Computational Economics
https://doi.org/10.1007/s10614-025-10993-2

A Random Forest Approach to Detect and Identify Unlawful Insider Trading

Krishna Neupane[1] · **Igor Griva**[2]

Accepted: 3 May 2025

Abstract
Detecting Unlawful Insider Trading (UIT) remains a critical yet challenging task. The sheer volume of transactions, often numbering in the thousands per minute, necessitates the development of sophisticated algorithms to augment manual review. While previous studies have demonstrated the potential of Machine Learning (ML) methods, such as Random Forest (RF), for UIT detection, this study significantly improves upon existing approaches. We expand the feature set from 25 previously used features to 110, incorporating 85 novel features that prove highly informative for UIT identification. Features were result of merging of multiple database for the first time to best of our knowledge. Furthermore, we substantially increase the training dataset size to 3,984 transactions. Consequently, our RF model achieves a UIT detection accuracy over 95 percent, a substantial improvement over previously reported results. Our key findings also highlight the significant influence of corporate governance factors on UIT and underscore the importance of data quality and feature engineering for optimal model performance. This research demonstrates the effectiveness of ML tools in automating UIT detection, thereby reducing reliance on manual analysis and enhancing efficiency.

Keywords Unlawful insider trading · Ensemble methods · Decision trees · Random forests · Fraudulent behavior

1 Introduction

Insiders have access to Material Non-Public Information (MNPI), and the 1934 Securities Exchange Act prohibits its misuse. Unlawful Insider trading (UIT) undermines market integrity, even when it appears routine. UIT often involves complex strategies using factors like momentum, value, and profitability, making detection difficult. SEC enforcement of IT regulations, like Rule 10b-5, is challenging due to the ambiguity of

✉ Krishna Neupane
kneupan@gmu.edu

[1] Department of Computational and Data Sciences, George Mason University, Fairfax, VA, USA

[2] Department of Mathematical Science, George Mason University, Fairfax, VA, USA

Published online: 03 June 2025

unlawful trading. Regulators struggle with sophisticated insider trading, but machine learning (ML) offers a solution. ML techniques, like decision trees, excel at analyzing complex data, finding hidden patterns, and detecting anomalies better than traditional methods, effectively capturing intricate IT activities. Traditional IT detection uses rigid rules and models, prone to overfitting (Mayo & Hand, 2022). ML offers a more adaptable, data-driven approach, identifying previously unknown patterns in complex datasets.

The study of insider trading (IT) has a rich history within finance and economics. Seminal works, such as Kyle (1985), have significantly advanced our understanding of market microstructure and information asymmetry by modeling the interplay between informed traders and market makers. Notable contributions include (Amihud & Mendelson, 1987) introduced the influential ILLIQ measure, establishing a robust link between illiquidity and expected returns. John and Narayanan (1997), who highlighted the potential unintended consequences of regulations, and Fishman and Hagerty (1995), who emphasized the potential for mandatory disclosure rules to create unintended trading opportunities for insiders. Easley et al. (2002) demonstrated the crucial role of information risk in asset pricing. Ahern (2017) provided crucial empirical evidence on the validity of commonly used proxies for informed trading. Chang et al. (2022) examined algorithmic trading's impact on insider behavior, while Machan (1996) advocated for an objective economic analysis, potentially overlooking broader societal implications. Ali and Hirshleifer (2017) identified opportunistic trading based on pre-earnings announcement profitability, and Cumming et al. (2011)) linked such trading to firm misconduct.

Legal literature complements economic and financial perspectives on insider trading. US insider trading law, rooted in Section 10(b) of the 1934 Securities Exchange Act, has evolved through several key legal theories. The classical theory, emphasizing "equal access" to information (Bines, 1976), prohibits insiders with MNPI from trading on that advantage, establishing a fiduciary duty. Landmark cases like Cady, Roberts & Co. (1961), Texas Gulf Sulphur Co. (1968), and Chiarella v. United States (1980) solidified this theory. However, Langevoort (1999) argues that the current framework may be both too restrictive (missing some tippee situations) and too broad (encompassing legitimate activities). The tipper-tippee theory, exemplified by Dirks v. SEC (1983), addresses situations where insiders ("tippers") share confidential information with others ("tippees") who then trade on it. The Supreme Court ruled in Dirks that tippee liability requires a personal benefit to the tipper (Nagy, 2020). The misappropriation theory broadened liability to include those who breach a fiduciary duty to the source of confidential information, even if not to the company itself, as seen in United States v. Chiarella (1980). This theory also encompasses "Fraud on the Market." Cases like SEC v. Dorozhko (2009) introduced an "affirmative misrepresentation" theory, though this has been criticized (Odian, 2011). This evolution reflects a tension between preventing market abuse and protecting legitimate activity (Bondi & Lofchie, 2011). The complexity of these theories highlights the challenges of enforcement (Perino, 2019), and the impact of insider trading on investor confidence remains debated (Manne, 2005).

While foundational, prior research may not adequately reflect the complexities of today's markets. Several limitations are common across the studies reviewed.

Data quality, completeness, and availability are crucial, and issues like bias, missing data, and accuracy of labeled datasets can significantly impact model effectiveness. Industry-specific nuances require models to adapt to the unique characteristics of different sectors. The rise of high-frequency and algorithmic trading, coupled with rapid technological advancements in information dissemination, has dramatically reshaped market dynamics. Furthermore, evolving regulations, globalization, and increasing ethical concerns necessitate continuous research. This article addresses these evolving market realities by building on prior work and leveraging advanced data analytics and machine learning to better understand insider trading and its effect on market efficiency.

More recently, research has increasingly employed machine learning techniques to enhance the detection of the UIT such as Random Forests (RF) and Extreme Gradient Boosting (XGBoost), for example by Deng et al. (2021), Deng et al. (2019) hereinafter (DCZ), Lauar and Arbex Valle (2020). DCZ contribute to the field of IT detection by employing advanced ensemble methods. The innovation in each approach leverages the power of decision trees to identify potential instances of IT with greater accuracy and efficiency. Specifically, Deng et al. (2021) showed the effectiveness of an intelligent system combining PCA and Random Forest. Their system, analyzing 26 indicators, achieved high recall (up to 83.87 percent) within 60-day windows, identifying key features for insider trading detection. These methods offer advantages like identifying complex patterns in financial data. Ensemble methods, known for robust fraud detection, use techniques like parameter updates to improve accuracy (Sundarkumar & Ravi, 2015; Louzada & Ara, 2012). Despite challenges like the bias-variance trade-off, these methods efficiently integrate data mining and modeling, extracting predictive features within a unified framework, allowing for cost-effective data labeling (Iskhakov et al., 2020; Altmann et al., 2010; Strobl et al., 2007). By capturing short-run noise and using historical context, they generate reliable predictions and understandable inferences (Khademian, 2022; Agrawal et al., 2019; Christensen et al., 2016). Domain-agnostic methods like decision trees analyze data microstructure, independent of specific estimators, and automatically identify key variables (Malhotra et al., 2015; Iskhakov et al., 2020). These advancements show ML's potential to revolutionize UIT detection (Camerer, 2019; Fudenberg & Liang, 2019). Beyond accuracy, understanding model predictions is crucial for trust, bias detection, and responsible decisions. Ensemble models facilitate the interpretability.

This work, due to its simplicity, makes several contributions. First, it addresses limitations of traditional insider trading (IT) analysis, including reliance on manual feature engineering, omission of key variables, and difficulty handling data interdependencies. By training ensemble models, it aims to uncover hidden patterns in IT behavior, providing insights into insiders' decisions. Second, it uses machine learning to identify hidden IT strategies, including potentially fraudulent trades, that traditional methods might miss, mitigating "p-hacking" bias and enabling detection of subtle patterns. Third, it expands the feature space from 26 (in the benchmark study) to 110 by merging databases from multiple sources (see Section 3 for details) and compares results across multiple scenarios. Fourth, to our knowledge, this is the first study to apply this research to the US capital market, expanding the geographic scope. Finally,

the study implements a fully automated system for data labeling, preprocessing, and results analysis.

The study is organized as follows: Section 2 introduces the proposed methodology outlining the theory behind the PCA (Section 2.1) and random forest (Section 2.2), the tuning parameters (Section 2.3), feature selection criteria (Section 2.4) and performance measures (Section 2.5); Section 3 describes the experimental setup such as data-preprocessing steps, implementation of cross-validation, criteria of parameter control and integration of methods; Section 4 analyzes the results that begin with data description (Section 4), steps of dimensionality reduction (Section 4.1), results from components of confusion matrix (Section 4.2), ranking of important and relevant variables (Section 4.3), and ranking of variables (Section 4.3.1 and Section 4.3.2). In the final two sections- we discuss results (Section 5) and provide our conclusions, recommendations and future course (Section 6).

2 Proposed Methodology

Multicovariate and high-dimensional financial and trade data often exhibit underlying hidden structures. To effectively detect irregularities within such complex datasets, various machine learning techniques have proven successful, including Principal Component Analysis (PCA) , gradient boosting, RF (RF) and neural networks . These methods serve as crucial intermediary steps in empirical economic research (Wager & Athey, 2018). Building upon the pioneering work of DCZ, which integrated PCA and RF to detect UIT, this study extends these powerful modeling techniques to a broader dataset of US securities, encompassing a significantly larger feature set (110 features compared to 26 in DCZ). Similar to previous study, the study incorporates PCA to reduce dimensionality and removal of correlation, and subsequently integration to the RF extract significant features by leveraging the strengths of decorrelated trees to enhance the accuracy and robustness of UIT detection. Furthermore, this study emphasizes the importance of addressing potential limitations of machine learning models, such as uncertainty, quality decline, bias, and lack of transparency, interpretability, and reproducibility (Stoyanovich et al., 2017; O'Neil, 2016). By expanding the feature set and rigorously evaluating the performance of RF, it seeks to mitigate aforementioned risks and improve the overall reliability and generalizability of the results. Prior to implementing these techniques, the study undergoes a series of data preprocessing steps, as detailed in Section 3.

2.1 Principal Component Analysis

To enable comparison with the DCZ study, the paper implements PCA (PCA), a dimensionality reduction technique that filters noise and removes correlation between features to extract the most relevant information from the data. It achieves this by computing principal components (PCs) - linear combinations of the original variables (Wang et al., 2018; Abdi & Williams, 2010). The first principal component captures the largest variance in the data, while subsequent components are orthogonal to the pre-

ceding ones, sequentially explaining decreasing amounts of variance. PCA has found widespread application in various domains, including finance and anomaly detection (Aït-Sahalia & Xiu, 2019). It has been used to identify exceptional volatilities (Egloff et al., 2010), analyze investor sentiments (Baker & Wurgler, 2006), discern policy uncertainties (Baker et al., 2016), and detect aberrant stock and bond returns (Pasini, 2017; Pérignon et al., 2007; Driessen et al., 2003; Feeney, 1967). More recently, PCA has been applied in the market cross-correlation analysis and systemic risk measurement (Billio et al., 2012; Zheng et al., 2012; Kritzman et al., 2011).

In a reduced dimensional space, DCZ successfully implemented PCA with 26 features to identify UIT. Following a similar approach, in this paper, PCA is implemented to obtain comparable results. Furthermore, the analysis is extended by incorporating an additional 110 features. Formally, given a data matrix $\mathbf{D} \in \mathbf{R}^{n \times m}$ is matrix with n rows and m features and m features, PCA aims to project the p-dimensional vectors onto a q-dimensional subspace, where PCs are sequentially organized in the real coordinate space. The first component, the i-th vector in the sequence, is orthogonal to the preceding (i-1) components and maximizes variance, minimizing the average squared perpendicular distance from the data points to the line. This process effectively reduces random variability by identifying and utilizing the most informative dimensions (Aluja-Banet & Nonell-Torrent, 1991). The core of PCA involves estimating eigenvalues and the covariance matrix, with the eigenvalues of the covariance matrix exhibiting consistent asymptotic normal estimators (Anderson, 1963). The steps of the PCA algorithm are outlined in Algorithm 1.

Algorithm 1 PCA.

Input : $\mathbf{D}^{n \times m}$ Data matrix
1. Obtain the empirical mean of each column
2. Center the data by subtracting off the mean of each column of $p \times 1$-dimensions, represented as $\mathbf{B}$
3. Compute the covariance matrix $C = \frac{1}{N}\mathbf{B}^T\mathbf{B}$ from above Step 2
4. Compute the eigenvalues and eigenvectors of C so $\mathbf{V}^{-1}\ \mathbf{CV} = \mathbf{E}$, $\mathbf{V}$ holds eigenvectors of $\mathbf{C}$ and $\mathbf{E}$ is the diagonal $M \times M$ diagonal eigenvalue matrix
5. Sort the columns of $\mathbf{D}$ into order of decreasing eigenvalues
6. Obtain the cumulative values of eigenvalues by summing up eigenvalues for each row (represents the overall predictive power)
7. Obtain the projections with eigenvalues

PCA, despite its widespread success and acceptance, faces several limitations (Aït-Sahalia & Xiu, 2019). Firstly, the "curse of dimensionality" poses a significant challenge. As the number of dimensions (variables) in the data increases, the required sample size to maintain consistent results grows exponentially (Bellman, 1958). This can lead to shallow consistency of eigenvalues, particularly in high-dimensional datasets. Secondly, the assumption of parameter constancy can be problematic. PCs are linear combinations derived from the data, which may not adequately capture nonlinear patterns within the data. Thirdly, the assumption of data independence can be violated in time series data, where dependencies and non-stationarity are common.

This can lead to inappropriate inferences (Zhang & Tong, 2022; Brillinger, 2001). Despite these limitations, PCA offers valuable insights. By removing noise and correlation features, PCA highlights the contribution and relevance of prominent individual variables. In the context of identifying UIT, integrating PCA with RF allows for the recalibration of feature weights, minimizing the risk of misclassification and the identification of falsely unlawful activity.

2.2 Random Forest

DCZ integrates PCA with RF, a data-driven non-parametric ensemble method pioneered by Breiman (2001). RF utilizes bootstrap resampling (bagging) to create training sets for individual decision trees, offering the advantage of Out-of-Bag (OOB) error prediction. OOB error estimates are calculated based on the prediction errors of trees that were not involved in the bootstrap sampling of the corresponding data points (Friedman et al., 2000). By aggregating numerous de-correlated trees, RF effectively controls variance and improves accuracy (Chen et al., 2022). This ensemble approach allows RF to capture non-linearities and interactions within the data. In essence, RF aims to group observations with similar predictors by sequentially growing a series of decision trees. RF has found widespread application in various fields, including chemical informatics (Svetnik et al., 2003), ecology (Prasad et al., 2006), 3-D object recognition (Shotton et al., 2011), epidemiology (Azar et al. 2014), and remote sensing (Pal 2005). The steps of RF are outlined in Algorithm 2.

Definition

A RF is a classifier consisting of a collection of tree-structured classifier $h(x, \Theta_k), k = 1, ..., k$ where the Θ_k are independently distributed random vectors and each tree casts a unit vote for the most popular class at input x (Breiman (2001)). The RF algorithm is shown in Algorithm 2[1].

Algorithm 2 RF.

Input : $\mathbf{D} := (x_1, y_1), ..., (x_n, y_n)$ with Features $\mathbf{F}$, and number of trees in forest $\mathbf{B}$
begin
  RANDOMFOREST($\mathbf{S}$, $\mathbf{F}$)
  $H \leftarrow \emptyset$
  for $i \in 1, ...,$ **B do**
    $\mathbf{S}^i \leftarrow$ A bootstrap sample from $\mathbf{S}$
    $\mathbf{h}_i \leftarrow$ RANDOMIZEDTREELEARN ($\mathbf{S}^{(i)}$, $\mathbf{F}$)
    $H \leftarrow H \cup H\{h_i\}$
  return H
begin
  RANDOMIZEDTREELEARN($\mathbf{S}$, $\mathbf{F}$)
  At each node:
    $\mathbf{f} \leftarrow$ very small subset of $\mathbf{F}$
    Split on best feature in f
  return The Learned Tree

[1] https://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/ensembles/RandomForests.pdf

In supervised learning settings, RF employs randomness to accurately map input values (x) to their corresponding targets (y), represented by the functional relationship $y = f(x)$ for an unknown joint distribution $P(x, y)$. Decision trees within the RF minimize the correlation (ρ) among the collection of p-dimensional variables. The core of RF involves generating a collection of k decision trees. For each tree (k -th tree), a random vector (Θ_k) is generated independently of previous vectors ($\Theta_1, \ldots, \Theta_1$) but with the same distribution. This results in a classifier ($h(x, \Theta_k)$) that utilizes input vectors (x) with N training examples. By constructing this ensemble of trees, RF learns the functional relationship between x and y by producing a prediction model $\hat{f}(X, \theta)$, controlled by the k-dimensional hyper-parameter configuration from the search space. This process aims to estimate the expected risk of the inducing algorithm with respect to on new data, also sampled from P. The results obtained from Algorithm 2 are further refined by employing k-fold cross-validation. This technique involves repeatedly resampling the data to minimize empirical risk and approximate the generalization error (Afendras & Markatou, 2015; Bergstra & Bengio, 2012; Witten et al., 2011; Arlot & Celisse, 2010; Shao, 1993; Efron, 1983). This iterative process helps to optimize parameter values and improve the model's performance on unseen data.

2.3 Parameter Tuning

Parameter tuning plays a crucial role in optimizing RF (RF) performance, specifically by reducing the overall Out-of-Bag (OOB) prediction error. Key RF parameters that significantly impact accuracy, minimize overfitting, and improve the relative contribution to correct prediction include (Eggensperger et al., 2018; Friedman, 2001):

- m_{try}: Controls the number of features randomly considered for each tree node, balancing correlation and tree independence. Lower values generally improve stability and performance (Probst et al., 2018; Bernard et al., 2009). It is calculated as square root of number of features $\sqrt{n_{features}}$.
- $ntrees$: The number of trees in the forest. While not technically a hyper-parameter, increasing $ntrees$ stabilizes the model, though it increases computational time. Probst et al. (2018); Scornet (2017).
- $max\ depth$: Limits the depth of each tree. While unpruned trees were initially favored, recent findings suggest that proper tuning can significantly improve performance, albeit with increased computational cost (Belkin et al., 2019).
- $sample\ rate$: Controls the fraction of data used for each tree, with a default value of 1.0. Adjusting this parameter can improve generalization and enhance predictive performance on validation and/or test sets.

2.4 Feature Importance

The multi-covariate input during RF tree construction aims to extract, compare, and rank significant features, enhancing model interpretability, explainability, and predictive accuracy. This process also contributes to reduced time and space complexity, as well as improved generalization error (Zhou, 2022; Xu et al., 2014; Duchi et al.,

2008). Various methods have been proposed to identify interactive non-linearities and incorporate known sparsity structures (Qian et al., 2022; Xu et al., 2014; Guyon & Elisseeff, 2003; Genuer et al., 2010; Strobl et al., 2007). Originally, Breiman (2001) introduced Mean Decrease of Impurity (MDI), based on Gini scores, to assess feature importance. MDI measures the decrease in node impurity, which represents the probability of misclassifying a randomly chosen transaction if labeled randomly according to the class distribution (Nembrini et al., 2018). As a byproduct of the RF splitting process, node splits are chosen to maximize impurity reduction (largely homogeneous). Therefore, nodes with larger impurity decreases are ranked higher. However, MDI has limitations. Firstly, it ranks features during training, lacking access to test data. Secondly, it tends to favor features with high cardinality, potentially diluting the significance of correlated features. Subsequent work by Fisher et al. (2019) introduced a permutation-based approach to address the limitations of MDI. This method calculates the contribution of each covariate by permuting its values in out-of-bag examples while keeping other predictors fixed, and then evaluating the impact on the model's predictions (Nembrini et al., 2018). This approach mitigates bias and addresses the tendency of MDI to favor high-cardinality features.

2.5 Performance Measure

To evaluate the performance of binary supervised classification problems, a 2×2 confusion matrix is commonly used. This matrix schematically represents the actual and predicted class labels, as shown in Table 1 (Hastie et al., 2009). The details of the components of confusion matrix is presented in Table 2.

3 Experimental Setup

The Securities Exchange Act of 1934 mandates that individual insiders file Form 4 within 48 hours of any transaction. For this study, 9.6 million publicly available Form 4 filings from the Securities and Exchange Commission's (SEC) Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system between 2003 and 2022 was collected. These filings were preprocessed using a Python text crawler and stored in a MongoDB collection. By using Central Index Key (CIK), the Form 4 data was merged with the Center for Research in Security Prices and Compustat-CapitalIQ. The merge links individual insider transactions with corresponding daily trade and financial data. The corresponding quarterly financial data was matched to the transaction date given both fall in the same quarter. To ensure data completeness, missing quarterly financial

Table 1 Confusion Matrix for Binary Classification of IT Transactions: A Schematic Representation of Model Performance

Tot. Pop. (P+N)	Predicted Labels (PP+PN)	
	Lawful(P)	Unlawful(N)
Lawful(+)	True Lawful	False Unlawful
Unlawful(-)	False Lawful	True Unlawful

Table 2 Computing Confusion Matrix Components for IT Transaction Classification: Definitions and Formulas derived from Table 1

Definition	Equation
True Lawful (True Positives) are lawful transactions classified as lawful, that is, (+ as +). True Positive Rate (TPR, Recall or Sensitivity), calculated with Equation 1, is the proportion of positive instances that are correctly classified as positive by the model.	$\text{TPR} = \frac{TP}{TP+FN}$ (1) $= \frac{TP}{P} = 1 - FNR$
False Lawful (False Positives) are unlawful transactions but the model wrongly classifies as lawful, that is, (− as +). False Positive Rate (FPR, False Alarm or Fall Out) calculated with Equation 2, is the probability that incorrect classification of unlawful as lawful.	$\text{FPR} = \frac{FP}{FP+TN}$ (2) $= \frac{FP}{N} = 1 - TNR$
True Unlawful (True Negatives) represent unlawful transactions correctly identified as unlawful, that is, (− as −). True Negative Rate (TNR, Specificity) is calculated as in Equation 3.	$\text{TNR} = \frac{TN}{TN+FP}$ (3) $= \frac{TN}{N} = 1 - FPR$
False Unlawful (False Negatives) are lawful transactions, but model wrongly classifies as unlawful(negative), that is, (+ as −). False Negative Rate (FNR, Miss Rates) calculated with Equation 4 is the probability of true negatives (lawful) missed equation.	$\text{FNR} = \frac{FN}{TP+FN}$ (4) $= \frac{FN}{P} = 1 - TPR$
Accuracy (ACC) measures effectiveness of classifier to the correctly identify transactions across all lawful(+) and unlawful(−) transactions. In unbalanced dataset the presence of large negative class and the measure favors of the true negatives, the outcomes become inconsequential Equation (see 5).	$\text{ACC} = \frac{TP+TN}{P+N}$ (5) $= \frac{TN+TP}{TP+FP+FN+TN}$
Precision (PRE) is proportion of correct to the sum of the correct and incorrect classification that the class agreement of the positive labels classified by the classifier (see Equation 6).	$\text{PRE} = \frac{TP}{TP+FP}$ (6)

results were replaced with data from the subsequent quarter. The resulting MongoDB documents contain 110 features encompassing ownership, corporate governance, profitability, financial performance, risk, and market returns. Unlawful transactions were identified by matching Form 4 filers with defendants listed in publicly available SEC court complaints. A text crawler with Levenshtein distance was used for this matching process, with a score of 85 percent or higher indicating a potential match. The database ultimately contained 1992 identified unlawful transactions. The feature set includes 20 numerical covariates that closely resemble those used in the DCZ study (Table 4). Since there was lack of one-one comparison to remaining six features, here the settings incorporated five additional categorical features: Acquisition and Disposition, IsDirector, IsOfficer, and IsTenPercentOwner. In summary, the feature set that closely resemble DCZ study include variables commonly used in prominent financial models, such as the Black-Scholes Model (Black & Scholes, 1973), Capital Asset Pricing Model (Sharpe, 1964), Efficient Market Hypothesis (Fama, 1970), Arbitrage Pricing

Theory (Ross, 1978), and models within Behavioral Finance (Shiller, 1989; Tversky & Kahneman, 1974). As a part of data preprocessing, numerical features were standardized using z-scores that yields faster convergence and improved interpretability (Gelman, 2008). Categorical features were one-hot encoded.

The experiments are ran on two datasets: one with 320 transactions and another with 3984 transactions, both with a balanced 50:50 ratio of lawful and unlawful transactions. The smaller dataset was randomly selected from the larger pool. Each experiment was further subdivided based on feature selection and the inclusion of PCA. The 1992 unlawful transactions represent the maximum available for analysis. Lawful transactions were randomly sampled from the remaining 9.6 million filings. For example, to create the 320-transaction dataset, 120 unlawful transactions were randomly selected, and the remaining 200 were randomly sampled from the pool of lawful transactions. Hyper-parameter tuning was performed using 5-fold cross-validation with a randomized search space. The hyperparameters included $ntrees$, m_{try}, max_depth, and $sample_rate$. An 80:20 training-testing split was maintained, and each experiment was repeated 100 times. The results are summarized in Table 5. Feature importance was assessed using both Gini impurity and permutation importance. Permutation importance addresses the bias of Gini impurity towards high-cardinality features. To further account for feature correlations, hierarchical clustering based on Spearman rank-order correlation was applied with that allows to select single representative feature was selected from each cluster. These adjustments allowed us to relax assumptions DCZ study made, such as:

- Qualitative features, often prefixed with terms like "significant," can be subjective and introduce ambiguity into the analysis. The DCZ study utilized "significance" criteria without providing a clear quantitative definition. To mitigate this, in this study the use of qualitative descriptors is minimized, enhancing the empirical justification of the results.
- Confining data selection to a single industry group can introduce selection bias. Unlike the DCZ study, the data in this study is randomly selected transactions from diverse industrial groups to enhance the generalizability and objectivity of findings.
- Recognizing the dynamic and complex nature of the US securities market, authors acknowledge that numerous unlawful transactions may go undetected due to limited investigative resources. Therefore, authors adopt a cautious approach, treating each transaction as potentially unlawful for the purpose of the analysis.

To ensure reliable performance evaluation, the code iterates 100 times by randomly selecting positive (lawful) transactions on every run. For each iteration, the performance metrics derived from the confusion matrix in a separate spreadsheet is recorded. The average performance across these 100 iterations provides a more robust estimate compared to a single run. The final results are benchmarked against performance metrics obtained from Artificial Neural Networks, Support Vector Machines, Adaboost, and RF models as reported in DCZ.

4 Results

The results are based on a balanced dataset comprising 3984 transactions (1992 unlawful) with 110 dimensions, as detailed in Section 2. For illustrative purposes, Table 3(b) presents a random subset of Table 3(a), designed to align with the dataset used in Deng et al. (2021). Notably, both tables demonstrate a higher frequency of "purchases" among insider transactions. This observation can be attributed to executive compensation structures, which often include restricted stock options and bonuses tied to performance targets.

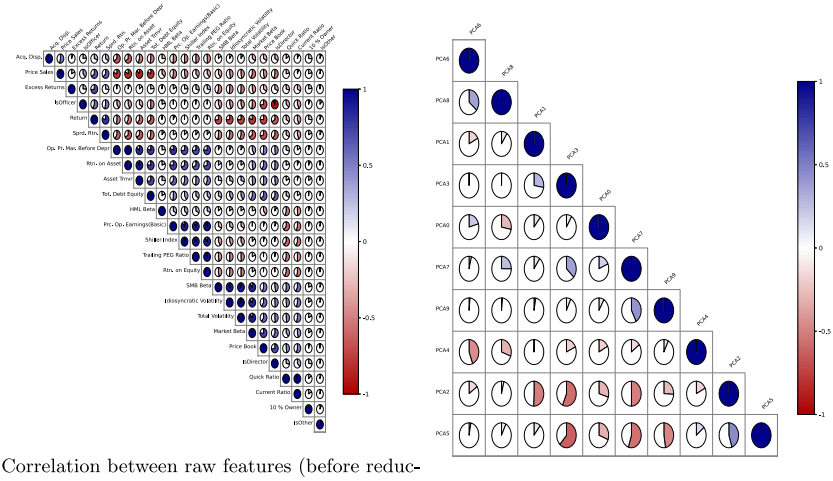
4.1 Results of Dimensionality Reduction

This section presents the results of PCA, a technique that transforms a set of correlated variables into a smaller set of uncorrelated components. These PCs (PCs) are linear combinations of the original variables, capturing the directions of maximum variance within the data. Each PC is represented by an eigenvector, whose corresponding eigenvalue indicates the amount of variance explained by that component. The first PC captures the most variance, followed by subsequent PCs in descending order of explained variance. The component loadings, calculated as the product of the eigenvector and the square root of the eigenvalue, represent the correlation between each original variable and the corresponding PC. To determine the optimal number of PCs, analysis performed and analyzed the cumulative explained variance ratio (EVR). The EVR for a given PC is the ratio of its eigenvalue to the sum of all eigenvalues, indicating its relative contribution to the total variance. The analysis revealed that, on average, 10 PCs were sufficient to explain 94.76 percent of the cumulative variance in the data. Figure 1b illustrates the relationships between these 10 PCs, based on a representative subset of 100 experimental runs. Each subplot depicts the correlation between two PCs, reflecting the linear reduction of multi-dimensional data into uncor-

Table 3 Illustrative Distribution of Labeled IT Transactions

Label	Sell	Purchases	Total
Lawful	405	1587	1992
Unlawful	318	1674	1992
Lawful	27	133	160
Unlawful	26	134	160

The sub-tables presents the distribution of labeled IT transactions used in the analysis. A maximum of 1992 labeled unlawful transactions were available. The lawful transactions were randomly selected from a pool of approximately 9.6 million transactions. A balanced dataset was created by randomly selecting an equal number of lawful transactions, resulting in a 50/50 ratio of lawful to unlawful instances. **Panel (a)** displays the distribution of the full dataset, comprising 1992 unlawful and 1992 lawful transactions. **Panel (b)** presents a subset of 320 transactions randomly selected from the full dataset in Panel (a). This subset was chosen to match the sample size used in the benchmark studies by Deng et al. (2021) and Deng et al. (2019)



(a) Correlation between raw features (before reduction)

(b) Correlation between PCs (after reduction)

Fig. 1 Correlation matrices illustrating the impact of dimensionality reduction. (a) Heat-map of Spearman correlation coefficients between original features. Red indicates negative, blue positive correlations; circle size reflects correlation magnitude (ρ). Clusters of highly correlated variables indicate multicollinearity. (b) Heatmap of Pearson correlation coefficients between PCs. Reduced dimensionality and decorrelation are evident in the decreased number of strong correlations

related features. The PCs are ordered based on their explained variance. It is important to note that the direction of a PC is not unique; therefore, the interpretation of the plot remains unchanged regardless of the axis on which a particular PC is plotted.

To visually represent the interrelationships between variables, a heatmap (Fig. 1a) is employed. This intuitive visualization utilizes a color gradient to depict the strength and direction of Spearman correlations between variables. Blue and red gradients represent positive (perfect at +1) and negative (perfect at -1) correlations, respectively. The heatmap effectively summarizes the relationships between variables, revealing several key insights. A strong correlation was observed between idiosyncratic volatility and total volatility, highlighting the interplay between unsystematic and systemic risk. This suggests that IT activities may be strategically influenced by overall market volatility (Koumou, 2020). Furthermore, a linear relationship was observed between the quick ratio and current ratio, indicating that insiders' short-term liquidity access may influence their trading behavior (Cohen et al., 2012).Finally, a negative correlation was observed between the director and officer variables. This suggests that while an individual may serve as a director, overseeing the company's overall strategy, they may not necessarily be actively involved in daily operations. This distinction can be nuanced and often not clearly defined (Gregory et al., 1997).

4.2 Results of Classification of Insider Trading Transactions

This section presents the performance of the RF model, evaluated using a confusion matrix. The results, summarized in Table 5, are based on an average of 100

experiments, each with 5-fold cross-validation. This iterative approach significantly reduces performance variability. For instance, the standard deviation of accuracy for 320 transactions and 25 features decreased from 10.6 to 0.31 and from 6.91 to 0.45 when the number of experiments was increased from 10 to 100, respectively. Similar improvements were observed for other performance metrics, justifying the use of 100 experimental runs for the analysis. Table 5 presents the results organized by transaction count (320 and 3984), feature count (25 and 110), and the inclusion of PCA. This study employed a random search strategy within a 5-fold cross-validation framework to optimize RF hyperparameters. Key findings include an optimal m_{try} range of 0.35 to 0.95, demonstrating the importance of balancing feature diversity and model complexity. The number of trees ($ntrees$) significantly impacted performance, with optimal values exceeding 100. It was evident that increasing the number of trees initially improved accuracy as the model gains better generalization. However, diminishing returns and even slight performance decreases can occurred due to overfitting or other factors. The highest accuracy was achieved with 940 trees. Regarding the number of trees ($ntrees$), while not strictly a hyperparameter, it is crucial to use a sufficiently high value to ensure model stability and convergence (Probst & Boulesteix, 2017; Scornet, 2017). The analysis revealed the expected trade-offs between model complexity and generalization, with max_depth influencing overfitting. While varying the fraction of samples used to train each tree impacts model performance, the experiments indicated that growing number of samples improved classification accuracy. Interestingly, average run training time does not always exhibit a strong correlation with either average fit time. This suggests that prediction time can be influenced by factors beyond model complexity, such as the number of trees in the ensemble or the depth of individual trees, which may not always directly correlate with accuracy. These findings highlight the trade-offs inherent in hyperparameter tuning: optimizing for accuracy often involves increased training time, while balancing computational efficiency during both training and prediction phases requires careful consideration of model complexity and other hyperparameters.

The experiment consistently outperformed the benchmark, achieving an accuracy of at least 80.12 percent in all scenarios, exceeding the benchmark accuracy of 77.88 percent for the PCA-RF model. Overall, the models demonstrated high accuracy, with values consistently exceeding 80 percent across all scenarios. The full dataset (3984 transactions) consistently achieved higher accuracy compared to the subset of 320 transactions, indicating the benefit of larger datasets for model performance (rows 5-8 of Table 5). The inclusion of 110 features generally led to higher accuracy compared to using only 25 features. PCA integration had a mixed impact on accuracy, with some scenarios showing slight improvements and others exhibiting slight decreases. This improvement can be attributed to two factors: (1) the increased diversity of lawful transactions sampled from a larger pool, and (2) the broader scope of our analysis, which extended beyond the time-window constraints of previous studies. These findings demonstrate the potential applicability of our approach to real-world scenarios within the SEC. While our results demonstrate a significant improvement in accuracy, it is crucial to acknowledge that accuracy alone should not be the sole determinant of model performance. Relying solely on accuracy can lead to overgeneralization and potentially misleading conclusions (Table 4).

Table 4 Performance of Benchmark Machine Learning Models

	ANN	SVM	Adaboost	RF	
				No PCA [a]	With PCA*
ACC	69.57	75.33	74.75	77.15	77.88
FNR	19.21	21.42	26.62	20.14	22.70
FPR	34.07	27.75	24.42	25.48	21.56
PRE	NA	NA	NA	NA	78.94
TNR	65.93	72.75	75.58	74.52	78.44
TPR	80.79	78.58	73.38	79.86	77.30

The table presents the performance metrics of various machine learning models reported in Deng et al. (2021) and Deng et al. (2019) on a dataset of 320 transactions with 26 features, used to identify and detect UIT. Performance of Benchmark Methods on 320 Transactions with 26 Features, used to identify and detect UIT

[a] Deng et al. (2019), No PCA

* Deng et al. (2021), With PCA

This research emphasizes the importance of Total Positive Rate as a critical performance metric, particularly for decision-making in the context of IT detection. Compared to overall accuracy, TPR provides a more nuanced assessment by focusing on the model's ability to correctly identify all instances of lawful transactions. A higher TPR indicates that the model reliably detects lawful transactions, minimizing the number of false negatives and ensuring effective detection with minimal "spillage". Among the benchmark methods, Artificial Neural Networks achieved a TPR of 80.79 percent, marginally higher than the 77.30 achieved by the DCZ method. In contrast, the results presented here consistently demonstrated superior TPR performance across all scenarios, with the lowest value observed being 84.79 percent (second column of Table 5). The model demonstrated high TPR, indicating a low rate of missing true positives (failing to identify actual unlawful transactions). TPR generally increased with an increase in dataset size and the number of features. It is crucial to note that TPR does not consider instances where lawful transactions are incorrectly classified

Table 5 Average performance metrics of 100 experiments with 5-fold cross-validation

	Subset of 320 Randomly Selected Transactions				Full Dataset (3984 Transactions)			
	25 Features		110 Features		25 Features		110 Features	
	No PCA	With PCA	No PCA	With PCA	No PCA	With PCA	No PCA	With PCA
ACC	82.95	80.12	90.54	83.42	97.87	97.14	99.13	98.13
FNR	14.23	15.21	7.29	13.66	1.53	1.41	0.70	1.07
FPR	19.88	24.55	11.64	19.49	2.72	4.31	1.03	2.66
PRE	81.61	77.99	89.14	81.94	97.31	95.81	98.96	97.38
TNR	80.12	75.45	88.36	80.51	97.28	95.69	98.97	97.34
TPR	85.77	84.79	92.71	86.34	98.47	98.59	99.30	98.93

Results are presented for two dataset sizes (320 and 3984 transactions) and two feature sets (25 and 110 features), with and without the integration of PCA with the RF algorithm

as unlawful (false positives). Therefore, TPR is particularly valuable for "ruling out" the possibility of unlawful activity, as it minimizes the risk of misidentifying lawful transactions. In situations where a data point is ambiguous or indeterminate, TPR may not be the most informative metric for identifying true negatives. PCA integration had a minimal impact on TPR, with slight variations across different scenarios.

This study demonstrates the effectiveness of the proposed model in minimizing false positives (FPR). FPR represents the rate at which unlawful transactions are incorrectly classified as lawful. The benchmark PCA-RF model exhibits an FPR of 21.56 percent. In contrast, the proposed model, in its initial configuration with 25 features and without PCA, exhibits an FPR of 24.55 percent (Table 5, second column). However, across all scenarios presented in the first four columns of Table 5, the proposed model consistently outperforms the benchmark methods in terms of FPR. These findings suggest that dimensionality reduction through PCA may not be the most effective strategy for minimizing false positives in this context. Furthermore, the proposed model exhibits remarkably low FPR rates when utilizing all 110 features. With all features considered, the FPR drops to a minimum of 1.03 percent (Table 5, seventh column). While a slight increase in FPR is observed with the larger dataset of 3984 transactions, this is expected and likely attributable to the increased complexity of the data. Importantly, the proposed model consistently achieves lower FPR values compared to all benchmark methods across all experimental scenarios. This consistently low FPR provides strong evidence of high true positive rates and low false negative rates, demonstrating the model's effectiveness in accurately identifying and classifying UIT. PCA integration sometimes led to an increase in FPR, suggesting that dimensionality reduction may not always be beneficial in minimizing false alarms.

Specificity, known as the true negative rate (TNR), measures the model's ability to correctly identify truly unlawful transactions. By focusing solely on unlawful instances, TNR excludes lawful transactions, leading to a higher proportion of true negatives and a lower number of false positives (unlawful transactions incorrectly classified as lawful). This indicates the model's effectiveness in "ruling in" genuinely unlawful transactions. Compared to benchmark methods, the proposed model demonstrates superior TNR performance across most scenarios. Notably, the model achieves higher TNR even when using fewer features (25 features) than the benchmark model (fifth column of Table 5), highlighting its ability to effectively identify unlawful transactions. While the integration of PCA generally leads to a slight decrease in TNR, the performance improves significantly with the inclusion of more transactions. For instance, with the full dataset (3984 transactions), the TNR increases by 18.9 percent when PCA is integrated (from 78.44 percent to 97.34 percent), as shown in Table 5. These results suggest that the random selection of lawful transactions plays a crucial role in defining model performance. While these metrics indicate a substantial ability to identify unlawful transactions, it is crucial to exercise caution. Overfitting and the presence of unaccounted-for noise variables may influence model performance. Furthermore, the focus on TNR primarily assesses the model's ability to "rule out" lawful transactions, limiting its ability to definitively identify false positives. Nevertheless, achieving high TNR is crucial for minimizing false accusations and reducing the administrative burden associated with investigating potentially unlawful transactions.

The model consistently demonstrated high precision, indicating a low rate of false positives among the predicted unlawful transactions. This signifies that when the model flagged a transaction as unlawful, it was generally accurate, minimizing the risk of incorrectly classifying lawful transactions as suspicious. Precision generally increased with an increase in dataset size and the number of features, suggesting that larger datasets and more comprehensive feature sets improved the model's ability to accurately identify true positives. The impact of PCA integration on precision was mixed, with some scenarios showing slight improvements and others exhibiting slight decreases, indicating that dimensionality reduction through PCA did not consistently enhance precision in this context.

The performance of the model in distinguishing between true legal and false legal transactions (i.e., true positives versus false positives) was generally higher when using the selected features compared to using all features. This suggests that the selected features may be more effective in identifying true positives than the full set of features. While the Area Under the Curve metric generally exceeds 0.5 (indicating better performance than random guessing) for both feature sets, a significant variation in AUC values is observed between the two datasets: 13.3 percent for the full dataset (3984 transactions) and 98 percent for the subset of 320 transactions. This substantial difference suggests that the performance may be influenced by the specific selection of features in the smaller dataset. Further investigation is necessary to fully understand the impact of feature selection on model performance and to determine the optimal feature set for accurate and reliable IT detection.

This study demonstrates a ensemble are powerful to learn from data and correctly identify UIT across all scenarios. The superior performance observed with the selected features in the smaller dataset may be attributed to a higher level of "curation" within this subset. The selection process may have inadvertently resulted in a more informative and less noisy subset of features, potentially leading to inflated identification rates. The performance of the proposed model are consistent with the findings of the DCZ study. The results are superior, potentially because the proposed data model ignores emphasis on qualitative features and discards time window framework proposed by DCZ. Interestingly, DCZ do not explicitly describe how temporal properties were incorporated or justified. This raises concerns regarding potential biases or misclassifications introduced by the time window approach. Despite these limitations, the DCZ study effectively utilized a carefully selected set of explainable features, grounded in economic theory. This approach demonstrated strong performance even with a reduced number of indicators.

4.3 Variable Importance

By design the RF is able to perform implicit variable selection by effectively identifying and utilizing a subset of strong covariates. By leveraging the inherent structure of the data, RF can eliminate irrelevant features while maintaining or even improving classification accuracy (Genuer, Poggi, and Tuleau-Malot, 2010). To illustrate this concept, experiment utilized inbuilt Gini impurity scores proposed by Breiman (2001) as a measure of feature importance. While Gini impurity scores provide a valuable

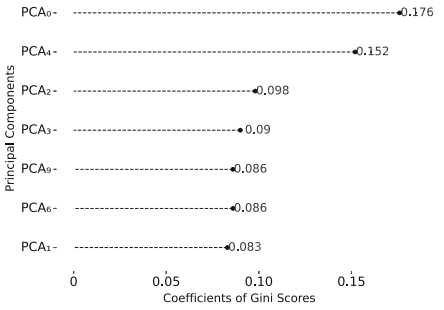
insights, it is crucial to consider alternative approaches for comparative reasons and assess the impact of feature interactions and potential noise within the data. By comparing the results obtained using Gini impurity scores with those obtained through permutation importance, one can gain deeper insights into the relative importance of different features and the potential impact of feature correlations and noise on model performance. This section provides detailed results from both methods.

4.3.1 Impurity Based Variable Importance

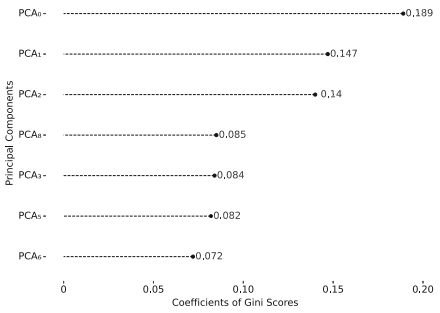
To illustrate the feature ranking process, we selected a representative experiment and compared the feature rankings obtained using Mean Decrease in Impurity (MDI) with those reported in Deng et al. (2021). Figure 2a and b visualize the ranked PCs for datasets with 25 and 110 features, respectively. The horizontal axis in both figures represents the Gini scores, which indicate the frequency with which a particular feature was selected for a split in the decision tree. Higher Gini scores correspond to greater feature importance.

In Fig. 2a, PCA_0 exhibits the highest rank, followed by PCA_4,PCA_2, and so on, down to PCA_1. For PCA_0, Table 6(a) reveals that features related to returns and acquisition/disposition activities exhibited the highest contributions. This observation aligns with established financial theories, as returns are a key determinant of investor behavior in the face of uncertainty (Foresi & Peracchi, 1995). Additionally, risk-related features, such as those associated with fair market value, were found to play a significant role in influencing IT decisions.

Similarly, PCA_0, the highest-ranking principal component in Table 6(b) for the full feature set (110 features), is significantly influenced by variables related to risk-taking behavior, such as returns, alpha, and beta. In Fig. 2b, PCA_0 and PCA_6 represent the



(a) 25 Features and 320 Transactions Match Previous Study



(b) 110 Features and 3984 Transactions

Fig. 2 This figure presents the ranking of PCs based on their Gini impurity scores. Gini impurity measures the probability of misclassifying a randomly chosen data point if it were labeled according to the class distribution. Mathematically, Gini impurity (G) is defined as: $G = \sum_{i=1}^{C} p(i) \times (1 - p(i))$ where C is the number of classes and $p(i)$ is the probability of belonging to class i

Table 6 An illustration of the contribution of individual covariates to the PCs (PCs)

Features	PCA_0	PCA_4	PCA_2	Features	PCA_0	PCA_1	PCA_2
Return	0.32	0.02	-0.11	Return	0.14	-0.14	0.03
Acq. Disp.	0.12	-0.39	-0.32	Alpha	-0.02	0.04	-0.02
Sprd. Rtn.	-0.34	-0.16	-0.16	Market Beta	-0.10	0.06	0.03
Market Beta	-0.31	0.15	-0.17	SMB Beta	-0.06	0.13	-0.05
SMB Beta	-0.02	0.40	0.16	HML Beta	-0.02	-0.02	-0.04
HML Beta	-0.37	0.07	-0.34	Idiosyncratic Volatility	-0.05	0.10	-0.05
Idiosyncratic Volatility	-0.39	0.06	-0.34	Total Volatility	-0.06	0.11	-0.05
Total Volatility	-0.01	-0.26	0.09	R Squared	-0.17	0.03	0.07
Prc. Op. Earnings(Basic)	-0.16	-0.49	0.04	Excess Returns	0.01	-0.06	0.02
Price Book	-0.29	0.02	0.28	IsDirector	-0.01	0.00	0.00

Panel (a) shows the contributions of covariates to PCs for the model trained on a subset of 25 features and 320 transactions. Panel (b) shows the contributions of covariates to PCs for the model trained on the full set of 110 features and 3984 transactions. PCs are ordered in descending order of their Gini scores, with the most influential PC appearing first

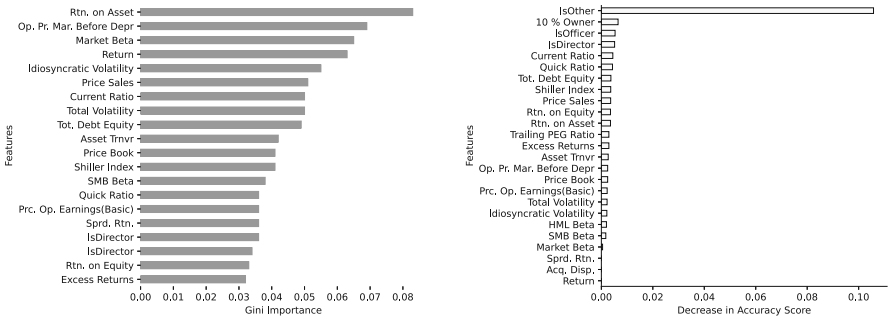
highest and lowest ranked PCs, respectively. The strong association between returns and PCA_0 aligns with the established notion that returns are a key determinant of investor behavior and asset pricing. This suggests that insiders may leverage their information advantage to influence asset prices, potentially leading to unlawful trading activities driven by short-term profit motives. However, it is important to note that highly cardinal and correlated features, such as those related to market risk, may exhibit inflated importance scores. While these features appear to have a strong influence, their actual impact on IT behavior may be less pronounced than initially suggested. Conversely, factors related to executive insider activity may have a more significant impact on unlawful trading than variables solely related to market risk and returns.

4.3.2 Permutation Based Variable Importance

While impurity-based feature importance provides valuable insights into the relative importance of PCs (as shown in Table 6), directly inferring the rank of individual features based on these scores can be challenging. Gini impurity scores vary across different components, making direct comparisons difficult. Furthermore, MDI methods are known to exhibit a bias towards high-cardinality features and their rankings are primarily based on the training data, potentially limiting their generalizability to unseen data. Moreover, the presence of highly correlated covariates within our financial dataset presents a unique challenge. When one of two correlated features is permuted, the model may still have access to the information contained within the other correlated feature. This can lead to an underestimation of the importance of both features, even if both are individually significant for the classification task (Meinshausen, 2008).

To address these limitations, we employed permutation importance, a model-agnostic method that assesses the importance of each feature by evaluating the impact of permuting its values while keeping all other predictors fixed. This approach provides a more robust and unbiased estimate of feature importance compared to traditional impurity-based methods. To further refine analysis, the process utilized hierarchical clustering based on Spearman rank correlation to identify and group highly correlated features. Spearman rank correlation is particularly suitable for this task as it measures the strength of the monotonic relationship between two variables based on their ranks. Within each cluster, a single representative feature was selected. The clustering process was performed using Ward's minimum variance method, which minimizes the variance within each cluster. The distance matrix for the clustering analysis was derived from the correlation matrix. This approach effectively addresses the challenges posed by correlated features and provides a more robust and reliable assessment of feature importance in the analysis.

Figure 3a illustrates the ranking of features based on MDI. Notably, features such as Return on Assets, Operating Profit Margin, and Market beta, which tend to exhibit a high degree of variability, consistently rank high in terms of Gini impurity. This observation suggests a potential bias towards high-cardinality features within the Gini impurity framework. Furthermore, the presence of high correlations between certain features, such as Return on Assets, Profit Margins, and market beta, can complicate the interpretation. To address these limitations, permutation importance is employed to assess feature importance. Figure 3b presents the feature rankings based on permutation importance. Interestingly, permutation importance reveals that low-cardinality categorical features, such as "IsOther," "Ten Percent Owner," and "IsOfficer," exhibit higher levels of importance compared to the MDI based ranking. A key observation is that permuting even the most highly ranked features based on Gini impurity resulted



(a) Gini Scores

(b) Permutation Values

Fig. 3 Feature Importance Rankings before removal of correlation in Training Data. This figure illustrates the ranking of features based on two different methods: (a) Gini impurity scores (Breiman, 2001) and (b) permutation importance. Both plots are derived from a randomly selected experiment with 3984 transactions and 110 features. For the comparative purposes between methods training data based plots are displayed

in only a 0.10 percent decline in accuracy, suggesting that the individual importance of these features may be overestimated. This finding contradicts the high test accuracy observed in Table 5 (99.13 percent with PCA and 98.13 percent without PCA), which suggests that the model's performance is not solely dependent on the importance of any single feature. To further investigate the impact of feature correlations, hierarchical clustering was employed to group highly correlated features. Subsequently, by selecting a representative feature from each cluster and re-running the permutation importance analysis on the reduced feature set. This approach aimed to mitigate the potential influence of correlated features on the feature importance rankings.

The right-hand side of Fig. 4 presents a heat-map illustrating the correlation matrix of the selected features. The diagonal of the heat-map represents perfect correlation. The left-hand side of Fig. 4 displays a dendrogram depicting the hierarchical clustering of the features. The dendrogram visually represents the relationships and similarities between different groups of variables, which are interconnected by "clades." For example, "Price Earnings (Basic)" and "Return on Equity (RoE)" are grouped together within a single clade, which then merges with "Trailing PEG Ratio" to form a larger clade. Similarly, "Acquisition and Disposition" and "Ten Percent Ownership" are grouped together in a separate clade. The height of each clade in the dendro-

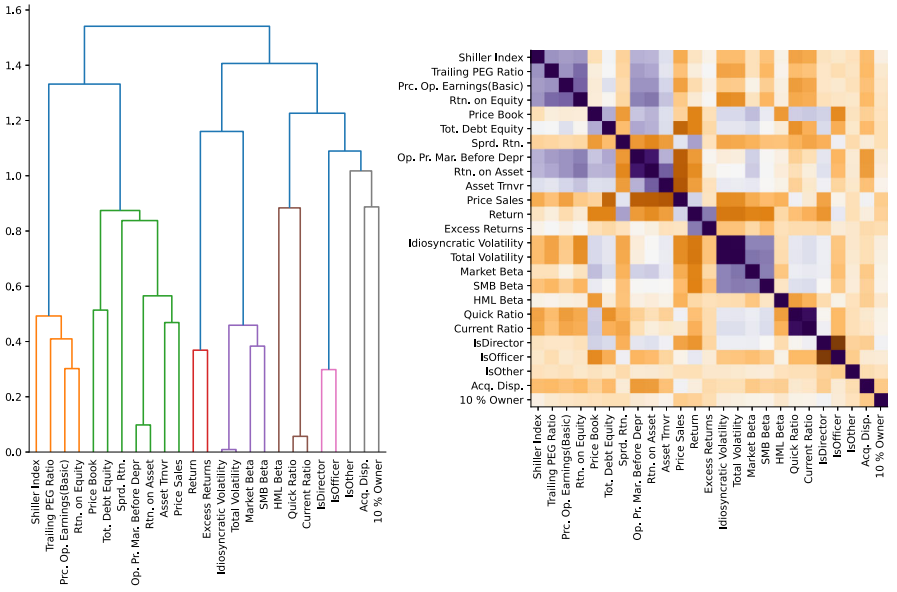


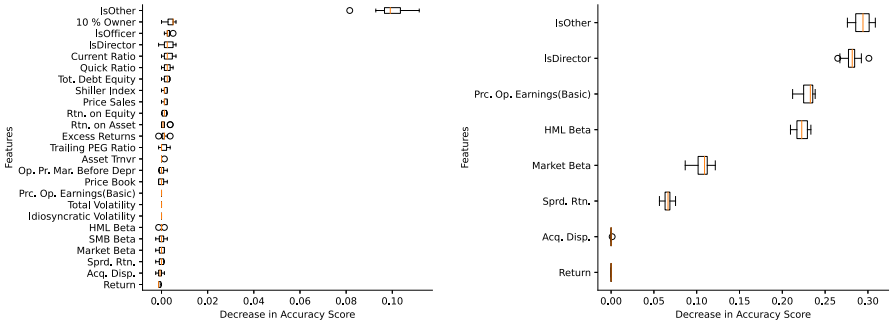
Fig. 4 Hierarchical Clustering and Correlation Matrix of Features. The left panel displays a dendrogram illustrating the hierarchical clustering of features based on Spearman rank-order correlations. The dendrogram visually represents the relationships between features, with similar features grouped together. The right panel presents a heat-map depicting the Spearman rank-order correlation matrix between a subset of selected features

gram reflects the distance between the clustered groups. Higher clade heights indicate greater dissimilarity between the grouped features.

The removal of correlated features can significantly impact model performance. The analysis revealed a potential accuracy drop of up to 0.3 percent when dealing with correlated features (Fig. 5b), compared to a 0.10 percent drop when dealing with uncorrelated features (Fig. 5a). Interestingly, while the relative importance of features such as "IsOther" and "Ten Percent Owner" remained consistent, their prominence became more apparent after decorrelation. These findings align with the observations of Strobl et al. (2008), who emphasized that permutation importance can be affected by correlations between predictor variables, particularly in high-dimensional datasets. Despite the impact of removal of correlation, the analysis consistently demonstrates the importance of financial statement variables in predicting UIT. Furthermore, variables related to market risk, such as market beta and the value premium (the spread between high and low book-to-market ratio companies), were found to be significant predictors of the UIT. These findings highlight the influence of market dynamics and institutional factors on IT decisions, as suggested by prior research on dividend policy, investor behavior, and the impact of institutional investors (Henry et al., 2017; Campbell & Shiller, 1988; Grinblatt et al., 1984).

5 Discussion

This study proposes two variations of the RF (RF) algorithm for the detection of UIT. These models leverage detailed trade and financial data, incorporating interpretable statistical scores to enhance feature selection and improve model performance. To evaluate the effectiveness of the proposed models, we conducted rigorous performance assessments, focusing on both balanced accuracy and sensitivity. Furthermore, we



(a) Ranks before removal of multicollinearity

(b) Ranks after removal of multicollinearity

Fig. 5 Ranking of the covariates based on descending order of the permutation values. The Fig. 5a is the ranking before the hierarchial clustering of covariates based on the Spearman rank-order correlation as shown in Fig. 4. The ranking shown in Fig. 5b is based on permutation values obtained after removing collinearity

replicated the methodology employed by DCZ to establish a meaningful benchmark for comparison. The replication of the DCZ method involved several enhancements, including semi-manual data labeling, the inclusion of transactions from multiple industries, and the removal of undefined qualitative features. With refined approach, the experiments were conducted in the dataset was then fed into an automated pipeline for model training and testing. The results demonstrate that the proposed models effectively uncover latent structures within the data, achieving superior performance compared to the original DCZ approach. Importantly, these models enhance interpretability by identifying a limited set of key features that are most relevant for detecting unlawful trading activity. The highlights of the results are:

- Low Fall-Out Rate:In the context of IT detection, minimizing false positives (i.e., incorrectly classifying unlawful transactions as lawful) is crucial. This is analogous to acquitting an individual who is actually guilty of a crime. Compared to benchmark methods, our proposed model demonstrates a consistently low FPR, as evident in Table 5. This indicates that the model effectively minimizes false alarms, reducing the risk of incorrectly classifying unlawful activities as legitimate. The low FPR of our model is particularly significant because it minimizes the reliance on arbitrary judgments and subjective thresholds often associated with traditional econometric studies. By minimizing false positives, our model can potentially assist regulatory bodies like the SEC in streamlining their investigative processes, reducing the burden of manual investigations, and redirecting resources towards more effective surveillance and the detection of previously not captured instances of the UIT.
- Low Miss Rates: A critical aspect of IT detection is minimizing the misclassification of lawful transactions as unlawful. This is analogous to the situation where an innocent individual is wrongfully incarcerated, which carries significant personal and societal costs. The proposed model exhibits a low miss rate, indicating that it rarely misclassifies lawful transactions as unlawful. This minimizes the risk of unnecessary investigations and the associated costs. While our model excels at minimizing false positives, it is important to acknowledge that further research is needed to further refine its ability to accurately identify true negatives.
- High Sensitivity: The proposed method exhibits high sensitivity, demonstrating a strong ability to correctly identify lawful transactions. This indicates a low rate of false negatives, minimizing the risk of incorrectly classifying lawful activities as unlawful. These findings provide strong evidence of high true positive rates and low false negative rates, demonstrating the model's effectiveness in accurately identifying and classifying lawful transactions. These results have significant implications for regulatory agencies. By leveraging the insights gained from this research, enforcement agencies can develop more robust systems and refine their investigative methodologies. This will enable them to focus their resources on investigating truly suspicious activities, enhancing the efficiency and effectiveness of their enforcement efforts.
- High Specificity: The results exhibits high specificity, demonstrating a strong ability to correctly identify unlawful transactions while minimizing the misclassification of lawful transactions as unlawful. This indicates a high proportion of true

negatives (correctly identified unlawful transactions) and a low number of false positives (unlawful transactions incorrectly classified as lawful).

- High Accuracy: Across all scenarios, the proposed method demonstrates superior classification accuracy compared to all benchmark methods. The analysis revealed that increasing the size of the dataset and removing time-window constraints significantly improved model performance. These findings have practical implications for regulatory agencies like the SEC, as they highlight the potential benefits of utilizing larger datasets and exploring alternative data sources in their enforcement efforts. Furthermore, even when maintaining a constant number of unlawful transactions and randomly sampling 50 percent of lawful transactions, the RF)model consistently outperformed all benchmark methods. These results underscore the effectiveness of the proposed approach in accurately classifying IT activities. In comparison to the method proposed by Deng et al. (2021), our model consistently demonstrates meaningful improvements in classification accuracy across all experimental scenarios.

In addition to the algorithm-specific feature importance rankings obtained using Mean Decrease in Impurity (MDI), the experiments employed a model-agnostic approach based on permutation importance. Unlike MDI, which is primarily based on training data, permutation importance can be extended to assess feature importance on unseen test data. This allows for a more robust evaluation of feature significance and provides insights into the potential impact of data leakage. By comparing feature rankings obtained using MDI and permutation importance, the observed a notable shift in the prominence of certain feature categories. Specifically, categorical features related to corporate governance gained significantly more prominence in the permutation importance rankings compared to the MDI-based rankings (Fig. 5a vs. Figure 5b). This observation highlights a key limitation of MDI: when dealing with correlated features, the importance of individual variables can be distorted. In such cases, one variable may contain information that is already captured by another correlated variable, leading to an underestimation of its true importance (Avanzi et al., 2023). Permutation importance, by contrast, provides a more robust assessment of feature significance by isolating the contribution of each individual feature while controlling for the influence of correlated variables. The analysis confirms the significance of governance, financial, and trading-related features in predicting IT behavior. Moreover, the introduction of additional features did not lead to overfitting, indicating that the model effectively learned to distinguish relevant from irrelevant information. This robustness was achieved through rigorous hyper-parameter tuning using 5-fold cross-validation, which minimized generalization error. In summary, this study extends the findings of the DCZ study by demonstrating the effectiveness of our proposed RF models in detecting UIT activities with improved accuracy, interpretability, and robustness.

6 Conclusions and Future Directions

This research demonstrates the significant potential of data-driven approaches to enhance unlawful insider trading (UIT) detection with high accuracy over 95 per-

cent, directly achieving its objectives. A fully automated system was developed for data labeling, preprocessing, and results analysis. Leveraging a comprehensive dataset of financial statements, market data, and corporate governance information, a Random Forest model was created that outperforms existing methods, achieving high accuracy, precision, and recall while minimizing false positives and negatives. This model uncovered hidden patterns in UIT behavior, providing insights into insider decision-making. The study expanded the feature space from 26 (in the benchmark study) to 110 by merging several databases , comparing results across multiple scenarios and sectors. To our knowledge, this is the first such application to the US capital market, broadening the geographic scope of analysis. Feature importance analysis identified corporate governance, financial performance, and market risk as key UIT predictors. These findings offer valuable insights for regulators, investors, and policymakers, enabling enhanced surveillance, streamlined investigations, and more effective regulations.

While this study offers valuable insights into unlawful insider trading (UIT) detection, several avenues for future research remain. Although the current Random Forest (RF) model demonstrates high accuracy, explaining the causal effects of individual features on predictions is crucial for regulatory and policy purposes (Athey et al., 2019). Future research should develop methods to address this. Exploring advanced hyper-parameter optimization techniques, such as Bayesian optimization, evolutionary algorithms, and grid search, could further improve model performance and computational efficiency. Incorporating a broader range of features, like the 447 anomaly-related features identified by Hou (2017), may provide deeper insights into IT drivers, potentially mitigating the curse of dimensionality. Furthermore, addressing the class imbalance problem inherent in real-world UIT occurrences is essential. While this study used a balanced dataset, future research should investigate model performance on imbalanced datasets. Finally, while this research highlights the potential of data-driven methods for fairer and more efficient financial markets by improving UIT detection and prevention, future work should also explore alternative machine learning algorithms, incorporate new data sources, and conduct more in-depth feature importance analyses to further refine UIT detection models, including addressing ethical considerations and model explainability. In conclusion, this research demonstrates the effectiveness of data-driven approaches, specifically advanced data mining and modeling techniques, in detecting UIT, paving the way for more efficient and effective methods of identifying and mitigating the risks associated with it.

Appendix

Selected Financial Indicators. A compiled list of financial indicators extracted from company financial statements, including balance sheets, income statements, and cash flow statements. * denotes variables that align with those utilized in the studies by Deng et al. (2021) and Deng et al. (2019). The indicators are categorized to assess various aspects of company performance, such as: profitability, valuation, liquidity, capitalization, financial soundness, growth, leverage, risk, returns, and corporate governance.

Group	Variables
Activity/Efficiency Ratios	Asset Turnover*, Inventory Turnover, Payables Turnover, Receivables/Current Assets
Annual Valuation Ratios	Shiller's P/E, Dividend Yield, Dividend Payout Ratio, Enterprise Value Multiple, Price-to-Cash Flow, Price-to-Earnings, excl. EI (diluted)*, Price-to-Earnings, incl. EI (diluted)*, Forward P/E to 1-year Growth (PEG) ratio*, Forward P/E to Long-term Growth (PEG) ratio*, Trailing PEG Ratio, Price-to-Sales Ratio*
Capitalization Ratios	Capitalization Ratio, Long-term Debt/Invested Capital, Common Equity/Invested Capital, Total Debt/Invested Capital
Financial Soundness Ratios	Cash Flow to Debt, Cash balance to Total Liabilities, Current Liabilities as Percentage of Total Liabilities, Total Debt as percentage of Total Assets, Gross debt to EBITDA, Long-term Debt/Book Equity, Free Cash Flow/Operating Cash Flow, Interest as Percentage of Average Long-term Debt, Interest as Percentage of Average Total Debt, Inventory/Current Assets, Long-term Debt/Total Liabilities, Total Liabilities/Total Tangible Assets, Operating Cash Flow to Current Liabilities, Profit before D&A to current liabilities, Receivables Turnover, Short-Term Debt/Total Debt
Liquidity Ratios	Cash Conversion Cycle, Cash Ratio, Current Ratio*, Quick Ratio (Acid Test)*, Quoted Spread
Miscellaneous Ratios	Accruals/Average Assets, Advertising as Percent of Sales, Market Capitalization, Price-to-Book Ratio*, Research and Development as percent of Sales, Sales per Dollar Total Stockholders' equity, Sales per Dollar Invested Capital, Sales per Dollar Working Capital, Labor Expenses/Sales
Ownership /Governance	Acquisition Disposition, Derivatives Held, Adjusted Derivatives Held, IsDirector, IsOfficer, IsOther, Ten Percent Ownership
Profitability Ratios and Rates of Return	After Tax Return on Average Common Equity, After Tax Return on Total Stock Holder's Equity, After Tax Return on Invested Capital, Alpha (Excess Return), Cash Flow Margin, Effective Tax Rate, Trailing PEG Ratio, Gross Profit Margin*, Gross Profit/Total Assets, Net Profit Margin, Operating Profit Margin After Depreciation*, Operating Profit Margin Before Depreciation*, Pre-tax Return on Total Earning Assets, Pre-tax return on Net Operating Assets, Pretax Profit Margin, Return on Assets*, Return on Capital Employed, Return on Equity*
Risk	Ask, Ask or High Price, Beta (High Minus Low)*, Market Beta*, Small-minus-big Size factor*, Bid, Bid Ask Spread, Bid or Low Price, Effective Spread, Excess Return from Risk Model*, Idiosyncratic volatility from the q-factor model, Kyle Lambda, Number of Derivatives, Number of Derivatives after Trade, Number of Trades, Price Impact, Market R-Squared, Realized Spread, Return, Returns without Dividend, Underlying Market Equity Volume, Underlying Shares Adjust, Outstanding Shares, Underlying Market Price, Underlying Market Price Adjust, Spread of Return, Total Volatility*, Volume, Exercise Price, Exercise Price Adjust
Shareholder's Equity, Invested Capital and Operating Cash Flow	Book-to-Market
Solvency Ratios	Debt-to-equity Ratio, Debt-to-assets, Debt-to-Capital, After Tax Interest Coverage, Interest Coverage Ratio
Valuation Ratios	Price-to-Operating EPS, excl. EI (basic), Price-to-Operating EPS, excl. EI (diluted)

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews. Computational statistics, 2*(4), 433–459.

Afendras, G., & Markatou, M. (2015). Optimality of training/test size and resampling effectiveness of cross-validation estimators of the generalization error.

Agrawal, A., Gans, J., & Goldfarb, A. (2019). *The economics of artificial intelligence: an agenda*. University of Chicago Press.

Ahern, K. R. (2017). Information networks: Evidence from illegal insider trading tips. *Journal of Financial Economics, 125*(1), 26–47.

Aït-Sahalia, Y., & Xiu, D. (2019). Principal component analysis of high-frequency data. *Journal of the American Statistical Association, 114*(525), 287–303.

Ali, U., & Hirshleifer, D. (2017). Opportunism as a firm and managerial trait: Predicting insider trading profits and misconduct. *Journal of Financial Economics, 126*. https://doi.org/10.1016/j.jfineco.2017.09.002

Altmann, A., Toloşi, L., Sander, O., et al. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics, 26*(10), 1340–1347.

Aluja-Banet, T., & Nonell-Torrent, R. (1991). *Local principal component analysis. Qüestioó, 3*, 267–278.

Amihud, Y., & Mendelson, H. (1987). Trading mechanisms and stock returns: An empirical investigation. *The Journal of Finance, 42*. https://doi.org/10.2307/2328369

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics, 34*(1), 122–148.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40–79.

Athey, S., Tibshirani, J., Wager, S. (2019). Generalized random forests. *The Annals of Statistics, 47*(2). https://doi.org/10.1214/18-AOS1709

Avanzi, B., Taylor, G., Wang, M., & Wong, B. (2023). *Machine learning with high-cardinality categorical features in actuarial applications*

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance, 61*(4), 1645–1680.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics, 131*(4)

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences, 116*(32), 15849–15854. https://doi.org/10.1073/pnas.1903070116

Bellman, R. (1958). Dynamic programming and stochastic control processes. *Information and Control, 1*(3), 228–239.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research, 13*(null), 281–305

Bernard, S., Heutte, L., & Adam, S. (2009). Influence of hyperparameters on random forest accuracy. *Multiple classifier systems: 8th international workshop, mcs 2009, reykjavik, iceland, june 10-12, 2009. proceedings 8* (pp. 171–180)

Billio, M., Getmansky, M., Lo, A. W., et al. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics, 104*(3), 535–559.

Bines, H. E. (1976). Modern portfolio theory and investment management law: Refinement of legal doctrine. *Columbia law review, 76*(5), 721–798.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy, 81*(3), 637–654.
Bondi, B. J., & Lofchie, S. D. (2011). The law of insider trading: legal theories, common defenses, and best practices for ensuring compliance. *NYU Journal of Law and Business, 8*(1)
Breiman, L. (2001). Random forests. *Machine Learning*, 5–32
Brillinger, D. R. (2001). *Time series: data analysis and theory*. SIAM.
Camerer, C. F. (2019). *Artificial Intelligence and Behavioral Economics*. The Economics of Artificial Intelligence: An Agenda. University of Chicago Press. https://doi.org/10.7208/chicago/9780226613475.003.0024
Campbell, J. Y., & Shiller, R. J. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *The Review of Financial Studies, 1*(3), 195–228.
Chang, M., Gould, J., Huang, Y., et al. (2022). Insider trading and the algorithmic trading environment. *International Review of Finance, 22*(4), 725–750.
Chen, X., Cho, Y. H., Dou, Y., & Lev, B. (2022). Predicting future earnings changes using machine learning and detailed financial data. *Journal of Accounting Research, 60*(2), 467–515.
Christensen, H. B., Hail, L., & Leuz, C. (2016). Capital-market effects of securities regulation: Prior conditions, implementation, and enforcement. *The Review of Financial Studies, 29*(11), 2885–2924.
Cohen, L., Malloy, C., & Pomorski, L. (2012). Decoding inside information. *The Journal of Finance, 67*(3), 1009–1043.
Cumming, D., Johan, S., & Li, D. (2011). Exchange trading rules and stock market liquidity. *Journal of Financial Economics, 99*(3), 651–671.
Deng, S., Wang, C., Fu, Z., et al. (2021). An intelligent system for insider trading identification in chinese security market. *Computational Economics, 57*(2), 593–616.
Deng, S., Wang, C., Li, J., et al. (2019). Identification of insider trading using extreme gradient boosting and multi-objective optimization. *Information (Basel), 10*(12), 367.
Driessen, J., Melenberg, B., & Nijman, T. (2003). Common factors in international bond returns. *Journal of International Money and Finance, 22*(5), 629–656.
Duchi, J., Shalev-Shwartz, S., Singer, Y., et al. (2008). Efficient projections onto the l 1 -ball for learning in high dimensions. *Proceedings of the 25th international conference on machine learning* (pp. 272–279). Acm.
Easley, D., Hvidkjaer, S., & O'Hara, M. (2002). Is information risk a determinant of asset returns? *The Journal of Finance, 57*. https://doi.org/10.1111/1540-6261.00493
Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association, 78*(382), 316–331.
Eggensperger, K., Lindauer, M., Hoos, H. H., et al. (2018). Efficient benchmarking of algorithm configurators via model-based surrogates. *Machine Learning, 107*(1), 15–41.
Egloff, D., Leippold, M., & Wu, L. (2010). The term structure of variance swap rates and optimal variance swap investments. *Journal of Financial and Quantitative Analysis, 45*, 1279–1310.
Fama, E. F. (1970). Multiperiod consumption-investment decisions. *The American Economic Review, 60*(1), 163–174.
Feeney, G. J. (1967). *Risk aversion and portfolio choice*. Wiley.
Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*
Fishman, M. J., & Hagerty, K. M. (1995). The mandatory disclosure of trades and market liquidity. *The Review of Financial Studies, 8*(3), 637–676.
Foresi, S., & Peracchi, F. (1995). The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association, 90*(430), 451–466.
Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics, 28*(2), 337–407.
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.
Fudenberg, D., & Liang, A. (2019). Predicting and understanding initial play. *The American Economic Review, 109*(12), 4112–4141.
Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine, 27*(15), 2865–2873.

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters, 31*(14), 2225–2236.
Gregory, A., Matatko, J., & Tonks, I. (1997). Detecting information from directors' trades: Signal definition and variable size effects. *Journal of Business Finance and Accounting, 24*(3), 309–342.
Grinblatt, M. S., Masulis, R. W., & Titman, S. (1984). The valuation effects of stock splits and stock dividends. *Journal of Financial Economics, 13*(4), 461–490.
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(Mar), 1157–1182
Hastie, T., Tibshirani, R., & Friedman, J. (2009). Statistics the elements of statistical learning. *The Mathematical Intelligencer, 27*
Henry, D., Nguyen, L., & Pham, V. H. (2017). Institutional trading before dividend reduction announcements. *Journal of Financial Markets, 36*. https://doi.org/10.1016/j.finmar.2017.07.003
Hou, K. (2017). *Replicating anomalies*. National Bureau of Economic Research.
Iskhakov, F., Rust, J., & Schjerning, B. (2020). Machine learning and structural econometrics: contrasts and synergies. *The Econometrics Journal, 23*(3), S81-s124.
John, K., & Narayanan, R. (1997). Market manipulation and the role of insider trading regulations. *The Journal of Business (Chicago, Ill.), 70*(2), 217–247
Khademian, A. M. (2022). *Sec and capital market regulation: The politics of expertise*. University of Pittsburgh Press
Koumou, G. B. (2020). Diversification and portfolio theory: a review. *Financial Markets and Portfolio Management, 34*(3), 267–312.
Kritzman, M., Li, Y., Page, S., et al. (2011). Principal components as a measure of systemic risk. *Journal of Portfolio Management, 37*(4), 112–126.
Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica, 53*. https://doi.org/10.2307/1913210
Langevoort, D. (1999). Rereading "cady, roberts": The ideology and practice of insider trading regulation. *Columbia Law Review, 99*(5), 1319–1343.
Lauar, F., & Arbex Valle, C. (2020). Detecting and predicting evidences of insider trading in the brazilian market. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 241–256)
Louzada, F., & Ara, A. (2012). Bagging k-dependence probabilistic networks: An alternative powerful fraud detection tool. *Expert Systems with Applications, 39*(14), 11583–11592.
Machan, T. R. (1996). What is morally right with insider trading. *Public Affairs Quarterly, 10*(2), 135–142.
Malhotra, P., Vig, L., Shroff, G. M., et al. (2015). Long short term memory networks for anomaly detection in time series. *Esann.*
Manne, H. G. (2005). Insider trading: Hayek, virtual markets, and the dog that did not bark. *The Journal of Corporation Law, 31*(1), 167.
Mayo, D. G., & Hand, D. (2022). Statistical significance and its critics: practicing damaging science, or damaging scientific practice? *Synthese, 200*(3), 220.
Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika, 95*(2), 265–278.
Nagy, D.M. (2020). Chiarella v. united states and its indelible impact on insider trading law. *Tennessee Journal of Law & Policy*, *15*(1), 6
Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics, 34*(21), 3711–3718. https://doi.org/10.1093/bioinformatics/bty373, https://academic.oup.com/bioinformatics/article-pdf/34/21/3711/48920785/bioinformatics_34_21_3711.pdf
Odian, E. A. (2011). Sec v. dorozhko's affirmative misrepresentation theory of insider trading: an improper means to a proper end. *Marquette Law Review, 94*(4), 1313
O'Neil, C. (2016). *Weapons of math destruction : how big data increases inequality and threatens democracy*. Crown.
Pasini, G. (2017). Principal component analysis for stock portfolio management. *International Journal of Pure and Applied Mathematics : IJPAM, 115*(1)
Perino, M. A. (2019). The lost history of insider trading. *University of Illinois Law Review, 2019*(3), 951.
Prasad, A., Iverson, L., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems (New York), 9*(2), 181–199.
Pérignon, C., Smith, D. R., & Villa, C. (2007). Why common factors in international bond returns are not so common. *Journal of International Money and Finance, 26*(2), 284–304.

Probst, P., Bischl, B., & Boulesteix, A-L. (2018). *Tunability: Importance of hyperparameters of machine learning algorithms.* arXiv:1802.09596

Probst, P., & Boulesteix, A. L. (2017). *To tune or not to tune the number of trees in random forest?*

Qian, H., Wang, B., Yuan, M., et al. (2022). Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree. *Expert Systems with Applications, 190*, Article 116202.

Ross, S. A. (1978). The current status of the capital asset pricing model (capm). *The Journal of Finance, 33*(3), 885–901.

Scornet, E. (2017). Tuning parameters in random forests. *ESAIM. Proceedings and Surveys, 60*, 144–162.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association, 88*(422), 486–494.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance (New York), 19*(3), 425.

Shiller, R. (1989). Market volatility mit press. *Cambridge Mass*

Shotton, J., Fitzgibbon, A., Cook, M., et al. (2011). Real-time human pose recognition in parts from single depth images. *Cvpr 2011* (pp. 1297–1304). Ieee

Stoyanovich, J., Howe, B., Abiteboul, S., et al. (2017). Fides: Towards a platform for responsible data science. *Proceedings of the 29th international conference on scientific and statistical database management* (pp. 1–6). Acm.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*(1), 307–307.

Strobl, C., Boulesteix, A. L., Zeileis, A., et al. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics, 8*(1), 25–25.

Sundarkumar, G. G., & Ravi, V. (2015). A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. *Engineering Applications of Artificial Intelligence, 37*, 368–377.

Svetnik, V., Liaw, A., Tong, C., et al. (2003). Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences, 43*(6), 1947–1958.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science, 185*(4157), 1124–1131.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Wang, Y., Sun, K., Yuan, X., et al. (2018). A novel sliding window pca-ipf based steady-state detection framework and its industrial application. *IEEE Access, 6*, 20995–21004.

Witten, I. H., Frank, E., Hall, M. A., et al. (2011). *Data mining: Practical machine learning tools and techniques*. San Francisco: Elsevier Science and Technology.

Xu, Z., Huang, G., Weinberger, K. Q., et al. (2014). Gradient boosted feature selection. *Proceedings of the 20th acm sigkdd international conference on knowledge discovery and data mining* (pp. 522–531).

Zhang, X., & Tong, H. (2022). Asymptotic Theory of Principal Component Analysis for Time Series Data with Cautionary Comments. *Journal of the Royal Statistical Society Series A: Statistics in Society, 185*(2), 543–565. https://doi.org/10.1111/rssa.12793

Zheng, Z., Podobnik, B., Feng, L., et al. (2012). Changes in cross-correlations as an indicator for systemic risk. *Scientific Reports, 2*(1), 888–888.

Zhou, S. (2022). *Random forests and regularization.* ProQuest Dissertations Publishing.