Journal of Economics, Finance and Accounting Studies
ISSN: 2709-0809
DOI: 10.32996/jefas
Journal Homepage: al-kindipublisher.com/index.php/jefas/index

JEFAS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

RESEARCH ARTICLE

Forward-Looking Financial Reporting Risk Screening in Saudi Listed Firms: Evidence from the Beneish M-Score and Explainable Machine Learning

Ibrahim Sweiti
Department of Accounting, College of Business and Economics, Qassim University, Qassim, Saudi Arabia
ORCID: 0000-0002-0014-3381
Corresponding Author: Ibrahim Sweiti, **E-mail:** im.alswaiti@qu.edu.sa

ABSTRACT

Most applications of the Beneish M-Score are contemporaneous: they screen financial statements after a reporting year has ended. This study asks whether the eight Beneish indicators measured in year t contain information about the following year's Beneish screening classification. The analysis uses a balanced panel of 66 Saudi listed non-financial firms from 2018 to 2025, yielding 528 annual financial statements and 462 M-Score observations for 2019-2025. Under the primary M > -1.78 rule, 66 of 462 firm-years (14.3%) generate a screening signal. On the untouched 2025 holdout, logistic regression achieves ROC-AUC = 0.718 and average precision = 0.371, while Random Forest achieves ROC-AUC = 0.684 and average precision = 0.406 and produces the higher precision and F1 at the pre-specified 0.50 decision-score threshold. A rolling-origin stability analysis across 2021-2025 uses the continuous current-year M-Score as the benchmark: pooled ROC-AUC is 0.537 for M_t, 0.661 for logistic regression, and 0.653 for Random Forest. Firm-cluster bootstrap inference indicates that both fitted models improve discrimination relative to M_t, whereas their difference is not statistically distinguishable. Unweighted refits produce similar ranking performance. SHAP identifies TATA as the leading Random Forest feature, followed by DEPI, AQI, and DSRI. The evidence supports explainable machine learning as a complementary prioritization tool rather than a calibrated fraud probability or a substitute for professional review.

KEYWORDS

Beneish M-Score; financial reporting risk; screening; Random Forest; explainable artificial intelligence; SHAP; Saudi Arabia; IFRS; forensic accounting

ARTICLE INFORMATION

ACCEPTED: 16 September 2026 **PUBLISHED:** 16 October 2026 **DOI:** 10.32996/jefas.2026.8.10.3

1. Introduction

Financial reports are expected to reduce information asymmetry, yet they are also produced in an environment where managers exercise judgment over estimates, classifications, timing, and recognition. That tension is familiar in accounting research. Agency theory explains why managers and outside capital providers may not have identical incentives (Jensen & Meckling, 1976), while the earnings-management literature shows that reporting discretion can range from defensible judgment to aggressive reporting and, in more serious cases, material misstatement (Healy & Wahlen, 1999; Dechow et al., 1996). The practical challenge is therefore not to treat every unusual accounting outcome as misconduct, but to identify cases that deserve closer examination.

The Beneish M-Score is attractive for that purpose because it is transparent and inexpensive to implement. Its eight components summarize changes in receivables, margins, asset quality, sales growth, depreciation, selling and administrative costs, accruals, and leverage (Beneish, 1999). A high score is not proof of fraud; it is a screening signal. Beneish (1999) explicitly frames the model as a way to identify statements whose accounting relationships resemble those observed among known manipulators, leaving the underlying economic explanation to subsequent analysis. Later evidence shows that the score also contains information relevant to capital-market outcomes (Beneish et al., 2013).

Fraud-prediction research has since moved well beyond fixed scoring rules. Statistical learning and machine-learning models can capture interactions and nonlinearities that a single linear score cannot. At the same time, the literature is clear that apparent predictive gains can be misleading when fraud is rare, class proportions are artificial, or future information leaks into model development (Perols, 2011; Perols et al., 2017; Shahana et al., 2023). Large-sample accounting studies show that machine

AL-KINDI

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

learning can add useful predictive structure (Bao et al., 2020; Bertomeu et al., 2021), but its value depends on the decision setting and on the cost of false positives and false negatives (Beneish & Vorst, 2022).

Interpretability is a separate concern. A model that labels a company as risky without indicating which accounting relationships drove that judgment has limited professional value. SHAP provides a principled way to decompose a prediction into feature-level contributions (Lundberg & Lee, 2017). Recent fraud-detection studies increasingly pair predictive models with explanations that can be understood by auditors, regulators, and investors (Zhou et al., 2023; Cai & Xie, 2024). In a forensic setting, this matters because an explanation can turn a risk score into a concrete review question—for example, whether the concern arises from accruals, receivables, depreciation behavior, or asset composition.

Saudi Arabia is a useful setting in which to examine this combination of forensic screening and predictive analytics. The Capital Market Authority required listed companies to prepare financial statements under IFRS as endorsed in Saudi Arabia from 1 January 2017 (CMA, 2016). Research on the transition documents important implementation challenges and changes in the information environment (Nurunnabi et al., 2020; Alomair et al., 2022; Nurunnabi et al., 2022). More recent Saudi evidence, based on 111 listed non-financial firms, reports stronger financial-reporting quality after mandatory IFRS adoption and shows that audit-committee effectiveness is positively associated with the reporting-quality consequences of the new regime (Alaskar, 2026). The present sample begins in 2018, so the analysis is conducted entirely within the post-adoption reporting regime rather than mixing local-GAAP and IFRS observations.

The design addresses a problem that is easy to overlook in AI applications of the Beneish model. Because the M-Score is an algebraic function of the eight Beneish indicators, training a machine-learning model to reproduce the same-year M-Score class from those same indicators is largely a reconstruction exercise. This study instead asks whether the indicators carry information beyond the year in which they are measured. Current-year Beneish variables are used to predict whether the firm crosses the screening boundary in the following year, and the final year remains untouched until evaluation. The task is therefore harder than same-year classification, but it is closer to the practical question faced by an auditor, regulator, or analyst deciding where to direct limited attention before the next reporting cycle is known.

The findings are deliberately not presented as a simple 'AI wins' result. On the final 2025 holdout, logistic regression ranks observations better by ROC-AUC, whereas Random Forest attains the higher average precision and a more selective precision-F1 trade-off at the pre-specified operating threshold. A rolling-origin stability analysis also shows that both fitted models improve pooled discrimination relative to the continuous current-year M-Score benchmark, but the difference between logistic regression and Random Forest is not statistically distinguishable. This pattern is more informative than a single winner: the value of a screening model depends on the review objective, the cost of missed cases, and the burden created by false alerts.

The next section develops the accounting and prediction literature. Section 3 explains the data, M-Score construction, temporal validation, and SHAP analysis; Sections 4 and 5 report and interpret the results; the final sections set out the implications, limitations, and conclusions.

2. Literature Review and Conceptual Development

2.1 Financial-reporting incentives and forensic screening

Agency conflicts create the setting in which reporting incentives arise, but they do not imply that discretion is inherently improper. Financial reporting requires estimates for revenue recognition, impairment, provisions, useful lives, capitalization, and accrual measurement. Problems arise when discretion is used to alter the economic impression conveyed to outsiders. Healy and Wahlen (1999) place managerial judgment at the center of earnings management, while enforcement evidence links serious reporting problems to financing incentives and governance weaknesses (Dechow et al., 1996). A useful screening model therefore needs to be sensitive to unusual accounting patterns without treating those patterns as proof of intent.

Forensic models operationalize that idea by translating reporting patterns into observable signals. Dechow et al. (2011) show that combinations of accounting variables can help predict material misstatements. Beneish (1999) takes a compact and highly interpretable route: eight year-over-year indices summarize the relationships most likely to change when reported performance becomes unusually aggressive. The appeal of the M-Score is not that it settles the fraud question, but that every component can be traced back to a financial-statement relationship that an accountant can investigate.

2.2 The Beneish M-Score and the meaning of a screening signal

The Beneish framework developed in stages. Beneish (1997) showed that published financial-statement information could help identify accounting profiles associated with GAAP violations among firms with extreme performance. The 1999 study formalized the familiar eight-variable specification, combining DSRI, GMI, AQI, SGI, DEPI, SGAI, TATA, and LVGI into one screening score. This study uses $M > -1.78$ as its primary screening boundary. Crossing that boundary identifies an accounting profile that merits additional review; it does not establish fraud, intent, or a regulatory violation.

Subsequent evidence shows why that distinction matters. Beneish et al. (2013) document capital-market information in the manipulation probability captured by the score. Repousis (2016) demonstrates that the model can be implemented at scale outside the United States, while Marais et al. (2023) find much weaker sensitivity and precision in South Africa. Taken together, these studies suggest that the M-Score is portable as a screening framework, but its empirical performance is not institution-free. That is the basis for treating Saudi results as context-specific evidence rather than as a universal fraud classification rule.

2.3 Statistical learning and machine learning in fraud detection

The financial-fraud literature has compared statistical and machine-learning methods for more than a decade. Cecchini et al. (2010) show that nonlinear classifiers can identify management-fraud cases from financial data. Perols (2011) demonstrates that rankings among algorithms change with fraud prevalence and misclassification costs, and Perols et al. (2017) show that data preparation and variable selection can be as consequential as the classifier itself. These findings are especially relevant here because the positive class is small and the intended use is screening rather than retrospective explanation.

More recent accounting studies provide stronger benchmarks for prospective prediction. Bao et al. (2020) combine accounting knowledge with machine learning in a large U.S. fraud sample. Bertomeu et al. (2021) examine one- and two-year-ahead misstatement prediction and show that interactions among accounting, market, governance, and audit variables can matter. Beneish and Vorst (2022) shift attention from statistical fit to the economic cost of prediction errors. Recent applications also show continued interest in ensemble methods: Özari et al. (2025) combine Beneish- and Altman-related variables with Random Forest and compare classifications with Borsa Istanbul fines, while Steingen and Löw (2025) assess machine-learning fraud detection in a cross-country setting. These studies motivate the use of a transparent logistic benchmark, a nonlinear Random Forest, several complementary metrics, and a future-period holdout.

2.4 Explainable artificial intelligence and professional usability

Prediction alone is not enough in a high-stakes accounting setting. Auditors and regulators need to know why a model regards an observation as risky, both to design follow-up procedures and to avoid treating a statistical score as authoritative evidence. Lundberg and Lee (2017) develop SHAP as a unified feature-attribution framework. It can summarize which variables matter most across a sample and, when needed, explain an individual prediction.

That link between prediction and explanation is increasingly visible in fraud research. Zhou et al. (2023) explicitly design Shapley-based explanations around stakeholder needs, and Cai and Xie (2024) connect explainable fraud detection to the internal relationships among financial-statement variables. The present study keeps that logic close to the accounting model: the machine-learning stage is not introduced as a replacement for forensic judgment, but as a test of whether familiar Beneish signals interact in a way that is useful for next-year risk prioritization.

2.5 Saudi Arabia, IFRS, and the research gap

Saudi listed companies entered the IFRS reporting regime in 2017 under a market-wide transition supported by the CMA (CMA, 2016). Nurunnabi et al. (2020) document implementation challenges among listed firms; Alomair et al. (2022) study changes in value relevance around adoption; and Nurunnabi et al. (2022) examine the reporting effects of mandatory IFRS. Alaskar (2026) adds recent evidence from 111 Saudi non-financial firms, finding improved reporting quality after mandatory IFRS adoption and a positive moderating role for audit-committee effectiveness. Together, these studies show that the institutional setting changed materially after 2017, but they do not remove the need for forensic screening: IFRS still relies on estimates and managerial judgment, and the meaning of unusual accounting ratios remains shaped by firm economics, governance, and enforcement.

What remains less developed is a Saudi test that links an interpretable forensic score to genuinely prospective machine learning. Regional evidence also cautions against assuming that one screening model performs uniformly across settings. Using 365 GCC companies and 1,095 firm-year observations, Hakami et al. (2020) compare Beneish, Dechow, and Altman-based signals with audit opinions and show that detection performance varies by model and audit context. Existing reviews likewise emphasize class imbalance, target definition, and validation design as recurring weaknesses in financial-statement fraud research (Shahana et al., 2023). Emerging applications show that Random Forest and other ensemble methods can be useful, but performance depends on the target and the institutional setting (Özari et al., 2025; Steingen & Löw, 2025). The gap addressed here is therefore methodological as well as geographic: the model must predict a future outcome, preserve a transparent benchmark, and explain the accounting content of its predictions.

Accordingly, the study addresses four research questions:

RQ1. How frequently do Saudi listed non-financial firms generate Beneish screening signals during 2019-2025, and how stable are those signals over time?

RQ2. Do the eight Beneish indicators in year t improve one-year-ahead discrimination relative to the continuous current-year M-Score benchmark?

RQ3. How stable are the logistic-regression and Random Forest results across rolling origins and on the untouched 2025 holdout when ranking and classification metrics are considered together?

RQ4. Which Beneish indicators contribute most to Random Forest predictions according to SHAP?

3. Methodology

3.1 Research setting, data, and analytical sample

The empirical setting is the Saudi equity market after mandatory IFRS adoption. The financial-statement dataset was compiled by the author from publicly available audited annual reports and Saudi market disclosures for listed non-financial firms from 2018 through 2025. The starting cleaned analytical file contained 81 symbol labels and 648 firm-year observations. Identity normalization reduced these labels to 80 unique companies; one unresolved duplicate identity with conflicting financial values was then removed. The final balanced panel additionally requires continuous eight-year coverage and finite values for all eight Beneish indicators in every scored year. Thirteen companies fail that model-specific requirement because AQI is undefined in at least one year, leaving 66 firms and 528 annual financial statements. No statistical imputation is used in the final sample.

Beneish indicators require a prior-year comparison, so 2018 serves as the base year and the first M-Score is calculated for 2019. The forensic panel therefore contains 462 M-Score observations (66 firms x 7 years, 2019-2025). Financial firms are outside the study because their balance-sheet structure, regulatory capital requirements, and operating model make the standard Beneish relationships less comparable with those of non-financial companies. Table 1 reports both the sample-construction flow and the final analytical structure.

Table 1. Study design and analytical structure

Item	Analytical structure
Panel A. Sample construction	
Starting cleaned analytical file	81 symbol labels; 648 firm-years
After identity normalization	80 unique companies
After unresolved duplicate-identity removal	79 companies
After requiring finite eight-variable Beneish inputs in every scored year	66 firms; 528 firm-years
Panel B. Final design	
Market	Saudi listed non-financial firms
Financial-statement window	2018-2025
M-Score years	2019-2025
M-Score observations	462
Forecasting pairs	396 (2019 predictors through 2024 predictors; targets 2020-2025)
Final temporal test	2024 indicators predicting the 2025 screening class

3.2 Beneish indicators and M-Score specification

The eight-variable Beneish model is calculated from annual sales, receivables, cost of goods sold, current assets, total assets, net property, plant and equipment, depreciation expense, selling/general/administrative expense, current liabilities, long-term debt, net income, and operating cash flow. Table 2 reports the operational formulas. All ratios are calculated firm by firm in chronological order so that each observation uses only information from the current and immediately preceding financial year.

Table 2. Beneish M-Score indicators

Code	Indicator	Operational formula	Economic interpretation
DSRI	Days Sales in Receivables Index	(Receivables_t / Sales_t) / (Receivables_t-1 / Sales_t-1)	Unusual growth in receivables relative to sales
GMI	Gross Margin Index	Gross margin_t-1 / Gross margin_t	Deteriorating margins may increase pressure to manage earnings
AQI	Asset Quality Index	[1 - (CA_t + PPE_t)/TA_t] / [1 - (CA_t-1 + PPE_t-1)/TA_t-1]	Growth in less tangible/non-current asset components
SGI	Sales Growth Index	Sales_t / Sales_t-1	High growth can increase expectations and financing pressure

Code	Indicator	Operational formula	Economic interpretation
DEPI	Depreciation Index	Dep. rate_t-1 / Dep. rate_t	Changes consistent with slower depreciation recognition
SGAI	SG&A Index	(SG&A_t / Sales_t) / (SG&A_t-1 / Sales_t-1)	Changes in operating cost intensity
TATA	Total Accruals to Total Assets	(Net income_t - CFO_t) / TA_t	Accrual intensity: divergence between accounting earnings and operating cash flow
LVGI	Leverage Index	Leverage_t / Leverage_t-1	Changes in debt pressure and covenant incentives

Notes: CA = current assets; PPE = net property, plant and equipment; TA = total assets; CFO = operating cash flow. For TATA, net income is used consistently because income from continuing operations is not separately available for every firm-year in the source data. This is a practical deviation from the original continuing-operations formulation and is applied uniformly across the sample. A sensitivity test using the original earnings definition cannot be implemented from the present dataset without introducing a non-equivalent proxy; operating profit, for example, is not a valid substitute for income from continuing operations. The prominence of TATA in the later SHAP analysis is therefore interpreted with this measurement limitation in mind.

The composite score is calculated as:

M = -4.84 + 0.920(DSRI) + 0.528(GMI) + 0.404(AQI) + 0.892(SGI) + 0.115(DEPI) - 0.172(SGAI) + 4.679(TATA) - 0.327(LVGI)

The coefficients above are those of the published eight-variable Beneish specification (Beneish, 1999). The primary classification in this study is M > -1.78. Throughout the paper, an observation above the cutoff is called a 'Beneish screening signal' rather than a case of manipulation or fraud. The label refers to an accounting profile that warrants follow-up; it does not observe managerial intent, audit findings, restatements, or regulatory enforcement.

3.3 Descriptive analysis and robustness treatment

Descriptive analysis reports annual medians, means, standard deviations, counts, and screening rates. Medians receive particular attention because ratio-based indices can become extreme when a prior-year denominator is small. The main analysis uses the raw finite ratios: no trimming, clipping, or winsorization is applied before the primary M-Score or prediction results are produced. No statistical imputation is applied to the final analytical panel. Distributional diagnostics include skewness and excess kurtosis, and sector results are descriptive and are shown only where at least three firms are represented. As a robustness check, each of the eight indicators is winsorized at the pooled 1st and 99th percentiles and the M-Score is recomputed; robustness is assessed by the number of classifications that change.

3.4 Forward-looking prediction design

The predictive design deliberately avoids same-year circularity. If the eight Beneish indicators were used to predict the same-year M-Score class, the target would be an algebraic function of the predictors. Instead, the indicators in year t are matched to the risk class in year t+1. With six predictor years (2019–2024), this produces 396 one-year-ahead firm-year pairs and turns the exercise into a genuine forecasting problem.

Observations with target years 2020-2024 form the development period (330 pairs). The 2025 target year, predicted from 2024 indicators, is held out entirely until final evaluation (66 firms). Hyperparameter selection follows expanding-window validation: validation years are 2022, 2023, and 2024, and each is predicted using only earlier target years (2020-2021 -> 2022; 2020-2022 -> 2023; 2020-2023 -> 2024). Mean validation ROC-AUC is used for selection. The final specifications are then refit on the full 2020-2024 development period before the untouched 2025 evaluation. Logistic regression uses standardized predictors, class-weight balancing, and C = 0.10. Random Forest uses 500 trees, maximum depth = 3, minimum leaf size = 10, square-root feature sampling, balanced-subsample class weights, and random_state = 42.

Three ranking approaches are compared. The primary benchmark is the continuous current-year M-Score (M_t), which retains more information than a binary persistence rule and requires no estimation. Logistic regression provides an interpretable linear benchmark, while Random Forest provides a nonlinear alternative capable of capturing interactions without imposing a functional form (Breiman, 2001). The fitted models use class weights to reduce majority-class dominance. Because class weighting changes the probability scale, the resulting outputs are treated as decision scores rather than calibrated event probabilities. The paper therefore does not use Brier scores to claim probability-calibration superiority. At the pre-specified 0.50 decision-score threshold, precision, recall, F1, specificity, accuracy, and balanced accuracy are reported; thresholds of 0.30, 0.40, and 0.60 are shown as descriptive operating-point checks and are not selected using the 2025 outcomes.

Evaluation emphasizes ROC-AUC and average precision (AP) for ranking, together with class-sensitive threshold metrics. AP is particularly useful because the positive class is uncommon. Uncertainty on the 2025 holdout is assessed with 5,000

nonparametric bootstrap resamples, including intervals for ROC-AUC, AP, precision, recall, and F1. A supplementary rolling-origin stability analysis then refits model parameters on expanding historical windows for target years 2021-2025 while keeping the final hyperparameter architecture fixed. This diagnostic is used to assess whether the 2025 result is an isolated year effect; it is not presented as a second independent hyperparameter-selection exercise. The pooled rolling predictions are evaluated with a 5,000-resample firm-cluster bootstrap so that repeated observations from the same company remain together. Finally, unweighted refits are reported as a robustness check on whether class weighting drives the discrimination results.

3.5 Explainability

TreeSHAP values are calculated for all 66 observations in the untouched 2025 Random Forest holdout. Global importance is the mean absolute SHAP value for each Beneish indicator. Feature-SHAP correlations are reported only as descriptive direction checks. Neither SHAP magnitude nor its sign is interpreted causally: the analysis describes how the fitted model uses the accounting inputs, not why a company reports a particular accounting pattern.

4. Results

4.1 Descriptive properties of the Beneish indicators

Table 3. Descriptive statistics for the forensic indicators (N = 462)

Variable	Mean	SD	Q1	Median	Q3
DSRI	1.035	0.424	0.852	1.000	1.141
GMI	1.665	13.669	0.803	0.992	1.124
AQI	2.266	13.144	0.924	1.011	1.154
SGI	1.740	12.541	0.933	1.046	1.177
DEPI	1.055	0.571	0.919	0.979	1.074
SGAI	1.023	0.308	0.885	0.996	1.139
TATA	-0.040	0.083	-0.080	-0.041	0.004
LVGI	1.139	1.899	0.935	1.005	1.117
M-Score	-1.157	18.800	-2.932	-2.548	-2.182

The forensic indices are strongly non-normal. GMI, AQI, SGI, and LVGI have means well above their medians because a small number of firm-years produce very large ratios. The typical observation is much less extreme: most median indices lie close to one, and median TATA is -0.041. Distributional diagnostics reinforce this point. Skewness is 19.31 for GMI, 20.97 for SGI, 20.87 for LVGI, and 18.66 for the M-Score; the corresponding excess kurtosis values are 392.57, 445.84, 444.03, and 378.60. The median M-Score is -2.548, below the -1.78 screening boundary. These heavy tails are why the analysis emphasizes medians, classification robustness, and out-of-sample ranking rather than interpreting arithmetic means in isolation.

4.2 Annual Beneish screening signals

Table 4. Annual Beneish M-Score results

Year	N	Median M	Mean M	Signals (M > -1.78)	Signal rate
2019	66	-2.555	5.352	17	25.8%
2020	66	-2.649	-2.251	9	13.6%
2021	66	-2.545	-1.075	9	13.6%
2022	66	-2.548	-1.970	10	15.2%
2023	66	-2.551	-2.504	5	7.6%
2024	66	-2.537	-3.259	6	9.1%
2025	66	-2.425	-2.389	10	15.2%

Across 462 firm-years, 66 observations (14.3%) cross the -1.78 screening threshold. The rate is 25.8% in 2019, 13.6% in both 2020 and 2021, 15.2% in 2022, 7.6% in 2023, 9.1% in 2024, and 15.2% in 2025. The pattern is not monotonic, and annual medians move far less than annual means. At the firm level, 43 of the 66 companies cross the screening boundary at least once, but only 17 do so in more than one year; no company is flagged in more than three of the seven scored years. Persistence is also limited: among 56 year-to-year transitions that begin with a screening signal, only 11 (19.6%) remain above the threshold in the following year. The evidence therefore points to episodic screening signals rather than a fixed firm label.

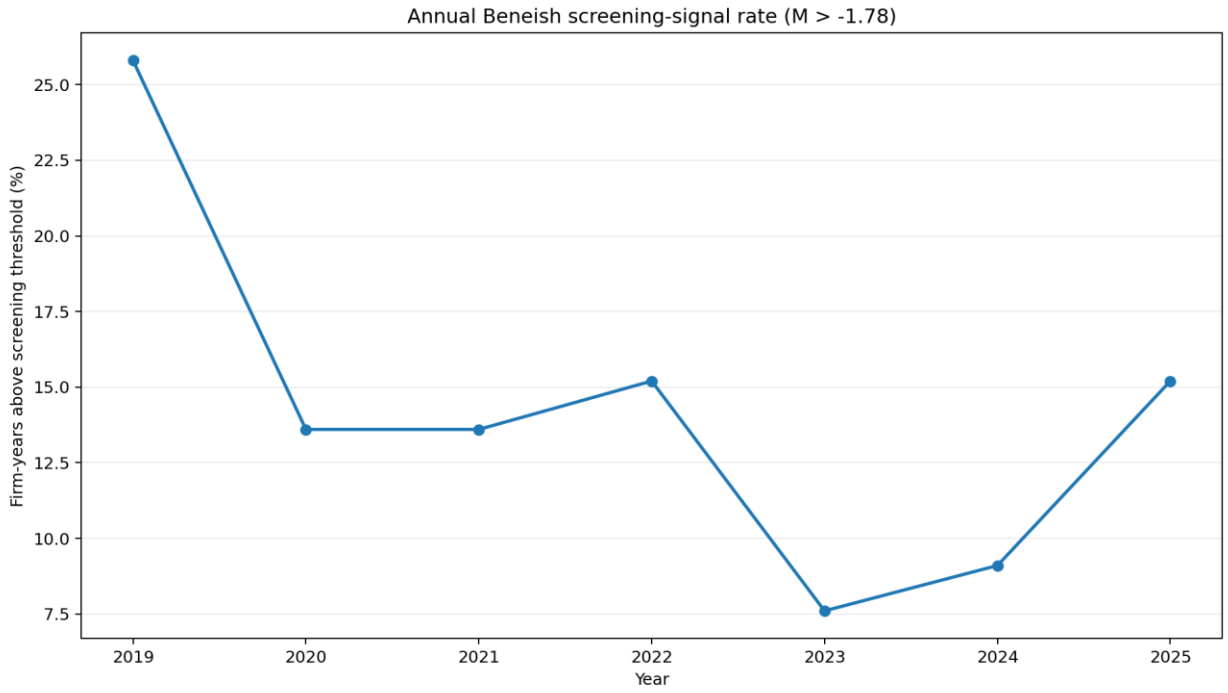

Figure 1. Annual prevalence of Beneish screening signals (M > -1.78).

These annual percentages should not be read as estimates of fraud prevalence. They show how often the accounting profile of a firm-year crosses the Beneish screening boundary. The fall after 2019 and the rise in 2025 are descriptive facts in this sample; the research design does not identify the macroeconomic, sectoral, or firm-specific causes of those changes.

4.3 Sector-level descriptive patterns

Table 5. Sector-level Beneish screening signals for sectors with at least three firms

Sector	Firms	Firm-years	Signals	Rate
Consumer Durables & Apparel	3	21	5	23.8%
Health Care Equipment & Services	3	21	4	19.0%
Consumer Staples Distribution & Retail	4	28	5	17.9%
Transportation	5	35	6	17.1%
Real Estate Management & Development	6	42	7	16.7%
Food & Beverages	8	56	8	14.3%
Consumer Services	6	42	6	14.3%
Materials	20	140	15	10.7%
Telecommunication Services	3	21	1	4.8%

Note: Sector comparisons are descriptive; cell sizes are not large enough to support strong cross-sector inference.

Among sectors represented by at least three firms, Consumer Durables & Apparel has the highest descriptive signal rate (23.8%), followed by Health Care Equipment & Services (19.0%) and Consumer Staples Distribution & Retail (17.9%). Materials, the largest sector in the sample, records 10.7%. These figures are useful for describing where signals appear, but the cell sizes are too small for causal or ranking claims. Sector economics, working-capital cycles, investment intensity, and accounting estimates can all affect the component ratios.

4.4 Untouched 2025 predictive performance

The untouched 2025 holdout contains 66 firms and ten positive screening cases (15.2%). The continuous current-year M-Score provides a more informative benchmark than carrying forward a binary class: its 2025 ROC-AUC is 0.509 and average precision is 0.197, essentially showing little one-year-ahead ranking power on its own. An always-negative majority-class rule would achieve 84.8% accuracy while detecting none of the ten positive cases (recall = 0; balanced accuracy = 0.500), illustrating why raw accuracy is not a meaningful primary criterion in this setting. Both fitted models produce materially more informative rankings than the continuous M_t benchmark.

Table 6. Predictive performance on the untouched 2025 temporal holdout

Model	ROC-AUC	Average precision	Accuracy	Bal. Acc.	Precision	Recall	F1
Continuous M_t score	0.509	0.197	—	—	—	—	—
Logistic Regression	0.718	0.371	0.667	0.639	0.250	0.600	0.353
Random Forest	0.684	0.406	0.803	0.679	0.385	0.500	0.435

Note: The continuous M_t benchmark is evaluated as a ranking score and therefore has no 0.50 classification metrics. Logistic regression and Random Forest use the pre-specified 0.50 decision-score threshold for the classification columns. Higher values indicate better performance. The average-precision baseline for a random ranking equals the positive-class prevalence in the 2025 holdout (10/66 = 0.152).

The two fitted models behave differently. Logistic regression has the higher ROC-AUC (0.718 versus 0.684), indicating better overall ranking across thresholds, whereas Random Forest has the higher average precision (0.406 versus 0.371). At the 0.50 decision-score threshold, Random Forest is more selective: precision is 0.385 versus 0.250, F1 is 0.435 versus 0.353, and balanced accuracy is 0.679 versus 0.639. Logistic regression has higher recall (0.600 versus 0.500). In counts, logistic regression identifies six of the ten positive cases but produces 18 false positives; Random Forest identifies five positives with eight false positives. The practical distinction is therefore broader detection versus a shorter follow-up list, not an unqualified ranking of one algorithm over the other.

Sampling uncertainty is substantial because the holdout contains only ten positive cases. In 5,000 percentile-bootstrap resamples, the 95% interval for logistic ROC-AUC is 0.526-0.886 and the interval for Random Forest is 0.455-0.886; the paired interval for the ROC-AUC difference (Random Forest minus logistic regression) is -0.127 to 0.055. Uncertainty is similarly wide for threshold-based metrics. For logistic regression, the bootstrap intervals are 0.080-0.438 for precision, 0.250-0.900 for recall, and 0.129-0.550 for F1. The corresponding Random Forest intervals are 0.111-0.667, 0.167-0.833, and 0.133-0.667. These intervals caution against fine-grained claims of model superiority.

Table 6A. Rolling-origin stability of one-year-ahead discrimination

Target year	N	Positive cases	Continuous M_t AUC	Logistic AUC	Random Forest AUC
2021	66	9	0.628	0.614	0.569
2022	66	10	0.493	0.596	0.570
2023	66	5	0.705	0.843	0.816
2024	66	6	0.367	0.569	0.633
2025	66	10	0.509	0.718	0.684
Pooled	330	40	0.537	0.661	0.653

Note: This is a stability diagnostic of the final frozen model architectures. At each origin, model parameters are refit using only observations from earlier target years; hyperparameters are not re-optimized retrospectively. The continuous M_t benchmark requires no fitting.

Across the pooled rolling-origin predictions, the continuous M_t benchmark has ROC-AUC = 0.537 and AP = 0.184. Logistic regression reaches ROC-AUC = 0.661 and AP = 0.263, while Random Forest reaches ROC-AUC = 0.653 and AP = 0.236. In a 5,000-resample firm-cluster bootstrap, the AUC improvement over M_t is 0.124 for logistic regression (95% CI 0.020 to 0.242; p = 0.018) and 0.115 for Random Forest (95% CI 0.027 to 0.211; p = 0.010). The Random-Forest-minus-logistic difference is -0.009 (95% CI -0.078 to 0.064; p = 0.755). Thus the stronger conclusion is that the eight-indicator models add one-year-ahead discrimination beyond the continuous current-year M-Score; the data do not support a claim that one fitted algorithm dominates the other.

The decision-score threshold changes the operational meaning of the screen even though it does not alter the underlying ranking. At a 0.30 cutoff, logistic regression captures all ten positive cases but flags 59 of the 66 firms, creating a heavy review burden. At 0.60, logistic regression generates 12 alerts and Random Forest only four; the Random Forest screen is highly selective at that point (precision = 0.750), but recall falls to 0.300. These cutoffs are not interpreted as calibrated event probabilities. They are operating points on class-weighted model scores and are not used to identify an ex post 'optimal'

threshold. The 0.50 cutoff remains the headline specification because it was fixed before the 2025 outcomes were evaluated. Figure 2 and Table 6B show the precision-recall trade-off across several plausible operating points.

Table 6B. Sensitivity of 2025 classification performance to the decision-score threshold

Model	Threshold	Alerts	Precision	Recall	F1	Specificity
Logistic regression	0.30	59	0.169	1.000	0.290	0.125
	0.40	46	0.196	0.900	0.321	0.339
	0.50	**24**	**0.250**	**0.600**	**0.353**	**0.679**
	0.60	12	0.333	0.400	0.364	0.857
Random Forest	0.30	50	0.180	0.900	0.300	0.268
	0.40	28	0.214	0.600	0.316	0.607
	0.50	**13**	**0.385**	**0.500**	**0.435**	**0.857**
	0.60	4	0.750	0.300	0.429	0.982

Note: Alerts are the number of observations classified as positive among 66 holdout firms. Specificity is the true-negative rate. The 0.50 decision-score threshold remains the pre-specified headline cutoff; alternative thresholds are descriptive operating-point checks and were not chosen using holdout outcomes.

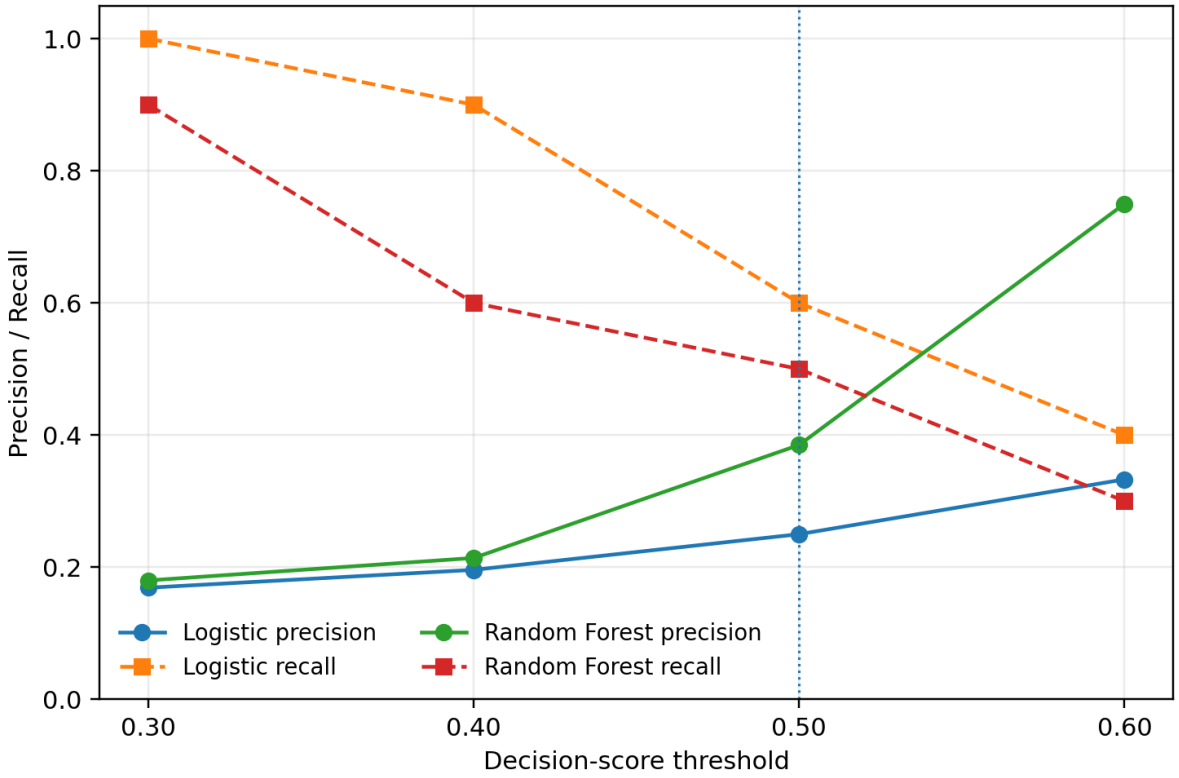

Figure 2. Precision-recall trade-off across decision-score thresholds on the 2025 temporal holdout.

Note: The vertical reference line marks the pre-specified 0.50 operating threshold. The scores are not interpreted as calibrated event probabilities.

4.5 Explainability of the Random Forest predictions

Table 7. Random Forest SHAP importance on the 2025 holdout

Indicator	Mean |SHAP|	Feature-SHAP correlation
TATA	0.07415	0.840
DEPI	0.04434	0.442
AQI	0.03680	-0.408
DSRI	0.02432	0.156

Indicator	Mean |SHAP|	Feature-SHAP correlation
SGAI	0.02275	-0.412
LVGI	0.01959	0.040
SGI	0.01513	0.403
GMI	0.01486	-0.506

Note: Mean |SHAP| summarizes contribution magnitude; the correlation column is descriptive and not causal.

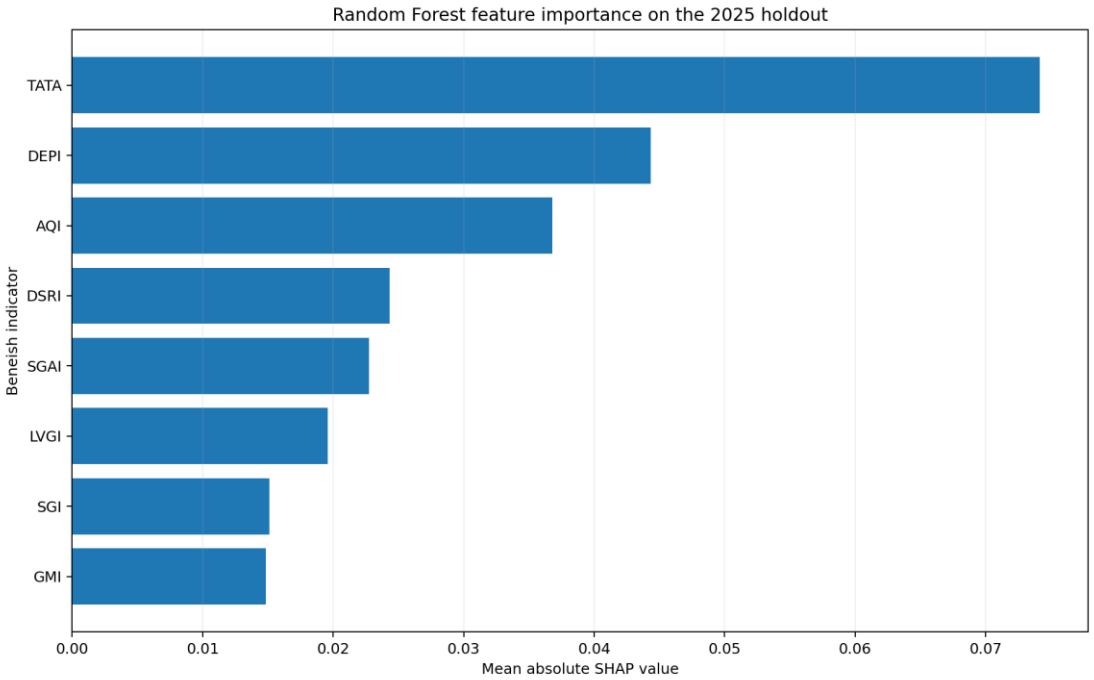

Figure 3. SHAP feature importance for Random Forest predictions on the 2025 holdout.

TATA is the largest Random Forest contributor on the 2025 holdout, followed by DEPI, AQI, and DSRI. The ranking is economically interpretable but should not be treated as causal. TATA captures the gap between accounting earnings and operating cash flow; DEPI and AQI reflect depreciation and asset-composition changes; DSRI captures receivable growth relative to sales. Some feature-SHAP correlations run opposite to the contemporaneous signs embedded in the Beneish equation, notably AQI, SGAI, and GMI. This is not a contradiction. The Random Forest predicts a next-year screening outcome, whereas the Beneish coefficients describe the same-year score. Conditional interactions, denominator effects, and year-to-year normalization of extreme ratios can therefore produce an inverse next-year contribution. SHAP directions are model-conditional descriptions, not structural accounting effects.

4.6 Robustness checks

Winsorizing each Beneish indicator at the 1st and 99th percentiles changes only three of the 462 classifications at the -1.78 threshold (0.65%). This is reassuring because the raw ratios are markedly heavy-tailed. The primary classification pattern is therefore not an artifact of a handful of extreme observations, even though some raw means are sensitive to them. Importantly, winsorization is a robustness exercise only; the main reported results use the untrimmed finite ratios.

Two additional checks address model specification. First, a supplementary regularized logistic specification adds log total assets and sector indicators to the eight Beneish variables. On the 2025 holdout, this augmented specification produces ROC-AUC = 0.655 and AP = 0.334, below the corresponding 0.718 and 0.371 for the parsimonious eight-indicator logistic model. Second, the fitted models are re-estimated without class weights. The purpose is not to replace the primary screening specification, but to test whether the ranking conclusions are an artifact of reweighting the minority class.

Table 8. Sensitivity of discrimination to class weighting

Model	Class weighting	2025 AUC	2025 AP	Rolling pooled AUC	Rolling pooled AP
Logistic regression	Balanced	0.718	0.371	0.661	0.263
Logistic regression	None	0.725	0.360	0.657	0.248
Random Forest	Balanced-subsample	0.684	0.406	0.653	0.236
Random Forest	None	0.675	0.465	0.675	0.262

Note: Rolling pooled results use target years 2021-2025 and the same frozen model architecture as the rolling-origin stability analysis.

Removing class weights changes the level and operational interpretation of model scores, but it does not overturn the discrimination results. On the 2025 holdout, unweighted logistic regression has ROC-AUC = 0.725 and AP = 0.360, while unweighted Random Forest has ROC-AUC = 0.675 and AP = 0.465. Across the pooled rolling origins, the corresponding AUCs are 0.657 and 0.675. This check indicates that the ranking conclusions do not depend on class weighting: unweighted logistic regression is essentially unchanged, and unweighted Random Forest is, if anything, slightly stronger in average precision and pooled AUC. The class-weighted configuration is retained as the primary specification because it was fixed in advance, and the check reinforces the decision not to describe the weighted outputs as calibrated probabilities.

The primary Beneish classification also remains sensitive, as expected, to the M-Score cutoff. Replacing -1.78 with the more permissive -2.22 boundary raises the number of flagged firm-years from 66 of 462 (14.3%) to 125 (27.1%). In the 2025 target year the positive class increases from 10 to 19 firms. Under this alternative target, logistic regression records ROC-AUC = 0.698 and AP = 0.493, whereas Random Forest records ROC-AUC = 0.638 and AP = 0.511. The exact prevalence changes, but the substantive model comparison remains mixed: logistic regression ranks better by ROC-AUC while Random Forest yields the higher average precision.

5. Discussion

5.1 What the Beneish results show in a post-IFRS Saudi setting

The first result is that Beneish screening signals remain observable in a post-IFRS Saudi reporting environment. The 14.3% rate is a screening frequency, not an estimate of misconduct. A company may cross the threshold because of rapid growth, working-capital changes, restructuring, major investment, or unusual but legitimate accounting estimates. International evidence reaches a similar conclusion: the Beneish model can be useful as a first-pass screen, but its sensitivity and precision vary across institutional settings (Repousis, 2016; Marais et al., 2023). The firm-level evidence in this study reinforces that interpretation. Most signals are episodic rather than persistent, which argues against treating the score as a permanent label attached to a company.

The year-to-year pattern also argues against treating M-Score status as a permanent firm characteristic. Only 11 of 56 transitions that begin above the primary screening boundary remain above it in the following year (19.6%), and the continuous M_t benchmark produces only modest pooled rolling-origin discrimination (ROC-AUC = 0.537). This weak serial persistence is consistent with substantial year-to-year normalization in ratio-based signals, especially when an extreme denominator or transitory operating condition drives the initial observation. It should not be interpreted as proof of economic mean reversion in the underlying business process; rather, it explains why a next-year model need not preserve the contemporaneous direction of every Beneish component. Screening therefore needs to be refreshed each reporting cycle, and isolated annual signals require contextual follow-up rather than permanent labeling.

5.2 What machine learning adds—and what it does not

The main methodological point is the shift from reconstruction to forecasting. Same-year classification would give a machine-learning algorithm access to the very ratios used to calculate the target and would therefore overstate what "AI" adds. Predicting the following year is more demanding. It asks whether current accounting patterns contain persistent information that survives beyond the period in which they are measured.

The answer is qualified rather than one-sided. The continuous current-year M-Score is a weak one-year-ahead benchmark, and both fitted models improve pooled rolling-origin discrimination over that score. The firm-cluster bootstrap indicates that these gains in AUC are unlikely to be sampling noise at conventional levels, whereas the difference between logistic regression and Random Forest is not statistically distinguishable. On the final 2025 holdout, logistic regression provides the stronger ROC-AUC and recall, while Random Forest provides higher average precision, precision, and F1 at the pre-specified operating threshold. Unweighted refits lead to the same broad conclusion. In practical terms, logistic regression favors broader ranking performance,

whereas Random Forest can produce a shorter review list at selected operating points. The evidence therefore supports model complementarity rather than algorithmic superiority.

This modest result is important for how accounting AI should be evaluated. Studies based on large U.S. samples show substantial gains from richer machine-learning specifications (Bao et al., 2020; Bertomeu et al., 2021), while cross-country work demonstrates that transferability cannot be assumed (Steingen & Löw, 2025). A smaller, single-market panel with a strict temporal holdout is unlikely to produce spectacular accuracy. It does, however, provide a more realistic estimate of how a model might behave when asked to score an unseen reporting year.

5.3 Explainability and accounting meaning

The SHAP results add accounting meaning to the forecasting exercise. TATA is the leading contributor, followed by DEPI, AQI, and DSRI. TATA is distinctive because it summarizes a cross-statement tension: how far accounting earnings are supported by operating cash flow. That tension is economically relevant in an accrual system, where reported performance necessarily incorporates estimates about timing, collectability, capitalization, depreciation, and other judgments. At the same time, the negative feature-SHAP correlations observed for AQI, SGAI, and GMI deserve caution. They do not reverse the accounting interpretation of the Beneish coefficients. The fitted model is predicting a later-year screening class and can use nonlinear interactions, so a high current-year feature may contribute negatively to next-year risk after conditioning on the other indicators. Limited serial persistence and normalization of extreme ratios make such conditional sign reversals plausible.

The Saudi setting makes that interpretation especially important. The non-financial market contains firms with long operating cycles, substantial fixed assets, inventory commitments, and material working-capital requirements. Those features can generate legitimate differences between earnings and cash realization, particularly around periods of investment or growth. The result should therefore be read narrowly: TATA appears to be the most informative starting point for follow-up in this sample, not a stand-alone indicator of wrongdoing. Its value lies in directing attention to the quality and cash support of reported earnings, while DEPI, AQI, and DSRI indicate which asset, depreciation, or receivable relationships may deserve a closer look. This cautious use is consistent with evidence that forensic-model performance depends on institutional and audit context (Hakami et al., 2020) and with recent Saudi evidence linking post-IFRS reporting quality to governance effectiveness (Alaskar, 2026).

That interpretability is arguably more useful than a small difference in a summary accuracy statistic. A reviewer can move from a high model decision score to a set of questions: Is cash conversion weakening relative to earnings? Have receivables grown faster than sales? Has the depreciation rate changed? Has the composition of non-current assets shifted? SHAP does not answer those questions, but it makes clear where the follow-up should begin. This human-in-the-loop use is consistent with the direction of explainable fraud research (Zhou et al., 2023; Cai & Xie, 2024).

5.4 Practical implications

For auditors and audit committees, the framework is best viewed as a triage tool rather than an automated verdict. A TATA- or DSRI-driven alert points toward a different line of inquiry from one dominated by DEPI or AQI. The former may justify closer reconciliation of earnings and cash flow, receivable aging, cut-off, and revenue recognition; the latter may shift attention toward useful lives, depreciation methods, capitalization, impairment, and the composition of long-lived assets. Figure 2 and Table 6B make the operating decision explicit. Lower decision-score thresholds are appropriate when missing a potentially concerning case is especially costly, but they consume more review capacity; higher thresholds conserve that capacity by concentrating attention on fewer observations, at the price of lower recall. Because the weighted model scores are not calibrated probabilities, the threshold should be treated as a resource-allocation rule rather than a statement about an absolute probability of misstatement.

For investors and creditors, the same output can supplement traditional ratio analysis, while regulators can use it to prioritize scarce review resources. In all cases, the score should remain a risk-ranking device rather than a public accusation. The temporal holdout matters because it tests the model under a condition closer to actual use: the outcome year is genuinely unseen when the model is built.

6. Conclusion

This paper examines whether the Beneish framework contains useful information for one-year-ahead financial reporting risk screening in Saudi listed non-financial firms. The balanced sample covers 66 firms from 2018 to 2025 and yields 462 M-Score observations for 2019-2025. Sixty-six of the 462 firm-years (14.3%) cross the $M > -1.78$ screening boundary. The rate is highest in 2019, falls to its low point in 2023, and rises again in 2025. At the firm level, the signals are mainly episodic rather than persistent.

Current Beneish indicators contain information about the following year's screening class, but the comparison between models is not one-dimensional. Across the pooled rolling-origin stability analysis, logistic regression and Random Forest improve ROC-AUC from 0.537 for the continuous M_t benchmark to 0.661 and 0.653, respectively; firm-cluster bootstrap intervals support

both improvements over the benchmark but not a difference between the fitted models. On the untouched 2025 holdout, logistic regression has the higher ROC-AUC (0.718) and recall (0.600), while Random Forest has the higher AP (0.406), precision (0.385), and F1 (0.435) at the fixed decision-score threshold. Unweighted refits and the alternative -2.22 Beneish cutoff do not overturn the broader conclusion that the two algorithms offer different screening trade-offs rather than a clear winner.

SHAP identifies TATA as the leading Random Forest feature, followed by DEPI, AQI, and DSRI. This makes the model operationally interpretable because its output can be connected to accrual intensity, cash-flow support, depreciation, asset composition, and receivables. The paper's contribution is the combination of prospective temporal validation, transparent accounting signals, and explicit uncertainty rather than a claim of algorithmic superiority or confirmed fraud detection.

More broadly, the findings argue for restraint in accounting applications of artificial intelligence. A useful model should be tested on information that was genuinely unavailable during development, compared with a transparent benchmark, and explained in accounting terms. Under those conditions, machine learning can extend forensic analysis without displacing professional judgment.

7. Limitations and Future Research

The outcome in this paper is a future Beneish screening signal, not regulator-confirmed fraud or misstatement. The predictive results therefore concern the persistence or emergence of an accounting-risk profile. A stronger external-validation design would link the same predictors to a complete, independently assembled event set such as Saudi CMA enforcement actions, restatements, modified audit opinions, or other documented reporting failures. A selective lookup of only model-flagged firms would create ascertainment bias and could overstate apparent validation, so a partial ad hoc check is not reported here. Özari et al. (2025) provide a useful methodological precedent by validating Beneish- and Altman-related Random Forest features against administrative fines in the Turkish market. A Saudi replication requires a systematic event dataset constructed independently of the present model outputs.

The evidence also comes from one market and a moderate number of firms. The balanced panel supports clean temporal comparison, but it constrains model complexity and leaves only ten positive observations in the final holdout. The wide bootstrap intervals for AUC, precision, recall, and F1 make that limitation visible. A conventional post-hoc power calculation would not remove this uncertainty and could create false precision; replication with a larger positive class is more informative. Extending the design across Gulf Cooperation Council markets would also permit a direct test of external validity and comparison with GCC evidence showing that Beneish, Dechow, and Altman screening models behave differently across audit settings (Hakami et al., 2020).

Two measurement issues deserve attention in future work. First, ratio-based forensic variables can become extreme when lagged denominators are small. The winsorization check and the reported skewness and kurtosis show that the main classifications are not driven by a handful of outliers, but the underlying distributions remain heavy-tailed. Second, net income is used as the consistent earnings input for TATA because income from continuing operations is not separately available for all observations. The original TATA definition therefore cannot be reproduced exactly from the current dataset. Replacing the missing line item with EBIT or operating profit would change the construct rather than provide a valid sensitivity test, so no such proxy is introduced. Future datasets with the continuing-operations line item can test whether the TATA ranking and SHAP importance are preserved under the original specification.

Declarations

Data availability. The study uses financial-statement information compiled by the author from publicly available audited annual reports and Saudi market disclosures. The analytical dataset and reproducible code can be made available subject to the requirements of the target journal and any source-use restrictions.

Ethics Approval and Consent to Participate. Not applicable. This study relies exclusively on publicly available secondary corporate financial information and does not involve human participants, personal data, surveys, interviews, experiments, or interventions. Accordingly, institutional ethics approval and informed consent to participate were not required.

Conflicts of Interest: The author declares no conflict of interest.

Funding: No specific external funding was received for this study.

AI Use Disclosure. Generative AI tools were used to support language polishing, proofreading, drafting assistance, and code review. No unverified references were retained. All data, analytical choices, numerical results, interpretations, and conclusions were reviewed and validated by the author, who remains fully responsible for the content and accuracy of the manuscript.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

Alaskar, M. Z. (2026). The impact of audit committee effectiveness on the mandatory adoption of IFRS and the financial reporting quality: Empirical evidence from an emerging market. Cogent Business & Management, 13(1), 2654790. https://doi.org/10.1080/23311975.2026.2654790

Alomair, A., Farley, A., & Yang, H. H. (2022). The impact of IFRS adoption on the value relevance of accounting information in Saudi Arabia. Accounting & Finance, 62(2), 2839–2878. https://doi.org/10.1111/acfi.12902

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

Beneish, M. D. (1997). Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. Journal of Accounting and Public Policy, 16(3), 271–309. https://doi.org/10.1016/S0278-4254(97)00023-9

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

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

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

Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2021). Using machine learning to detect misstatements. Review of Accounting Studies, 26(2), 468–519. https://doi.org/10.1007/s11142-020-09563-8

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Cai, S., & Xie, Z. (2024). Explainable fraud detection of financial statement data driven by two-layer knowledge graph. Expert Systems with Applications, 246, 123126. https://doi.org/10.1016/j.eswa.2023.123126

Capital Market Authority (CMA). (2016, December 8). The Capital Market Authority assures implementation of IFRS on 1/1/2017, and adopts several initiatives to support listed entities. https://cma.gov.sa/en/MediaCenter/PR/Pages/IFRS.aspx

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

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. Contemporary Accounting Research, 28(1), 17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1996). Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC. Contemporary Accounting Research, 13(1), 1–36. https://doi.org/10.1111/j.1911-3846.1996.tb00489.x

Hakami, T. A., Rahmat, M. M., Yaacob, M. H., & Mohd Saleh, N. (2020). Fraud detection gap between auditor and fraud detection models: Evidence from Gulf Cooperation Council. Asian Journal of Accounting & Governance, 13, 1–13. https://doi.org/10.17576/AJAG-2020-13-01

Healy, P. M., & Wahlen, J. M. (1999). A review of the earnings management literature and its implications for standard setting. Accounting Horizons, 13(4), 365–383. https://doi.org/10.2308/acch.1999.13.4.365

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of Financial Economics, 3(4), 305–360. https://doi.org/10.1016/0304-405X(76)90026-X

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Marais, A., Vermaak, C., & Shewell, P. (2023). Predicting financial statement manipulation in South Africa: A comparison of the Beneish and Dechow models. Cogent Economics & Finance, 11(1), 2190215. https://doi.org/10.1080/23322039.2023.2190215

Nurunnabi, M., Donker, H., & Jermakowicz, E. K. (2022). The impact of mandatory adoption of IFRS in Saudi Arabia. Journal of International Accounting, Auditing and Taxation, 49, 100509. https://doi.org/10.1016/j.intaccaudtax.2022.100509

Nurunnabi, M., Jermakowicz, E. K., & Donker, H. (2020). Implementing IFRS in Saudi Arabia: Evidence from publicly traded companies. International Journal of Accounting & Information Management, 28(2), 243–273. https://doi.org/10.1108/IJAIM-04-2019-0049

Özari, Ç., Can, E. N., & Demirkale, Ö. (2025). Financial fraud detection with Altman Z-Score and Beneish M-Score via Random Forest: Verified by Borsa Istanbul fines (2018–2022). SAGE Open, 15(4), 1–17. https://doi.org/10.1177/21582440251386174

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

Perols, J. L., Bowen, R. M., Zimmermann, C., & Samba, B. (2017). Finding needles in a haystack: Using data analytics to improve fraud prediction. The Accounting Review, 92(2), 221–245. https://doi.org/10.2308/accr-51562

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

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

Steingen, L., & Löw, E. (2025). Using machine learning to detect financial statement fraud: A cross-country analysis applied to Wirecard AG. Journal of Risk and Financial Management, 18(11), 605. https://doi.org/10.3390/jrfm18110605

Zhou, Y., Li, H., Xiao, Z., & Qiu, J. (2023). A user-centered explainable artificial intelligence approach for financial fraud detection. Finance Research Letters, 58, 104309. https://doi.org/10.1016/j.frl.2023.104309

Appendix A. Reproducibility details for the predictive analysis

Table A1. Predictive-analysis reporting checklist

Component	Specification
Target	Beneish screening signal in t+1: 1 if M > -1.78, 0 otherwise
Predictors	DSRI, GMI, AQI, SGI, DEPI, SGAI, TATA, LVGI measured in t
Development targets	2020-2024 for primary 2025 model development
Untouched test target	2025 (using 2024 predictors)
Logistic regression	Standardized predictors; class_weight = balanced; C = 0.10
Random Forest	500 trees; max_depth = 3; min_samples_leaf = 10; max_features = sqrt; class_weight = balanced_subsample; random_state = 42
Hyperparameter selection	Three expanding validation folds inside 2020-2024 development targets: 2020-2021 -> 2022; 2020-2022 -> 2023; 2020-2023 -> 2024. Mean ROC-AUC used for selection.
Classification threshold	0.50 decision-score threshold for headline classification metrics; 0.30, 0.40, and 0.60 as operating-point sensitivity checks
Explainability	TreeSHAP calculated on all 66 Random Forest observations in the untouched 2025 holdout
Robustness	1st/99th percentile winsorization; continuous M_t benchmark; rolling-origin stability analysis; unweighted model refits; alternative M-Score cutoff at -2.22; 5,000 bootstrap resamples
Bootstrap seed	20261012
Primary ranking baseline	Continuous current-year M-Score (M_t)
Probability interpretation	Class-weighted model outputs are treated as decision scores, not calibrated event probabilities; Brier score is not used
Rolling-origin stability	Target years 2021-2025; expanding training windows; final hyperparameter architecture held fixed; firm-cluster bootstrap for pooled AUC differences
Sample construction	81 symbol labels / 648 firm-years -> 80 normalized companies -> 79 after unresolved duplicate removal -> 66 balanced firms after finite-indicator requirement
TATA sensitivity	Original continuing-operations formulation cannot be reproduced from the available line items; no non-equivalent proxy is substituted