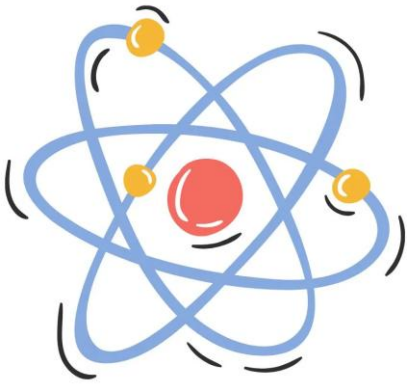


ISSN: 1672 - 6553

JOURNAL OF DYNAMICS AND CONTROL

VOLUME 10 ISSUE 03: P14-48

A TIME-ORDERED APPROACH TO PROFITABILITY INDEXING AND PREDICTION: EVIDENCE FROM VIETNAM

Tuyen Le Nam, Tam Phan Huy

University of Economics and Law, Ho Chi Minh City, Vietnam and Vietnam National University, Ho Chi Minh City, Vietnam.

JOURNAL OF DYNAMICS AND CONTROL
https://doi.org/10.71058/jodac.v10i03002

A TIME-ORDERED APPROACH TO PROFITABILITY INDEXING AND PREDICTION: EVIDENCE FROM VIETNAM

Tuyen Le Nam[1], Tam Phan Huy[1]*

[1]University of Economics and Law, Ho Chi Minh City, Vietnam and Vietnam National University, Ho Chi Minh City, Vietnam.

**Corresponding author:tamph@uel.edu.vn*
The research is funded by the University of Economics and Law, Vietnam National University, Ho Chi Minh City, Vietnam.
Manuscript Received on: 01/08/2025; Accepted on: 15/02/2026; Published: 03/03/2026

Abstract: *This study revisits a common issue in corporate finance measurement and forecasting: profitability is fundamentally multi-dimensional, yet research and practical screening often rely on single ratios that may deliver incomplete or even conflicting signals. To address this, the paper constructs an objective composite profitability index by combining correlated profitability proxies, and then tests whether a firm's profitability class in year t+1 can be predicted using only information observed in year t within a strictly time-ordered design that guards against information leakage. The analysis uses firm-year financial statement data for Vietnamese listed non-financial firms on HOSE and HNX from 1998–2025, provided by the Institute for Development and Research in Banking Technology (Vietnam National University, Ho Chi Minh City). The empirical pipeline applies outlier treatment and feature scaling solely on the training sample, builds the composite index through principal component analysis (PCA), and performs out-of-sample classification with nonlinear machine learning models. Training is restricted to targets up to 2020, and performance is evaluated on a holdout window spanning 2021–2024. Results indicate that the profitability proxies share sufficient common structure to yield a stable composite measure, and that next-year profitability status remains meaningfully predictable beyond the training period. In the baseline specification, an RBF-kernel SVM achieves the highest F1-score (0.7635), while ensemble models deliver the strongest AUC (Random Forest 0.8788; XGBoost 0.8779). Overall, the paper proposes a practical end-to-end framework that links measurement to prediction and argues that composite profitability indicators—implemented with leakage-safe preprocessing—should be considered a minimum standard for credible forecasting, screening, and monitoring.*

https://doi.org/10.71058/jodac.v10i03002

1. Introduction

Profitability is widely used as a key marker of firm performance because it reflects how effectively a company converts its operational activities and resource base into economic returns (Fama & French, 2000; Nissim & Penman, 2001). Despite its importance, profitability is not directly observable as a single variable in financial statements. Instead, it is typically approximated through a set of accounting-based indicators, including return on assets (ROA), return on equity (ROE), return on capital (ROC), earnings per share (EPS), and profit margins. Each measure captures a distinct aspect of performance: ROA highlights asset utilization, ROE embeds financing decisions and leverage, EPS expresses earnings on a per-share basis, and margin measures relate profitability to pricing power and cost structure. Consequently, any single ratio can offer only a partial—and potentially unstable—assessment of firm performance.

Beyond interpretation, profitability measurement also raises statistical concerns. These proxies often move together because they share common accounting components, particularly net income, which creates redundancy and can generate multicollinearity in empirical applications. In addition, financial ratios frequently exhibit skewness, heavy-tailed distributions, and sensitivity to denominators and temporary shocks, making them vulnerable to extreme values (Deakin, 1976; Frecka & Hopwood, 1983; McLeay & Omar, 2000). If such properties are not handled properly, both dimensionality reduction and predictive modeling can become fragile.

A coherent response is to conceptualize profitability as a latent trait and summarize multiple observed proxies using a composite index. Principal component analysis (PCA) is especially suitable because it converts correlated variables into orthogonal components that capture their dominant shared variation (Jolliffe, 2002; Wold et al., 1987). A PCA-based composite profitability index reduces overlap among proxies, avoids arbitrary weighting schemes, and facilitates comparisons across firms and across time. However, measurement alone does not fully meet practical needs. Investors, analysts, and managers often seek a forward-looking judgment—specifically, whether a firm's profitability is likely to remain above or drop below a meaningful threshold in the next period.

Crucially, credible forecasting depends on how evaluation is conducted. In financial panel settings, look-ahead bias can arise not only from obvious target leakage but also from "hidden" leakage

https://doi.org/10.71058/jodac.v10i03002

embedded in preprocessing. For example, winsorization cutoffs, scaling parameters, or PCA loadings estimated using the full dataset may unintentionally incorporate information from later periods, artificially inflating predictive performance (Bergmeir & Benítez, 2012). This risk is particularly relevant in emerging markets, where firm heterogeneity and data quality can vary more sharply across time and across entities.

The literature has largely advanced along two separate tracks. One stream develops composite indices for profitability or performance, primarily for descriptive ranking and benchmarking. Another stream focuses on predicting profitability-related outcomes with econometric or machine learning methods, sometimes without enforcing time-consistent preprocessing. This separation leaves an applied methodological gap: what is needed is a unified framework that connects multidimensional measurement to deployable out-of-sample forecasting under an information-set-consistent pipeline.

This study addresses that gap for Vietnamese listed non-financial firms (HOSE and HNX) over 1998–2025. It introduces an integrated approach in which profitability proxies are first combined into a PCA-based composite index, which is then used to define a next-year profitability status label. Prediction is performed from year t information to year $t+1$ outcomes, with all transformations estimated strictly on the training sample and applied unchanged to the holdout period. This design strengthens both measurement consistency and forecasting validity.

The research has two linked aims. First, it constructs a compact, low-noise composite profitability index from a screened set of proxies using PCA. Second, it evaluates whether next-year profitability status (defined via the index) is predictably out of sample under a strict leakage-free protocol. The contribution is especially relevant for emerging-market corporate analytics, where decision support requires both robust measurement and realistic forecasting performance.

The remainder of the paper is structured as follows. The literature review motivates the integrated framework. The methodology section describes the data, variable construction, leakage-safe preprocessing, PCA index development, and forecasting design. The results section reports empirical findings and robustness checks. The concluding section summarizes implications and outlines directions for future research.

https://doi.org/10.71058/jodac.v10i03002

2. Theoretical background and literature review

2.1. Profitability as a multidimensional construct

In corporate finance and accounting, profitability is commonly defined as a firm's ability to generate earnings relative to the resources deployed over an accounting period. Importantly, profitability is not treated as a single directly observed statistic; rather, it is understood as a multidimensional performance domain reflected across financial statement components (Penman, 2010). Capital markets research further emphasizes that accounting numbers summarize several underlying economic processes, including operational efficiency, financing structure, and competitive positioning (Kothari, 2001).

Return-based measures such as ROA and ROE both assess earnings efficiency, but they differ materially in their denominator structures. ROA captures the productivity of total assets, while ROE incorporates leverage and therefore blends operating outcomes with capital structure choices (Fama & French, 2000). This implies that firms with identical net income can appear stronger or weaker depending on financing decisions. Empirical evidence aligns with this distinction: leverage often amplifies ROE volatility relative to ROA (Nissim & Penman, 2001).

EPS introduces a shareholder-based scaling by expressing earnings per share outstanding. Because EPS is affected by equity issuance, repurchases, and other capital structure changes, it reflects per-share value creation but may not directly represent asset-level operating efficiency (Biddle et al., 2009).

Margin indicators add another dimension by linking earnings to revenues and therefore capturing price–cost outcomes. Net profit margin (NPM) reflects the final residual of revenue after costs, non-operating items, taxes, and financing effects, making it informative about pricing outcomes and cost discipline. Accounting research suggests that margins may respond differently than asset-based returns during macroeconomic volatility, since margins react directly to revenue conditions and cost shocks, while return measures also embed balance-sheet structure and investment intensity (Lev & Thiagarajan, 1993).

From a distributional standpoint, financial ratios rarely behave like normal variables; they often exhibit heavy tails and distortions driven by denominator effects (Deakin, 1976; McLeay & Omar, 2000). These features reinforce the view that no single ratio can adequately represent profitability.

https://doi.org/10.71058/jodac.v10i03002

Instead, profitability is better modeled as a latent construct inferred from a system of correlated indicators, where each proxy captures a distinct economic channel but also contains proxy-specific noise (Jolliffe, 2002).

In panel contexts where firms differ in capital intensity and financing structure, relying on one ratio can lead to unstable rankings across firms and shifting conclusions over time (Gujarati, 2012). A multi-proxy measurement approach that combines return-based (ROA, ROE), per-share (EPS), and margin-based (NPM) indicators is therefore consistent with both theory and empirical evidence that profitability is inherently multidimensional.

2.2. Composite profitability index construction

Because profitability proxies often share common accounting ingredients—especially net income and revenue—they tend to be strongly correlated. High correlation can create multicollinearity, weakening interpretability and increasing estimation variance in regression-based analysis (Gujarati, 2012). Dimensionality reduction methods, particularly PCA, are commonly applied to reduce this redundancy (Jolliffe, 2002).

PCA re-expresses correlated variables as orthogonal components that maximize explained variance in sequence (Wold et al., 1987). In financial settings, the first component frequently captures the dominant return-related signal, while subsequent components may isolate additional dimensions such as margin behavior or per-share outcomes (Sabău Popa et al., 2021).

Prior to extracting components, it is standard to assess whether the data are appropriate for PCA. The Kaiser–Meyer–Olkin (KMO) statistic evaluates sampling adequacy and shared covariance (Kaiser, 1974), while Bartlett's test checks whether the correlation structure departs significantly from an identity matrix (Bartlett, 1954). These diagnostics help ensure that PCA is used only when a meaningful common structure exists.

PCA is also sensitive to scale differences and extreme observations. Because financial ratios can show large dispersion due to denominator issues and accounting irregularities (Deakin, 1976), robust preprocessing—such as winsorization and standardization—helps stabilize variance-based extraction (Hair et al., 2010).

https://doi.org/10.71058/jodac.v10i03002

Accordingly, building a PCA-based profitability index should be viewed as a structured measurement strategy rather than a purely mechanical step: it combines economic reasoning about profitability channels with careful preprocessing and disciplined variance extraction.

2.3. Theoretical foundations

The framework underlying this study draws on three complementary principles: (i) latent-construct reasoning, (ii) profitability persistence and mean reversion, and (iii) information-set-consistent forecasting.

First, latent-variable theory argues that many economic attributes are not directly observable and must be inferred from multiple imperfect indicators (Bollen, 1989). Profitability fits this logic because no single accounting ratio fully captures firm performance (Penman, 2010). PCA provides a practical mechanism for extracting dominant shared variation in a way that aligns with latent-factor interpretation (Jolliffe, 2002).

Second, theories of persistence and mean reversion suggest that firms can sustain superior profitability when they possess competitive advantages, but competitive pressures gradually erode abnormal performance over time (Fama & French, 2000). Empirical work shows profitability is partly persistent across years yet tends to converge toward industry averages (Nissim & Penman, 2001), implying that current profitability contains information relevant for predicting future profitability status.

Third, forecasting theory emphasizes the information-set principle: predictions should be formed using only what is known at the forecast origin (Shmueli, 2010). In time-ordered panel data, violating chronology leads to look-ahead bias and undermines validity (Kapoor & Narayanan, 2022). Bergmeir and Benítez (2012) demonstrate that improper validation procedures in time-series contexts can meaningfully overstate predictive performance.

A major practical source of bias is preprocessing leakage, where scaling parameters or PCA loadings are estimated using the full dataset rather than being fitted on the training sample only (Belesis et al., 2023). Such leakage can inflate metrics like AUC (Fawcett, 2006) while reducing real-world deployability. For this reason, strict time-based splitting and training-only transformations are best understood as theoretical necessities, not optional technical refinements.

https://doi.org/10.71058/jodac.v10i03002

2.4. Empirical studies

Related empirical evidence can be grouped into three strands. First, a measurement-focused literature develops composite profitability or performance indices to synthesize correlated accounting indicators. PCA-based indices are widely used because they compress overlapping measures into orthogonal components. For example, Sabău Popa et al. (2021) construct a PCA-derived performance index from multiple indicators and show distinct loading patterns across components, consistent with the view that firm performance is multi-dimensional.

Second, a forecasting-focused literature compares predictive models and increasingly highlights the strengths of ensemble methods. Meta-analytic evidence from Kutub Uddin et al. (2022) suggests that ensembles such as random forests and gradient boosting often outperform standalone models in accuracy and stability. However, reported gains depend heavily on rigorous validation. Kapoor and Narayanan (2022) argue that data leakage is common and can drive reproducibility failures in applied machine learning. Typical mistakes include estimating normalization parameters or PCA loadings on the full sample rather than restricting them to the training data. Consequently, time-ordered validation with leakage-safe preprocessing is essential for ensuring that reported accuracy and AUC reflect feasible forward-looking prediction (Bergmeir & Benítez, 2012; Kapoor & Narayanan, 2022).

Third, work in emerging markets—particularly across Asia—often combines PCA-based feature extraction with nonlinear classifiers. Zhang et al. (2015), for instance, build a PCA-based composite profitability measure for Chinese listed construction firms and apply an SVM classifier, reporting strong performance on held-out data. This illustrates how unsupervised dimensionality reduction paired with nonlinear modeling can be effective in developing-economy settings.

In Vietnam, by contrast, much of the existing evidence relies on simpler empirical setups. Many studies examine profitability determinants using cross-sectional or panel regressions with a single proxy (often ROA); Nguyen and Nguyen (2020) follow this type of design for Vietnamese listed firms. Such work typically does not test forward-looking predictive validity under strict time splits. Similarly, Vietnamese research on financial distress frequently emphasizes in-sample discrimination rather than genuine out-of-sample forecasting. Tran et al. (2023) revisit Altman's Z-score using logistic regression and ROC analysis for Vietnamese firms, identifying

https://doi.org/10.71058/jodac.v10i03002

discriminative ratios but without implementing a strictly time-ordered holdout protocol. More broadly, Vietnam-focused studies rarely document leakage-safe pipelines (e.g., holding out final years) or systematically test robustness to alternative labeling thresholds. As a result, the deployable forecasting performance of profitability models for Vietnamese firms remains insufficiently established.

Overall, prior research points to a gap between rigorous end-to-end pipelines—combining PCA-based composite measurement with machine learning under strict time-consistent validation—and the approaches commonly used in Vietnam-specific studies. Closing this gap requires multidimensional index construction, objective forward-looking label definitions, and leakage-proof preprocessing and training procedures so that reported performance aligns with real-world forecasting feasibility.

3. Methodology

3.1. Data

This study examines next-year profitability status for publicly listed non-financial firms in Vietnam using a firm–year panel dataset. The sample includes non-financial companies traded on the Ho Chi Minh Stock Exchange (HOSE) and the Hanoi Stock Exchange (HNX). Firm-level financial statement data are sourced from the Institute for Development and Research in Banking Technology (VNUHCM-IBT), Vietnam National University, Ho Chi Minh City, covering 1998–2025 and organized at the firm–year level. The empirical setup is explicitly forward-looking: predictors are measured in year t, while the outcome of interest is profitability status in year $t + 1$. To maintain chronological validity, the last observed year for each firm is removed because the next-year label cannot be formed. The final panel therefore contains one observation per firm i in year t. The proxy set is designed to capture multiple profitability channels, spanning return efficiency (ROA, ROE, ROC), per-share performance (EPS), and margin performance (NPM). Table 1 summarizes the notation for panel indices, profitability proxies, standardized variables, principal component scores, the composite profitability index, and the next-year binary label used in the forecasting task.

https://doi.org/10.71058/jodac.v10i03002

Table 1. Variable definition

Variable	Symbol	Description
Firm index	i	Cross-sectional unit (each listed firm).
Year index	t	Time unit (year).
Firm-year observation	(i,t)	Listed-firm identifier used as the cross-sectional dimension of the panel
Return on assets	$X_{1,i,t}$ (ROA)	Calendar year serving as the time dimension
Return on equity	$X_{2,i,t}$ (ROE)	A single firm–year observation corresponding to firm *i* in year *t*
Return on capital	$X_{3,i,t}$ (ROC)	Parent-shareholder net income scaled by average total assets; when lagged assets are unavailable, total assets in year *t* are used as the denominator
Earnings per share	$X_{4,i,t}$ (EPS)	Parent-shareholder net income divided by average equity attributable to parent shareholders; if lagged equity is missing, equity in year *t* is used instead
Net profit margin	$X_{5,i,t}$ (NPM)	Parent-shareholder net income normalized by paid-up capital
Standardized proxy	Z_k	Basic earnings per share (excluding extraordinary items) as reported in the source dataset
Principal component score	PC_m	Parent-shareholder net income divided by revenue from core business activities; employed as an operating-margin proxy when operating profit is not reported

https://doi.org/10.71058/jodac.v10i03002

Variable	Symbol	Description
Composite index	$P_{i,t}$	Standardized proxy computed using the training-sample mean and standard deviation only
Next-year label	$Label_{i,t+1}$	Score of principal component m, obtained as a PCA-loading-weighted linear combination of standardized proxies
Forecasting task	$X_{i,t} \rightarrow Label_{i,t+1}$	Composite profitability index formed as an explained-variance-weighted aggregation of principal component scores

Source: By author

The notation highlights a deliberate separation between a measurement layer—where a latent profitability signal is extracted via PCA—and a prediction layer—where next-year profitability status is forecast with machine learning classifiers. This structure is essential for transparency because it allows the validity of the constructed measure and the usefulness of the forecast to be evaluated separately. The next subsection describes the leakage-safe preprocessing and forecasting design used to ensure credible out-of-sample evaluation.

3.2. Empirical models and leakage-safe forecasting design

Profitability ratios frequently exhibit heavy-tailed distributions, extreme observations, and scale heterogeneity, which can distort covariance-based methods (including PCA) and reduce the robustness of nonlinear classifiers when untreated (Deakin, 1976; Frecka & Hopwood, 1983; McLeay & Omar, 2000). In addition, preprocessing can create information leakage if trimming cutoffs, scaling moments, or transformation parameters are computed using the full sample rather than being estimated solely from the training period (Bergmeir & Benítez, 2012; Kapoor & Narayanan, 2022). To mitigate both problems, the study adopts a leakage-safe pipeline in which all data-driven transformations are fitted using training data only and then applied unchanged to the holdout period.

https://doi.org/10.71058/jodac.v10i03002

The preprocessing pipeline has three steps. First, observations are aligned so that features measured at year t correspond to labels defined at year $t+1$. Second, each proxy is winsorized at the 1% tails using percentile bounds estimated from the training sample, which reduces the influence of extreme values while retaining economically meaningful interior variation. Third, winsorized proxies are standardized using the training-period mean and standard deviation before PCA construction and classifier estimation. For each proxy $X_{k,i,t}$, winsorization uses training-only cutoffs:

$$X_{k,i,t}^{(w)} = \min\{\max(X_{k,i,t}, q_{k,0.01}^{train}), q_{k,0.99}^{train}\}.$$

The standardized variable is then computed using training moments:

$$Z_{k,i,t} = \frac{X_{k,i,t}^{(w)} - \mu_k^{train}}{\sigma_k^{train}},$$

where μ_k^{train} and σ_k^{train} are calculated exclusively from the training sample. This design prevents future distributional information from entering model construction through preprocessing, ensuring that winsorization bounds, scaling parameters, and subsequent PCA loadings are all learned strictly within the training period, consistent with time-ordered evaluation principles (Bergmeir & Benítez, 2012).

3.3. Proposed Research Framework

The proposed framework forecasts next-year profitability status in a way that is both economically motivated and empirically disciplined. Instead of predicting a single observed profitability ratio, the approach first aggregates multiple proxies into a composite profitability signal and then predicts the next-year binary status of this latent construct. The framework consists of three stages: (i) leakage-safe preprocessing, (ii) PCA-based labeling, and (iii) machine learning forecasting. Stage 1 applies training-only preprocessing to control outliers and eliminate look-ahead contamination. Stage 2 uses PCA to summarize correlated profitability proxies into a small number of orthogonal latent dimensions, which are combined into a composite profitability index; the next-year binary label is then defined by the sign of this index. Stage 3 estimates nonlinear machine learning classifiers that map current-year profitability proxies to the next-year label.

https://doi.org/10.71058/jodac.v10i03002

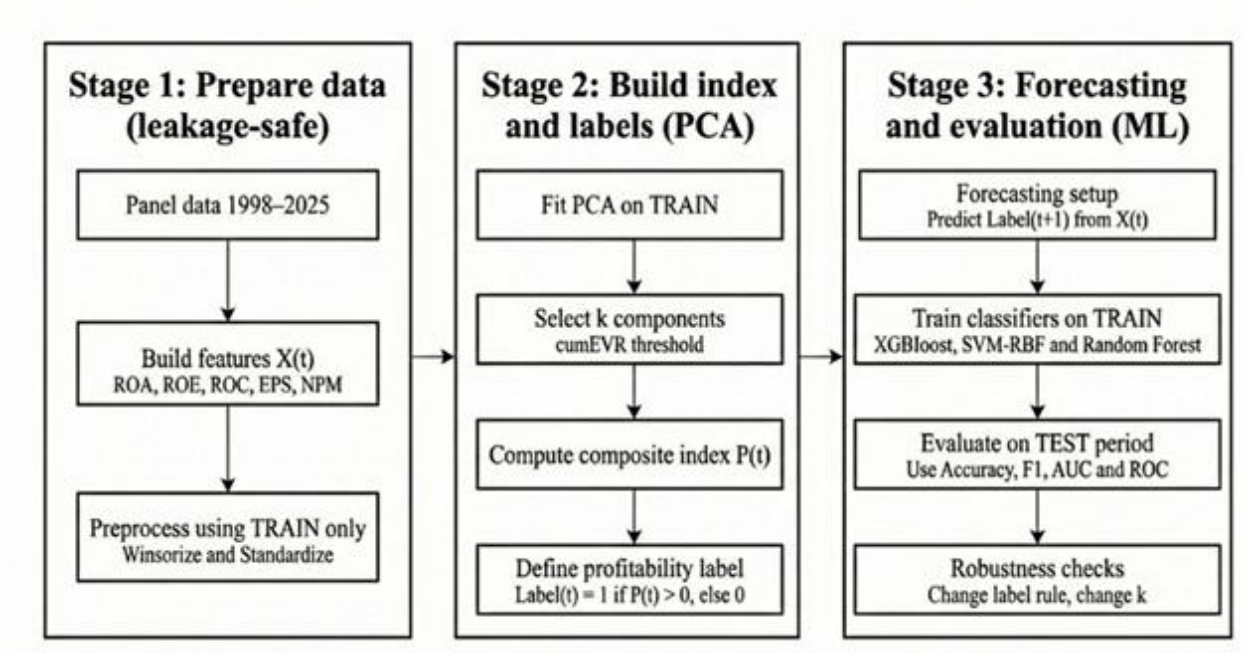


Figure 1. Integrated measurement-to-prediction framework

Source: By author

3.4. Empirical Models and Leakage-Safe Forecasting Design

To produce interpretable and credible forecasts of next-year profitability status, the analysis couples a PCA-based composite measurement stage with nonlinear classifiers under a strict time-ordered evaluation design that avoids look-ahead bias (Bergmeir & Benítez, 2012). Let $Label_{i,t+1} \in \{0,1\}$ denote next-year profitability status for firm i, defined by the sign of the composite profitability index:

$$Label_{i,t+1} = 1(P_{i,t+1} > 0).$$

Predictors observed at year t are the five profitability proxies:

$$X_{i,t} = (ROA_{i,t}, ROE_{i,t}, ROC_{i,t}, EPS_{i,t}, NPM_{i,t}).$$

To ensure comparability across models, the feature set is held fixed. All transformations that can learn from data are estimated on the training period only and applied unchanged to the holdout period (2021–2024). Each proxy is winsorized at the 1% tails using training-only bounds, standardized using training-only moments, and then mapped through PCA fitted on standardized

https://doi.org/10.71058/jodac.v10i03002

training data (years ≤ 2020). In the main specification, PCA retains $M = 3$ components, targeting cumulative explained variance of at least 0.85 in line with standard dimensionality-reduction practice (Jolliffe, 2002; Wold et al., 1987). Component scores are computed as:

$$PC_{m,i,t} = \sum_{k=1}^{K} a_{m,k} Z_{k,i,t},$$

and the composite index is constructed as a variance-share weighted aggregation:

$$P_{i,t} = \sum_{m=1}^{M} \omega_m PC_{m,i,t}, \qquad \omega_m = \frac{\lambda_m}{\sum_{j=1}^{M} \lambda_j},$$

where λ_m denotes the eigenvalue (explained variance) of component m. Weighting retained components by explained variance preserves their relative contribution to the composite profitability signal.

Three nonlinear classifiers are estimated on the same feature set. An RBF-kernel SVM models nonlinear decision boundaries after standardization (Cortes & Vapnik, 1995; Vapnik, 2013). Random forest improves stability by aggregating decorrelated trees under nonlinear interactions (Breiman, 2001). XGBoost builds shallow trees sequentially and relies on regularization and subsampling to limit overfitting (Friedman, 2001; Chen & Guestrin, 2016). Forecasting is evaluated under a strict time split: training targets are restricted to years up to 2020, while the holdout test period covers 2021–2024. Hyperparameters are selected via validation procedures that preserve temporal ordering. Model selection prioritizes the F1-score due to mild class imbalance in the constructed label, while AUC is reported as a threshold-free measure of ranking performance (Fawcett, 2006).

https://doi.org/10.71058/jodac.v10i03002

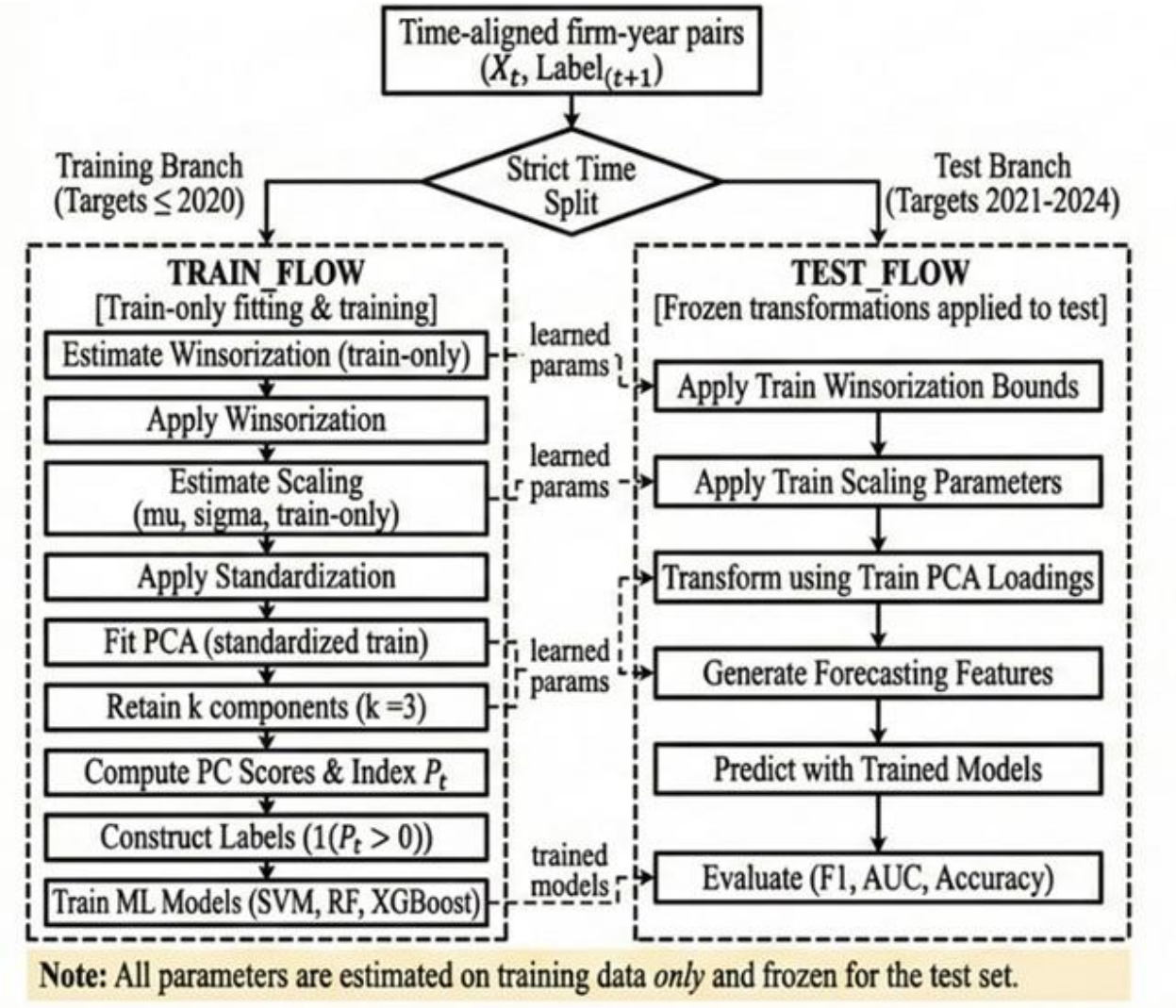


Figure 2. Leakage-safe PCA–ML pipeline for next-year profitability forecasting

Notes: Winsorization bounds, scaling parameters, and principal component analysis loadings are estimated from the training period only and then applied unchanged to the 2021–2024 test set.

Source: By author

The time-ordered split is designed to reflect a realistic decision setting in which only information available at the end of year t is used to predict outcomes in year $t + 1$. Each firm-year contributes a feature vector $X_{i,t}$ measured at t, while the target label is defined at $t + 1$ from $P_{i,t+1}$. Observations are aligned so that predictors and labels are separated by exactly one year, and the last firm-year is removed because $Label_{i,t+1}$ cannot be observed. The chronological split is applied

https://doi.org/10.71058/jodac.v10i03002

on the label year: training targets satisfy $t + 1 \leq 2020$, while holdout evaluation uses $2021 \leq t + 1 \leq 2024$. This implies training predictors are drawn from $t \leq 2019$ and testing predictors from $2020 \leq t \leq 2023$. Crucially, the split is implemented prior to any data-driven preprocessing so that winsorization bounds, standardization parameters, and PCA loadings are learned strictly from the training window and carried forward unchanged, preventing look-ahead bias and ensuring feasible out-of-sample prediction under an information-set-consistent pipeline (Bergmeir & Benítez, 2012).

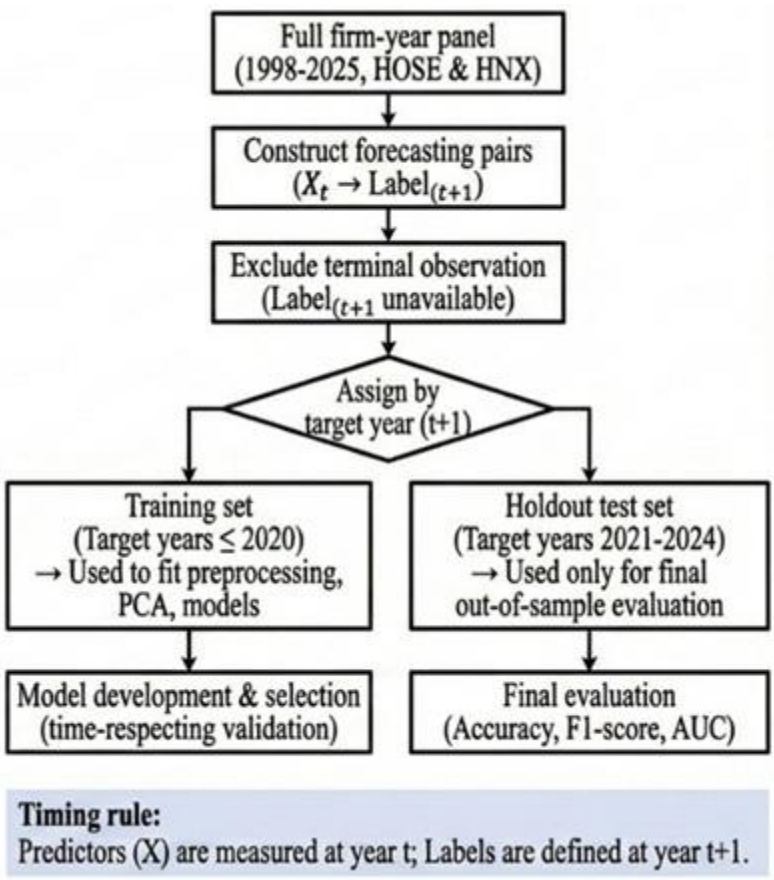


Figure 3. Time-ordered train–test split for next-year profitability prediction

Notes: Predictors are measured at year t, labels at year $t + 1$; training targets are ≤ 2020, and the holdout test period is 2021–2024.

Source: By author

https://doi.org/10.71058/jodac.v10i03002

For transparency and reproducibility, the main specification can be summarized as follows. The input feature set comprises five profitability proxies measured at year t:

$$X_{i,t} = \{X1\_ROA, X2\_ROE, X3\_ROC, X4\_EPS, X5\_NPM\}.$$

The target is the next-year binary profitability status $Label_{i,t+1}$, defined as $1(P_{i,t+1} > 0)$, where $P_{i,t+1}$ is the PCA-based composite index computed at year $t+1$. The pipeline is executed under strict leakage-safe preprocessing: (i) chronological splitting by label year (training targets ≤ 2020; test targets 2021–2024), (ii) training-only winsorization at the 1% tails $(q = 0.01)$ with fixed bounds applied to the test sample, (iii) training-only standardization with fixed parameters applied to the test set, and (iv) PCA fitted on standardized training data retaining $M = 3$ components with cumulative explained variance ≥ 0.85. Finally, SVM (RBF), RF, and XGBoost are trained to learn the mapping $X_{i,t} \rightarrow Label_{i,t+1}$, with hyperparameter tuning that respects temporal ordering; models are compared primarily using F1-score, with AUC reported for threshold-free ranking performance (Fawcett, 2006).

3.5. Evaluation Metrics

Out-of-sample predictive performance is assessed using accuracy, precision, recall, F1-score, and AUC. While accuracy captures the overall share of correct classifications, it may be uninformative when the class distribution is imbalanced. Accordingly, the F1-score serves as the primary criterion for model selection because it summarizes precision and recall within a single threshold-based metric. In addition, AUC is reported to reflect threshold-invariant ranking ability through receiver operating characteristic (ROC) analysis (Fawcett, 2006).

3.6. Robustness Designs

Robustness checks are conducted along two dimensions to determine whether the main findings are sensitive to the labeling rule and the PCA dimensionality choice. First, an alternative labeling scheme is considered in which the year-median of the composite index is used as the cutoff, replacing the zero-threshold applied in the baseline specification. Second, the PCA stage is re-estimated with a different number of retained components (M=4M=4M=4) instead of the main setting. Together, these tests help distinguish the stability of the predictive results from specific measurement-layer design choices and evaluate whether modest adjustments in label construction or retained dimensionality materially affect out-of-sample forecasting performance.

https://doi.org/10.71058/jodac.v10i03002

4. Results and discussion

4.1. Tail behavior, outliers, and co-movement among profitability proxies

Profitability indicators derived from accounting ratios and per-share measures are typically heavy-tailed, sensitive to denominator effects, and prone to extreme observations. If untreated, these characteristics can weaken both covariance-based extraction and nonlinear classification. Table 2 highlights pronounced tail risk in the raw inputs ($N = 9{,}727$), with the most severe dispersion observed in EPS and NPM. Prior to winsorization, EPS ranges from $-16{,}889.77$ to $407{,}572.91$, while NPM spans -114.11 to $1{,}285.78$, indicating variability far beyond the central portion of the distributions.

Table 2. Outlier impact (before vs. after winsorization at 1% tails)

Panel A. Before winsorization

Variable	N	Mean	SD	Min	P1	P99	Max
ROA	9727	0.07	0.08	-1.59	-0.11	0.33	0.78
ROE	9727	0.12	0.26	-12.61	-0.27	0.51	2.01
ROC	9727	0.17	0.26	-1.69	-0.22	1.08	8.45
EPS	9727	1,741.77	4,963.19	-16,889.77	-2,228.87	10,492.79	407,572.91
NPM	9727	0.26	14.44	-114.11	-0.37	0.77	1,285.78

Panel B. After winsorization

Variable	Mean	SD	Min	P1	P99	Max
ROA	9727	0.07	0.07	-0.09	-0.09	0.33
ROE	9727	0.13	0.12	-0.24	-0.24	0.51
ROC	9727	0.16	0.18	-0.18	-0.18	0.95
EPS	9727	1,636.75	1,854.60	-1,908.94	-1,908.94	10,076.08
NPM	9727	0.10	0.14	-0.29	-0.29	0.70

Notes: Winsorization bounds are computed from training data only. Source: By author

Applying 1% tail winsorization with training-only cutoffs markedly compresses dispersion and truncates extremes while leaving interior quantiles largely preserved. This supports the use of robust preprocessing prior to standardization and PCA, consistent with evidence that financial

https://doi.org/10.71058/jodac.v10i03002

ratios are often non-normal and outlier-sensitive (Deakin, 1976; Frecka & Hopwood, 1983; McLeay & Omar, 2000).

Co-movement across proxies is also substantial. Table 3 reports selected correlations within the TRAIN fitting window. The return-based indicators are strongly correlated (e.g., ROA–ROE = 0.81; ROE–ROC = 0.65), reflecting shared accounting information and motivating PCA as a structured approach to reduce redundancy when multicollinearity is present (Jolliffe, 2002). NPM is also positively related to ROA and ROE, indicating that margin variation partly aligns with the common profitability factor while still preserving channel-specific content.

Table 3. Key correlations (evidence of co-movement)

Pair	Corr
ROA–ROE	0.81
ROA–ROC	0.64
ROE–ROC	0.65
ROA–NPM	0.52
ROE–NPM	0.43

Source: By author

4.2. PCA adequacy and component structure

PCA is conducted on the five screened proxies after training-only winsorization and training-only standardization. Suitability diagnostics indicate that the correlation matrix contains sufficient common covariance for stable extraction. Table 4 reports the Kaiser–Meyer–Olkin (KMO) measure and Bartlett's sphericity test. The overall KMO is 0.77 (minimum KMO across variables: 0.72), suggesting adequate sampling adequacy. Bartlett's test rejects the null of an identity matrix ($p < 0.001$), supporting PCA implementation (Kaiser & Rice, 1974; Bartlett, 1951).

https://doi.org/10.71058/jodac.v10i03002

Table 4. Principal component analysis suitability tests (compact)

Test	Statistic	Value
Kaiser–Meyer–Olkin overall	KMO	0.77
Kaiser–Meyer–Olkin minimum (among variables)	KMO min	0.72
Bartlett's test of sphericity	chi-square (degrees of freedom)	14,265.07 (10)
Bartlett's test of sphericity	p-value	< 0.001

Source: By author

The retained-component variance profile is summarized in Table 5. Three components are kept because cumulative explained variance reaches 0.89, exceeding the pre-specified 0.85 threshold. This achieves meaningful dimensionality reduction while retaining most systematic variation in the proxy set (Jolliffe, 2002; Wold et al., 1987).

Table 5. Principal component analysis explained variance (retained components only)

Component	Explained variance ratio	Cumulative
PC1	0.58	0.58
PC2	0.18	0.76
PC3	0.13	0.89

Source: By author

The variance profile is visualized in Figure 4, which corroborates retaining three components to exceed the 0.85 cumulative threshold while avoiding negligible incremental variance from later components.

https://doi.org/10.71058/jodac.v10i03002

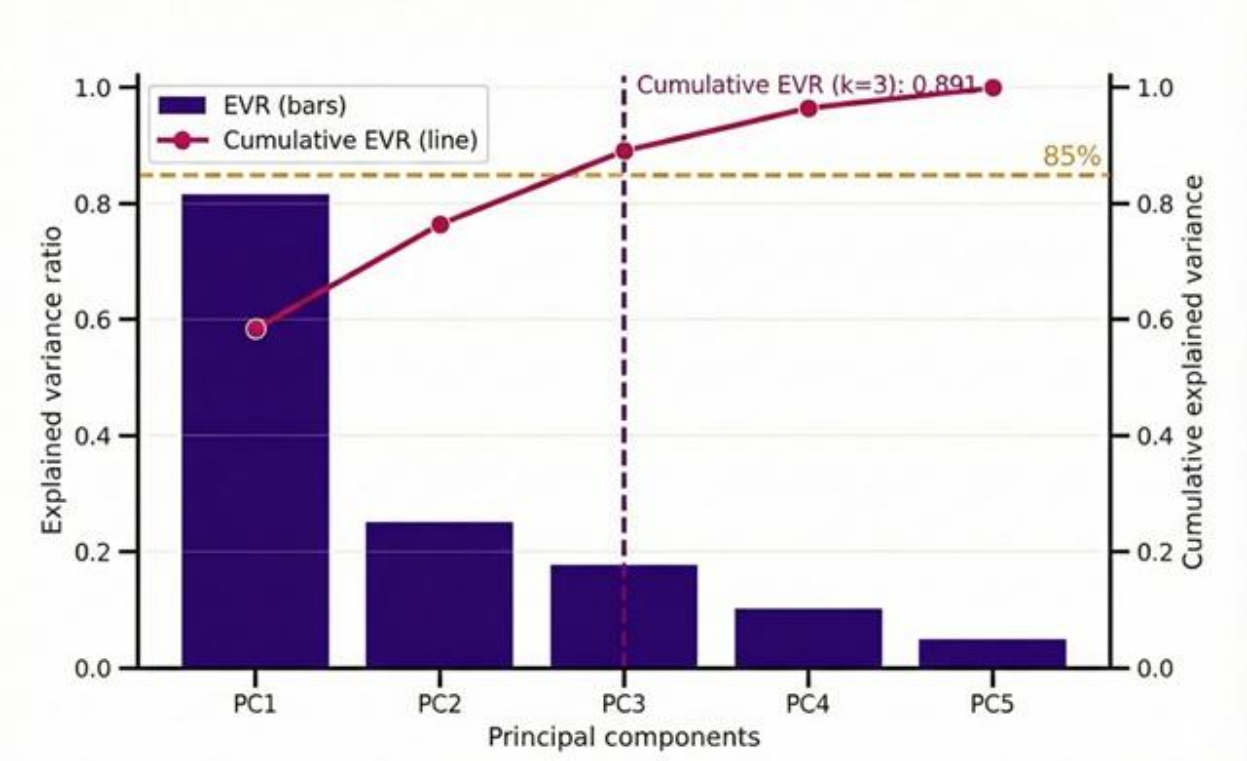


Figure 4. Principal component analysis variance explained (Pareto)

Notes: Three components reach cumulative explained variance 0.891 (> 0.850), while later components contribute little incremental variance and are excluded to avoid adding noise (Zhang et al., 2015).

Source: By author

The loading patterns are economically interpretable. Table 6 presents dominant loadings for the retained components (|loading| ≥ 0.35). PC1 loads positively on ROA, ROE, and ROC and also on NPM, reflecting a broad profitability factor combining return efficiency and margin information. PC2 is primarily driven by EPS with a negative contribution from NPM, capturing a per-share dimension contrasted against margin variation. PC3 is dominated by NPM and also loads positively on EPS, indicating an orthogonal margin-oriented component that still contains some per-share variation. Overall, PCA separates correlated profitability measures into orthogonal latent dimensions aligned with distinct economic channels (Jolliffe, 2002).

https://doi.org/10.71058/jodac.v10i03002

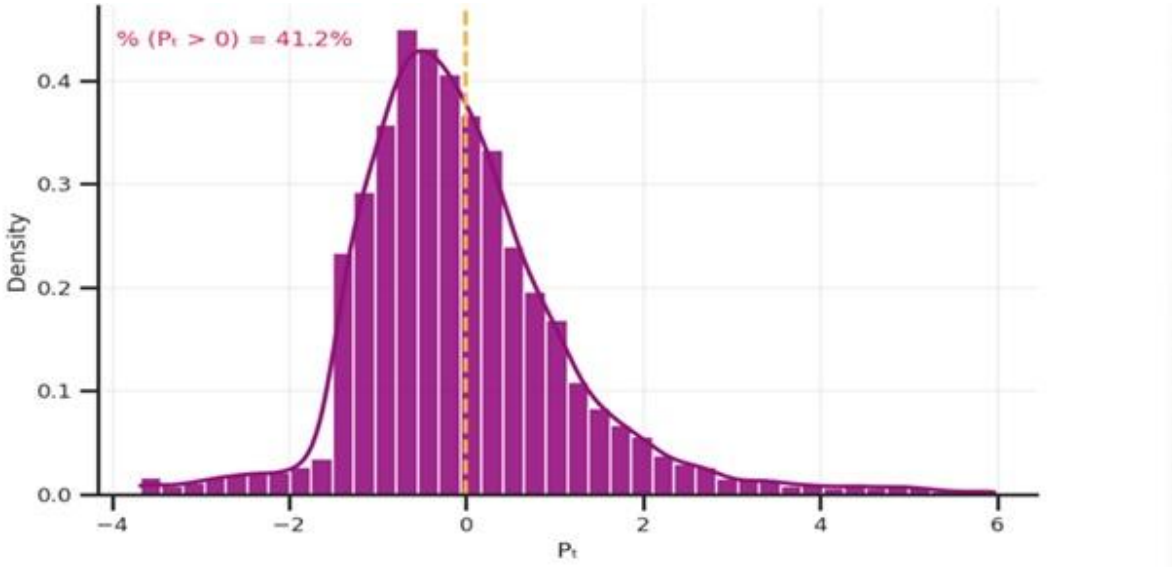
Table 6. Dominant principal component analysis loadings (|loading| ≥ 0.35, k = 3)

PC	Variable	Loading
PC1	NPM	0.373
PC1	ROA	0.532
PC1	ROC	0.482
PC1	ROE	0.520
PC2	EPS	0.851
PC2	NPM	-0.470
PC3	EPS	0.419
PC3	NPM	0.773

Source: By author

4.3. Composite index distribution and labeling implications

Figure 5 displays the distribution of the composite profitability index. The mass of observations clusters near zero, implying that many firm-years lie close to the decision boundary; small shifts in underlying proxies can therefore change the sign of the index and flip the assigned label. This feature, combined with mild class imbalance, reinforces the use of F1-score as the primary selection metric because it balances precision and recall (Martikainen et al., 1995).



https://doi.org/10.71058/jodac.v10i03002

Figure 5. Composite profitability index distribution

Notes: Mild right-skew; vertical line at zero defines the labeling threshold; share above zero is 41.2% (mild class imbalance).

Source: By author

Class separation in PCA space is further illustrated in Figure 6, which plots three-dimensional component scores by the next-year label. The substantial overlap across classes in $(PC1, PC2, PC3)$ indicates weak linear separability, supporting the use of nonlinear learners and ensembles. Such overlap suggests that the label structure is governed by nonlinear interactions and conditional patterns rather than a single linear direction in PCA space (Lam, 2004; Cortes & Vapnik, 1995; Breiman, 2001).

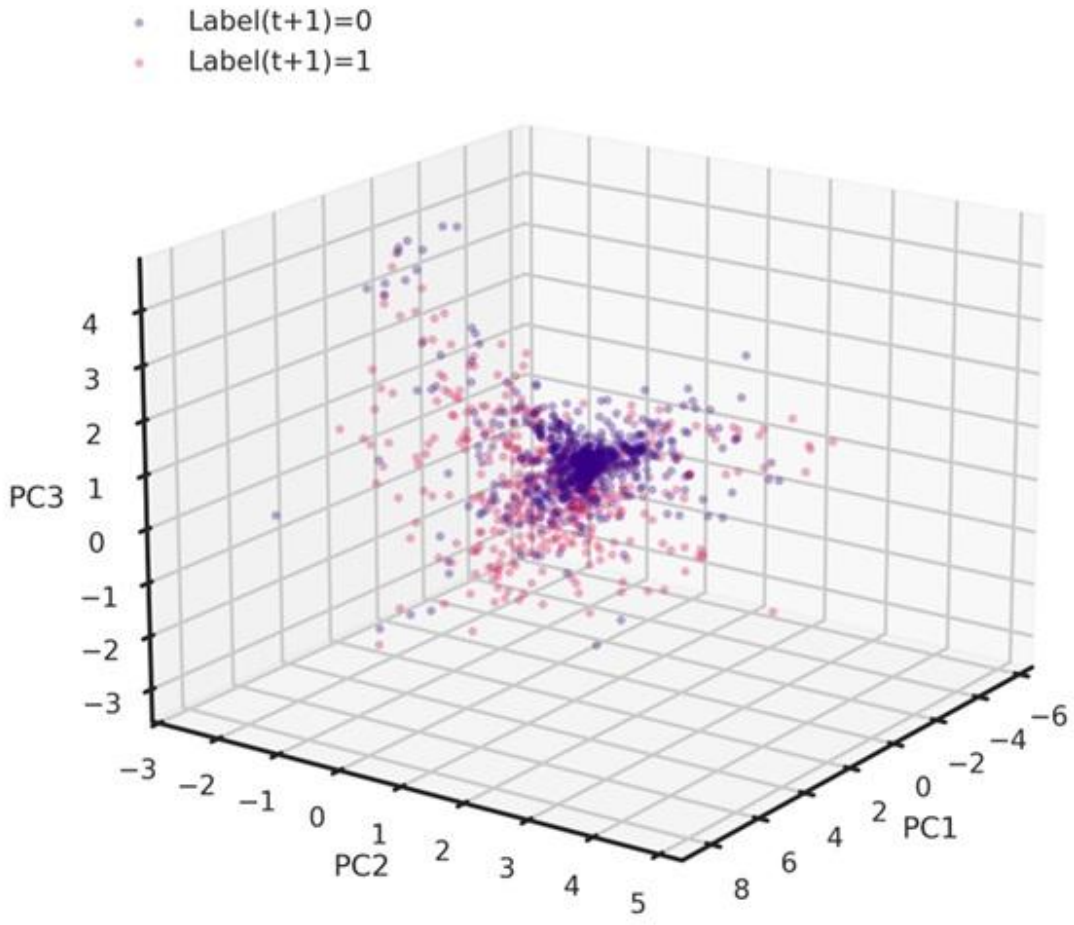


Figure 6. Three-dimensional principal component scores by next-year label

https://doi.org/10.71058/jodac.v10i03002

Notes: Substantial overlap between classes in (PC1, PC2, PC3), implying weak linear separability and motivating nonlinear classifiers and ensembles.

Source: By author

4.4. Holdout forecasting performance

Forecast accuracy is evaluated on the strict holdout period (2021–2024) under training-only preprocessing and a time-ordered split. Table 7 reports accuracy, F1-score, and AUC for SVM (RBF), Random Forest, XGBoost, and a majority-class baseline. All three models outperform the baseline by a wide margin, indicating that current-year proxies contain meaningful information for predicting next-year profitability status under a deployable leakage-safe pipeline.

Table 7. Out-of-sample performance (2021–2024, minimal)

Model	Accuracy	F1-score	AUC
SVM (RBF)	0.831	0.763	0.860
XGBoost	0.832	0.759	0.878
Random forest	0.828	0.754	0.879
Baseline (majority)	0.644	0.000	0.500

Notes: Baseline is majority class. Test set $n = 2471$ *(positive = 879, negative = 1592).*

Source: By author

SVM (RBF) delivers the strongest threshold-based performance (highest F1), whereas XGBoost records the highest accuracy. In contrast, AUC is largest for Random Forest (0.878774) and XGBoost (0.877932), both exceeding the SVM's AUC. The tight clustering of performance across model families suggests that predictability is not specific to one algorithm but reflects stable signal content in the profitability proxy system. These findings align with maximum-margin nonlinear classification (Cortes & Vapnik, 1995), boosted-tree learning for nonlinear interactions and conditional splits (Friedman, 2001; Chen & Guestrin, 2016), and averaging-based ensembles in random forests (Breiman, 2001).

https://doi.org/10.71058/jodac.v10i03002

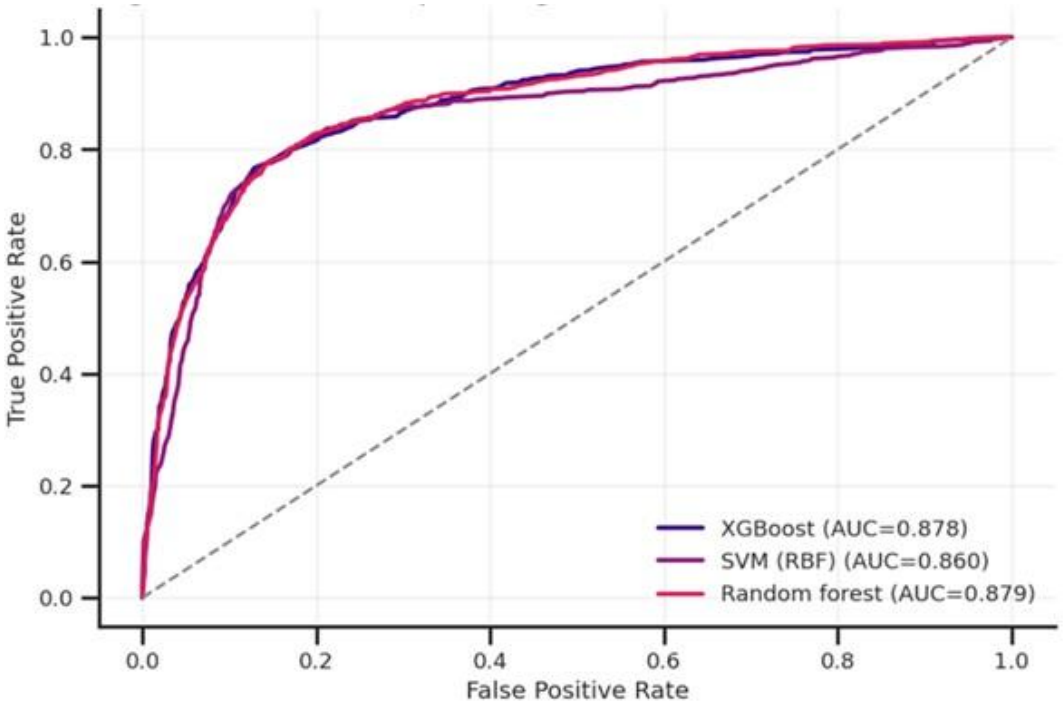


Figure 7. Receiver operating characteristic curves (test set)

Notes: Curves lie well above the random diagonal; Random forest and XGBoost achieve AUC about 0.878, above the support vector machine.

Source: By author

To complement ROC evidence, Table 8 reports the confusion matrix for XGBoost. The model correctly classifies most negative observations while retaining non-trivial recall for the positive class.

Table 8. Confusion matrix (XGBoost, test set)

	Pred 0	**Pred 1**
Actual 0	1405	187
Actual 1	227	652

Source: By author

https://doi.org/10.71058/jodac.v10i03002

4.5. Model interpretation and robustness

Feature-importance diagnostics provide a consistent narrative across tree-based models. Figure 8 indicates that Random Forest places greatest weight on ROA and ROE, with EPS, ROC, and NPM contributing secondary information—consistent with the dominance of return-efficiency proxies in the primary PCA factor and with the ensemble-tree structure of random forests (Breiman, 2001).

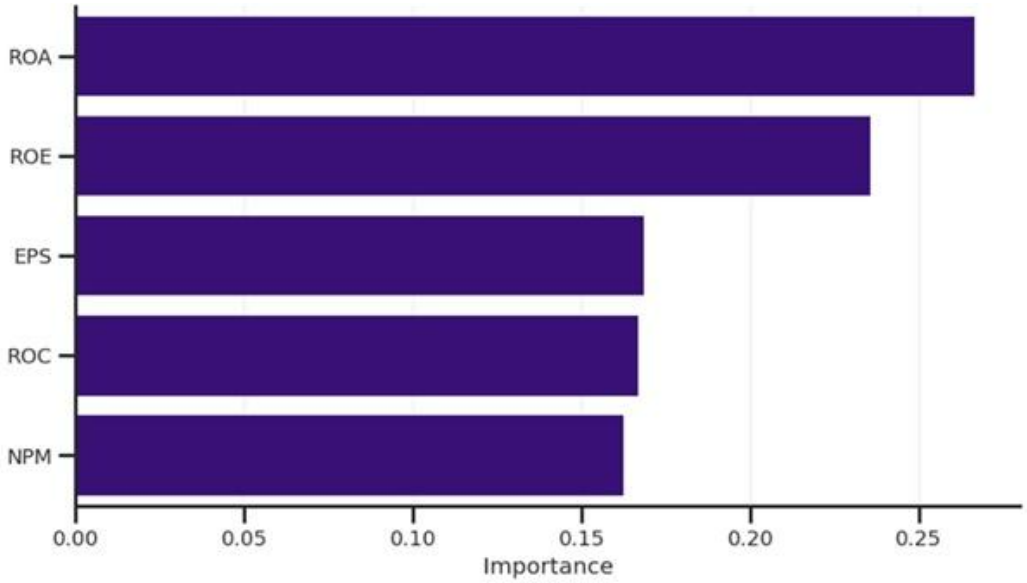


Figure 8. Random forest feature importance

Notes: ROA and ROE dominate; EPS, ROC, and NPM contribute moderately.

Source: By author

Similarly, Figure 9 shows that XGBoost is most sensitive to ROA, followed by ROE and EPS, while NPM and ROC appear mainly as conditional refinements. This pattern is consistent with gradient boosting, where early splits concentrate on the most discriminative predictors and subsequent trees capture residual structure (Friedman, 2001; Chen & Guestrin, 2016).

https://doi.org/10.71058/jodac.v10i03002

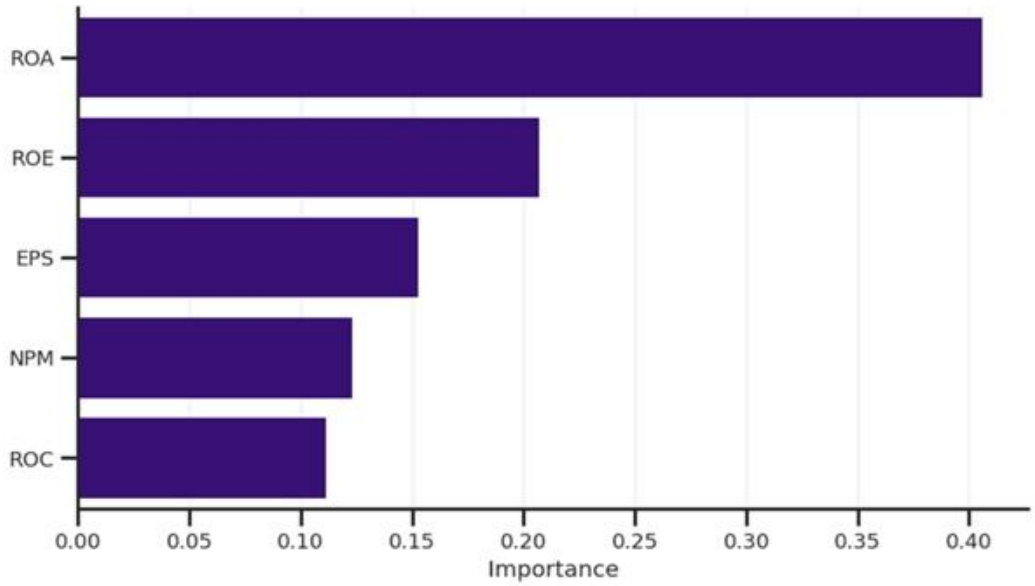


Figure 9. XGBoost feature importance

Notes: ROA dominates strongly, followed by ROE and EPS; NPM and ROC contribute as conditional refinements.

Source: By author

Robustness is assessed without changing the screened proxy set, focusing instead on two design choices that can affect conclusions: the label threshold and the number of retained PCA components. Table 9 summarizes the main and robustness specifications. Performance remains stable under (i) an alternative labeling rule based on the within-year median of the composite index and (ii) a higher retained dimensionality ($k = 4$). Although the top-performing model can vary across specifications (SVM in MAIN versus XGBoost in ROB A/ROB B), predictive quality remains consistently strong, implying that results are not driven by a single threshold choice or a single PCA dimensionality setting.

https://doi.org/10.71058/jodac.v10i03002

Table 9. Robustness summary (compact)

Spec	k	Label rule	CumEVR(k)	KMO min	Best model	F1	AUC
MAIN	3	P_t > 0	0.888	0.717	SVM (RBF)	0.763	0.860
ROB A	3	P_t > median(P|YEAR)	0.888	0.717	XGBoost	0.769	0.859
ROB B	4	P_t > 0	0.964	0.717	XGBoost	0.765	0.878

Source: By author

Overall, the evidence supports an integrated measurement-to-prediction approach to profitability. First, the descriptive statistics confirm heavy tails and sensitivity to outliers in raw profitability proxies, validating the use of training-only winsorization prior to standardization and PCA (Deakin, 1976; Frecka & Hopwood, 1983; McLeay & Omar, 2000). Second, strong co-movement across proxies and favorable PCA diagnostics indicate that a small number of components can capture dominant shared variation while separating per-share and margin-related dimensions (Jolliffe, 2002; Wold et al., 1987). Third, strict out-of-sample results demonstrate that next-year profitability status is meaningfully predictable under leakage-safe preprocessing and time-ordered evaluation, with nonlinear classifiers producing strong F1 and AUC on the 2021–2024 holdout set (Bergmeir & Benítez, 2012; Fawcett, 2006). Robustness checks further show that these conclusions persist under reasonable variations in label definition and retained PCA dimensionality.

5. Conclusions and recommendations

5.1. Conclusions

This study assessed whether profitability—viewed as a multidimensional construct—can be measured objectively and forecast credibly in an emerging-market context through an integrated measurement-to-prediction framework. Using firm-year financial statement data for publicly listed non-financial firms in Vietnam (HOSE and HNX) over 1998–2025, the analysis combines leakage-safe preprocessing, PCA-based composite index formation, and strict time-ordered out-of-sample classification of next-year profitability status. Evidence from the 2021–2024 holdout evaluation

https://doi.org/10.71058/jodac.v10i03002

indicates that the proposed end-to-end design can simultaneously deliver (i) a statistically well-defined composite profitability index and (ii) practically meaningful predictive performance under deployment-consistent testing conditions.

First, the data confirm that profitability is both multidimensional and highly sensitive to extreme observations. The proxy set exhibits heavy tails and large outliers—most notably in EPS and NPM—making robust treatment essential before any variance-based extraction or nonlinear modeling. Training-only winsorization at the 1% tails materially compresses dispersion and truncates extremes while leaving interior quantiles broadly intact, supporting robust preprocessing as a prerequisite for stable estimation and model training. Correlation patterns within the TRAIN fitting window further reveal strong co-movement across return-based measures (e.g., ROA–ROE ≈ 0.81; ROE–ROC ≈ 0.65), reinforcing the rationale for compressing overlapping information rather than relying on a single indicator.

Second, PCA is empirically appropriate and produces a compact yet information-preserving representation. Suitability tests indicate sufficient shared covariance in the screened five-proxy system (KMO overall = 0.7749; KMO minimum = 0.7175), and Bartlett's test strongly rejects an identity correlation matrix ($\chi^2 = 14{,}265.07$, df = 10, $p < 0.001$) (Kaiser & Rice, 1974; Bartlett, 1951). In the main specification, three components are retained because cumulative explained variance reaches 0.8881 (> 0.85), achieving substantial dimensionality reduction without discarding most systematic variation. The loading structure also admits a clear economic interpretation: PC1 captures a broad profitability factor driven by ROA, ROE, and ROC with an additional contribution from NPM; PC2 primarily reflects per-share performance via EPS with a countervailing NPM loading; PC3 is margin-centered (NPM) with remaining EPS content. This structure strengthens the substantive meaning of the composite index and supports its use for screening and ongoing monitoring.

Third, next-year profitability status derived from the composite index is predictably out of sample under strict, leakage-safe evaluation. The composite-index distribution implies mild imbalance, with approximately 41.16% of observations above zero, motivating the use of F1-score as the primary selection criterion. Under a strict chronological split (training targets through 2020; holdout testing over 2021–2024), all nonlinear classifiers substantially outperform the majority baseline (baseline accuracy ≈ 0.6443; baseline F1 = 0). In the baseline specification, SVM (RBF)

https://doi.org/10.71058/jodac.v10i03002

yields the strongest threshold-based performance (F1 = 0.7635; accuracy ≈ 0.8312), whereas tree ensembles provide the strongest ranking ability by AUC (Random Forest ≈ 0.8788; XGBoost ≈ 0.8779), both exceeding the SVM's AUC (≈ 0.8600). Consistently strong F1 and AUC across model families indicates that predictability reflects stable signal content in year-t proxies for year-$t+1$ profitability status, consistent with profitability persistence and partial mean-reversion dynamics.

Fourth, forecasting reliability is driven by the integrity of the full pipeline, not by the choice of classifier alone. Winsorization bounds, standardization parameters, and PCA loadings are estimated strictly on the training period and then propagated forward unchanged to the 2021–2024 holdout window. This end-to-end chronological discipline ensures that reported performance corresponds to feasible real-time prediction rather than inflated backtest metrics caused by preprocessing leakage. The results therefore reinforce a practical methodological requirement: credible financial forecasting must apply time-ordered evaluation and training-only estimation for every data-dependent transformation.

Fifth, interpretability patterns are consistent with the composite index structure. Feature-importance evidence indicates that ROA and ROE account for the largest predictive contributions, with EPS, ROC, and NPM serving as secondary refinements. This aligns with the PCA decomposition in which the first component explains the largest variance share and is dominated by return-efficiency signals, strengthening coherence between the measurement layer (index construction) and the prediction layer (classification).

Sixth, the main conclusions remain stable under reasonable variations in labeling and PCA dimensionality. When the label cutoff is replaced by a within-year median threshold (ROB A), XGBoost becomes the top performer and predictive strength remains high (F1 ≈ 0.7690; AUC ≈ 0.8590). When the number of retained components increases to $k = 4$ (ROB B), XGBoost again performs best with comparable accuracy (F1 ≈ 0.7654; AUC ≈ 0.8783). Across specifications, the identity of the best model may change, but overall predictability remains consistently strong, suggesting that findings are not driven by a single threshold rule or a single PCA retention choice.

Overall, the study meets its objectives: profitability can be summarized by an objective and economically interpretable PCA-based composite index constructed under leakage-safe

https://doi.org/10.71058/jodac.v10i03002

preprocessing, and next-year profitability status derived from that index can be predicted credibly out of sample under strict time ordering.

5.2. Recommendations

For investors and portfolio managers. Equity screening in Vietnam should move beyond single-ratio filters toward composite profitability assessment. The evidence suggests that a PCA-based index preserves the dominant shared profitability signal while retaining orthogonal per-share and margin information. Practically, the index can serve as (i) a primary screening input for watchlists and portfolio maintenance and (ii) a monitoring gauge for early detection of profitability deterioration. Forecast outputs should be used as decision-support prioritization rather than a standalone trading rule, and classification thresholds should be calibrated to the investor's tolerance for false positives versus false negatives—particularly given the concentration of observations near the zero cutoff.

For corporate managers and internal performance monitoring teams. Internal dashboards should not rely on a single profitability metric for evaluation, early warning, or resource allocation. A tiered monitoring structure is recommended: use the composite profitability index as a headline KPI and supplement it with component-level diagnostics (return efficiency vs. per-share vs. margin channels) to identify the source of changes. Given the consistent dominance of ROA/ROE signals, asset efficiency and equity returns should be treated as first-order indicators, while EPS, ROC, and NPM can be used as conditional diagnostics to interpret shifts and support targeted corrective actions.

For regulators, exchanges, and data infrastructure providers. Market institutions should promote reproducible, leakage-safe standards when publishing analytics or evaluating model-based claims. Results show that leakage can arise not only through target contamination but also through preprocessing steps (winsorization cutoffs, scaling parameters, PCA loadings). Formal guidance that requires strict chronological splits and training-only transformations would strengthen transparency and reduce the risk of overstated backtest performance in applied research and practitioner reporting.

For analysts and quantitative research teams. Teams building profitability forecasting systems should treat measurement integrity and predictive usefulness as connected but distinct design

https://doi.org/10.71058/jodac.v10i03002

goals. The framework demonstrated here supports an explicit pipeline: (i) training-only outlier handling, (ii) training-only standardization, (iii) PCA-based index construction and label definition, and (iv) time-ordered out-of-sample evaluation. Model selection should emphasize F1-score when threshold decisions matter under imbalance, while AUC should be reported to quantify ranking strength for monitoring and prioritization tasks.

5.3. Limitations and future research

Several limitations warrant attention. First, external validity may be limited because evidence is drawn from publicly listed non-financial firms in Vietnam. Future studies should replicate the framework in other emerging and developed markets, and in private-firm settings, to evaluate whether PCA structure, class balance, and forecastability vary across institutional environments.

Second, the outcome is binary and threshold-based. Defining profitability status by the sign of the composite index supports classification and decision support but compresses information about magnitude. Future work could extend the framework to (i) continuous prediction of the composite index level or its change, (ii) ordinal multi-class labeling (e.g., low/medium/high profitability regimes), or (iii) cost-sensitive objectives aligned more directly with investment or managerial loss functions.

Third, the feature set is deliberately narrow. The main design focuses on profitability proxies to preserve a clean measurement-to-prediction link. Future research could incorporate broader firm characteristics (size, leverage, cash flow, growth), sector structure, and macro-financial variables to test incremental predictive value while maintaining the same leakage-safe discipline. Additional validation designs (e.g., rolling-origin evaluation; expanding versus rolling windows) and richer interpretability approaches (e.g., time-consistent SHAP analyses) would further enhance deployability and decision transparency.

Taken together, these extensions would refine both the measurement and forecasting components of the framework and broaden the applicability of composite-profitability forecasting for investors, firms, and market-supporting institutions.

https://doi.org/10.71058/jodac.v10i03002

References

Apu, K. U., Rahman, M. M., Hoque, A. B., & Bhuiyan, M. (2022). Forecasting future investment value with machine learning, neural networks, and ensemble learning: A meta-analytic study. *Review of Applied Science and Technology, 1*(02), 01–25.

Bartlett, M. S. (1951). The effect of standardization on a χ^2 approximation in factor analysis. *Biometrika, 38*(3/4), 337–344.

Bartlett, M. S. (1954). A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 296–298.

Belesis, N. D., Papanastasopoulos, G. A., & Vasilatos, A. M. (2023). Predicting the profitability of directional changes using machine learning: Evidence from European countries. *Journal of Risk and Financial Management, 16*(12), 520.

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences, 191*, 192–213.

Biddle, G. C., Hilary, G., & Verdi, R. S. (2009). How does financial reporting quality relate to investment efficiency? *Journal of Accounting and Economics, 48*(2–3), 112–131.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Deakin, E. B. (1976). Distributions of financial accounting ratios: Some empirical evidence. *The Accounting Review, 51*(1), 90–96.

https://doi.org/10.71058/jodac.v10i03002

Fama, E. F., & French, K. R. (2000). Forecasting profitability and earnings. *The Journal of Business, 73*(2), 161–175.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Frecka, T. J., & Hopwood, W. S. (1983). The effects of outliers on the cross-sectional distributional properties of financial ratios. *The Accounting Review, 58*(1), 115–128.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Gujarati, D. N. (2012). *Basic econometrics* (4th ed.).

Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Springer.

Jolliffe, I. (2025). Principal component analysis. In *International Encyclopedia of Statistical Science* (pp. 1945–1948). Springer Berlin Heidelberg.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*(1), 31–36.

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement, 34*(1), 111–117.

Kapoor, S., & Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science. *arXiv preprint arXiv:2207.07048*.

Kothari, S. P. (2001). Capital markets research in accounting. *Journal of Accounting and Economics, 31*(1–3), 105–231.

Lam, M. (2004). Neural network techniques for financial performance prediction: Integrating fundamental and technical analysis. *Decision Support Systems, 37*(4), 567–581.

https://doi.org/10.71058/jodac.v10i03002

Lev, B., & Thiagarajan, S. R. (1993). Fundamental information analysis. *Journal of Accounting Research, 31*(2), 190–215.

Martikainen, T., Perttunen, J., Yli-Olli, P., & Gunasekaran, A. (1995). Financial ratio distribution irregularities: Implications for ratio classification. *European Journal of Operational Research, 80*(1), 34–44.

McLeay, S., & Omar, A. (2000). The sensitivity of prediction models to the non-normality of bounded and unbounded financial ratios. *The British Accounting Review, 32*(2), 213–230.

Nguyen, T. N. L., & Nguyen, V. C. (2020). The determinants of profitability in listed enterprises: A study from Vietnamese stock exchange. *Journal of Asian Finance, Economics and Business, 7*(1), 47–58.

Nissim, D., & Penman, S. H. (2001). Ratio analysis and equity valuation: From research to practice. *Review of Accounting Studies, 6*(1), 109–154.

Penman, S. H. (2010). *Financial statement analysis and security valuation*. McGraw-Hill/Irwin.

Popa, D. C. S., Popa, D. N., Bogdan, V., & Simut, R. (2021). Composite financial performance index prediction—a neural networks approach. *Journal of Business Economics and Management, 22*(2), 277–296.

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-STS330

Tran, T., Nguyen, N. H., Le, B. T., Thanh Vu, N., & Vo, D. H. (2023). Examining financial distress of the Vietnamese listed firms using accounting-based models. *PLoS One, 18*(5), e0284451.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1–3), 37–52.

https://doi.org/10.71058/jodac.v10i03002

Zhang, H., Yang, F., Li, Y., & Li, H. (2015). Predicting profitability of listed construction companies based on principal component analysis and support vector machine—Evidence from China. *Automation in Construction, 53*, 22–28.