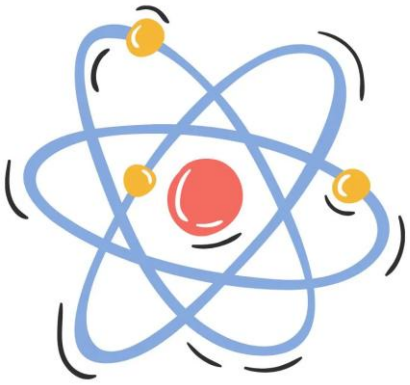

ISSN: 1672 - 6553

JOURNAL OF DYNAMICS AND CONTROL

VOLUME 9 ISSUE 7: 40 - 59

FREE CASH FLOW AND OVERINVESTMENT: MACHINE LEARNING PERSPECTIVES IN VIETNAM

Phong Nguyen Anh, Tam Phan Huy, Thanh Ngo Phu

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

JOURNAL OF DYNAMICS AND CONTROL
DOI: https://doi.org/10.71058/jodac.v9i7004

FREE CASH FLOW AND OVERINVESTMENT: MACHINE LEARNING PERSPECTIVES IN VIETNAM

Phong Nguyen Anh, Tam Phan Huy*, Thanh Ngo Phu
University of Economics and Law and Vietnam National University, Ho Chi Minh City, Vietnam
Corresponding author: tamph@uel.edu.vn

Funding: This research is funded by Vietnam National University, Ho Chi Minh City (VNU-HCM) under grant number DM2024-34-01.

Abstract: *This study develops and evaluates a comprehensive machine learning framework to classify overinvestment among Vietnamese listed firms using firm-level financial and governance variables. Drawing on agency theory and prior empirical research, overinvestment is defined as firms with free cash flow above the sample median and Tobin's Q below the median. The dataset includes 6,561 firm-year observations from 2000 to 2024, covering companies listed on the Hanoi and Ho Chi Minh Stock Exchanges. Seven classification algorithms were compared: Logistic Regression, Random Forest, XGBoost, LightGBM, Support Vector Machine, K-Nearest Neighbors, and Artificial Neural Networks. Performance was assessed via 10-fold cross-validation across multiple metrics including accuracy, precision, recall, F1 score, AUC-ROC, MCC, Brier score, and computational efficiency. Results show that ensemble methods, particularly Random Forest and XGBoost, achieved the highest predictive performance, while simpler classifiers like SVM and KNN consistently underperformed. The findings confirm that advanced machine learning techniques can effectively model the nonlinear and heterogeneous determinants of overinvestment. This study contributes to the literature by demonstrating the applicability of modern predictive analytics in an emerging market context and providing evidence that supports agency theory perspectives on investment inefficiency.*

Keywords: **Overinvestment, Machine Learning, Ensemble Methods, Investment Efficiency, Vietnam.**

JEL codes: G32, C45, M41

1. Introduction

Investment efficiency has long been a central concern in corporate finance, as suboptimal allocation of resources can have profound implications for firm performance, valuation, and stakeholder welfare. Overinvestment arises when firms commit capital to projects whose returns do not justify their costs, often driven by managerial incentives, agency conflicts, or behavioral biases (Biddle et al., 2009; Jensen, 1986). This problem is especially pronounced in emerging economies, where governance mechanisms are weaker, disclosure practices less transparent, and monitoring by external stakeholders more constrained (Chen et al., 2011; Le & Tannous, 2016). In Vietnam, the rapid development of capital markets over the past two decades has been accompanied by increased concerns about inefficient investment behaviors among listed firms. However, despite the growing relevance of this issue, empirical research specifically focused on systematically detecting and classifying overinvestment in Vietnamese firms remains limited.

Recent studies have made significant advances in understanding the determinants of overinvestment, including the role of free cash flow, growth opportunities, ownership structures, and disclosure quality (Al Dah et al., 2023; Le et al., 2024; Richardson, 2006). While conventional econometric approaches have yielded valuable insights, they often rely on linear specifications and may not fully capture the complex, nonlinear interactions among financial, governance, and market variables that drive overinvestment behavior (Hastie et al., 2009). Consequently, there has been growing interest in leveraging machine learning techniques to improve predictive accuracy and uncover hidden patterns in firm-level data (Hao et al., 2021; Lakhal et al., 2021; Tam et al., 2023). To date, few studies have systematically applied a combination of traditional classifiers and advanced deep learning models to the overinvestment context within Vietnam's unique institutional environment.

The significance of addressing this research gap is considerable. Improving the detection of overinvestment has practical implications for investors, regulators, and policymakers who aim to enhance the efficiency

DOI: https://doi.org/10.71058/jodac.v9i7004

and credibility of Vietnam's capital markets. More reliable identification of firms prone to overinvestment can facilitate better capital allocation, inform corporate governance reforms, and support regulatory oversight (Le et al., 2024). Additionally, applying advanced machine learning models alongside interpretable methods contributes to the methodological advancement of investment efficiency research (Arik & Pfister, 2021; Goodfellow et al., 2016). As the Vietnamese economy continues to integrate into regional and global markets, the need for rigorous approaches to evaluate firms' investment behavior has become increasingly important.

Accordingly, this study has the objective to develop and evaluate a comprehensive machine learning framework to classify overinvestment in Vietnamese listed firms based on firm-level financial and governance characteristics. The methodology involves constructing a dataset covering the period from 2000 to 2024, labeling overinvestment using the free cash flow approach, and comparing the predictive performance of logistic regression, random forest, gradient boosting machines, support vector machines, K-nearest neighbors, artificial neural networks, and TabNet. The remainder of this paper is structured as follows. Section 1 provides the introduction. Section 2 reviews the background theories and relevant empirical literature. Section 3 describes the research methodology, including data collection, variable measurement, and model development. Section 4 presents and discusses the empirical results. Section 5 concludes the study and offers implications for future research and practice.

2. Literature review

2.1 Background Theories

Investment decision-making within firms has long been a central concern of corporate finance. A fundamental question is whether managers allocate resources efficiently or whether agency conflicts and information asymmetries lead to suboptimal investment outcomes such as overinvestment. Several foundational theories contribute to understanding this phenomenon. One of the most influential frameworks is the Free Cash Flow Hypothesis proposed by Jensen (1986), which argues that managers have incentives to invest excess cash flow into projects that may not maximize shareholder value. According to this view, when firms generate substantial free cash flows and lack profitable growth opportunities, agency costs rise because managers may pursue investments that expand their control or prestige rather than create economic value. This theory predicts that overinvestment is more likely in firms with abundant internal funds coupled with limited investment opportunities. Consequently, free cash flow becomes a critical indicator of potential inefficiencies in capital allocation.

Complementing this perspective, Agency Theory provides a broader framework for understanding managerial behavior in settings where ownership and control are separated (Fama & Jensen, 1983). Agency Theory posits that managers, as agents, may act in their own interest rather than in the interest of principals (shareholders), particularly when monitoring mechanisms are weak or when performance-based incentives are insufficient to align interests. Overinvestment can be conceptualized as an agency cost arising from the discretion managers have over free cash flow. This theoretical lens helps explain why even profitable firms may engage in value-destroying investments if governance structures do not effectively discipline managerial decision-making.

Another relevant perspective is derived from Tobin's Q Theory, which emphasizes that firms should invest when the marginal value of assets exceeds their replacement cost (Tobin, 1969). Under this framework, the ratio of market value to book value (Q) serves as a proxy for growth opportunities. When Q is low, implying limited profitable projects, the justification for additional investment diminishes. Integrating Tobin's Q with agency considerations, Richardson (2006) operationalizes overinvestment as investment levels exceeding those predicted by growth opportunities and firm characteristics. This approach underscores the interaction between investment incentives (captured by Q) and managerial discretion (captured by free cash flow),

DOI: https://doi.org/10.71058/jodac.v9i7004

providing a rigorous basis for empirical measurement. Moreover, Resource-Based Theory suggests that managers may overinvest in internal projects to build slack resources or protect their domain, especially when external discipline is weak (Barney, 1991; Penrose, 2009). This perspective emphasizes the strategic behavior of managers in accumulating assets, which can manifest in excessive capital expenditures even when expected returns are low. Such behavior reinforces the prediction that free cash flow without robust governance may lead to persistent inefficiencies.

These theoretical foundations collectively inform the premise of this research. The free cash flow hypothesis and agency theory explain why overinvestment arises from managerial incentives and weak governance. Tobin's Q provides a market-based benchmark for expected investment given growth prospects. Resource-based views offer complementary explanations for why managers may intentionally accumulate resources beyond efficient levels. Integrating these perspectives justifies the empirical approach of classifying overinvestment based on the coexistence of high free cash flow and low growth opportunities. In this study, these theories establish a coherent framework for identifying firms that are likely to overinvest and for developing machine learning models to classify such behavior systematically. By leveraging insights from agency costs, market valuation, and strategic resource allocation, the research aims to contribute to the understanding of how financial characteristics and firm incentives interact to shape investment decisions.

2.2 Empirical Studies

Empirical research on overinvestment has expanded substantially over the past two decades, integrating diverse methods and perspectives. One prominent stream investigates how corporate governance mechanisms influence investment efficiency. Studies consistently show that firms with weak governance structures tend to allocate resources inefficiently, with higher tendencies toward overinvestment. For example, Wang et al. (2016) find that board independence is negatively associated with overinvestment, reflecting more effective monitoring of managerial decisions. Similarly, Tahir et al. (2019) document that ownership concentration constrains managerial discretion, reducing excessive capital expenditures. Complementing these findings, Al Dah et al. (2023) demonstrate that CEO power amplifies agency problems, increasing the likelihood of overinvestment in low-growth contexts. Collectively, this evidence supports the notion that governance arrangements play a crucial role in shaping investment decisions.

Another important body of work examines the role of financial constraints and cash holdings in driving investment inefficiency. Ullah et al. (2020) reveal that firms with abundant internal funds are more prone to overinvest when external financing frictions are minimal, while Chen et al. (2011)show that free cash flow interacts with country-level investor protection to predict excess investment. In line with these arguments, Benlemlih and Bitar (2018) find that firms with weaker creditor rights are more likely to overinvest, suggesting that legal environments significantly moderate investment behavior. Moreover, Huang et al. (2015) highlight that high cash reserves can lead to persistent overinvestment when combined with limited market discipline.

Research has also investigated how information asymmetry and disclosure quality affect overinvestment. Studies such as Francis et al. (2013) and Hou et al. (2016) emphasize that higher transparency mitigates agency costs by reducing information asymmetry between managers and investors. For instance, Hammami and Hendijani Zadeh (2020) show that voluntary disclosure reduces overinvestment by clarifying the firm's true investment opportunities to external stakeholders. Complementing these results, Le et al. (2024) provide evidence from Vietnam that improved disclosure practices and stronger accounting standards are associated with greater investment efficiency, especially among listed firms. Similarly, Dinh et al. (2023) show that earnings quality in Vietnam reduces the incidence of overinvestment, supporting the view that credible financial reporting is an important disciplinary mechanism in emerging markets.

DOI: https://doi.org/10.71058/jodac.v9i7004

An additional strand of empirical work explores how managerial characteristics and behavioral factors shape investment decisions. Cai (2013) find that CEO overconfidence is positively associated with overinvestment, especially when growth opportunities are limited. Ben-David et al. (2013) document that managerial miscalibration, where managers systematically overestimate their ability to generate returns, can lead to systematic overinvestment. Along related lines, Ho et al. (2016) demonstrate that managerial entrenchment increases the persistence of excessive investments over time. These behavioral findings resonate with recent evidence from Vietnam, where Dinh Nguyen et al. (2021) observe that CEO overconfidence significantly predicts overinvestment and tends to be more pronounced among state-owned firms and firms with concentrated ownership structures.

Cross-country analyses further highlight how institutional environments and cultural norms shape overinvestment patterns. Chen et al. (2015) show that firms in economies with lower investor protection and higher political interference face higher overinvestment risk. Similarly, Bae et al. (2012) provide evidence that firms operating in emerging markets are more likely to engage in excessive investment due to weaker enforcement and pervasive state ownership. This pattern is also evident in Vietnam, where Le and Tannous (2016) document that state ownership is associated with greater overinvestment and lower investment efficiency, reflecting political objectives and agency conflicts.

More recently, scholars have begun incorporating machine learning techniques to improve the measurement and prediction of overinvestment. Hao et al. (2021) apply ensemble learning models to classify overinvestment episodes, achieving higher accuracy than conventional regression approaches. In a related study, Lakhal et al. (2021) use support vector machines and random forest classifiers to detect overinvestment based on financial ratios and governance indicators. Tam et al. (2023) demonstrate the potential of machine learning for the Vietnamese context, applying gradient boosting methods to classify overinvestment among listed manufacturing firms, achieving substantial improvements in predictive accuracy.

Finally, a growing literature has linked overinvestment to firm performance and market valuation. Empirical evidence consistently indicates that excessive investment is associated with declining profitability and negative abnormal returns. For instance, Shen and Ruan (2022) show that overinvestment leads to lower subsequent return on assets and return on equity. Likewise, Duygan-Bump et al. (2015) document that persistent overinvestment negatively affects firm value, particularly in low-growth industries. In parallel, Chen et al. (2023) find that investors penalize firms exhibiting signs of overinvestment, resulting in discounted valuations. In the Vietnamese setting, Nghia (2022) provide evidence that overinvestment significantly undermines firm profitability and market value, with the effect being more severe among firms with limited board independence and weaker investor protection. This body of work reinforces the economic relevance of understanding and detecting overinvestment behavior in both developed and emerging economies.

3. Methodology

3.1 Data

This study utilizes firm-level panel data collected from companies listed on the Vietnam stock exchange markets, specifically the Ho Chi Minh Stock Exchange (HOSE) and the Hanoi Stock Exchange (HNX). The initial sampling frame comprises all publicly listed firms on these two exchanges, totaling 649 firms over the period from 2000 to 2024. The data were obtained from the secondary data source Refinitiv. To ensure the accuracy and consistency of the dataset, observations with missing or incomplete records were excluded to enhance the reliability of the analysis. In addition, outlier values were identified and removed using the Z-score method, whereby any observation exceeding ±3 standard deviations from the mean were considered an outlier and excluded from the final sample. This approach is consistent with established

DOI: https://doi.org/10.71058/jodac.v9i7004

practices in empirical corporate finance research and helps mitigate the influence of extreme values on estimation results. After applying these procedures, the final dataset includes 6,561 firm-year observations covering a 25-year period. This sample provides a balanced representation of firms across sectors and sizes in Vietnam's stock markets, offering sufficient variability to examine the patterns and determinants of overinvestment.

Table 1: Variable Definitions and Measurements

Variable Name	**Symbol**	**Measurement / Definition**
Overinvestment	OVER	Dummy variable, 1 if FCF > FCF median and TobinQ < tobinQ median; 0 otherwise.
Log of Total Assets	LTA	Natural logarithm of total assets at fiscal year-end
Firm Age	AGE	Number of years since the firm's listing on the exchange
Current Ratio	CURRATIO	Current Assets ÷ Current Liabilities
Operating Cash Flow / Total Assets	OCF_TA	Operating Cash Flow ÷ Total Assets
Total Debt / Total Assets	DEBT_TA	Total Debt ÷ Total Assets
Return on Assets	ROA	Net Income ÷ Total Assets
Return on Equity	ROE	Net Income ÷ Shareholders' Equity
Net Profit Margin	NPM	Net Income ÷ Sales Revenue
Sales Growth Rate	SGROWTH	(Current Year Sales − Prior Year Sales) ÷ Prior Year Sales
Capital Expenditures / Total Assets	CAPEX_TA	Capital Expenditures ÷ Total Assets
Net Investment / Total Assets	NETINV_TA	(Change in Net Property, Plant, and Equipment) ÷ Total Assets
Board Size	BOARDSIZE	Total number of directors on the board
CEO Duality	CEODUAL	Dummy variable: 1 if CEO is also the board chair, 0 otherwise
Ownership Concentration (%)	OWNCONC	Percentage of shares held by the largest shareholder
State Ownership (%)	STATEOWN	Percentage of shares held by the government or state-related entities
Dividend Payout Ratio	DIVPAYOUT	Dividends Paid ÷ Net Income
Tobin's Q	TOBINQ	(Market Value of Equity + Book Value of Debt) ÷ Book Value of Total Assets

Source: by author

DOI: https://doi.org/10.71058/jodac.v9i7004

Table 1 presents the list of input variables selected for the machine learning classification of overinvestment among Vietnamese listed firms. First, to classify overinvestment, this study adopts a widely used approach that combines firms' free cash flow and growth opportunities, reflecting theoretical foundations proposed by (Jensen, 1986) and operationalized in subsequent empirical research (Al Dah et al., 2023; Richardson, 2006). Specifically, firms are identified as overinvesting when they simultaneously exhibit high free cash flow, defined as operating cash flow exceeding capital expenditures and ranking above the median threshold of the sample, and low growth opportunities, proxied by Tobin's Q falling below the median threshold. This criterion captures the tendency of managers with abundant internal funds and limited profitable projects to allocate resources inefficiently, thereby aligning the measurement of overinvestment with established practices in the literature.

Firm size (log of total assets) and firm age are included to account for scale effects and life cycle stages, which can shape both financial flexibility and managerial incentives. Liquidity and cash flow indicators, such as the current ratio and operating cash flow relative to total assets, reflect firms' capacity to finance investments internally. Measures of leverage (total debt to total assets) capture the role of external financing constraints and monitoring by creditors, which can either discipline or enable overinvestment. Profitability metrics, including return on assets, return on equity, and net profit margin, provide information on overall performance and are often associated with managerial confidence and the perceived availability of funds for discretionary investments.

In addition to financial indicators, the model incorporates variables related to growth opportunities, governance, and payout policy. Sales growth and Tobin's Q are widely used proxies for investment prospects, helping to distinguish firms likely to overinvest because of managerial discretion rather than genuine expansion needs. Capital expenditures and net investment relative to total assets serve as benchmarks for investment intensity. Governance characteristics, including board size, CEO duality, ownership concentration, and state ownership, are critical given the extensive evidence linking governance structures to agency problems and investment efficiency, particularly in emerging markets like Vietnam. Finally, the dividend payout ratio is included as an indicator of cash distribution discipline, which may constrain managers' ability to channel free cash flow into suboptimal projects. Together, these variables offer a comprehensive representation of the financial, governance, and market factors relevant for predicting overinvestment behavior.

3.2 Models

This study applies a combination of traditional machine learning classifiers and advanced deep learning models to classify overinvestment in Vietnamese listed firms. The rationale for using multiple algorithms is to capture both linear and complex nonlinear relationships among financial and governance variables, while benchmarking predictive performance across diverse methodological approaches. The following section describes how each algorithm works, the parameters adopted in this research, and the justification for its inclusion.

Logistic Regression is a widely used linear classification technique that estimates the probability of a binary outcome through the logistic function (Hosmer Jr et al., 2013). Specifically, it models the log-odds of overinvestment as a linear combination of predictors, as expressed by Equation (1):

$$log\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta_0 + \sum_{j=1}^{k} \beta_i X_j \qquad (1)$$

where $P(y = 1)$ denotes the probability of overinvestment and X_j are the explanatory variables. In this study, the logistic regression model will be implemented with L2 regularization (ridge penalty) to mitigate multicollinearity and avoid overfitting. The regularization parameter C will be tuned via cross-validation.

DOI: https://doi.org/10.71058/jodac.v9i7004

Logistic regression is included as a baseline model because it provides interpretability and enables direct comparison with prior econometric studies on investment efficiency (Biddle et al., 2009).

Random Forest is an ensemble learning method that constructs a large number of decision trees during training and outputs the mode of their predictions (Breiman, 2001). Each tree is built using a bootstrap sample of the data, and a random subset of features is selected at each split, which enhances robustness to noise and reduces overfitting. The prediction for an observation is the majority vote among individual trees, as shown in Equation (2):

$$\hat{y} = mode\{T_1(x), T_2(x), \dots, T_b(x)\} \qquad (2)$$

where $T_b(x)$ denotes the prediction from the bbb-th tree. The random forest classifier in this study will be trained with 500 trees (n_estimators = 500), a maximum tree depth set to none (allowing trees to grow fully), and the Gini impurity criterion. This algorithm is employed because it consistently delivers strong performance on tabular data and generates variable importance rankings that can reveal the most influential factors underlying overinvestment (Hastie et al., 2009).

Gradient Boosting Machines, specifically XGBoost and LightGBM, are advanced ensemble techniques that build trees sequentially to minimize classification errors (Chen & Guestrin, 2016; Ke et al., 2017). Unlike random forests, which grow trees independently, gradient boosting constructs each new tree to correct the residual errors from the previous ensemble. The model iteratively updates predictions using Equation (3):

$$F_m(x) = F_{m-1}(x) + \upsilon . h_m(x) \qquad (3)$$

where $h_m(x)$ is the new base learner trained to fit the residuals, and υ is the learning rate. For XGBoost, the parameters will include 500 boosting rounds (n_estimators = 500), a maximum depth of 4, a learning rate of 0.05, and subsampling at 0.8. LightGBM will be configured with similar settings. These algorithms are selected because they have demonstrated state-of-the-art performance in financial prediction tasks and efficiently capture complex feature interactions (Nielsen, 2016).

Support Vector Machine (SVM) is a supervised learning model that seeks to construct a hyperplane in a high-dimensional space to optimally separate classes (Cortes & Vapnik, 1995). The classifier solves the optimization problem presented in Equation (4):

$$\frac{min}{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \qquad (4)$$

subject to the constraints:

$$y_i(w.\phi(x_i + b)) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \qquad (5)$$

where C is the penalty parameter and $\phi(x)$ represents the kernel function. This study will use an RBF kernel to capture nonlinear relationships, with C and γ optimized through grid search. SVM is included because of its robustness in handling high-dimensional spaces and its ability to model complex decision boundaries (Hastie et al., 2009).

K-Nearest Neighbors (KNN) is a non-parametric algorithm that classifies an observation based on the majority class among its k nearest neighbors (Cover & Hart, 1967). Given a new observation x_0, the classifier identifies the k observations in the training set closest to x_0 (measured by Euclidean distance) and assigns the most frequent label. In this study, KNN will be implemented with $k = 5$. Although simpler than ensemble methods, KNN provides a useful benchmark and highlights local patterns in the data that more complex algorithms may smooth over.

DOI: https://doi.org/10.71058/jodac.v9i7004

Artificial Neural Network (ANN), implemented as a Multilayer Perceptron, is a deep learning model that learns nonlinear transformations of input data through interconnected neurons across layers (Goodfellow et al., 2016). The output of each neuron is computed as shown in Equation (6):

$$a^{(l)} = f\left(W^{(l)}.a^{(l-1)} + b^{(l)}\right) \qquad (6)$$

where f is an activation function, $W^{(l)}$ is the weight matrix, and $b^{(l)}$ is the bias vector. The ANN in this study will use three hidden layers with 64, 32, and 16 neurons respectively, ReLU activation, and dropout regularization (dropout rate = 0.2) to prevent overfitting. The model will be trained using the Adam optimizer with a learning rate of 0.001. ANN is included to capture deep nonlinear interactions among variables that traditional classifiers may overlook.

TabNet is a recent deep learning architecture specifically designed for tabular data (Arik & Pfister, 2021). Unlike conventional neural networks, TabNet employs sequential attention mechanisms to dynamically select which features to focus on at each decision step. This attention-based learning enables both high predictive accuracy and interpretability. The model in this study will be trained with a maximum of 200 epochs, a batch size of 256, and early stopping based on validation loss. TabNet is included because it has demonstrated superior performance on structured data and provides meaningful feature importance explanations through learned feature masks.

By combining these algorithms, this research aims to comprehensively assess overinvestment classification performance across models ranging from interpretable linear approaches to advanced deep learning architectures. This multi-method approach ensures robustness of findings and provides richer insights into the determinants of overinvestment in Vietnamese listed firms.

The predictive performance of the classification models will be assessed using a comprehensive set of evaluation metrics that capture various dimensions of model effectiveness. In addition to accuracy, precision, recall, F1 score, and AUC-ROC, the study will include specificity, which measures the true negative rate or the proportion of correctly identified non-overinvestment cases. This metric is especially important in contexts where false positives may lead to unnecessary scrutiny or misallocation of resources (Sokolova & Lapalme, 2009). To provide a balanced assessment of model performance when the dataset is imbalanced, the Matthews Correlation Coefficient (MCC) will be computed, as it combines all four confusion matrix elements into a single coefficient ranging from -1 to +1, with higher values indicating stronger agreement between predictions and actual labels (Chicco & Jurman, 2020). Moreover, Cohen's Kappa will be used to quantify the level of agreement between predicted and actual classifications beyond chance expectations, which offers a more robust evaluation in situations where class distributions are unequal (Cohen, 1960).

Additionally, the study will report the area under the precision-recall curve (PR-AUC), which provides insight into model performance in correctly identifying positive cases while minimizing false positives, particularly when the positive class is relatively rare (Davis & Goadrich, 2006). To assess probabilistic calibration, the Brier Score will be calculated, as it measures the mean squared difference between predicted probabilities and the actual binary outcomes, where lower scores indicate better-calibrated predictions (Brier, 1950). Finally, calibration plots will be produced to visually examine the alignment between predicted probabilities and observed frequencies across probability bins (Niculescu-Mizil & Caruana, 2005). Together, these evaluation metrics offer a multi-faceted and rigorous assessment of model discrimination, calibration, and reliability, thereby ensuring that the comparative analysis of machine learning algorithms reflects both predictive accuracy and practical usability in overinvestment detection.

DOI: https://doi.org/10.71058/jodac.v9i7004

4. Results & Discussion

4.1 Descriptive Analysis

Table 1 provides a comprehensive overview of the descriptive statistics for the main variables employed in this study. Firm size exhibits a moderate dispersion, with a mean logarithm of total assets of 27.5 and a relatively narrow standard deviation of 1.63, suggesting that most firms in the sample are of comparable scale. The average firm age is approximately 15 years, reflecting a diverse mixture of newly listed companies and more established entities, while the skewness of 0.59 indicates a slight concentration of younger firms. The current ratio displays considerable variability, with a mean of 2.24 and a maximum of 19.48, implying significant differences in liquidity positions across firms. Notably, the skewness (3.67) and kurtosis (17.59) of the current ratio reveal the presence of extreme values, consistent with prior evidence that liquidity metrics often display heavy tails due to episodic cash surpluses or liquidity constraints.

Table 1. Descriptive analysis

	Count	Mean	Std	Min	25%	50%	75%	Max	Skewness	Kurtosis
size	6,561	27.5	1.63	23.26	26.31	27.41	28.53	32.42	0.28	-0.15
age	6,561	15.38	7.33	0	10	15	20	39	0.59	0.12
current_ratio	6,561	2.24	2.13	0.09	1.16	1.56	2.42	19.48	3.67	17.59
ocf_ta	6,561	0.08	0.08	-0.27	0.03	0.07	0.12	0.45	0.44	1.78
debt_ta	6,561	0.48	0.21	0.02	0.32	0.49	0.64	0.97	-0.12	-0.85
roa	6,561	0.06	0.06	-0.16	0.02	0.05	0.08	0.29	0.85	1.86
roe	6,561	0.11	0.1	-0.69	0.05	0.1	0.16	0.76	-0.2	5.07
npm	6,561	0.09	0.22	-5.97	0.02	0.05	0.11	3.8	-3.24	196.26
capex_ta	6,561	0.04	0.05	0	0.01	0.02	0.06	0.26	1.83	3.17
div_ni	6,561	0	0.04	-0.24	0	0	0	1.47	24.38	751.16
tobinq	6,561	1.2	0.83	0.19	0.79	0.96	1.29	7.78	3.43	15.67
growth	6,561	0.13	0.47	-2.07	-0.08	0.08	0.24	4.79	3.18	19.58

Source: by author

Several variables demonstrate substantial skewness and kurtosis, underscoring heterogeneity in financial performance and investment behavior. For instance, net profit margin has a mean of 0.09 but exhibits extreme dispersion (standard deviation of 0.22) and highly negative skewness (-3.24), reflecting firms that reported substantial losses relative to sales. Similarly, the dividend-to-net-income ratio is concentrated around zero (mean=0, median=0) but shows a maximum value exceeding 1, indicating a small subset of firms distributing exceptionally high dividends relative to earnings, as captured by the kurtosis of 751.16. Tobin's Q has an average value above one (mean=1.2), suggesting that many firms are valued higher than the replacement cost of assets, yet the skewness (3.43) and kurtosis (15.67) point to a few firms with disproportionately high market valuations. Finally, the growth rate variable demonstrates a mean of 0.13 and high positive skewness (3.18), indicating that while most firms grew modestly, some experienced rapid expansion. These patterns highlight the importance of robust modeling techniques capable of handling outliers and distributional irregularities when classifying overinvestment behavior.

Table 2 reports the distribution of overinvestment across industries and stock exchanges, revealing substantial variation in investment behavior among Vietnamese listed firms. Overall, approximately 38.4%

DOI: https://doi.org/10.71058/jodac.v9i7004

of the total observations are classified as overinvestment cases, indicating that the phenomenon is relatively widespread in the sample. A consistent pattern emerges whereby certain industries, particularly Utilities and Information Technology, exhibit markedly higher proportions of overinvestment compared to others. For example, Utilities firms listed on the Ho Chi Minh Stock Exchange display an overinvestment rate exceeding 69%, and Information Technology firms in the same exchange record a similar pattern with 60.3% of observations categorized as overinvesting. These high proportions may reflect the capital-intensive nature of utilities and the rapid technological changes affecting IT firms, which can create both pressures to invest aggressively and challenges in aligning investments with sustainable returns.

Table 2. Overinvestment distribution

		No. non-over investment	No. over investment	% non-over investment	% over investment
Hanoi Stock Exchange	Communication Services	102	21	82.93	17.07
	Consumer Discretionary	109	45	70.78	29.22
	Consumer Staples	126	87	59.15	40.85
	Energy	129	102	55.84	44.16
	Financials	69	18	79.31	20.69
	Health Care	45	37	54.88	45.12
	Industrials	800	386	67.45	32.55
	Information Technology	70	33	67.96	32.04
	Materials	402	188	68.14	31.86
	Real Estate	57	31	64.77	35.23
	Utilities	42	54	43.75	56.25
Ho Chi Minh Stock Exchange	Communication Services	11	11	50	50
	Consumer Discretionary	243	181	57.31	42.69
	Consumer Staples	207	188	52.41	47.59
	Energy	58	56	50.88	49.12
	Financials	108	37	74.48	25.52
	Health Care	76	73	51.01	48.99
	Industrials	681	396	63.23	36.77
	Information Technology	23	35	39.66	60.34
	Materials	343	230	59.86	40.14
	Real Estate	271	152	64.07	35.93
	Utilities	70	158	30.7	69.3
Total		4,042	2,519	61.61	38.39

Source: by author

The results also highlight noteworthy differences between the two stock exchanges. While the overall rates of overinvestment are broadly comparable, several sectors show pronounced contrasts in their distribution. For instance, in the Energy sector, overinvestment accounts for 44.2% of firms listed on the Hanoi Stock Exchange but nearly 49.1% on the Ho Chi Minh Stock Exchange, suggesting that regional market dynamics

DOI: https://doi.org/10.71058/jodac.v9i7004

or disclosure practices could influence investment decisions. Consumer Discretionary and Consumer Staples sectors also display relatively high rates of overinvestment, often exceeding 40% on both exchanges, which may be linked to competition-driven expansion strategies and volatile consumer demand. Conversely, the Financials sector shows a consistently lower incidence of overinvestment, with only 20.7% and 25.5% of observations classified as such on the Hanoi and Ho Chi Minh Stock Exchanges, respectively, likely reflecting tighter regulatory oversight and more conservative investment policies. Overall, these findings underscore the importance of accounting for industry-specific and market-specific factors when modeling overinvestment, as sectoral and institutional contexts appear to shape firms' investment patterns in distinctive ways.

These patterns further illustrate the heterogeneity that motivates the application of machine learning classification techniques in this study. The marked differences in overinvestment prevalence across industries and between the Hanoi and Ho Chi Minh Stock Exchanges highlight the limitations of relying solely on linear or aggregate models to capture firms' investment behavior. Instead, the diverse distributions shown in Table 2 suggest that predictive models must be flexible enough to accommodate sectoral characteristics, firm-specific governance practices, and varying market environments. Accordingly, the use of ensemble and deep learning approaches in this research is intended to better capture these complex, nonlinear relationships and improve the identification of firms prone to overinvestment.

4.2 Results

Figure 1 provides a detailed comparison of the performance of different machine learning algorithms across four key metrics: accuracy, precision, recall, and F1 score. Overall, ensemble methods such as Random Forest and XGBoost consistently achieved the highest accuracy, clustering tightly around 0.83 to 0.85 with minimal variability across folds. This suggests strong stability and predictive reliability in detecting overinvestment cases. Logistic Regression also performed reasonably well, with accuracy levels averaging around 0.75, albeit lower than the ensemble methods. In contrast, SVM and KNN displayed notably weaker results, with SVM in particular exhibiting the lowest median accuracy close to 0.63 and a narrow interquartile range, indicating consistently underwhelming performance. ANN showed a broader distribution but achieved median accuracy comparable to Random Forest and XGBoost in some folds, demonstrating its potential to capture complex nonlinear relationships.

DOI: https://doi.org/10.71058/jodac.v9i7004

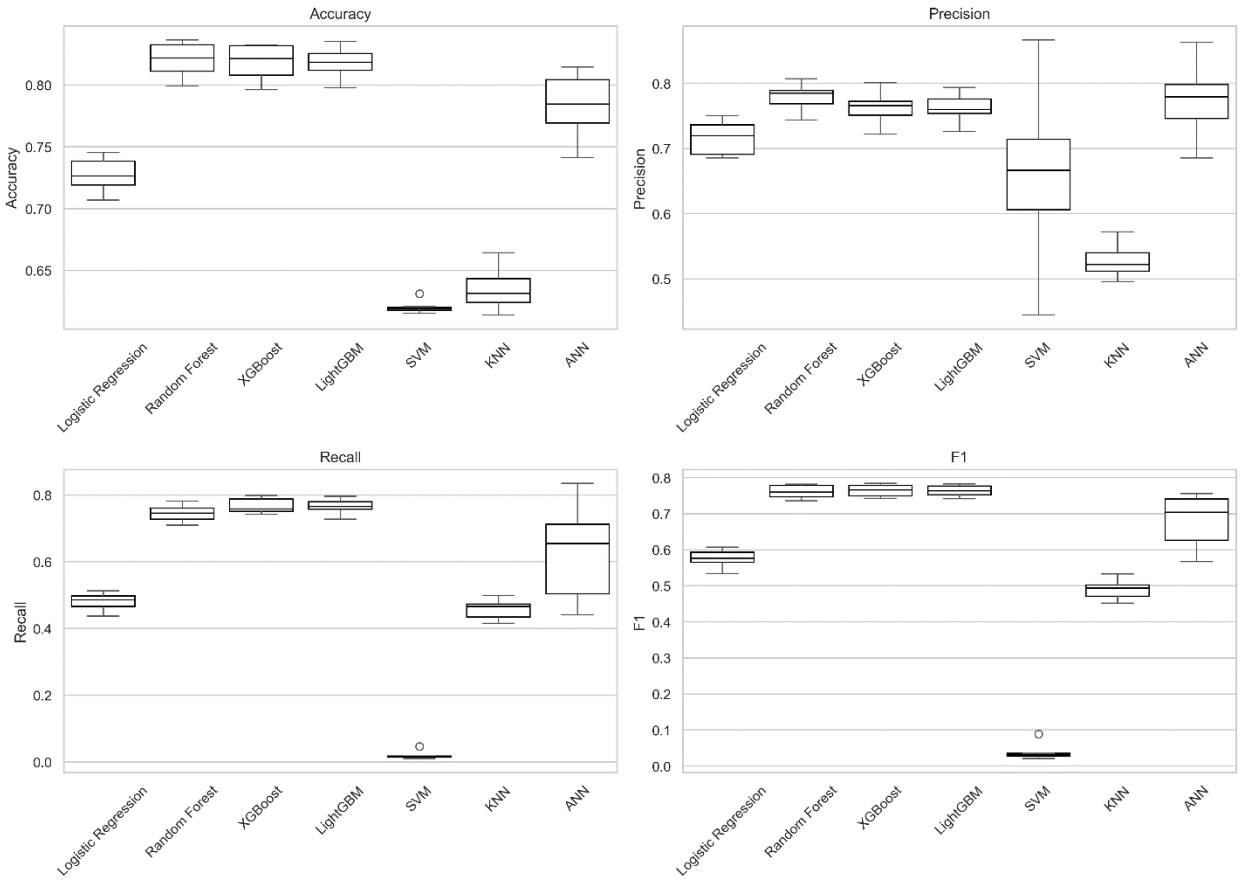

Figure 1. Accuracy, Precision, Recall & F1 between algorithms

Source: by authors

The precision and recall boxplots further highlight important trade-offs among the classifiers. Random Forest and XGBoost not only maintained high precision, with median values approaching or exceeding 0.8, but also delivered strong recall, indicating that these models effectively balanced minimizing false positives and false negatives. Logistic Regression demonstrated moderate precision and recall, reflecting its simpler linear structure and more conservative predictions. Interestingly, ANN yielded high variability in recall, ranging from approximately 0.5 to 0.8, suggesting sensitivity to data partitions in the cross-validation procedure. SVM exhibited a pronounced weakness in recall, with most folds recording values near zero, which significantly undermines its utility in correctly identifying overinvestment cases. This disparity underscores that while some classifiers maintain consistent predictive capacity, others struggle to generalize across validation folds.

The F1 scores reinforce the superiority of ensemble methods and highlight the practical limitations of simpler models for this task. Random Forest and XGBoost again achieved the highest median F1 scores, reflecting their balanced precision-recall profiles and confirming their suitability for overinvestment classification in the Vietnamese market context. ANN also performed competitively in terms of F1 score, despite the wider dispersion observed in recall, suggesting that neural networks can offer valuable predictive capabilities when properly tuned. In contrast, SVM and KNN recorded the lowest F1 scores, indicating persistent challenges in capturing the underlying patterns associated with overinvestment behavior. Overall, the results presented in Figure 1 demonstrate that ensemble approaches and ANN provide the most robust classification performance, while simpler methods such as SVM and KNN appear ill-suited to the complexity of the dataset.

DOI: https://doi.org/10.71058/jodac.v9i7004

Figure 2 presents the comparative performance of the algorithms across four additional evaluation metrics: Brier score, Cohen's Kappa, Matthews Correlation Coefficient (MCC), and PR_AUC. The Brier score results indicate that Random Forest and XGBoost achieved the lowest error rates, with median scores around 0.12, suggesting superior calibration of predicted probabilities relative to other classifiers. LightGBM also performed strongly on this metric, whereas SVM and KNN displayed substantially higher Brier scores exceeding 0.22, reflecting their limited ability to produce accurate probability estimates in this context. ANN exhibited intermediate Brier scores but with higher variability across folds, indicating sensitivity to data partitions and potential overfitting in some iterations. Logistic Regression maintained a stable Brier score near 0.18, consistent with its simpler, linear predictive capacity.

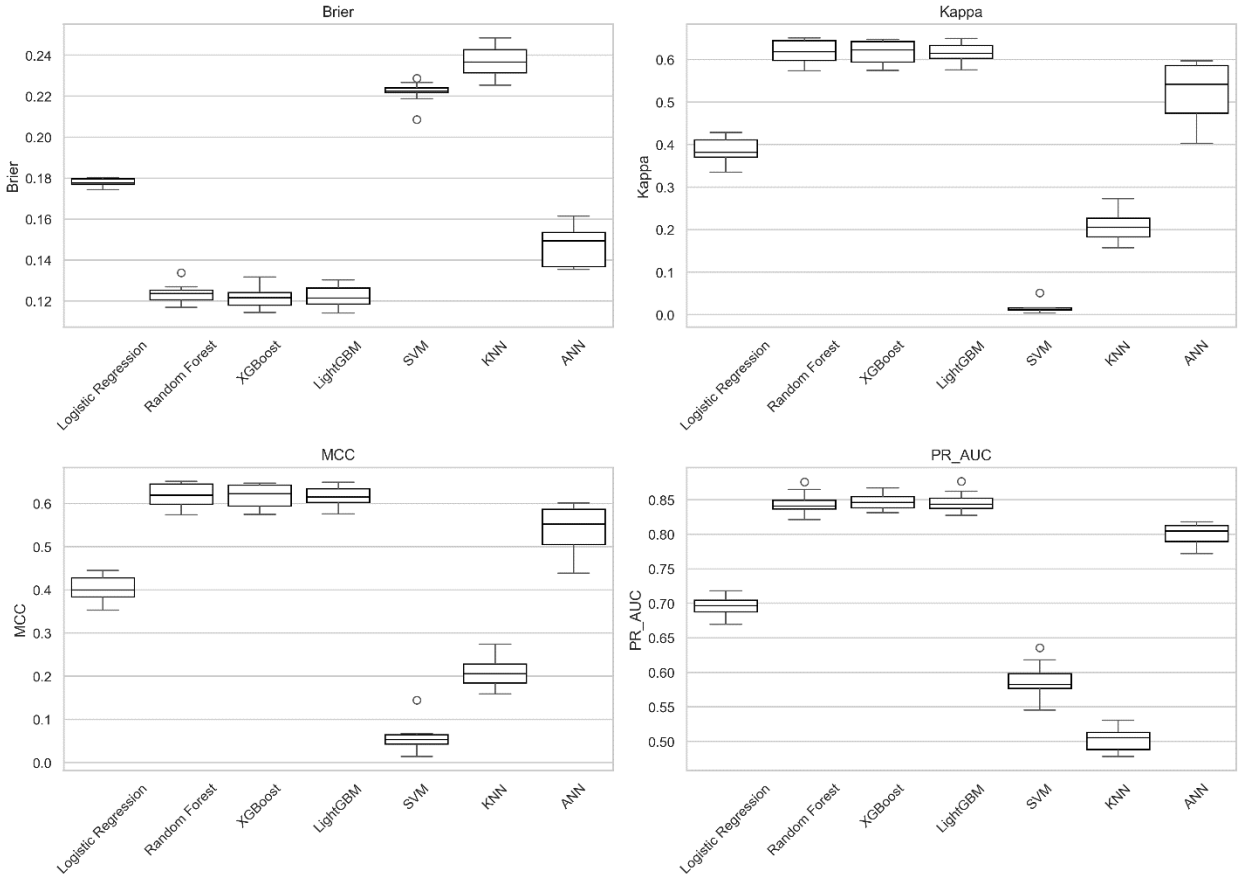

Figure 2. Brier, Kappa, MCC & PR_AUC between algorithms

Source: by authors

The Kappa, MCC, and PR_AUC metrics further underscore the dominance of ensemble methods and the relative underperformance of SVM and KNN. Random Forest, XGBoost, and LightGBM achieved consistently high Kappa and MCC values, exceeding 0.60, demonstrating strong agreement between predicted and actual labels and balanced handling of positive and negative cases. By contrast, SVM displayed near-zero Kappa and MCC, confirming its inability to distinguish overinvestment reliably. In terms of PR_AUC, which emphasizes the classifier's ability to identify positive (overinvestment) cases amid class imbalance, Random Forest and XGBoost again led, with median scores above 0.85. ANN performed competitively on PR_AUC, exceeding 0.80 in most folds, while Logistic Regression was moderate at approximately 0.75. The pronounced gap between ensemble models and simpler classifiers reinforces the conclusion that more sophisticated, non-linear techniques are better suited for modeling

DOI: https://doi.org/10.71058/jodac.v9i7004

overinvestment in Vietnamese listed firms, where heterogeneity and complex interactions between predictors are prevalent.

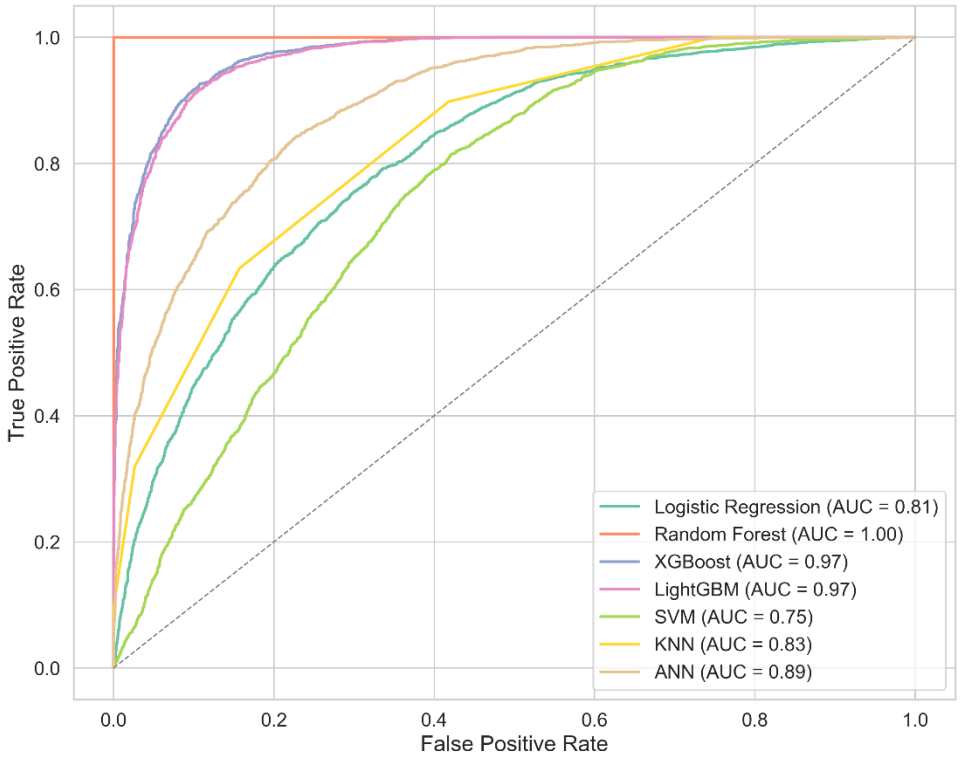

Figure 3. ROC-AUC between algorithms

Source: by authors

Figure 3 illustrates the receiver operating characteristic (ROC) curves for all classifiers, highlighting their discriminative power in distinguishing overinvestment cases. Random Forest achieved a perfect AUC of 1.00, indicating flawless separation between classes on the training data, which underscores its exceptional fit but may also reflect potential overfitting. Both XGBoost and LightGBM performed nearly as well, each attaining an AUC of 0.97, demonstrating their strong ability to capture complex patterns and interactions among predictors. ANN followed closely with an AUC of 0.89, suggesting robust predictive performance and effective modeling of non-linear relationships. In comparison, KNN and Logistic Regression recorded more moderate AUC values of 0.83 and 0.81, respectively, reflecting adequate but less optimal discriminative capabilities. SVM exhibited the weakest performance, with an AUC of 0.75, indicating limited ability to distinguish between overinvestment and non-overinvestment cases. Overall, the ROC curves reaffirm the superiority of ensemble methods and neural networks for this classification task, consistent with the other evaluation metrics presented in the study.

DOI: https://doi.org/10.71058/jodac.v9i7004

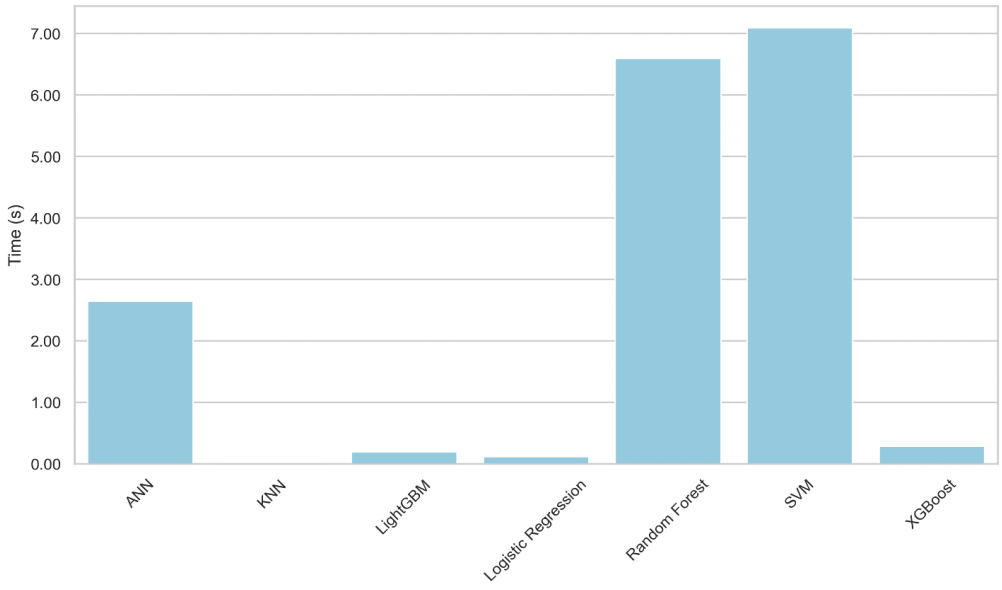

Figure 4. Average time consuming between algorithms

Source: by authors

Figure 4 compares the average computational time required by each algorithm, highlighting notable differences in training efficiency. Random Forest and SVM were the most time-consuming methods, each requiring approximately 6 to 7 seconds per fold on average, reflecting the higher computational complexity of tree ensemble construction and the kernel optimization process in SVM, respectively. In contrast, LightGBM and XGBoost demonstrated remarkable speed, completing training in less than half a second per fold, which underscores their design optimizations for handling large datasets efficiently. Logistic Regression and KNN also exhibited minimal processing time, further confirming their suitability when computational resources are constrained. The Artificial Neural Network required a moderate runtime of about 2.5 seconds per fold, balancing higher model complexity with acceptable execution time. Overall, these results suggest that while Random Forest and SVM deliver competitive accuracy in some settings, their higher computational costs may limit their practicality for time-sensitive applications, whereas boosting algorithms offer both speed and strong predictive performance.

5. Conclusion & Recommendation

5.1 Conclusion

This study set out with the objective of developing and evaluating a comprehensive machine learning framework to classify overinvestment among Vietnamese listed firms, using firm-level financial and governance indicators. The findings demonstrate that ensemble methods such as Random Forest, XGBoost, and LightGBM consistently outperformed other algorithms across nearly all evaluation metrics, including accuracy, precision, recall, F1 score, ROC-AUC, MCC, and PR_AUC. Artificial Neural Networks also showed strong predictive capability, albeit with greater variability across validation folds. In contrast, simpler methods like SVM and KNN consistently recorded lower performance, particularly in discriminative metrics and probabilistic calibration, underscoring their limited suitability for this classification task. These results collectively confirm that more advanced, nonlinear models are markedly better equipped to capture the complex relationships that characterize overinvestment behavior in the Vietnamese context.

DOI: https://doi.org/10.71058/jodac.v9i7004

Comparing these results with prior research reinforces the robustness and relevance of ensemble approaches. For example, Chen et al. (2023) highlighted the superiority of gradient boosting methods over traditional regressions in detecting corporate overinvestment, while Tam et al. (2023) reported similar advantages for tree-based ensembles when predicting investment efficiency. The current study's findings are consistent with these conclusions, further demonstrating that machine learning models provide significant improvements over classical econometric approaches in terms of classification accuracy and reliability. Additionally, the observed performance of Artificial Neural Networks aligns with the evidence presented by Goodfellow et al. (2016), who emphasized the capacity of deep learning architectures to model complex nonlinearities in structured financial data. These parallels suggest that the benefits of advanced machine learning methods are not confined to developed markets but are equally applicable in emerging economies like Vietnam.

The results also strongly resonate with the theoretical perspectives that underpin this research. The agency theory framework articulated by Jensen (1986) posits that firms with high free cash flow and limited growth opportunities are particularly susceptible to overinvestment, driven by managerial incentives and weak monitoring mechanisms. The consistently high classification performance achieved when combining financial ratios, governance indicators, and growth metrics lends empirical support to this theoretical proposition. Moreover, the findings validate the argument advanced by Biddle et al. (2009) that richer information environments and sophisticated modeling approaches can significantly enhance the detection of inefficient investment behaviors. By systematically integrating these theoretical considerations with data-driven techniques, this study bridges the gap between conceptual models of agency costs and practical tools for identifying overinvestment.

In addition to confirming and extending existing empirical evidence, the study contributes several novel insights. First, this study provides one of the first large-scale, multi-model comparative assessments of machine learning classifiers applied to overinvestment detection in the Vietnamese capital market, covering firms across both the Hanoi and Ho Chi Minh Stock Exchanges. Second, the results demonstrate that advanced ensemble and neural network models are capable not only of achieving high predictive accuracy but also of maintaining stable performance across multiple evaluation criteria, which is essential for robust classification in practice. Third, by incorporating a diverse set of firm-level variables, the study highlights the value of combining financial performance measures, governance structures, and market-based indicators to improve prediction.

Overall, the study makes a substantive contribution to the literature on corporate investment behavior and machine learning applications in finance. It offers empirical validation that sophisticated modeling techniques can materially improve the identification of overinvestment, thereby providing valuable tools for researchers, investors, and policymakers concerned with investment efficiency. In doing so, this study extends prior research on agency problems and investment inefficiency to the context of an emerging market and illustrates how modern data science methods can advance our understanding of longstanding corporate finance challenges.

5.2 Recommendation

Based on the findings of this study, several recommendations are proposed to enhance the detection and management of overinvestment among Vietnamese listed firms:

For corporate managers, it is essential to adopt more rigorous internal monitoring mechanisms that complement external governance structures. The results show that firms with high free cash flow and limited growth opportunities are particularly prone to overinvestment, consistent with agency theory. Management teams should therefore establish stricter capital budgeting protocols, including the application of hurdle rates that reflect risk-adjusted returns and regular post-investment reviews to assess project

DOI: https://doi.org/10.71058/jodac.v9i7004

performance. Incorporating data-driven tools, such as machine learning models similar to those evaluated in this research, can help firms proactively flag potential overinvestment scenarios before resources are committed.

For investors and analysts, the evidence suggests that relying solely on traditional financial statement analysis may be insufficient to identify overinvestment risks. The superior predictive performance of ensemble methods and artificial neural networks highlights the importance of integrating advanced analytics into investment due diligence processes. Investors should consider using machine learning classifiers to systematically screen firms with high free cash flow and low Tobin's Q, especially in sectors such as Utilities and Information Technology, where overinvestment prevalence was shown to be high. By incorporating these predictive tools, investors can better manage portfolio risks associated with inefficient capital allocation.

For regulators and policymakers, the results underscore the need to strengthen disclosure requirements and promote transparency in capital expenditure planning. The high incidence of overinvestment in certain industries suggests that existing reporting standards may not provide sufficient visibility into firms' investment rationales and project evaluation criteria. Regulators could consider mandating more granular disclosures of capital budgeting assumptions, internal rates of return, and project monitoring frameworks. Furthermore, supporting the development of centralized data repositories and platforms that enable market participants to access standardized firm-level information would facilitate broader adoption of predictive analytics for monitoring investment efficiency.

For auditors and corporate governance professionals, the findings indicate an opportunity to incorporate machine learning–based risk assessments into audit planning and board oversight functions. Given the demonstrated predictive accuracy of ensemble and neural network approaches, audit committees should encourage internal audit teams to experiment with these methods when evaluating the appropriateness of capital expenditures. Additionally, boards of directors should consider integrating predictive risk scoring into investment approval workflows, thereby aligning oversight practices more closely with modern data capabilities.

Finally, for professional training institutions and academic programs in finance and accounting, there is a strong rationale to update curricula to reflect the evolving role of machine learning in financial analysis. As this study shows, traditional econometric models may fail to capture the complex patterns associated with overinvestment. Equipping future finance professionals with practical skills in ensemble learning, deep learning, and model evaluation will be essential to improving the analytical capacity of firms and strengthening market discipline over capital allocation decisions.

5.3 Limitations & Further research

This study has several limitations that should be acknowledged. The analysis relies on historical financial data from Vietnamese listed firms, which may not capture unobservable factors such as managerial intent or qualitative aspects of corporate governance. Additionally, while advanced machine learning models demonstrated strong predictive performance, their interpretability can be limited compared to traditional econometric approaches, potentially complicating practical implementation and regulatory acceptance. Future research could build on these findings by exploring hybrid modeling approaches that combine machine learning classifiers with explainable AI techniques to improve transparency. Further studies might also examine the temporal dynamics of overinvestment behavior by incorporating time-series models or testing whether similar patterns hold in other emerging markets with different institutional environments.

DOI: https://doi.org/10.71058/jodac.v9i7004

References

[1] Al Dah, B., Dah, M., Harakeh, M., & Lobo, G. J. (2023). Director Excess Compensation and Investment Efficiency: The Economic Role of Accounting. *Journal of Accounting, Auditing & Finance*, 0148558X251347700.

[2] Arik, S. Ö., & Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. Proceedings of the AAAI conference on artificial intelligence,

[3] Bae, K.-H., Baek, J.-S., Kang, J.-K., & Liu, W.-L. (2012). Do controlling shareholders' expropriation incentives imply a link between corporate governance and firm value? Theory and evidence. *Journal of Financial Economics*, *105*(2), 412-435.

[4] Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of management*, *17*(1), 99-120.

[5] Ben-David, I., Graham, J. R., & Harvey, C. R. (2013). Managerial miscalibration. *The Quarterly journal of economics*, *128*(4), 1547-1584.

[6] Benlemlih, M., & Bitar, M. (2018). Corporate social responsibility and investment efficiency. *Journal of business ethics*, *148*, 647-671.

[7] 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.

[8] Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

[9] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1-3.

[10] Cai, J.-f. (2013). Does corporate governance reduce the overinvestment of free cash flow? Empirical evidence from China. *Journal of Finance and Investment Analysis*, *2*(3), 97-126.

[11] Chen, C.-C., Ho, K.-C., Li, H.-M., & Yu, M.-T. (2023). Impact of information disclosure ratings on investment efficiency: Evidence from China. *Review of Quantitative Finance and Accounting*, *60*(2), 471-500.

[12] Chen, D., Jiang, D., Ljungqvist, A., Lu, H., & Zhou, M. (2015). *State capitalism vs. private enterprise*. National Bureau of Economic Research.

[13] Chen, F., Hope, O.-K., Li, Q., & Wang, X. (2011). Financial reporting quality and investment efficiency of private firms in emerging markets. *The accounting review*, *86*(4), 1255-1288.

[14] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,

[15] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*, 1-13.

[16] Cohen, J. (1960). A cofficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

[17] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297.

[18] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

[19] Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning,

[20] Dinh Nguyen, D., To, T. H., Nguyen, D. V., & Phuong Do, H. (2021). Managerial overconfidence and dividend policy in Vietnamese enterprises. *Cogent Economics & Finance*, *9*(1), 1885195.

[21] Dinh, T. H. T., Nguyen, C. C., & Gan, C. (2023). Ownership concentration, financial reporting quality and investment efficiency: an empirical analysis of Vietnamese listed firms. *International Journal of Social Economics*, *50*(1), 111-127.

[22] Duygan-Bump, B., Levkov, A., & Montoriol-Garriga, J. (2015). Financing constraints and unemployment: Evidence from the Great Recession. *Journal of Monetary Economics*, *75*, 89-105.

[23] Fama, E. F., & Jensen, M. C. (1983). Separation of ownership and control. *The journal of law and Economics*, *26*(2), 301-325.

DOI: https://doi.org/10.71058/jodac.v9i7004

[24] Francis, B., Hasan, I., & Wu, Q. (2013). The benefits of conservative accounting to shareholders: Evidence from the financial crisis. *Accounting Horizons*, *27*(2), 319-346.

[25] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). MIT press Cambridge.

[26] Hammami, A., & Hendijani Zadeh, M. (2020). Audit quality, media coverage, environmental, social, and governance disclosure and firm investment efficiency: Evidence from Canada. *International Journal of Accounting & Information Management*, *28*(1), 45-72.

[27] Hao, W., Gao, H., & Liu, Z. (2021). An Evaluation Study on Investment Efficiency: A Predictive Machine Learning Approach. *Complexity*, *2021*(1), 6658516.

[28] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. In: Citeseer.

[29] Ho, P.-H., Huang, C.-W., Lin, C.-Y., & Yen, J.-F. (2016). CEO overconfidence and financial crisis: Evidence from bank lending and leverage. *Journal of Financial Economics*, *120*(1), 194-209.

[30] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

[31] Hou, Q., Jin, Q., Wang, L., & Zhang, G. (2016). Mandatory IFRS adoption, accounting quality, and investment efficiency: Evidence from China. *China Journal of Accounting Studies*, *4*(3), 236-262.

[32] Huang, C. J., Liao, T.-L., & Chang, Y.-S. (2015). Over-investment, the marginal value of cash holdings and corporate governance. *Studies in Economics and Finance*, *32*(2), 204-221.

[33] Jensen, M. C. (1986). Agency costs of free cash flow, corporate finance, and takeovers. *The American economic review*, *76*(2), 323-329.

[34] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

[35] Lakhal, N., Guizani, A., Sghaier, A., El ámine Abdelli, M., & Slimene, I. B. (2021). The impact of CSR performance on Efficiency of Investments using Machine Learning. International conference on business and finance 2021,

[36] Le, H. T. M., Cheng-Po, L., Phan, V. H., & Pham, V. T. (2024). Financial reporting quality and investment efficiency in manufacturing firms: The role of firm characteristics in an emerging market. *Journal of Competitiveness*, *16*(1), 62.

[37] Le, T. P. V., & Tannous, K. (2016). Ownership structure and capital structure: A study of Vietnamese listed firms. *Australian Economic Papers*, *55*(4), 319-344.

[38] Nghia, N. T. (2022). The impact of overinvestment on firm performance of Vietnam's listed companies. *Science & Technology Development Journal: Economics-Law & Management*, *6*(1), 2208-2219.

[39] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the 22nd international conference on Machine learning,

[40] Nielsen, D. (2016). *Tree boosting with xgboost-why does xgboost win" every" machine learning competition?* NTNU].

[41] Penrose, E. T. (2009). *The Theory of the Growth of the Firm*. Oxford university press.

[42] Richardson, S. (2006). Over-investment of free cash flow. *Review of accounting studies*, *11*, 159-189.

[43] Shen, Y., & Ruan, Q. (2022). Accounting conservatism, R&D manipulation, and corporate innovation: Evidence from China. *Sustainability*, *14*(15), 9048.

[44] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, *45*(4), 427-437.

[45] Tahir, S., Qamar, M. A. J., Nazir, M. S., & Usman, M. (2019). Does corporate governance reduce overinvestment? The mediating role of information asymmetry. *Pakistan Journal of Commerce and Social Sciences (PJCSS)*, *13*(4), 1068-1084.

[46] Tam, P. H., Tram, N. D. L., Anh, N. T. N., Nghia, N. Q. T., Linh, H. T., & Van Thanh, T. (2023). Application of machine learning in classification of overinvestment: Evidence from listed firms in Vietnam stock exchange market. *Science & Technology Development Journal: Economics-Law & Management*, *7*(4), 4814-4833.

DOI: https://doi.org/10.71058/jodac.v9i7004

[47] Tobin, J. (1969). A general equilibrium approach to monetary theory. *Journal of money, credit and banking*, *1*(1), 15-29.

[48] Ullah, I., Zeb, A., Khan, M. A., & Xiao, W. (2020). Board diversity and investment efficiency: evidence from China. *Corporate Governance: The international journal of business in society*, *20*(6), 1105-1134.

[49] Wang, Y., Chen, C. R., Chen, L., & Huang, Y. S. (2016). Overinvestment, inflation uncertainty, and managerial overconfidence: Firm level analysis of Chinese corporations. *The North American Journal of Economics and Finance*, *38*, 54-69.