, financial metrics (Fin), textual features (Text), and document embedding factors (Emb). Among the top 20 features, 12 are document embedding factors, denoted as DE-index, while the remaining 8 are financial variables. None of the textual features appear in the top 20. Although financial and textual features are distinct by construction, document embeddings capture not only deeper semantic information but also surface-level textual patterns. The dominance of embeddings among the most important predictors, combined with the absence of dictionary-based textual features, provides strong evidence that document embeddings

substantially improve fraud detection by capturing richer and more informative representations of disclosure content.

Among the top-ranked predictors, eight are financial variables: ceq (common equity), ppegt (property, plant and equipment), rect (receivables), sstk (sale of common and preferred stock), re (retained earnings), prcc (closing stock price), act (current assets), and sale (sales). These variables are closely linked to common misstatement categories identified in the AAERs, including revenue recognition, expense classification, shareholder equity accounts, capitalization of costs, accounts receivable, inventory, cost of goods sold, reserve accounts, liabilities, marketable securities, allowances for bad debt, and payables. Some of these misstatement categories are not reported as discrete line items in financial statements but are connected to or reflected in the top-ranked predictors. For instance, ceq (common equity) is affected by both revenues and expenses through retained earnings, and it is itself a core component of the broader shareholder equity account. ppegt (property, plant, and equipment) increases when firms capitalize costs that might otherwise have been expensed, a common misstatement practice. sstk (sale of common and preferred stock) affects shareholder equity by raising common stock and paid-in capital; it may also reduce reliance on debt, and equity proceeds can be redirected to less transparent uses, such as financing capital expenditures or covering reserve accounts that mitigate litigation risks. Finally, prcc (closing stock price) is not a financial statement item per se, but it reflects market valuation pressures and is strongly associated with incentives to manipulate key performance-related accounts to influence equity values.

Taken together, the SHAP rankings provide evidence that document embeddings provide substantial incremental explanatory power by capturing disclosure-based signals overlooked by dictionary-based textual features, and that financial variables most closely tied to known misstatement categories remain central predictors, reflecting the enduring importance of accounting fundamentals.

4.3 Why Document Embeddings Matter?

Bao et al. (2020) group the 28 financial variables into ten categories defined by Dechow et al. (2011), who link each category to a distinct type of accounting fraud. For each category, we compute the SHAP value as the sum of the SHAP values of its constituent financial features. If all financial features within a category are informative, the summed mean absolute category SHAP value for that category should exceed the individual SHAP values reported in Figure 3; otherwise, the correlated features within a category would introduce noise, reducing their usefulness for detection. Figure 4 illustrates this pattern. Only the first category, financial features that affect net income but are not tied to specific income, expense, or equity accounts, shows a relatively high mean absolute SHAP value (0.011). In contrast, the nine remaining financial categories all have mean absolute values below 0.007, with an average of 0.0033, which is even lower than the average absolute SHAP value of document embedding elements included among the top 20 features (0.0040). The red bar represents the combined SHAP value of the twelve most important embedding dimensions presented in Figure 3, reaching 0.022, more than twice that of the strongest financial category (0.011). This comparison shows that embeddings capture substantially more informative and discriminative signals of misstatement than traditional financial variables. Taken together, the results suggest that financial variables on their own provide limited predictive power, whereas document embeddings supply distinctive, non-redundant information that delivers stronger and more reliable indicators of fraud.

We acknowledge that financial variables contain useful information. Yet giving equal weight to, or relying exclusively on, financial features within the same category often introduces redundancy and noise, limiting their value for fraud detection. Document embeddings, by contrast, capture the semantic content of the entire 10-K can reveal signals linked to specific accounting categories that financial variables alone overlook. Our results show that certain embedding dimensions play a decisive role in identifying fraud. To explore this further, we map these dimensions to the categories defined by Dechow et al. (2011) using canonical correlations.

For each important embedding dimension identified in Figure 3, we compute its canonical correlation with each financial account category. This procedure generates, for each category, an optimal linear combination of the SHAP values of its constituent financial variables. A high correlation between an embedding dimension and this category-level signal suggests that the embedding captures semantic information relevant to that account category. Figure 5 shows that several embedding dimensions exhibit substantial correlations with specific categories. For example, DE-380 and DE-236 display correlations above 0.45 with income/expense, revenue, and asset categories, while DE-112 is strongly associated with assets and receivables. These results indicate that certain embedding dimensions systematically align with the economic content of financial variables, suggesting that document embeddings not only capture textual signals at the document level but also encode information tied to specific financial account categories. This provides further evidence that embeddings provide complementary, category-relevant signals that enhance fraud detection in ways traditional financial variables cannot.

5 Conclusion

This study introduces document embeddings into accounting fraud detection and shows their effectiveness when combined with XGBoost. Document embeddings are unsupervised, computationally efficient, and static vectors that summarize the semantic content of entire 10-K filings. They do not require fine-tuning or a forward-looking problem. By capturing semantic patterns of the whole document, embeddings provide signals that traditional financial variables and dictionary-based textual measures fail to detect.

Our empirical analyses demonstrate the value of this approach. Nearly half of the embedding dimensions differ systematically between fraudulent and nonfraudulent firm-years. Incorporating embeddings into XGBoost increases PR-AUC from 0.84% with financial variables alone to 1.87%, and further to 4.48% with contrastive learning. SHAP analysis confirms that embeddings dominate the set of important predictors, while canonical correlation tests show that certain embedding

dimensions align closely with economic categories such as revenue, assets, and receivables. These results indicate that embeddings capture both document-level semantics and category-specific information relevant to fraud, decisively strengthening prediction models.

The study makes three key contributions. First, it establishes document embeddings as a novel and effective method for representing the semantics of disclosure content in accounting research. Second, it demonstrates that embeddings, when combined with XGBoost, substantially improve fraud prediction in the presence of extreme class imbalance, a long-standing challenge in this literature. Third, it provides evidence that embeddings generate unique and category-relevant signals of misstatement that outperform dictionary-based textual features and enhance traditional financial predictors.

More broadly, this is the first study in accounting to apply document embeddings to corporate disclosures. The evidence shows that embeddings offer a powerful new approach for uncovering signals of fraud, providing regulators and practitioners with a scalable and replicable detection tool and establishing document embeddings as an effective tool for disclosure research in accounting.

References

Agrawal, A., Cooper, T., 2015. Insider trading before accounting scandals. Journal of Corporate Finance 34, 169–190.

Bao, Y., Hilary, G., Ke, B., 2022. Artificial intelligence and fraud detection. In: Babich, V., Brige, J. R., Hilary, G. (eds.), *Modern Approaches to Financial Research*, Springer, pp. 223–247.

Bao, Y., Ke, B., Li, B., Yu, Y. J., Zhang, J., 2020. Detecting accounting fraud in publicly traded u.s. firms using a machine learning approach. Journal of Accounting Research 58, 199–235.

Beasley, M. S., 1996. An empirical analysis of the relation between the board of director composition and financial statement fraud. The Accounting Review 71, 443–465.

Beneish, M. D., 1997. Detecting gaap violation: implications for assessing earnings management among firms with extreme financial performance. Journal of Accounting and Public Policy 16, 271–309.

Beneish, M. D., 1999. Incentives and penalties related to earnings overstatements that violate gaap. The Accounting Review 74, 425–457.

Brown, N. C., Crowley, R. M., Elliott, W. B., 2020. What are you saying? using topic to detect financial misreporting. Journal of Accounting Research 58, 237–291.

Buda, M., Maki, A., Mazurowski, M. A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks 106, 249–259.

Burns, N., Kedia, S., 2006. The impact of performance-based compensation on misreporting. Journal of Financial Economics 79, 35–67.

Call, A. C., Martin, G. S., Sharp, N. Y., Wilde, J. H., 2018. Whistleblowers and outcomes of financial misrepresentation enforcement actions. Journal of Accounting Research 56, 123–171.

Cecchini, M., Aytug, H., Koehler, G. J., Pathak, P., 2010. Detecting management fraud in public companies. Management Science 56, 1146–1160.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, NY, USA, KDD '16, p. 785–794.

Cheynel, E., Cianciaruso, D., Zhou, F. S., 2024. Fraud power laws. Journal of Accounting Research 62, 833–876.

Dechow, P., Ge, W., Schrand, C., 2010. Understanding earnings quality: A review of the proxies, their determinants and their consequences. Journal of Accounting and Economics 50, 344–401.

Dechow, P. M., Ge, W., Larson, C. R., Sloan, R. G., 2011. Predicting material accounting misstatements. Contemporary Accounting Research 28, 17–82.

Dechow, P. M., Sloan, R. G., Sweeney, A. P., 1996. Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the sec. Contemporary Accounting Research 13, 1–36.

Dong, W., Liao, S., Zhang, Z., 2018. Leveraging financial social media data for corporate fraud detection. Journal of Management Information Systems 35, 461–487.

Dyck, A., Morse, A., Zingales, L., 2010. Who blows the whistle on corporate fraud? The Journal of Finance 65, 2213–2253.

Efendi, J., Srivastava, A., Swanson, E. P., 2007. Why do corporate managers misstate financial statements? the role of option compensation and other factors. Journal of Financial Economics 85, 667–708.

Erickson, M., Hanlon, M., Maydew, E. L., 2006. Is there a link between executive equity incentives and accounting fraud? Journal of Accounting Research 44, 113–143.

EY Global Integrity Report, 2024.

Farber, D. B., 2005. Restoring trust after fraud: Does corporate governance matter? The Accounting Review 80, 539–561.

Green, B. P., Choi, J. H., 1998. Assessing the risk of management fraud through neural network technology. Auditing: A Journal of Practice Theory 16, 14.

Hoberg, G., Lewis, C., 2017. Do fraudulent firms produce abnormal disclosure? Journal of Corporate Finance 43, 58–85.

Hribar, P., Jenkins, N. T., 2004. The effect of accounting restatements on earnings revisions and the estimated cost of capital. Review of Accounting Studies 9, 337–356.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., Felix, W. F., 2011. Identification of fraudulent financial statements using linguistic credibility analysis. Decision Support Systems 50, 585–594.

Karpoff, J. M., Lee, D. S., Martin, G. S., 2008. The cost to firms of cooking the books. Journal of Financial and Quantitative Analysis 43, 581–611.

Kedia, S., Philippon, T., 2007. The economics of fraudulent accounting. The Review of Financial Studies 22, 2169–2199.

Kedia, S., Rajgopal, S., 2011. Do the sec's enforcement preferences affect corporate misconduct? Journal of Accounting and Economics 51, 259–278.

Larcker, D. F., Zakolyukina, A. A., 2012. Detecting deceptive discussions in conference calls. Journal of Accounting Research 50, 495–540.

Lau, J. H., Baldwin, T., 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation.

Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents. In: Xing, E. P., Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, PMLR, Bejing, China, vol. 32 of *Proceedings of Machine Learning Research*, pp. 1188–1196.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics 45, 221–247.

Lo, K., Ramos, F., Rogo, R., 2017. Earnings management and annual report readability. Journal of Accounting and Economics 63, 1–25.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance 66, 35–65.

Mikolov, T., Yih, W.-t., Zweig, G., 2013. Linguistic regularities in continuous space word representations. In: *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751.

Purda, L., Skillicorn, D., 2015. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. Contemporary Accounting Research 32, 1193–1223.

Schrand, C. M., Zechman, S. L., 2012. Executive overconfidence and the slippery slope to financial misreporting. Journal of Accounting and Economics 53, 311–329.

Summers, S. L., Sweeney, J. T., 1998. Fraudulently misstated financial statements and insider trading: An empirical analysis. The Accounting Review 73, 131–146.

Sun, Y., Wong, A. K. C., Kamel, M. S., 2009. Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence 23, 687–719.

U.S. Securities and Exchange Commission, 2015. Keynote address at the 2015 aicpa national conference: ”maintaining high-quality, reliable financial reporting: A shared and weighty responsibility” - Chair Mary Jo White. Tech. rep.

U.S. Securities and Exchange Commission, 2022. The auditor's responsibility for fraud detection - Paul Munter. Tech. rep.

U.S. Securities and Exchange Commission, Office of Inspector General, 2023. Management and performance challenges facing the securities and exchange commission – fiscal year 2023.

Walauskis, M. A., Khoshgoftaar, T. M., 2025. Unsupervised label generation for severely imbalanced fraud data. Journal of Big Data 12.

Table 1 Descriptive Statistics

The table presents the statistics of financial ratios and word frequency measures for nonfraudulent and fraudulent firms' 10-K filings. A two-sample t-test is conducted for each measure, and the corresponding significance levels are reported. The sample period covers from 1994 to 2019. The symbols, ***, **, and * indicate significance at the 1%, 5% and 10% levels, respectively.

	Nonfraudulent(N=84499)		Fraudulent(N=640)				Nonfraudulent(N=84499)		Fraudulent(N=640)		
	Mean	Std	Mean	Std	Diff		Mean	Std	Mean	Std	Diff
Panel A: Financial Ratios											
act	708.3	3417.6	1087.9	2542.6	-379.6***	re	427.6	5987.2	133.8	4776.3	293.8
ap	160.6	978.8	318.9	1039.1	-158.3***	rect	230.4	1171.5	367.3	973.0	-136.9***
at	2310.7	11164.4	2841.8	7691.5	-531.2	txp	17.2	165.0	23.4	132.5	-6.2
ceq	831.9	4553.7	1036.2	3022.1	-204.4	cogs	1263.5	7640.3	2004.5	6112.0	-741.1**
che	237.2	1741.1	299.6	959.5	-62.4	dp	98.5	566.5	101.8	330.0	-3.3
dlc	84.0	653.2	93.0	332.0	-9.0	ib	105.2	939.3	-27.1	2072.6	132.3***
dltt	599.3	2887.5	641.1	1966.5	-41.7	ni	106.3	963.2	-117.4	4087.1	223.7***
invt	160.3	945.7	267.4	747.1	-107.0***	sale	1890.4	10283.3	2780.7	7418.7	-890.3**
ivao	83.7	1077.8	121.9	655.1	-38.2	txt	50.8	456.6	56.5	249.8	-5.8
ivst	82.3	1322.7	77.6	509.4	4.7	xint	37.4	158.2	48.7	151.7	-11.3*
lct	495.1	2578.2	790.2	2228.7	-295.1***	dltis	216.0	1280.2	254.2	1112.5	-38.2
lt	1443.6	6997.2	1742.6	4854.9	-299.0	sstk	26.6	179.5	54.5	206.8	-28.0***
ppegt	1417.1	9196.0	1024.8	3016.6	392.3	csho	101.3	658.2	130.0	368.9	-28.7
pstk	7.6	100.9	14.1	103.7	-6.5	$prcc_f$	18.2	39.8	27.0	38.4	-8.8***
Panel B: Word Frequency Measures											
Complex	0.40	0.17	0.43	0.18	-0.038***	Strong	0.04	0.03	0.04	0.03	0.002*
Log(Size)	12.64	0.60	12.65	0.60	-0.008	Weak	0.07	0.05	0.06	0.04	0.006***
Negative	0.23	0.14	0.23	0.15	-0.003	Constrain	0.10	0.08	0.11	0.08	-0.001
Positive	0.07	0.04	0.07	0.03	-0.001	Complexity	0.06	0.04	0.06	0.04	-0.006***
Uncertainty	0.16	0.08	0.16	0.07	0.004	Negation	0.01	0.01	0.01	0.01	-0.001*
Litigious	0.25	0.27	0.28	0.30	-0.034***						

Figure 1. Histograms of t-Statistics for Document Embedding Vector Elements

The figure displays the histograms of t-statistics from the two-sample t-tests for each element of the document embedding vectors. The sample period covers from 1994 to 2019.

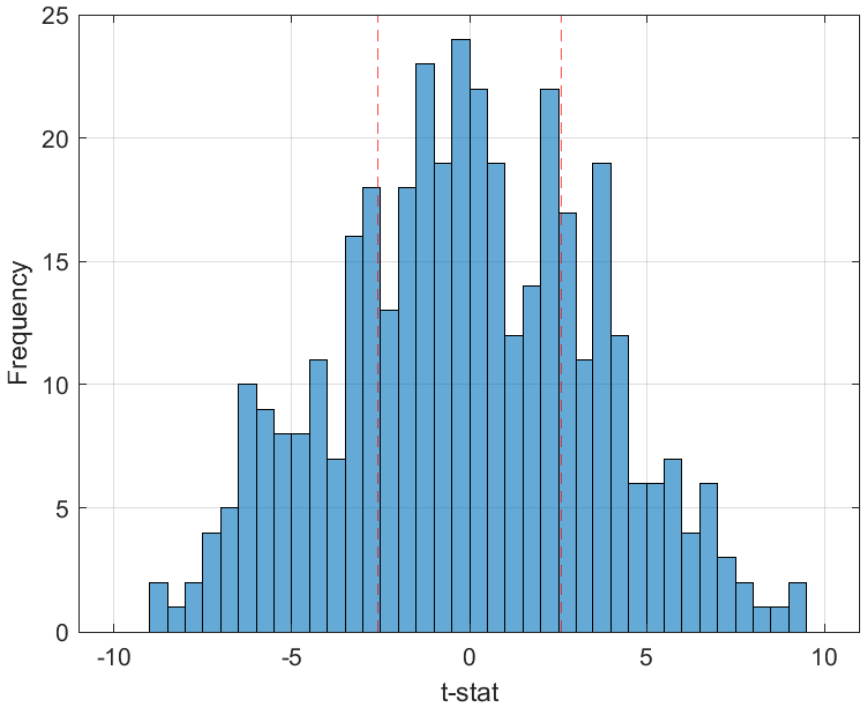


Figure 2. Out-of-Sample PR-AUC Performance of Fraud Detection Models

The figure reports the out-of-sample PR-AUC for three fraud detection models: a logistic model, an XGBoost model, and a fusion model that integrates XGBoost with contrastive learning. For each specification, we estimate the model using different combinations of three variable sets: financial metrics (Fin), word frequency measures (Text), and document embedding vectors (Emb). The models are trained on data from 1994-2008 and evaluated on a holdout sample from 2009–2019. The red line denotes 0.36%, the proportion of fraudulent firms in the test sample, and serves as the baseline PR-AUC benchmark.

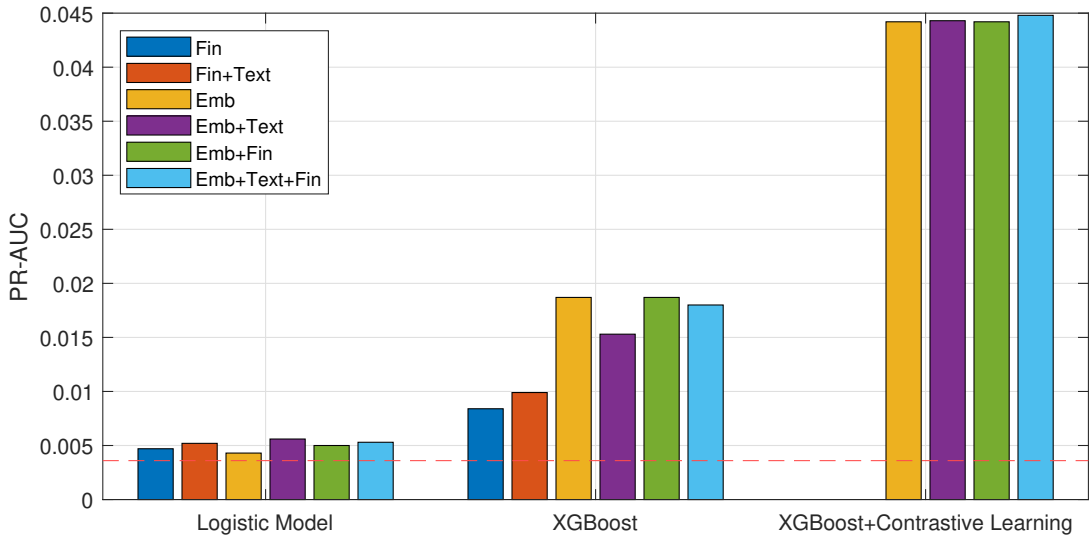


Figure 3. XGBoost Feature Importance Ranked by Absolute SHAP Values

The figure presents the top-20 features ranked by global importance (mean absolute SHAP value) in the testing sample over 2009-2019. The SHAP values are derived by training XGBoost model on the training sample, over 1994–2008, using three sets of predictors: financial metrics (Fin), textual features (Text), and document embedding vectors (Emb).

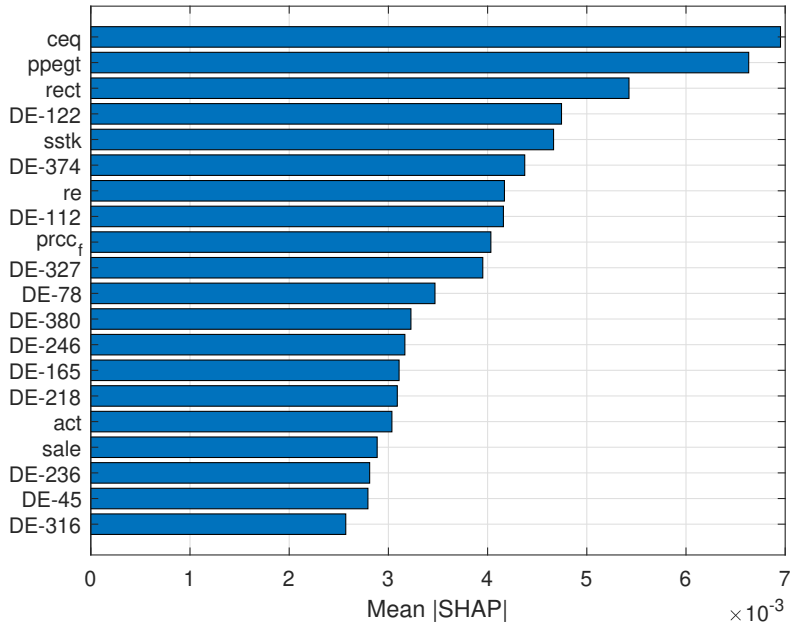


Figure 4. Absolute Category SHAP Values

The figure presents category global importance (mean absolute SHAP value) in the testing sample over 2009-2019. The SHAP values are derived by training XGBoost model on the training sample, over 1994–2008, using three sets of predictors: financial metrics (Fin), textual features (Text), and document embedding vectors (Emb). The 28 financial ratios are grouped into ten categories following Dechow et al. (2011), with category-level SHAP values computed as the sum of the SHAP values of their constituent ratios. The "12 DEs" represents the combined SHAP value of the twelve most important embedding dimensions reported in Figure 3.

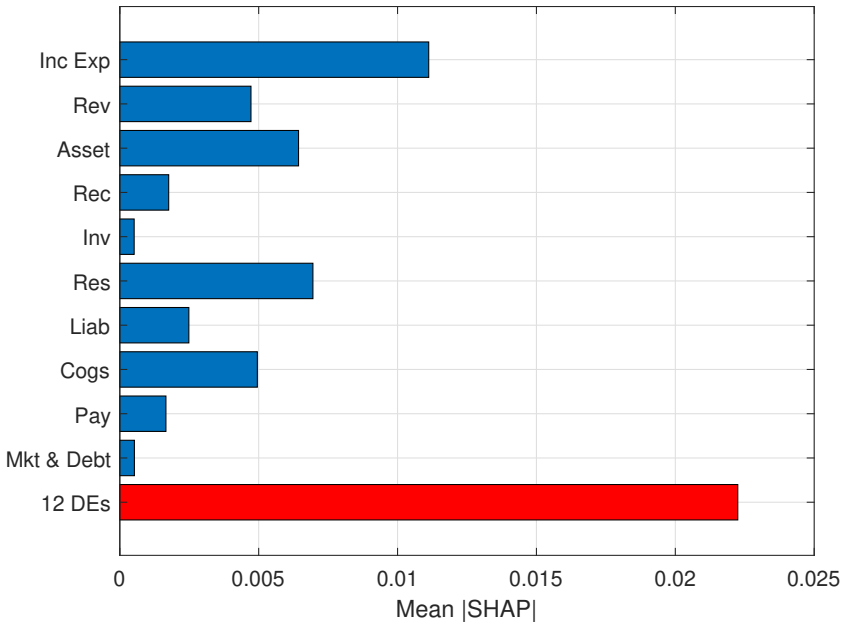


Figure 5. Canonical Correlation Matrix between Document Embedding and Category SHAP Values

The figure presents canonical correlation between document embedding SHAP values and category SHAP values in the testing sample over 2009-2019. The SHAP values are derived by training XGBoost model on the training sample, over 1994–2008, using three sets of predictors: financial metrics (Fin), textual features (Text), and document embedding vectors (Emb). The 28 financial ratios are grouped into ten categories following Dechow et al. (2011), with category-level SHAP values computed as the sum of the SHAP values of their constituent ratios. The "DE-" variables are SHAP value of the twelve most important embedding dimensions reported in Figure 3.

