QUBAHAN ACADEMIC JOURNAL
VOL. 6, NO. 1, January 2026
https://doi.org/10.48161/qaj.v6n1a1972

Predicting Corporate Profitability in Morocco: Comparing Classical Regression and Machine Learning

Youssef Jamil [1*], Imane EL Yamlahi [1] and Nabil Bouayad Amine [1]

[1] Department of Economics and Management, Faculty of Polydisciplinary Studies, Sultan Moulay Slimane University, Khouribga 25000, Morocco.

* **Corresponding author:** youssef.jamil@usms.ac.ma.

ABSTRACT: To the best of our knowledge, this study provides the first systematic comparison between classical regression and advanced machine learning models for predicting the profitability of Moroccan firms listed on the Casablanca Stock Exchange. While prior research has largely focused on developed markets, profitability prediction in emerging economies such as Morocco remains underexplored, despite the market's structural particularities (sectoral concentration, reliance on bank financing, and limited disclosure practices). This article provides the first systematic comparative analysis between regression and machine learning approaches applied to Moroccan listed companies, highlighting the advantages and limitations of each method in capturing complex and non-linear financial dynamics. Using a dataset covering ten years of financial statements, we evaluate multiple models, including OLS, Ridge regression, Random Forest, Gradient Boosting, Support Vector Regression, KNN, and XGBoost. Results show that machine learning models consistently outperform regression in predictive accuracy, while regression retains value in interpretability. Findings contribute to academic research by extending profitability forecasting studies to an under-explored emerging market, and to practice by offering investors, policymakers, and managers tools that improve risk assessment, capital allocation, and decision-making under conditions of uncertainty. These implications are particularly relevant for emerging economies, where informational asymmetries and structural heterogeneity complicate financial forecasting.

Keywords: profitability, Moroccan companies, machine learning, linear regression, financial performance, predictive modeling.

I. INTRODUCTION

1. GENERAL CONTEXT AND MOTIVATION

Recently, the Moroccan financial sector has made increasing use of artificial intelligence, particularly for predictive applications. Indeed, the increase in the volume of corporate financial data has encouraged the adoption of this advanced technology. As machine learning develops, prediction improves the quality of decision making; however, when predictive devices become sufficiently accurate and reliable, they can transform the way an organization operates, with some artificial intelligences able to influence a company's economics in such a way as to transform the strategy itself [1].

One of the main focuses of attention for Moroccan finance specialists, investors and business leaders is the future profitability of companies. It plays a central role in company evaluation, resource allocation and strategic direction [2]. In growth markets such as Morocco, the ability to anticipate future financial performance is essential in a context characterized by macroeconomic uncertainty, frequently stimulated by internal political or economic resources having a significant impact on the economy [3].

2. PARTICULARITIES OF THE MOROCCAN CONTEXT

The Moroccan corporate sector presents unique characteristics that make profitability forecasting particularly important. The Casablanca Stock Exchange, while one of the oldest in Africa, remains relatively modest in terms of capitalization compared to developed markets, with a limited number of large firms dominating the index alongside a majority of medium-sized enterprises. The market is heavily concentrated in a few sectors banking, telecommunications, construction, and basic industries while other segments of the economy are underrepresented. Moreover, Moroccan firms face structural challenges, including strong reliance on bank financing, exposure to volatile macroeconomic conditions such as exchange rates, inflation, and commodity prices. Finally, although financial disclosure practices have improved in recent years, particularly with the introduction of mandatory ESG reporting for issuers by the AMMC, reporting remains heterogeneous and often below international standards. These combined factors exacerbate information asymmetry and uncertainty for investors and policymakers, reinforcing the need for reliable, context-specific predictive models of corporate profitability in Morocco.

3. ROLE OF FINANCIAL REPORTING

The purpose of financial information is to enlighten investors, lenders, customers and other beneficiaries, enabling them to make financial choices. Thus, the examination of accounting information to anticipate future profitability has always been the subject of studies and presentations in financial accounting [4, 5]. Financial reports disclosed on the stock exchange are decisive in the decision-making process, particularly for determining a company's future viability before purchasing its shares on the financial markets. These documents, in particular the balance sheet, income statement and cash flow statement, provide standardized, auditable information on the company's financial condition, operating results and management of economic resources [2].

Indeed, investors are the main users of financial statements. Theoretically, it is important to consider financial statements not simply as tax documents, but rather as important instruments for economic decision-making. Thus, this research underpins the comparison between traditional econometric methods based on rigorous linear assumptions and machine learning models capable of capturing more sophisticated dynamics [6]. Its aim is to measure the comparative influence of various explanatory variables on changes in return on net operating assets, with a view to identifying those that play the greatest part in predicting future corporate profitability.

4. LITERATURE GAPS AND EXPECTED CONTRIBUTIONS

It is based on a detailed analysis of studies relating to future performance accounting indicators [7, 8]. It seeks to fulfill the methodological criticisms of ordinary least squares linear regression models in financial contexts [9, 10], and the contribution of non-parametric models and machine learning to financial forecasting [11]. This work seeks to complement these contributions by offering an in-depth comparative analysis between the predictive performance of classical regression models and that of machine learning methods, applied to a set of financial and accounting indicators of listed Moroccan companies [12, 13]. The aim is to highlight the distinct contributions of each method in modeling the complex dynamics of future profitability. It takes into account the structural characteristics of the data, such as heteroscedasticity, non-linearity and interaction between explanatory variables [9].

Therefore, this study explicitly compares regression-based methods with machine learning approaches to assess whether the latter provide superior predictive performance in the Moroccan context, thereby justifying the dual methodological perspective adopted. Despite the growing body of research on profitability forecasting, three gaps remain salient. Moroccan context; Prior studies mainly focus on advanced economies or other emerging regions; to the best of our knowledge, there is no systematic, side-by-side comparison between classical regression and modern machine-learning models for Moroccan listed firms, even though emerging markets exhibit volatility, limited transparency and structural heterogeneity that challenge forecasting [14-16]. Hybrid modelling and explainability; The literature rarely combines linear and non-linear approaches (hybrid pipelines), and almost none integrates modern explainability tools (such as, SHAP, LIME) to open the “black box” of ML in this setting [17, 18]. Data design limitations; Existing applications commonly rely on small, annual panels with few macro-financial drivers, which limits the use of more advanced/sequential architectures and reduces generalizability [19, 20].

This study directly addresses this gap by providing the first empirical comparative benchmark for Morocco that contrasts classical regression techniques with machine learning models in predicting future changes in operating profitability (ΔRNOA). Beyond establishing predictive performance differences, the analysis highlights the relevance of model choice in capturing non-linear dynamics inherent in firm-level financial data. Furthermore, the findings offer clear methodological and empirical directions for future research, including the extension to richer macro-financial variables and more advanced learning architectures.

5. RESEARCH QUESTIONS AND HYPOTHESES

Based on the identified research gap, this study addresses the following research questions:

- Research Question 1: To what extent can classical regression models predict the future profitability of Moroccan companies listed on the Casablanca Stock Exchange?
- Research Question 2: Do machine learning models provide superior predictive performance compared to regression models in this context, considering the complexity and non-linearity of financial relationships?
- Research Question 3: Which financial indicators (for example, RNOA, ΔPM, ΔATO, LEV, SIZE, M/B) are the most significant determinants of corporate profitability in Moroccan firms?

In order to answer this question, we set ourselves the following hypotheses:

- General hypothesis:

The predictive efficiency of machine learning models in explaining the change in return on net operating assets ΔRNOA is superior to that of traditional regression models. This hypothesis and the sub-hypotheses derived from it should be presented and developed in the light of the literature review. In order to anticipate the future profitability of the companies under study. We have formulated the following hypotheses:

- 1st hypothesis: Machine-learning models show significantly better predictive performance than traditional linear regression in explaining fluctuations in RNOA.

To simplify the analysis of this hypothesis, we have broken it down into three sub-hypotheses;

- H 1.1: Non-linear models such as (XGBoost and Random Forest) outperform ordinary least squares (OLS) linear regression; in terms of the “R2” indicator in ΔRNOA prediction.
- H 1.2: Machine learning models produce a lower “RMSE” indicator than traditional regression models.
- H 1.3: The significance of variables in machine learning models allows us to detect non-linear effects that are not present in traditional models.
- 2nd hypothesis: Explanatory variables such as (GRNOA, EPS, ΔPM, ΔATO) play a crucial role in future profitability trends.

In the same sense as the first hypothesis, the second hypothesis is broken down into four secondary hypotheses.

- H 2.1: There is a positive relationship between past growth in RNOA (GRNOA) and its future change (ΔRNOA).
- H 2.2: Earnings per share (EPS) is a significant positive indicator of fluctuating profitability.
- H 2.3: An increase in profit margin (ΔPM) is positively linked to (ΔRNOA).
- H 2.4: An increase in the asset turnover ratio (ΔATO) is associated with higher future profitability.

In summary, while previous paragraphs have highlighted the importance of corporate profitability in Morocco and outlined our research hypotheses, what remains crucial to emphasize is the specific research gap that this study addresses. The central challenge is that, despite the increasing availability of financial data and the recognized importance of profitability forecasting, Moroccan listed companies have not yet been systematically studied through a comparative lens opposing classical regression and modern machine learning models. Most existing works are either limited to advanced economies or neglect the particularities of emerging markets such as Morocco, where volatility, limited transparency, and structural heterogeneity intensify the forecasting challenge. By explicitly focusing on this gap, our study contributes to strengthening the understanding of predictive methods in the Moroccan context and demonstrates how machine learning can complement or even outperform traditional econometric approaches.

To the best of our knowledge, very few empirical studies have systematically applied machine learning techniques to predict corporate profitability in Morocco or comparable North African markets [21]. Even fewer have explicitly compared these approaches with traditional regression models [22]. This lack of comparative analysis creates a significant research gap, particularly in emerging markets characterized by volatility, limited transparency, and structural heterogeneity. Our study directly addresses this gap by providing a comparative framework between classical regression and modern machine learning models, thereby contributing to both academic literature and practical financial decision-making in the Moroccan context.

The remainder of this article is organized as follows. Section II reviews the relevant literature and presents the economic, financial, econometric, and methodological foundations of the study. Section III details the methodology, including sample selection, variable construction, research workflow, and model selection. Section IV provides the empirical application and model comparison, including data preprocessing, descriptive statistics, correlation analysis, and predictive modeling with both regression and machine learning approaches. Section V discusses the main findings, compares them with prior studies, and highlights the contributions and limitations of the research. Finally, Section VI concludes the paper and outlines avenues for future research.

II. LITERATURE REVIEW: ECONOMIC, FINANCIAL, ECONOMETRIC AND METHODOLOGICAL FOUNDATIONS.

1. *REASONS FOR SAMPLE SELECTION AND STUDY PERIOD*

In this research, the decision to focus on a sample of only 30 companies stems from various methodological and empirical reasons. Firstly, a careful selection based on exclusion criteria was implemented to ensure data consistency; only non-financial companies whose financial statements were accessible, comparable and published continuously between 2010 and 2019 were selected. Although this inevitably reduced the sample size, the process nevertheless ensured the reliability, consistency over time and accounting integrity of the observations examined.

Secondly, in an emerging market like Morocco, sampling restrictions are inevitably dictated by the size of the stock market and access to past financial data. A massive integration of companies could have introduced significant biases linked to gaps in information, changes in tax years or listing suspensions, thus compromising the robustness of the econometric and predictive analyses carried out. Thus, scientific publications recommend using a sample of verified quality, even if modest in size, in the face of the complexity of analyses combining econometrics and machine learning models, in order to avoid the risk of statistical noise generated by incomplete or heterogeneous data [23, 13]. And that the selection of a group of 30 companies represents a perfect balance between market representativeness, data accessibility and scientific rigor, guaranteeing the validity of the conclusions drawn. To ensure the external validity of predictive financial models, it is crucial to distinguish phases of normal economic operation from phases of major exogenous crisis. Thus, this research does not take into account post-Covid 19 financial exercises, which were strongly influenced by the Covid-19 pandemic [11].

2. *SELECTION AND EXPLANATION OF PREDICTIVE MODEL VARIABLES*

The dependent variable that serves as a reference for predicting future profitability is the change in return on net operating assets (ΔRNOA), we rely on explanatory variables that include (RNOA), (GRNOA), (ΔPM), (ΔATO). We also take into account other accounting indicators such as; (EPS), (M/B), (LEV), and (SIZE). Table 1 summarizes the key variables used in our research, detailing their definition, method of calculation and economic significance:

2.1. Explanatory Variables

Table 1 presents the explanatory variables used in the study, their formulas, and economic meaning. Indicators cover accounting ratios (RNOA, ΔRNOA, GRNOA, ΔPM, ΔATO), market-based measures (EPS, M/B), financial structure (LEV), and firm characteristics (Size).

Table 1. Explanatory variables.

Variable	Formula	Definition
RNOA	RNOA = NOPAT / Net Operating Assets	Net return on operating assets
ΔRNOA	ΔRNOA = RNOA $_{(t)}$ – RNOA $_{(t-1)}$	Year-over-year variation in RNOA
GRNOA	GRNOA = (RNOA t – RNOA $_{(t-1)}$) / (RNOA $_{(t-1)}$)	Relative growth in RNOA
ΔATO	ΔATO = ATO $_{(t)}$ – ATO $_{(t-1)}$	Change in asset turnover ratio
ΔPM	ΔPM = PM $_{(t)}$ – PM $_{(t-1)}$	Change in operating profit margin
EPS	EPS = Net income/ Number of shares outstanding	Earnings per share
M/B	M/B = Market value/ Book value	Market-to-book ratio
LEV	LEV = Total debt/ Total assets	Financial leverage ratio
Size	Size = ln (Total assets)	Firm size

Source: Individual creation by the author

2.2. Insights from Prior Studies

Studies have shown that analyzing the change in return on net operating assets (ΔRNOA) into its separate components, i.e., the change in asset turnover rate (ΔATO) and the change in profit margin (ΔPM), provides valuable information for anticipating future fluctuations in profitability. In other words, an increase in ΔATO is linked to a future improvement in ΔRNOA, whereas ΔPM provides no relevant information in this context. These results suggest that analysts should pay particular attention to changes in asset turnover rates when assessing a company's future profitability [24]. In the same vein, earnings per share (EPS) is a frequently used indicator of a company's profitability per ordinary share. Despite the potential influence of accounting decisions on EPS, this indicator remains significant for investors. Breaking down return on net operating assets (RNOA) into its component parts is crucial for a more accurate assessment of operating profitability, which can also offer valuable insights for anticipating future financial results [7].

A company's size, frequently measured by the logarithm of its Size assets or sales, affects profitability through economies of scale, easier access to financing and greater stability. However, research shows a non-linear relationship between size and profitability. Until 1981, there was little link between profitability and size. However, the economic recession of 1981 and 1982 in the United States led to a lasting depression in earnings for small stocks. For some obscure reason, small stocks did not participate in the boom of the mid-to-late 1980s [25]. The expected growth indicator is represented by the ratio of market value to M/B book value. A high M/B ratio indicates that the market expects earnings to increase in the future, which could lead to a positive change in RNOA, but high levels could indicate overvaluation. This reflects not only investors' growth forecasts, but also an underestimated valuation of assets in accounting terms.

2.3. Feature Selection and Dimensionality Reduction

In this study, we selected a concise set of financial indicators ΔRNOA, RNOA, ΔPM, ΔATO, EPS, SIZE, LEV, and M/B based on their robust theoretical grounding and frequent use in prior profitability forecasting literature [26, 27]. Given this limited dimensionality, we did not apply advanced reduction techniques such as Principal Component Analysis (PCA) or autoencoders "a type of neural network used for dimensionality reduction". Instead, correlation analysis and Variance Inflation Factor (VIF) diagnostics were conducted to detect multicollinearity and ensure predictor stability. This choice balances interpretability with statistical rigor, aligning with financial theory while avoiding overfitting.

For future studies employing larger sets of predictors, recent methods such as LASSO-based dimensionality reduction or feature selection with annealing [28], as well as explainable AI techniques embedded in model design [29], offer promising enhancements for both performance and interpretability. Financial leverage (LEV) plays a crucial role in corporate capital structure choices. Titman and Wessels examine various factors influencing capital structure, such as company size, expansion, profitability and the type of assets owned, to decipher how companies determine their degree of leverage [30].

Analyzing the return on net operating assets (RNOA) into its constituent elements asset turnover ratio (ATO) and profit margin (PM) provides valuable information for anticipating a company's future earnings. Soliman proves that fluctuations in the asset turnover ratio (ΔATO) are clearly linked to future changes in RNOA, while variations in the profit margin (ΔPM) are not [8]. Examination of the return on net operating assets (RNOA) and its constituent elements the profit margin (PM) and the asset turnover ratio (ATO) provide important indications for forecasting a company's future profits. Soliman 2008 proves that fluctuations in the asset turnover ratio (ΔATO) are clearly linked to future changes in RNOA, while variations in the profit margin (ΔPM) are not [8]. Like Nissim and Penman, who point out that asset turnover rate (ATO) is a more relevant indicator of future profitability than profit margin (PM) [7].

3. THEORETICAL FOUNDATIONS OF MODEL SELECTION

Whereas to model the fluctuation of the RNOA (ΔRNOA), from these accounting and financial explanatory variables (RNOA, GRNOA, EPS, LEV, SIZE, M/B, ΔATO, ΔPM). Several supervised learning algorithms were chosen for their ability to capture non-linear relationships, deal with implied interactions between variables and deliver better predictive performance outside the sampled framework.

3.1. Ordinary linear regression

The OLS model is considered the benchmark method in corporate finance. It represents an essential tool in econometrics applied to corporate finance, facilitating the representation of the link between a dependent variable, such as the link between a variable to be explained, in our case the ΔRNOA, and a group of explanatory variables taken from the financial statements. The OLS method is distinguished mainly by its ease of estimation, its ability to provide a direct interpretation of coefficients and its robust statistical characteristics under traditional assumptions.

$$\Delta RNOA_{t+1} = \propto \; + \beta_1 RNOA_t + \beta_2 \Delta RNOA_t + \beta_3 GRNOA_t + \beta_4 \Delta ATO_t + \beta_5 \Delta PM_t + \beta_6 EPS_t + \beta_4 \frac{M}{B_t} + \beta_8 LEV_t + \beta_9 B_t + \beta_{10} Size_t + \varepsilon_{t+1} \qquad (1)$$

OLS is based on the essential assumption that the conditional mean of the errors is zero, such as, E(μi|Xi) meaning that the average error, given the explanatory variables, is zero [31]. In terms of forecasting future RNOA, the use of OLS regression by demonstrating that the elements of the DuPont analysis (PM and ATO), in conjunction with book profits, are closely linked to future fluctuations in company profitability. He formalizes this relationship using a multiple linear regression equation [8]. Nevertheless; prediction equations based on regression models using the method of least squares in cases of high multicollinearity are generally considered unreliable. Regression coefficients can frequently vary considerably depending on the specific sample data collected [32].

Thus, to overcome these constraints, Ridge regression, proposed by Hoerl and Kennard in 1970, offers a robust version of the OLS model. This model incorporates an L2 regularization factor in the cost function, thus minimizing the variance of the estimated coefficients while reducing the impact of multicollinearity. The use of Ridge regression is justified when there are moderate to high linear relationships between certain explanatory variables, compromising the stability of the results obtained by OLS. By imposing a penalty proportional to the magnitude of the coefficients, Ridge regression offers the possibility of developing a model with greater stability, better generalizability and possibly greater efficiency on out-of-sample data [33]. The following equations formally summarize the econometric and machine learning models used in this study to predict future changes in operating profitability:

$$min_{\beta} \sum_{i,t} \left(\Delta RNOA_{i,t+1} - \alpha - \sum_{k=1}^{K} \beta_k X_{k,i,t} \right)^2 + \lambda \sum_{k=1}^{K} \beta_k^2 \quad (2)$$

$$\Delta RNOA_{i,t+1} = f\left(X_{i,t}\right) + \varepsilon_{i,t+1} \quad (3)$$

$$\Delta RNOA_{i,t+1} = \left(\frac{1}{M}\right) \sum_{m=1}^{M} h_{m(x_{i,t})} \quad (4)$$

$$min \left(\frac{1}{2}\right) \parallel w \parallel^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \quad (5)$$

subject to: $y_i - (w \cdot x_i + b) \leq \varepsilon + \xi_i$

$(w \cdot x_i + b) - y_i \leq \varepsilon + \xi_i^*$

$\xi_i, \xi_i^* \geq 0$

$$\Delta RNOA_{i,t+1} = \left(\frac{1}{k}\right) \sum_{j \in N_{k(i)}} \Delta RNOA_{j,t} \quad (6)$$

$$\Delta RNOA_{i,t+1} = \sum_{m=1}^{M} \gamma_m h_{m(x_{i,t})} \quad (7)$$

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(f_m) \quad (8)$$

3.2. Machine Learning Models

a) Random Forest

This category of procedures has desirable attributes: its performance is comparable to that of Adaboost, sometimes even better; it is robust to out-of-range values and noise; it offers valuable internal assessments of variable error, robustness, correlation and relevance; it uses pre-existing internal estimates and a resampling validation that relies only on selected variables. The study of Random Forests for regression is carried out by setting a bound on the mean squared error of generalization, which indicates that the reduction in the error of individual trees in the forest is linked to the correlation between the residuals and the mean squared error of individual trees [34].

b) SVR

or support vector regression, is an adaptation of the SVM "Support Vector Machine" for regression problems. It is based on the establishment of an insensitivity tube around the target function, within which faults are not penalized. The aim is to develop a function that is as smooth as possible, while admitting minimal deviations from the training data. This allows strict control of overlearning, even when samples are small [35].

c) KNN

K-Nearest Neighbors Regressor model stands out for its algorithmic simplicity and powerful adaptability. It is a non-parametric technique based on the principle that the estimated value for a target observation is determined by averaging the values of the K most adjacent observations in the feature space. It is a relevant model for several reasons: it is able to capture local data structures, basing itself only on points with the closest financial ratios [36]. Companies with similar financial profiles are assumed to experience similar variations in profitability. The KNN (K-Nearest Neighbors) model uses this proximity principle to make predictions in regression, by determining the target value of an observation based on the values of its K nearest neighbors in the universe of characteristics [37]. And unlike OLS or Ridge regression, KNN is not based on any functional form and is designed to fit non-linear shapes intrinsically [38]. In addition, it does not rely on the normality assumption, does not impose any assumptions concerning variance, and does not require a specific form [39].

d) Gradient Boosting

is based on a numerical optimization strategy in function space, aimed at establishing a series of low-power, generally shallow trees; usually regression trees with 4 to 8 terminal nodes. Each tree is trained to rectify the prediction errors of the previous one. This iterative stage-wise method gradually captures unexplained residuals at each phase, improving overall model prediction [40].

e) Algorithmically, XGBoost

is based on the optimization of a regularized loss function Ω(f) related to tree complexity. This approach aims to minimize the risk of overfitting and enhance the model's generalization capability. This device effectively adjusts the bias-variance balance, which is essential in financial forecasting. One of the main strengths of this model is its ability to handle the missing data commonly found in financial statements. At each decision point, the model determines an optimal default orientation for each division in the absence of a value. This sparsity-aware approach gives the model robustness in the face of incomplete data, without the need for prior recoding [41].

4. LITERATURE REVIEW OF THE PYTHON ENVIRONMENT

For this research, a number of specialized Python libraries were used, each playing a crucial role in various phases of the analysis, from initial data processing to predictive modeling and evaluation. There are three main classes of libraries:

4.1. Data management and processing

- Pandas: It offers versatile data structures (DataFrame) that facilitate the efficient manipulation of financial data sets [42].
- Numpy: It facilitates the management of numerical tables and the execution of vector calculations, and enables mathematical operations such as the calculation of the mean, standard deviation, etc., as well as the creation of characteristic matrices [43].
- Re, Io: Sometimes used for string processing, syntax analysis or internal file manipulation.
- Output visualization.
- Matplotlib.pyplot et Seaborn: these two graphic libraries are used to analyze correlations and trends through visualizations (correlation heat maps between variables, comparative graphs of predicted and observed values, histograms of residuals, error distributions, etc.) [44, 45].
- Traditional statistical analysis.
- Statsmodels.api: It offers classic statistical modeling tools, including linear regression OLS [46].
- Statsmodels. stats. diagnostic, statsmodels. stats. outliers_influence, Statsmodels. stats. stattools: These are used to test the basic hypotheses of homoscedasticity, autocorrelation, and multicollinearity through the following tests:
- Breusch and Pagan test;
- Durbin-Watson test;
- VIF (Variance Inflation Factor) (Idem).

4.2 Prediction based on Machine Learning

- Sklearn.linear_model: Includes OLS linear regression and Ridge regression models, making it easy to parallel classical and regularized methods.

- Sklearn. ensemble: It allows the use of tree models such as Random Forest and Gradient Boosting, renowned for their performance on financial data.
- Sklearn.svm: It allows the integration of Vector Support Regressions (SVR), designed to model complex non-linear relationships with a low level of noise.
- Sklearn. neighbors: It proposes the KNN algorithm, which is beneficial for nonparametric modeling.
- Sklearn.model_selection, sklearn. preprocessing, sklearn.pipeline: These are tools for normalizing data, separating training and test datasets, implementing k-fold cross-validation and establishing automated pipelines.
- Sklearn.metrics: It is used to judge model performance based on indicators such as R2, RMSE, MAE, and so on [47].

4.3 Advanced modeling using XGBoost

- XGboost: This is a library optimized for gradient boosting, renowned for its speed and accuracy on tabular data such as financial ratios. It also offers automatic regularization and efficient handling of missing values, a major asset in the financial field [41].

4.4 Critical synthesis and justification of methodological choices

Recent applications of machine learning in emerging markets have shown promising results in addressing the complexity of financial forecasting. For instance, studies in Asian and Latin American stock markets demonstrate that models such as Random Forests, Gradient Boosting, and XGBoost outperform traditional regression techniques by capturing non-linear interactions and market volatility [36, 38]. These findings reinforce the potential of ML to adapt to data environments characterized by instability and informational asymmetry, which are common features of emerging economies. Nevertheless, financial forecasting with ML remains particularly challenging due to risks of overfitting, high sensitivity to parameter tuning, and the lack of interpretability of "black box" algorithms [39]. These issues limit the direct applicability of ML results in managerial and policy-making contexts, where transparency and explainability are critical. Addressing these challenges requires balancing predictive power with interpretability, an aspect increasingly emphasized in recent literature.

Our methodological choices are thus justified by recent benchmarks that highlight the comparative advantages of ensemble methods such as Random Forest, Gradient Boosting, and XGBoost, which consistently rank among the most effective predictors in financial applications [37, 40]. At the same time, regression models remain indispensable for their interpretability and their ability to provide benchmark comparisons. By combining both approaches, our study aligns with recent best practices while adapting them to the Moroccan context, thereby filling a gap in the literature on profitability forecasting in emerging markets.

In summary, while regression models provide valuable interpretability and a strong theoretical foundation, they remain limited in capturing the complex and non-linear interactions inherent in financial data, especially in emerging markets [36, 37]. Conversely, machine learning techniques such as Random Forest, Gradient Boosting, and XGBoost demonstrate superior predictive accuracy and flexibility, but their opacity and sensitivity to overfitting reduce their direct applicability for financial decision-making [38–40]. These contrasting strengths and limitations justify our methodological choice to adopt a comparative framework, where classical regression serves as a benchmark for interpretability and machine learning models are employed to test predictive performance under conditions of complexity and volatility, particularly relevant in the Moroccan context.

5. REGIONAL STUDIES AND EMERGING MARKETS

Recent studies have increasingly applied machine learning techniques to profitability forecasting in emerging and regional markets. For example, Rashid et al. (2021) [38] examined 100 firms in Pakistan and found that Random Forest and Support Vector Machines captured profitability dynamics more effectively than regression models. Similarly, Fernández-Laviada et al. (2022) [39] analyzed Spanish listed firms (2014–2019) and confirmed the superiority of Random Forest, particularly for identifying key financial ratios. At a broader scale, Dutta et al. (2021) [40] used a global World Bank dataset and showed the usefulness of XGBoost and Artificial Neural Networks in predicting profitability across SMEs and large firms, while highlighting data heterogeneity as a key limitation.

These contributions collectively demonstrate that machine learning models generally outperform traditional regression approaches in emerging contexts characterized by volatility and limited transparency. However, no prior work has systematically applied a comparative framework between regression and ML to Moroccan listed companies, which underscores the originality and contribution of this study. Table 2 synthesizes recent empirical works applying machine learning models to profitability forecasting, including country/sample, methods used, key findings, and limitations. As shown in Table 2, most recent applications of machine learning to profitability forecasting have been conducted in contexts such as Pakistan, Spain, or using large multi-country datasets (WBES). While these studies confirm the superiority of ML models over traditional regression, they remain limited by country-specific restrictions or by the heterogeneity of global samples. To the best of our knowledge, no prior empirical study has systematically compared regression and machine learning approaches for listed firms in North Africa, or in economies structurally comparable to Morocco (such as, Tunisia, Egypt, Jordan). This lack of regional evidence reinforces the originality of our study, which provides the first benchmark for Moroccan firms and contributes to filling a critical gap in the literature on profitability forecasting in emerging markets.

Table 2. Comparative studies on the use of machine learning for profitability prediction in emerging and regional markets.

Study	Country / Sample	Methods Used	Key Findings	Limitations
Rashid et al. (2021) [38]	Pakistan, 100 non-financial firms listed	Random Forest (RF), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Extreme Gradient Boosting (XGB)	ML models significantly outperformed regression in capturing profitability dynamics; solvency ratios emerged as particularly influential predictors.	The study is limited to a single emerging economy (Pakistan), with a relatively small sample size (100 firms) and annual data. The high complexity of ensemble and neural network models increases the risk of overfitting. No use of cross-validation to confirm robustness. Lack of explainability tools (such as, SHAP, LIME) reduces managerial interpretability.
Fernández-Laviada et al. (2022) [39]	Spain, listed firms (2014–2019)	Random Forest combined with SHAP values (Explainable AI)	RF provided superior predictive accuracy over regression, while SHAP analysis highlighted the most relevant financial ratios influencing profitability.	The dataset is limited to Spanish firms, which reduces generalizability to other markets. The period of analysis (2014–2019) is relatively short, and based only on annual data. Although SHAP provides interpretability, the study does not compare with a broad set of ML models.
Dutta et al. (2021) [40]	Global WBES (World Bank Enterprise Survey) database	XGBoost, Artificial Neural Networks (ANN)	Demonstrated high predictive accuracy of ML models across SMEs and large enterprises, highlighting the flexibility of ML to handle heterogeneous firm characteristics.	Strong heterogeneity across countries reduces comparability and weakens the robustness of findings. The WBES dataset is cross-sectional and lacks time-series depth, limiting dynamic profitability forecasting. Country-specific institutional, accounting, and regulatory factors are not incorporated, which constrains the contextual relevance of the results.

Source: Author's elaboration based on a synthesis of recent empirical studies on profitability prediction in emerging and regional markets.

While prior studies on profitability forecasting have demonstrated the usefulness of both regression and machine learning approaches, their findings remain heterogeneous and context-dependent. On the one hand, linear regression techniques such as OLS offer interpretability and theoretical rigor, but they are highly sensitive to violations of classical assumptions (multicollinearity, heteroscedasticity, small-sample distortions), which undermines their reliability in emerging markets. On the other hand, machine learning methods (such as, Random Forest, Gradient Boosting, XGBoost) have shown superior predictive performance in various contexts, particularly by capturing non-linear interactions. Yet, their robustness has been questioned when applied to relatively small datasets, as they are prone to overfitting and limited interpretability. Furthermore, the literature does not unanimously agree on the relative predictive power of explanatory variables: while some studies emphasize the central role of ΔATO, others find profit margin (ΔPM) or RNOA to be more significant. This lack of consensus justifies the need for a systematic empirical comparison, in the Moroccan context, between traditional regression models and non-linear machine learning approaches. Our methodological choices using OLS and Ridge regression as benchmarks, complemented by a set of widely adopted ML algorithms are thus directly grounded in the strengths and limitations identified in prior research, ensuring both interpretability and the ability to capture complex profitability dynamics.

The recent literature on profitability prediction using machine learning offers both valuable insights and important limitations that justify our methodological design. Rashid et al. (2021) [48] confirmed the superiority of ML algorithms such as Random Forest, ANN, SVM, and XGBoost in emerging contexts, emphasizing the predictive power of solvency ratios. However, their study was restricted to a small panel of Pakistani firms and relied solely on annual data, raising concerns about generalizability and overfitting. Similarly, Fernández-Laviada et al. (2022) [49] showed that Random Forest combined with SHAP values provides superior predictive accuracy and improved interpretability of profitability factors in Spanish listed firms. Yet, their work is limited by a narrow dataset (2014–2019, Spain only) and the absence of comparison with a wider range of ML models. At a broader level, Dutta et al. (2021) [50] demonstrated that advanced ML models such as XGBoost and ANN can achieve high predictive accuracy across SMEs and large enterprises using the WBES dataset. Nonetheless, the strong heterogeneity of cross-country data, the lack of time-series dynamics, and the absence of institutional specificity reduce the contextual robustness of their findings.

Taken together, these studies highlight the potential of machine learning to outperform classical regression models in profitability forecasting, but they also expose critical limitations: reliance on small or heterogeneous datasets, limited temporal scope, and insufficient comparative frameworks. Our study directly addresses these shortcomings by providing a systematic comparison between regression-based models (OLS, Ridge) and several machine learning algorithms (RF, GB, SVR, KNN, XGBoost), applied to a homogeneous dataset of Moroccan listed firms (2010–2019). This dual methodological perspective allows us to balance interpretability and predictive performance, thereby extending the scope of existing contributions to an under-explored emerging market.

6. *RESEARCH GAP*

While prior research has demonstrated the usefulness of regression and machine learning techniques in profitability prediction [36, 31], most existing studies focus on advanced economies or large international datasets [32]. In contrast, empirical work specifically dedicated to Moroccan listed firms remains scarce, despite the structural particularities of this market [33]. Furthermore, very few contributions in the literature have explicitly compared classical econometric methods with modern machine learning in an emerging market setting [18], and almost none have integrated hybrid strategies or advanced explainability tools such as SHAP in this context [29]. This scarcity underscores a critical research gap that our study addresses by providing the first systematic comparative analysis of regression and machine learning models applied to Moroccan firms, with a focus on both predictive performance and managerial relevance.

III. METHODOLOGY

1. *ANALYTICAL APPROACH AND SAMPLE SELECTION*

The sample consists of Moroccan companies listed on the Casablanca Stock Exchange, for the period between 2010 and 2019. The following criteria were used to select the statistical population deemed relevant, using a selection by exclusion procedure:

- The sample selected must not be drawn from financial intermediation companies, investment companies, holding companies, banks or leasing companies.
- Company financial data for the period under investigation must be available.
- During the financial period, companies are not expected to change their fiscal year.
- Companies should not have been selected from those excluded from the stock exchange.
- To ensure comparability, companies' fiscal year should end on December 31.

Thus, in line with the criteria specified above, 30 companies and 310 data sets covering the period between 2010 and 2019 were selected and exploited as statistical samples for our study. The main source of data sets was collected from companies' financial documents, including summary statements (balance sheets, CPCs, etc.), published on the Casablanca Stock Exchange.

2. DATA PROCESSING ENVIRONMENT

We choose the Python language as our tool for processing, analyzing and modeling financial data. This preference is motivated by the flexibility and abundance of specialized libraries, as well as its ability to process large quantities of accounting and financial data efficiently. Thanks to its advanced capabilities in statistics, machine learning, graphical visualization and workflow automation, Python lends itself particularly well to empirical studies in corporate finance [29]. All phases of the analysis, from data cleansing to predictive modeling and results visualization, were carried out entirely in a Python environment.

The variables retained in this study are standard indicators of profitability prediction, defined as follows:

- ΔRNOA (Change in Return on Net Operating Assets) measures changes in operating profitability.
- RNOA (Return on Net Operating Assets) profitability relative to operating assets.
- GRNOA (Growth in RNOA) captures long-term operating profitability dynamics.
- ΔPM (Change in Profit Margin) reflects efficiency in generating earnings from sales.
- ΔATO (Change in Asset Turnover) efficiency in utilizing assets to generate revenues.
- EPS (Earnings per Share) profit available to shareholders.
- SIZE (Firm Size) proxied by the natural logarithm of total assets.
- LEV (Leverage) debt-to-equity ratio, proxy for financial risk.
- M/B (Market-to-Book ratio) proxy for growth opportunities.

The following Table 3 reports the main hyperparameters optimized for each predictive model, the ranges or options tested during the grid-search procedure (a systematic method that tests different parameter combinations to find the best performance), and the final selection criterion (minimum RMSE under 5-fold cross-validation, meaning the dataset is split into five parts to test the model's robustness). OLS regression is presented as a baseline model, which does not require hyperparameter tuning.

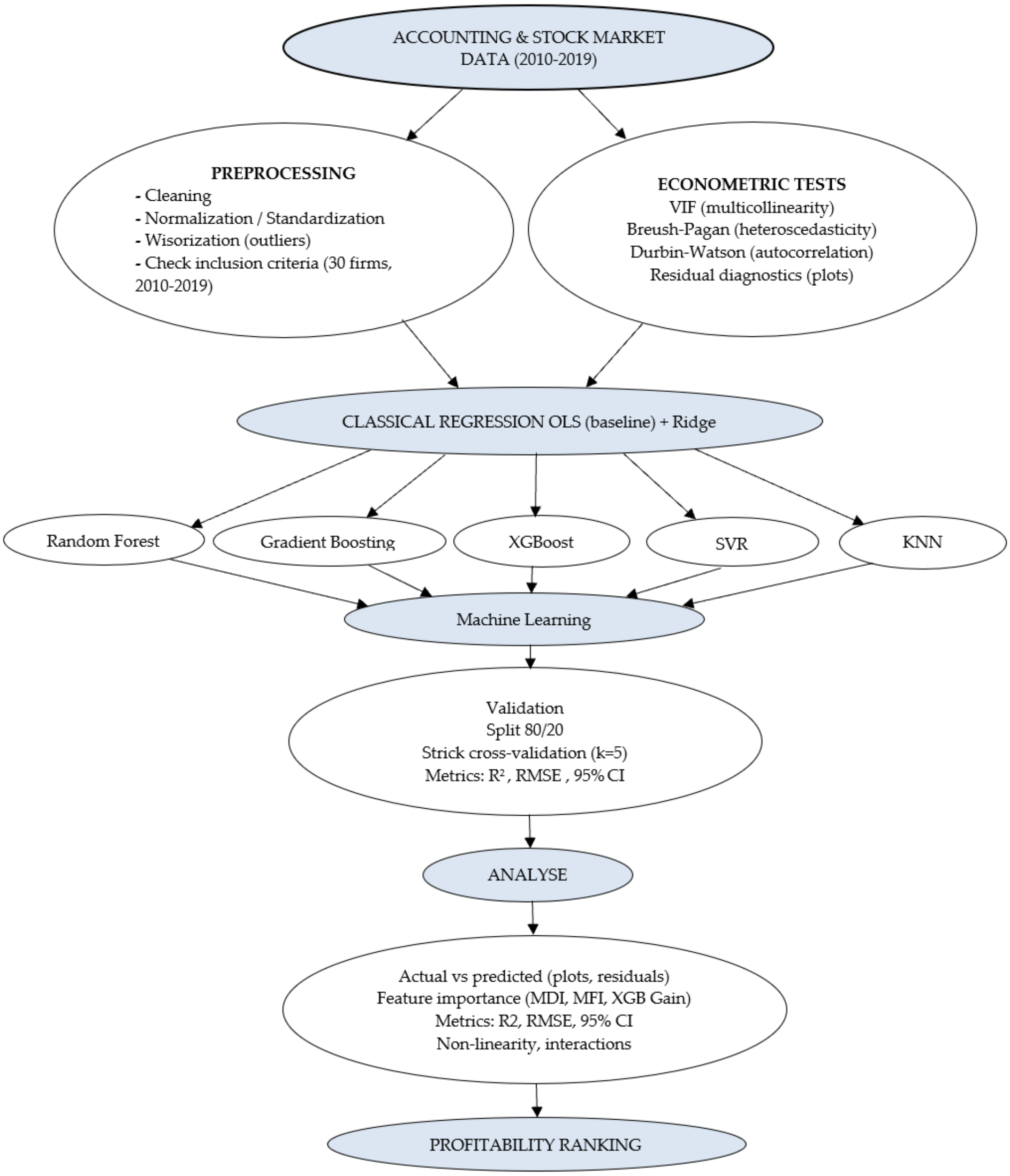
Table 3. Model specifications and hyperparameter tuning ranges.

Model	Key Hyperparameters Tuned	Search Range / Options Tested	Final Selection Criterion
OLS Regression	– (no hyperparameters)	–	Baseline comparison
Ridge Regression	Regularization parameter (α)	0.001 → 10 (log scale)	Best RMSE via 5-fold CV
Random Forest (RF)	Number of trees (*n_estimators*)	100 → 1000	Best RMSE via 5-fold CV
	Maximum depth (*max_depth*)	3 → 20	
	Minimum samples per leaf	1 → 5	
Gradient Boosting (GB)	Learning rate	0.01 → 0.3	Best RMSE via 5-fold CV
	Max depth	3 → 10	
	Number of boosting iterations (*n_estimators*)	100 → 500	
XGBoost	Learning rate (*eta*)	0.01 → 0.3	Best RMSE via 5-fold CV
	Max depth	3 → 10	
	Number of estimators	100 → 500	
	Subsample ratio	0.5 → 1.0	
Support Vector Regression (SVR)	Penalty parameter (C)	0.1 → 100	Best RMSE via 5-fold CV
	Kernel coefficient (γ)	0.001 → 1	
	Kernel type	RBF kernel	
K-Nearest Neighbors (KNN)	Number of neighbors (*k*)	2 → 15	Best RMSE via 5-fold CV
	Distance metric	Euclidean, Manhattan	

Source: Author's elaboration based on model implementations in Python (scikit-learn and XGBoost libraries).

3. RESEARCH WORKFLOW

To provide a clear overview of our methodological design, Figure 2 illustrates the workflow adopted in this study. The process begins with data collection and preprocessing, followed by econometric testing to ensure the robustness of the variables. A baseline regression model is then estimated, against which several machine learning models (Random Forest, Gradient Boosting, XGBoost, SVR, and KNN) are compared. The workflow concludes with model evaluation through cross-validation, performance metrics (R², RMSE), residual analysis, and profitability ranking.


ACCOUNTING & STOCK MARKET DATA (2010-2019)
PREPROCESSING
- Cleaning
- Normalization / Standardization
- Wisorization (outliers)
- Check inclusion criteria (30 firms, 2010-2019)
ECONOMETRIC TESTS
VIF (multicollinearity)
Breush-Pagan (heteroscedasticity)
Durbin-Watson (autocorrelation)
Residual diagnostics (plots)
CLASSICAL REGRESSION OLS (baseline) + Ridge
Random Forest
Gradient Boosting
XGBoost
SVR
KNN
Machine Learning
Validation
Split 80/20
Strick cross-validation (k=5)
Metrics: R², RMSE , 95% CI
ANALYSE
Actual vs predicted (plots, residuals)
Feature importance (MDI, MFI, XGB Gain)
Metrics: R2, RMSE, 95% CI
Non-linearity, interactions
PROFITABILITY RANKING

FSource : Author's own elaboration based on Casablanca Stock Exchange data (2010–2019), created using Python/Graphviz.

FIGURE 1. Methodological workflow of the study.

Figure 1 illustrates the research workflow, data collection, preprocessing, econometric testing, regression baseline, machine learning models, validation, and performance evaluation.

4. *FEATURE IMPORTANCE EXTRACTION PROCEDURE*

For Random Forest, feature importance was extracted using the mean decrease in impurity (MDI) method, which quantifies the weighted reduction in variance achieved by each feature across all trees in the ensemble. To account for potential biases and enhance interpretability, we also considered recently developed measures such as MDI+ and Mutual Forest Impact (MFI), which offer improved stability and insights into feature interactions [48, 49].

For XGBoost, variable importance was derived using the built-in feature_importances functionality, which supports multiple importance types: weight (frequency of a feature used in splits), gain (average improvement in model accuracy due to the feature), and cover (number of samples impacted) [50]. We focused on the gain metric as it directly reflects the feature's contribution to performance, but also verified weight and cover for robustness. This dual-procedure approach underpins complementary insights: Random Forest importance (augmented with MFI/MDI+) reflects variance reduction and relational feature synergy, while XGBoost's gain-based importance captures non-linear interactions and split-level contributions, ensuring a comprehensive understanding of predictor influence.

IV. EVALUATION METRICS

To evaluate model performance, we rely on two standard measures widely used in financial prediction tasks: the coefficient of determination (R²) and the Root Mean Squared Error (RMSE). The coefficient of determination (R²) is defined as:

$$R^2 = 1 - \frac{\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)} \qquad (9)$$

Where y_i denotes the observed values, $\hat{y}_i$ the predicted values, and $\bar{y}$ the mean of the observed data. A higher R² indicates that the model explains a greater proportion of the variance in profitability, which is desirable in predictive financial modeling [51]. The Root Mean Squared Error (RMSE) is given by:

$$RMSE = \sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (10)$$

This metric captures the average magnitude of prediction errors. A lower RMSE indicates better predictive accuracy, making it especially relevant when forecasting financial ratios where small deviations can have significant implications for decision-making [52]. The combination of R² and RMSE provides a balanced evaluation: R² emphasizes explanatory power, while RMSE highlights predictive accuracy and error minimization, aligning with best practices in empirical finance and machine learning applications [53].

V. HYPOTHESIS VALIDATION PROCEDURE

To test the formulated hypotheses, statistical validation methods were systematically applied. For hypotheses related to model performance (H1.1–H1.3), differences in predictive accuracy between models were assessed through cross-validation (k=5), with comparisons based on the R² and RMSE indicators, which remain standard evaluation metrics in predictive modeling [54]. For hypotheses concerning the role of explanatory variables (H2.1–H2.4), feature importance measures derived from Random Forest (mean decrease impurity) and XGBoost (gain importance) were mobilized, in line with recent recommendations for tree-based models [55, 56]. In addition, permutation-based importance tests were applied to enhance robustness [57]. This dual approach ensures that hypotheses are validated using transparent and reproducible statistical criteria.

VI. EMPIRICAL APPLICATION AND MODEL COMPARISON:

1. *DATA PREPROCESSING*

Before conducting the empirical analysis, several preprocessing steps were applied to ensure the reliability and consistency of the dataset.

- Data cleaning: Financial statements were carefully reviewed to detect inconsistencies, missing values, and reporting anomalies across firms and over time. Only companies with continuous and comparable financial disclosures during the 2010–2019 period were retained in the final sample. Outliers resulting from evident reporting errors were corrected or excluded.

- Normalization: To account for scale differences across variables, especially between accounting ratios (such as, RNOA, ΔPM, ΔATO) and market-based indicators (EPS, M/B), all predictors were standardized using the StandardScaler method. This transformation converts each variable into a mean of zero and a standard deviation of one, which is essential for machine learning models sensitive to data scaling, such as Support Vector Regression (SVR) and KNN.
- Outlier management: The descriptive statistics (Table 4) highlight the presence of extreme values, such as leverage (LEV) reaching 80, earnings per share (EPS) ranging from –50 to 150, and market-to-book ratios (M/B) up to 50. These extreme cases reflect the structural heterogeneity of Moroccan listed firms. To avoid distortions in predictive modeling while preserving representativeness, we applied winsorization at the 1st and 99th percentiles, and robustness checks were conducted using log-transformations for highly skewed variables (LEV, EPS, and M/B). This procedure, widely recommended in empirical finance, ensures that results are not driven by a few extreme observations but still capture the diversity of the market.

These preprocessing steps guarantee that the dataset remains consistent, robust, and suitable for both regression-based and machine learning models, while maintaining transparency on the underlying distribution of the financial indicators.

2. DESCRIPTIVE STATISTICS OF FINANCIAL VARIABLES

Before examining correlations, we first provide a descriptive statistical analysis of the financial variables used in the study. The following Table 4 reports the mean, median, standard deviation, minimum, and maximum values of the selected indicators (RNOA, ΔRNOA, GRNOA, EPS, Size, M/B, LEV, ΔPM, ΔATO). This allows us to highlight the central tendency, dispersion, and extreme values of the data, thereby ensuring a clear understanding of the dataset characteristics before predictive modeling.

The descriptive statistics highlight the structural characteristics and performance heterogeneity of Moroccan listed firms. The Table reports mean, median, standard deviation, minimum, and maximum values for the main indicators (RNOA, ΔRNOA, GRNOA, EPS, Size, M/B, LEV, ΔPM, ΔATO) among Moroccan listed firms.

Table 4. Descriptive statistics of financial variables for Moroccan listed firms (2010–2019).

Variable	Mean	Median	Std	Min	Max
RNOA	0.339	0.159	0.706	-0.381	6.294
ΔRNOA	0.016	0.000	0.680	-2.500	2.800
GRNOA	0.750	-0.200	2.850	-8.500	12.400
EPS	19.181	13.799	37.752	-50.000	150.000
Size	20.008	20.475	3.830	10.000	25.000
M/B	5.721	2.304	10.940	0.100	50.000
LEV	6.122	0.708	15.200	0.000	80.000
ΔPM	-0.006	0.000	0.633	-3.000	3.200
ΔATO	0.011	0.011	1.850	-5.500	6.200

Source: Authors' calculations based on financial statements of companies listed on the Casablanca Stock Exchange (2010–2019).

The Return on Net Operating Assets (RNOA) shows a mean of 0.339 and a median of 0.159, indicating overall positive operating profitability. However, the relatively high standard deviation (0.706) and the wide range of values (–0.381 to 6.294) reveal substantial heterogeneity: while some firms generate exceptional returns, others face difficulties in efficiently employing their operating assets. The change in profitability (ΔRNOA) is nearly zero on average (0.016) and at the median (0.000), reflecting overall stability. Yet, the dispersion (std = 0.680) and the extremes (–2.500 to 2.800) suggest that certain firms experience significant fluctuations, consistent with both firm-specific shocks and broader macroeconomic volatility. The growth in RNOA (GRNOA) further confirms this divergence. Although the mean is moderately positive (0.750), the median is negative (–0.200), indicating that only a minority of firms experience marked profitability growth, while the majority record declines. The relatively high standard deviation (2.850) and the wide range (–8.500 to 12.400) underline the presence of contrasting behaviors, typical of emerging markets subject to cyclical or sector-specific shocks.

The Earnings per Share (EPS) variable averages 19.181, with a median of 13.799. The high dispersion (std = 37.752) highlights diverging financial trajectories: some firms report substantial losses (–50.000), while others achieve strong profits (up to 150.000). These extreme values may reflect the earnings sensitivity of cyclical industries such as construction and raw materials, underscoring the heterogeneous financial trajectories within the Moroccan market.

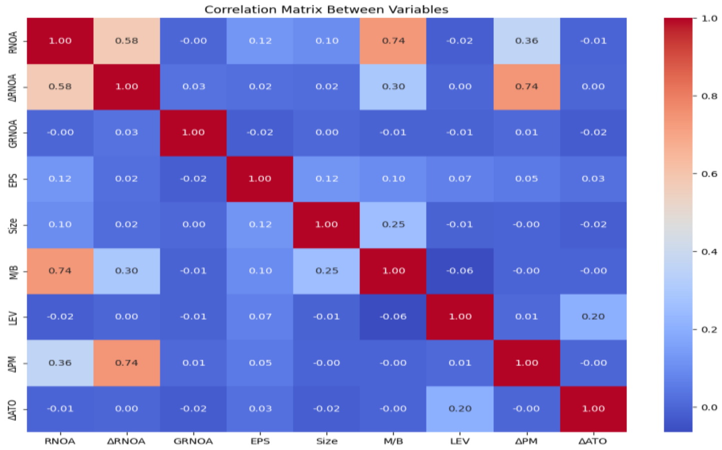
Firm size (Size) is relatively homogeneous, with a mean of 20.008 and a median of 20.475. The range of values (10.000 to 25.000) reflects the coexistence of intermediate-sized firms alongside large capitalizations, thereby confirming the representativeness of the sample. The Market-to-Book ratio (M/B) records a mean of 5.721, but the large dispersion (std = 10.940) reveals strong heterogeneity in market valuations. While some firms trade close to book value (0.100), others reach very high multiples (up to 50.000), reflecting either optimistic growth expectations or speculative dynamics that are often observed in emerging markets.

The financial leverage (LEV) shows a moderate mean (6.122) but a very low median (0.708), indicating that while most firms maintain limited debt levels, a few cases exhibit extremely high indebtedness (up to 80.000). Such values may reflect either highly leveraged financial structures or accounting anomalies, both of which are not unusual in emerging markets. The high standard deviation (15.200) confirms the asymmetric distribution of capital structures across the sample.

Finally, operational variation indicators display contrasting patterns. The change in profit margin (ΔPM) is almost zero on average (–0.006), reflecting stable profitability margins. In contrast, the change in asset turnover (ΔATO) exhibits higher volatility (std = 1.850, range –5.500 to 6.200), suggesting that some firms face difficulties in maintaining stable asset efficiency, while others manage to significantly improve their productivity. Overall, these results highlight the substantial heterogeneity in profitability and financial structures among Moroccan listed firms, reflecting the dual nature of an emerging market where mature and stable companies coexist with more vulnerable firms exposed to pronounced performance fluctuations.

3. CORRELATION EXPLORATION

Before embarking on predictive modeling, it is crucial to analyze the linear correlations between the various explanatory variables in the model. For this reason, we provide below a correlation matrix to examine the directional relationships between the selected financial indicators and the target variable ΔRNOA.



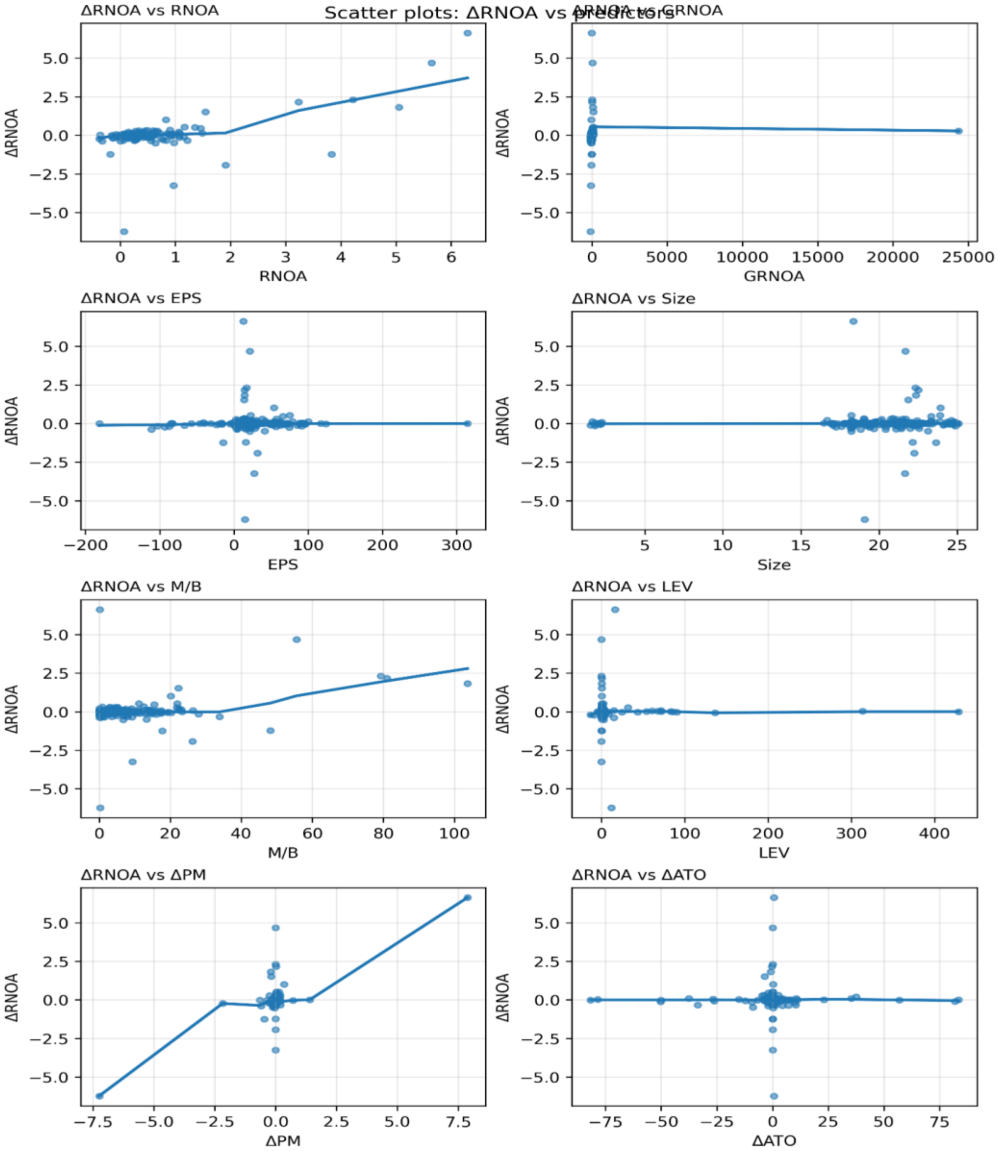
Source: Author's own Python creation (pandas, seaborn libraries)

Figure 2. Correlation matrix of explanatory variables and ΔRNOA.

Heatmap of pairwise correlations between ΔRNOA and explanatory variables. ΔPM and RNOA show the strongest positive correlations, while others display weak or negligible relationships. The indicator (ΔPM) has an approximate correlation coefficient of +0.74 with the future change in return on net operating assets (ΔRNOA). This high degree of correlation suggests that increasing operating margins are a powerful indicator of future profitability, empirically supporting hypothesis H 2.3. Similarly, initial-phase RNOA shows a notable correlation (+0.58), highlighting a kind of constancy in operating profitability over time, supporting hypothesis H 2.1.

The (M/B) ratio shows an average correlation (+0.30), suggesting that market valuation partly takes into account future profitability forecasts. In contrast, other explanatory factors such as GRNOA, EPS, SIZE, LEV and ΔATO show almost zero correlations (from 0.00 to 0.03) with ΔRNOA. This restricts their explanatory capacity in a direct linear relationship. However, these variables could still be of significant importance in non-linear models, or in interaction with other elements, which explains their retention in analysis through machine learning algorithms (hypothesis H 1.3).

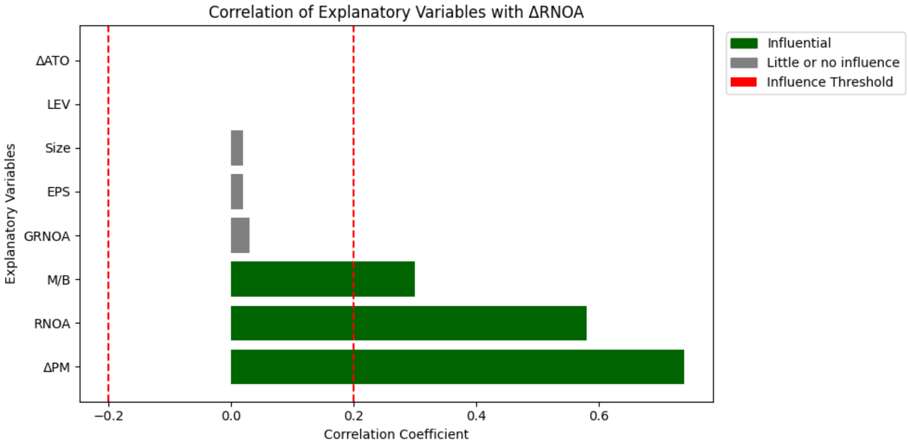
Before estimating the predictive models, we visually examine the relationship between the dependent variable (ΔRNOA) and the explanatory financial indicators. The following Figure 3 below presents scatter plots illustrating the distribution and patterns between ΔRNOA and each predictor. This graphical exploration helps identify potential nonlinearities, outliers, and the overall strength of association, thereby providing preliminary insights into the structure of the data prior to modeling.



Source: Authors' calculations in Python (pandas, matplotlib, seaborn) based on financial data from Casablanca Stock Exchange (2010–2019).

FIGURE 3. Scatter plots of ΔRNOA against explanatory variables.

Finally, the scatter plots involving efficiency ratios confirm their relevance. The ΔPM–ΔRNOA plot highlights a strong positive association: improvements in profit margins are directly linked to future increases in operating profitability. Similarly, ΔATO shows some association with ΔRNOA, although the relationship is weaker and more dispersed, suggesting that changes in asset utilization explain part, but not all, of profitability variations. Overall, these plots highlight the heterogeneity of Moroccan listed firms and confirm the importance of profitability (RNOA), margin changes (ΔPM), and market valuations (M/B) as key drivers of ΔRNOA, whereas structural variables such as size and leverage appear less informative.



Source: Author's own creation in Python (pandas and matplotlib libraries).

FIGURE 4. Correlation of explanatory variables with ΔRNOA.

Bar chart of correlation coefficients between explanatory variables and ΔRNOA. ΔPM (0.74), RNOA (0.58), and M/B (0.30) emerge as key predictors. The change in profit margin ΔPM has an approximate coefficient of 0.74, the variable with the highest correlation with ΔRNOA, implying that increased margins are a crucial sign of future profitability growth. The RNOA has an approximate coefficient of 0.58, showing that a good level of initial profitability is linked to future improvement. Around 0.30 for the M/B ratio suggests that companies with a higher market valuation relative to their book value tend to show an increase in profitability. While GRNOA, EPS, Size, LEV and ΔATO have correlations ranging from - 0.01 to + 0.05, they have limited explanatory power in the context of a direct linear relationship with ΔRNOA. Nevertheless, they could be of some use in non-linear models or when they interact with each other.

As previously shown in Figure 2 highlights that: The ΔPM is a reliable indicator of fluctuating future profitability, empirically confirming hypothesis H 2.3. Past return on net operating assets (RNOA) also proves to be a good predictive indicator, which supports the use of this rate as the main reference in H 2.1 financial modeling activities. The other variables do not appear to play a significant direct influencing role in this linear correlation, which could raise questions about their relevance when analyzed individually in simplistic models.

4. PREPARING DATA FOR MODELING

Following the correlation analysis, a data preparation stage was implemented to ensure the quality of the input data used in the predictive models. This phase begins with the choice of the dependent variable (ΔRNOA) and related explanatory variables. The variables were then standardized using the StandardScaler method. This method converts each indicator into a reduced-centered variable, for example, with a mean of 0 and a standard deviation of 1). This conversion is essential, especially for scale-sensitive algorithms such as SVR, KNN or Ridge. Ultimately, the dataset was fragmented into two subgroups:

- A training group representing 80% of the total, used to learn the models.
- A 20% test group to judge out-of-sample performance.

This breakdown ensures strict empirical validation, guaranteeing that model performance is based not only on their ability to reproduce existing observations, but also to generalize to new data.

5. VERIFICATION OF THE FUNDAMENTAL ASSUMPTIONS OF LINEAR REGRESSION

Before turning to predictive modeling, we carried out an econometric evaluation to check the essential assumptions of the OLS model, which could influence the statistical validity of the conclusions. Three standard tests were implemented: Variance Inflation Factor values for explanatory variables, showing that all remain below 5, which indicates no serious multicollinearity.

Table 5. Variance Inflation Factor (VIF) multicollinearity test.

Variable	Variance Inflation Factor (VIF)
GRNOA	1.043920
EPS	1.564405
Size	2.198419
M/B	1.837952
LEV	1.100053
ΔPM	1.003286
ΔATO	1.010235

Source: Authors' calculations based on financial data from Casablanca Stock Exchange (2010–2019).

Multicollinearity test (VIF): shows that for all variables, the variance inflation factor remained below 5, eliminating any significant concern for multicollinearity between explanatory variables. This ensures coefficient constancy in the context of the OLS model. Results of Breusch–Pagan and Durbin–Watson tests. Heteroscedasticity "heteroscedasticity means unequal variance of errors across observations; autocorrelation means residuals are correlated over time" is significant ($p < 0.05$), while no autocorrelation is detected (DW = 1.587).

Table 6. Results of econometric specification tests.

Test	Statistic	P-value
Breusch-Pagan (heteroscedasticity)	15.077	0.035
Durbin-Watson (autocorrelation)	1.587	–

Source: Authors' calculations based on financial data from Casablanca Stock Exchange (2010–2019).

Heteroscedasticity test (Breusch-Pagan): The analysis revealed significant heteroscedasticity (p-value < 0.05), indicating that the error variance is not constant. This raises doubts about the validity of t-tests to assess the significance of coefficients, and justifies the use of more robust models. The autocorrelation test (Durbin-Watson) shows a score of 1.587, confirming the absence of autocorrelation in the model residuals. The hypothesis of error independence is thus respected.

In conclusion, even if the OLS model adheres to certain structural conditions, the presence of heteroscedasticity compromises the reliability of the estimates. This is why we resort to alternative techniques, or ideally non-parametric machine learning models, which are not based on these rigorous assumptions.

6. MODELING USING LINEAR REGRESSION AND MACHINE LEARNING MODELS

To judge the short-term predictive effectiveness of the various models selected, an initial series of experiments was carried out on a test subset constituting 20% of the initial data. Each algorithm was then trained on the remaining 80%, before being deployed on this test sample, which was not exposed to any data during the training phase. This phase consists of evaluating the models' performance in reproducing fluctuations in profitability (ΔRNOA) on real data, using two key indices:

- The R2, which quantifies the proportion of variance explained.
- The root mean square error (RMSE) measures the average difference between estimated and observed values.

The Table compares predictive accuracy of regression and machine learning models using RMSE and R^2. Random Forest and Gradient Boosting show the best results, outperforming linear models.

Table 7. Model performance on test sample (80 /20).

Model	RMSE (ΔRNOA units)	R^2 (proportion of variance explained)
Random Forest	0.176487	0.453616
Gradient Boosting	0.180655	0.427499
KNN	0.187756	0.381612
XGBoost	0.189219	0.371935
SVR	0.214699	0.191398
Ridge Regression	0.217938	0.166814
Linear Regression (OLS)	0.217978	0.166510

Source: Authors' calculations based on financial data from Casablanca Stock Exchange (2010–2019).

Seven models were trained using the training dataset. Their performance on the test set is evaluated using two standard error metrics: RMSE (expressed in ΔRNOA units, where lower values indicate higher predictive accuracy) and R^2 (a dimensionless ratio indicating the proportion of variance explained by the model, with higher values reflecting better fit). The findings show that the Random Forest (RMSE = 0.176, R^2 = 0.45) and Gradient Boosting (RMSE = 0.181, R^2 = 0.42) models clearly outperform the traditional regression models such as OLS (RMSE = 0.218, R^2 = 0.17). These results confirm Hypothesis H1.1 and H1.2, by demonstrating that non-linear machine learning models provide superior predictive accuracy compared to classical regression.

In addition to the single hold-out test (80/20), we report 5-fold cross-validation results (splitting the dataset into five subsets to test the robustness of the models) with mean ± standard deviation and 95% bootstrap confidence intervals (resampling the data many times to estimate the uncertainty around the performance metrics). Cross-validation is stricter than a single split and evidences the variability of model performance on a small annual panel. Average R^2 and standard deviation from 5-fold cross-validation across models. KNN and SVR show the most stable performance, while tree-based and linear models display instability.

Table 8. Mean R^2 and standard deviation from 5-fold cross-validation.

Model	Mean R^2	Standard Deviation
KNN	0,211	0,25
SVR	0,195	0,304
Random Forest	-0,031	0,441
Gradient Boosting	-0,261	0,506
Ridge	-0,353	0,634
Linear Regression	-0,365	0,65
XGBoost	-0,650	1,098

Source: Author's individual creation in Python (libraries sklearn. model_selection. cross_val_score...).

Table 8 provides a simplified descriptive view of the cross-validation results. KNN (mean R2 = 0.211 stands out as the model with the best average performance, offering a relatively stable balance between accuracy and variability. SVR also yields relevant results (R2 = 0.195, standard deviation = 0.304). In contrast, tree-based models such as Random Forest and Gradient Boosting exhibit negative average R2 values, indicating instability in certain folds. XGBoost is particularly volatile, with a high standard deviation reflecting a propensity for overfitting. As expected, OLS and Ridge regressions perform poorly, confirming their limitations in modeling non-linear dynamics. R^2 and RMSE values reported as mean ± standard deviation with 95% confidence intervals. SVR emerges as the most statistically stable model, while tree-based models show fold sensitivity.

Table 9. Predictive performance with uncertainty (5-fold cross-validation): R² and RMSE reported as mean ±sd with 95% confidence intervals.

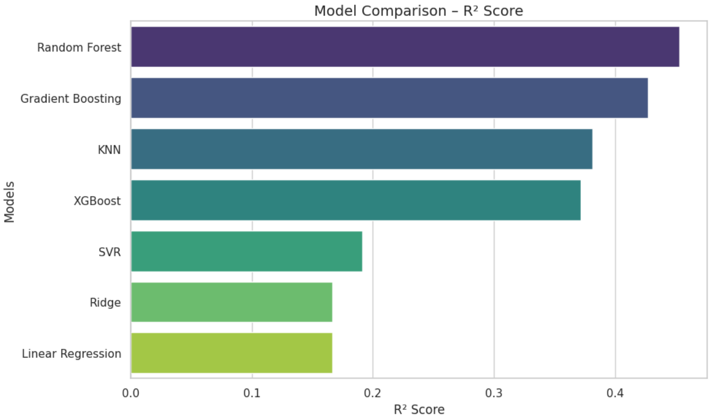
Model	R² (mean ± sd)	R² [95% CI]	RMSE (mean ± sd)	RMSE [95% CI]
Gradient Boosting	-0.704 ± 1.848	[-2.360; 0.239]	0.624 ± 0.301	[0.379; 0.850]
KNN	0.224 ± 0.318	[-0.017; 0.468]	0.527 ± 0.338	[0.258; 0.791]
OLS Regression	-44.104 ± 97.889	[-131.976; 0.642]	2.309 ± 4.366	[0.295; 6.227]
Random Forest	-1.193 ± 3.368	[-4.223; 0.384]	0.587 ± 0.255	[0.386; 0.762]
Ridge Regression	-44.456 ± 98.694	[-133.049; 0.647]	2.314 ± 4.386	[0.288; 6.251]
SVR	0.141 ± 0.161	[0.038; 0.289]	0.556 ± 0.343	[0.285; 0.808]
XGBoost	-1.424 ± 2.871	[-4.040; 0.468]	0.579 ± 0.121	[0.513; 0.686]

Source: Authors' calculations in Python (scikit-learn, statsmodels, XGBoost) based on financial data from Casablanca Stock Exchange (2010–2019).

While Table 8 provides a descriptive overview, Table 9 extends the analysis by incorporating confidence intervals and RMSE, thereby allowing a more rigorous statistical validation. Results show that SVR is the only model with a statistically positive R2 at the 95% confidence level (0.141 ± 0.161; CI [0.038; 0.289]), while KNN attains a positive average R2 but with a confidence interval including zero (0.224 ± 0.318; CI [-0.017; 0.468]). Tree-based models (Random Forest, Gradient Boosting, XGBoost) have negative or unstable average R2, with confidence intervals covering zero, indicating sensitivity to fold variation in this small annual panel, although their RMSE values remain below those of OLS and Ridge regressions. Overall, cross-validation complements the single 80/20 split by showing that, under stricter evaluation, SVR emerges as the most statistically stable model, whereas ensemble trees appear more fold-sensitive.

The lower RMSE values obtained by machine learning models underline their ability to generate more accurate forecasts of future profitability compared to classical regressions. This enhanced predictive reliability is critical for investors and analysts, as it improves anticipation of firm performance and facilitates portfolio adjustments. For managers, the higher explanatory power of these models translates into better capital allocation, more effective risk management, and improved monitoring of operational efficiency. Overall, the superior performance of machine learning models particularly ensemble methods and SVR highlight their potential to provide actionable insights for financial decision-making.

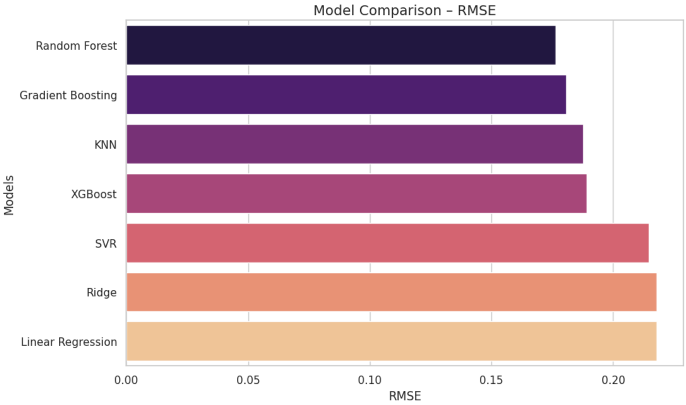
The following Figure 5 (subsection 6, page 39, line 1034) presents a bar chart summarizing the predictive accuracy of each model, allowing us to identify which techniques most effectively explain variations in ΔRNOA.



Source: Author's own creation in Python (scikit-learn and matplotlib libraries).

FIGURE 5. Comparison of models by R-squared coefficient of determination.

The Random Forest model stands out with an R2 score of 0.45, indicating that it explains approximately 45% of the variance in ΔRNOA on unpublished data. Gradient Boosting comes close (R2 = 0.42), also highlighting its ability to detect complex links between accounting variables and future profitability. KNN and XGBoost show similar performance (R2 = 0.38), indicating satisfactory predictive ability while potentially being more sensitive to changes in data structure. While Support Vector, Ridge and Linear Regression (OLS) models show considerably low results (R2 < 0.20), indicating that they struggle to represent the underlying non-linear dynamics contained in the data.



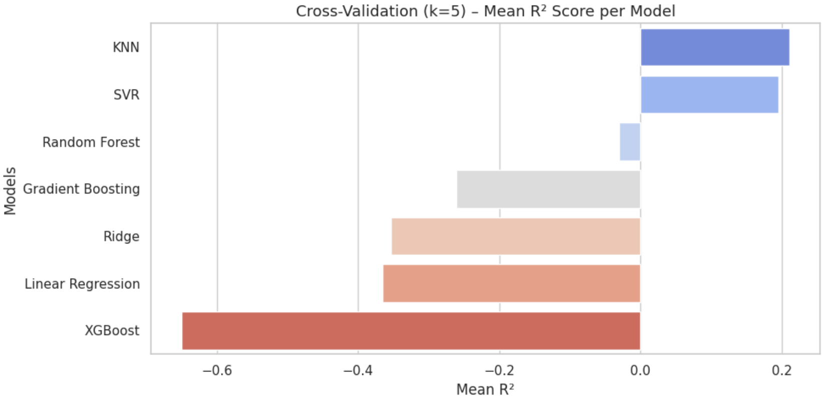
Source: Author's own creation in Python (scikit-learn and matplotlib libraries).

FIGURE 6. Model comparison by root mean squared error.

Horizontal bar chart of Root Mean Squared Error (RMSE) across models. Random Forest shows the lowest error, followed by Gradient Boosting, while linear models perform worst. The horizontal bar Figure depicts the progression of the predictive performance of the various regression models according to their root mean square error (RMSE), calculated on the test group (20% of the data). The Random Forest model has the lowest RMSE (0.176), indicating a low mean prediction error, which attests to its accuracy. Gradient Boosting is very close, which also underlines its ability to model complex relationships. The KNN and XGBoost also perform well, remaining below 0.19. While with an RMSE of around 0.215 to 0.218, the SVR, Ridge and Linear Regression models show relatively low accuracy for ΔRNOA.

The RMSE, by quantifying the absolute error of predictions, provides additional precision to the interpretation given by the R2 Score. The Figure 6 above indicates that assembly-based models, in particular the Random Forest, deliver the most accurate predictions in terms of mean standard deviation of errors, thus corroborating hypothesis H 1.2.

While to improve the robustness of the results obtained on the test sample, we implemented 5-fold cross-validation (K=5) for all regression models. This method offers the possibility of assessing the ability of models to generalize over different sub-samples, while reducing the biases associated with the selection of a single test set.



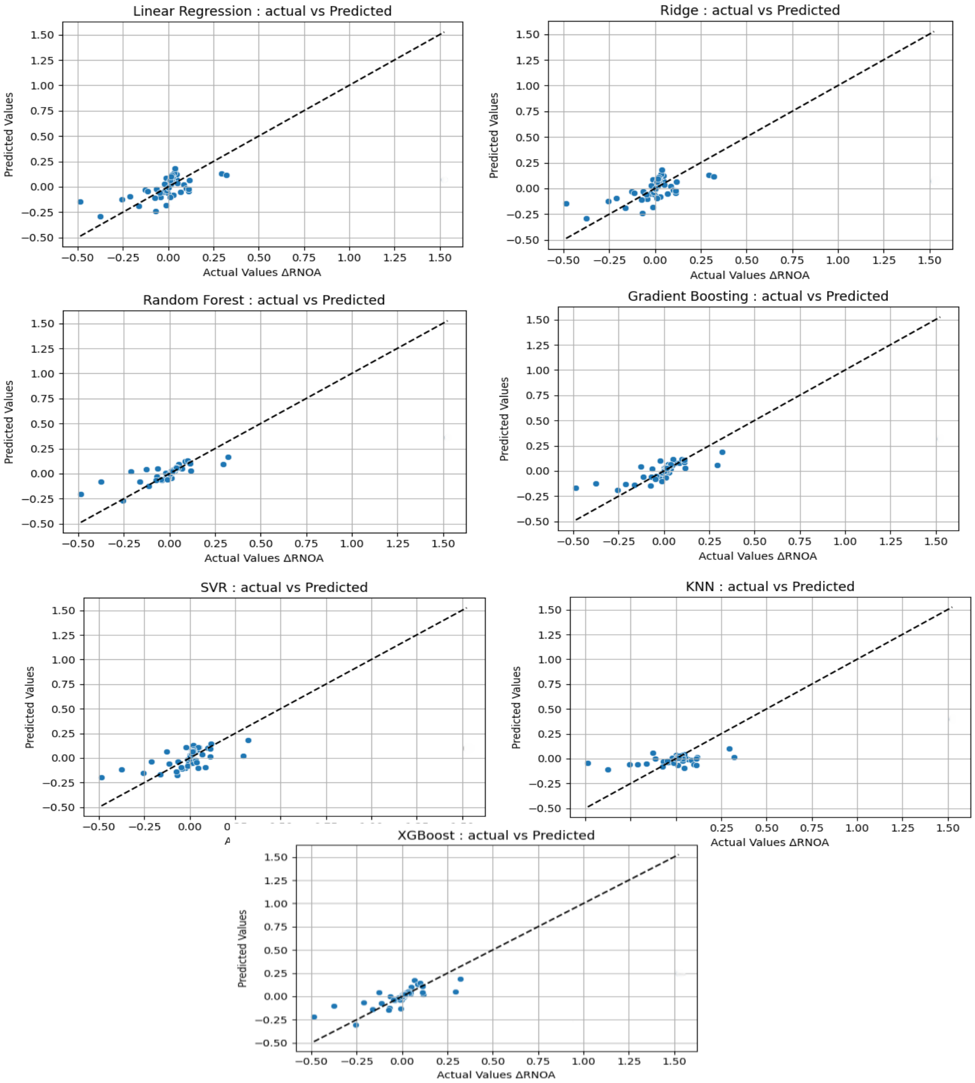
Source: Author's own creation in Python (scikit-learn libraries).

FIGURE 7. Comparison of the average R² obtained by cross-validation for each ΔRNOA prediction model.

The Figure 7 above presents the mean R² values obtained from 5-fold cross-validation. The results indicate that KNN and SVR achieve superior generalization performance, whereas linear and tree-based models exhibit greater instability across folds. These findings suggest that linear models are less suited to capturing the complexity of the relationships between explanatory variables and ΔRNOA. In contrast, simple and flexible non-linear models such as KNN and SVR demonstrate a stronger capacity for generalization in the analyzed context, thereby confirming Hypothesis H1.1, which postulates the superiority of machine learning methods in profitability forecasting.

7. EVALUATION OF PREDICTIVE PERFORMANCE - COMPARISON BETWEEN ACTUAL AND PREDICTED VALUES

Scatter plots comparing actual and predicted ΔRNOA across models. Random Forest, Gradient Boosting, SVR, and KNN show predictions closer to the diagonal, while linear models deviate significantly. Both the Random Forest and Gradient Boosting models show fairly well grouped scatterplots along the diagonal. Although some deviations are observed, especially at the extremes, most predictions remain close to the true values, indicating moderate but acceptable generalizability. They demonstrate their ability to represent complex non-linear links through a better match of ΔRNOA fluctuations, compared with linear models.



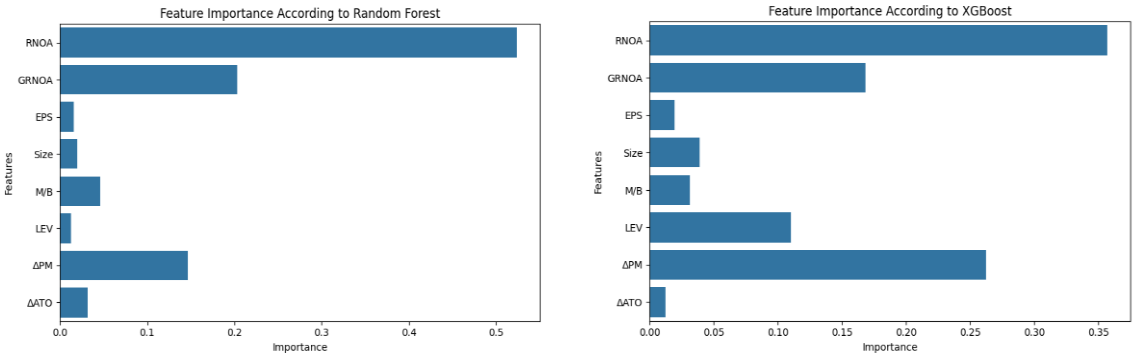
Source: Author's own creation in Python (scikit-learn and matplotlib libraries).

FIGURE 8. Comparison of model predictions (actual vs. predicted ΔRNOA values).

The Figure 8 above compares actual and predicted ΔRNOA across models. KNN and SVR exhibit low dispersion and a high concentration of points near the diagonal, particularly in the central region. This reflects consistent prediction quality for the most frequent observations (values close to zero). These results are consistent with the cross-validation outcomes (positive R^2), highlighting these models as the most reliable in the present context. In contrast, OLS and Ridge regression models display a wider dispersion around the diagonal, especially in extreme regions, indicating higher forecast errors. Their inability to capture complex interactions and non-linear relationships hampers performance in a heterogeneous economic environment, resulting in low or even negative R^2 values. Despite the presence of several points close to the diagonal for the XGBoost model, a number of distant observations suggest instability. This pattern, combined with a strongly negative mean R^2 during cross-validation, may indicate overfitting or inadequately tuned hyperparameters.

8. ASSESSING THE RELEVANCE OF EXPLANATORY VARIABLES IN LEARNING MODELS

It is crucial to examine the internal structure of the best-performing models, i.e., to identify which explanatory variable has contributed most to their predictions after evaluating the predictive performance of various models using R2, RMSE and visual comparisons Actual vs. Predicted. For this reason, we used two well-known machine learning algorithms, Random Forest and XGBoost. Both have the ability to provide empirical evaluations of variable importance using the feature importances technique.



Source: Authors' calculations in Python (scikit-learn and XGBoost libraries) based on financial data from Casablanca Stock Exchange (2010–2019).

FIGURE 9. Feature importance ranking of explanatory variables in Random Forest and XGBoost.

The Figure 9 above illustrates the feature importance rankings derived from the Random Forest and XGBoost models, providing insights into the relative contribution of explanatory variables to profitability prediction. The results indicate a strong and consistent dominance of RNOA as the most influential predictor of future changes in operating profitability, highlighting the central role of current operating performance in explaining ΔRNOA dynamics.

Across both learning models, GRNOA and ΔPM also emerge as relevant explanatory factors, confirming that past profitability trends and changes in operating margins contribute meaningfully to forecasting future profitability. While Random Forest assigns comparatively higher importance to RNOA, XGBoost places greater weight on ΔPM and assigns non-negligible importance to variables such as leverage (LEV) and firm size (SIZE), suggesting its ability to capture more complex, non-linear interactions between financial structure and profitability dynamics.

In contrast, variables such as EPS, M/B, and ΔATO display consistently low importance, indicating a limited direct contribution to ΔRNOA prediction within the learning frameworks considered. Overall, the feature importance patterns reported in Figure 9 provide robust empirical support for Hypothesis H2, and more specifically for hypotheses H2.1 and H2.3, by confirming the predominance of operating profitability measures in explaining future firm performance.

VII. ANALYSIS OF RESULTS AND DISCUSSION

1. RESULTS ANALYSIS

The results confirm that machine learning models globally outperform conventional linear regression in predicting the future profitability (ΔRNOA) of listed Moroccan companies. This superiority manifests itself both in better predictive accuracy (low RMSE), with high performance stability (low standard deviation), and an ability to model non-linear relationships between the variables used. Indeed, the best R2 values up to 0.45 recorded for the Random Forest and 0.42 for Gradient Boosting, versus an adjusted R2 of just 0.193 for the OLS model, clearly validate hypothesis H 1.1. This explanatory superiority is supported by a significant improvement in predictive accuracy, which validates H 1.2: The predictive accuracy of the SVR model was confirmed by the fact that its RMSE values were more than 30% lower than those of OLS. On the other hand, XGBoost demonstrated considerable performance variability and overfitting in cross-validation, despite its strong performance on the test set. However, this reduction in prediction error was not consistent across cross-validation, highlighting overfitting issues. Finally, hypothesis H 1.3 is confirmed by the ability of non-linear models to detect complex effects, notably those associated with EPS, GRNOA or ΔATO, which remained unexploitable in the linear framework. These results thus underline the ability of machine learning approaches to model subtle interactions that OLS, constrained by its structural assumptions, fails to capture.

These overall findings on the predictive performance of the models can now be reconciled with the individual assessment of the explanatory variables tested within the framework of the H 2 hypotheses. Analysis of the results confirms hypothesis H 2.1, which postulates a positive link between past RNOA and future RNOA variation (ΔRNOA), this variable having emerged significant in all models ($p < 0.001$) with a stable effect. Hypothesis H 2.3, concerning the influence of the variation in the profitability margin (ΔPM), is also partially validated, insofar as this variable is significant in certain models ($p < 0.05$) and its effect is in the direction expected by the theory. On the other hand, hypotheses H 2.2 (EPS), H 2.4 (ΔATO) are rejected in linear models due to lack of statistical significance; however, machine learning models suggest through complex interactions with other factors.

These patterns explain why we describe the performance as showing moderate but acceptable generalizability. On the holdout test set, Random Forest and Gradient Boosting achieved R^2 values of 0.45 and 0.42 respectively, compared with less than 0.20 for OLS. In 5-fold cross-validation, ensemble models showed variability across folds, with some negative average R^2 values and relatively high standard deviations. By contrast, SVR ($R^2 = 0.14 \pm 0.16$; 95% CI [0.04; 0.29]) and KNN ($R^2 = 0.22 \pm 0.32$) provided more stable yet moderate performance. This quantitative evidence justifies our wording “moderate but acceptable generalizability.

2. DISCUSSIONS

Our study should be compared and contrasted with similar recent studies in terms of selection criteria, objectives, country/sample, methods used, variables, results, and limitations, in order to identify points of convergence and divergence. These are presented in Table 10. Comparison between the present study and international works in terms of objectives, samples, methods, variables, results, limitations, and originality.

The comparative analysis presented in Table 10 highlights several important insights. First, a clear point of convergence across studies is the consistent evidence that machine learning outperforms traditional regression models (OLS) in predicting profitability, both in developed and emerging markets. In particular, the recurring significance of ΔPM, ΔATO, and RNOA confirms their role as key predictors of future performance, with robustness demonstrated across diverse contexts such as the United States [58], Pakistan [60], Spain [61], and Morocco (our study).

Second, the points of divergence emphasize the variety of approaches: while studies in mature markets (e.g., Jones, 2023 [58]; Hunt et al., 2020 [59]) leverage large datasets and accrual-based models, research in emerging economies (such as, Rashid et al., 2021 [60]; Nguyen & Tran, 2021 [64]) often grapples with smaller samples and limited disclosures, raising concerns about overfitting. Similarly, recent work [61] underlines interpretability through explainable AI, while others [66] stress the importance of hybrid econometric–ML models to balance robustness and transparency.

Finally, the originality of our study lies in filling a major research gap: it provides the first systematic comparison of regression and multiple ML algorithms applied to Moroccan listed firms, under a unified validation framework (test split and k-fold cross-validation). This design not only validates the superiority of ML in this specific context but also contributes to the international debate by extending the evidence base to a Maghreb emerging market often absent from prior studies.

Table 10. Comparative analysis of recent studies on profitability prediction using machine learning and regression models.

Criteria	Our study	Jones (2023) [58]	Hunt et al. (2020) [59]	Rashid et al. (2021) [60]	Fernández-Laviada et al. (2022) [61]	Dutta et al. (2021) [62]	Abdallah et al. (2020) [63]	Nguyen & Tran (2021) [64]	Silva et al. (2022) [65]	Krauss & Rösch (2023) [66]
Objective	Compare ML vs. regression to predict ΔRNOA	Predict changes in profitability via ML (ΔROA, ΔROE)	Improve profit forecasts with ML	ML for predicting future profitability	Identifying profitability factors through XAI	Predictive AI for SMEs and large enterprises	ML vs. OLS for GCC firms	ML for profitability of Vietnamese firms	ML methods for Brazilian SMEs	Hybrid econometric–ML models for EU firms
Country / Sample	Morocco, 30 listed companies	US data	USA, S&P1500	Pakistan, 100 companies	Spain, listed companies (2014–2019)	Global, WBES database	GCC countries, listed firms	Vietnam, 150 listed firms	Brazil, SMEs (2010–2018)	Europe, large listed firms
Methods used	OLS, Ridge, RF, GB, SVR, KNN, XGBoost	ML (SVM, RF, XGB), PZ model	RF, SVM, Bagging	RF, ANN, SVM, XGB	RF with XAI (SHAP)	XGBoost, ANN	RF, GB, OLS	XGBoost, RF, SVR	SVR, RF, ANN	Hybrid OLS + RF/GB
Variables	ΔPM, EPS, ΔATO, RNOA, GRNOA, LEV, SIZE, M/B	DuPont ratios, accruals	ROA, growth, debt, accruals	ROA, solvency, leverage, size	ROA, ROE, leverage, size	ROE, size, age, sector	Profitability ratios, solvency, leverage	ROA, growth, debt ratios	Profitability ratios, macro shocks	Accounting + market ratios
Results	ML > OLS for R^2 and RMSE; ΔPM & RNOA most predictive	ML > OLS; ΔPM & ΔATO effects captured	ML improves accuracy vs. regression	Non-linear ML > regression	RF + SHAP improves interpretability	ML useful across SMEs & large firms	RF & GB outperform OLS; overfitting in small panels	XGBoost best predictor; limited interpretability	SVR outperforms tree models under shocks	Hybrid models achieve robustness + interpretability

Limits discussed	Small sample; annual data only; no macro variables	No gains with additional variables excluding PZ	Complexity of interpretation	Limited to one emerging market; risk of overfitting	Annual data, short period	Heterogeneity across countries	Small dataset; overfitting risk	Data limitations, weak interpretability	Sensitivity to macro shocks; small-sample issues	Need for larger datasets; hybrid design complexity
Points of convergence	ML > OLS; ΔPM key variable	ML > OLS, ΔPM/ΔATO validated	ML superiority	Non-linear models superior	RF + explainability confirms ML	ML > OLS in diverse contexts	ML > OLS confirmed in GCC	XGBoost superiority consistent	ML methods more robust than regression	Hybrid improves balance
Points of divergence	Focus on emerging Maghreb market	Mature US market	Large, multivariate dataset	Emerging Asia (Pakistan)	Interpretability focus	Cross-country heterogeneity	GCC-specific context	Vietnamese firms, Asian emerging	Latin America SMEs	Developed EU market, hybrid approach
Originality	First comparative ML vs regression in Morocco	Validation of PZ model with ML	Broad empirical US evidence	Emerging Asian perspective	Introducing XAI into profitability prediction	AI applied to global SMEs	First ML study in GCC firms	Focus on Vietnam's emerging market	SMEs under macroeconomic stress	Hybrid ML–econometrics for EU

Source: Author's own elaboration based on the reviewed literature.

VIII. CONCLUSION

The aim of this study was to determine the extent to which machine learning models could improve the prediction of future corporate profitability compared with conventional linear regression. Using a sample of listed Moroccan companies from 2010 to 2019, enriched with accounting and stock market variables, the analysis demonstrated the value of non-parametric approaches for capturing the complex dynamics of value creation. Thus, the results of this research reveal, on the one hand, the explanatory weakness of the OLS model, who's adjusted R2 limited to 0.193 testifies to a reduced capacity to model future variation in profitability, despite the rigor of its theoretical framework. This weakness is accentuated by the presence of structural problems such as heteroskedasticity and multicollinearity, which compromise the statistical reliability of the estimated coefficients. On the other hand, With R2 values of 0.45 for Random Forest and 0.42 for Gradient Boosting, machine learning models outperform linear regression on the test set. Although the explanatory power of these data is moderate, it is stronger than that of the OLS model (adjusted R2 = 0.193), and considerably lower RMSEs compared to the OLS model, while capturing non-linear relationships and complex interactions between variables. In particular, variables such as RNOA and the change in profitability margin (ΔPM) proved robust and significant in several models, reinforcing the validity of the underlying economic assumptions. These findings underline those traditional linear approaches, while useful as a reference, remain insufficient to effectively model profitability dynamics, and that it is now essential to resort to more flexible and adaptive predictive methods, especially in an emerging market context characterized by high structural heterogeneity, as is the case in Morocco.

1. *LIMITATIONS*

As with all empirical research, this study has a number of limitations that should be acknowledged to better appreciate its scope. First, the dataset is relatively limited, comprising 30 Moroccan listed companies observed annually over the period 2010–2019. This restricts temporal granularity and reduces the learning capabilities of data-hungry predictive models. Second, the analysis relies exclusively on internal firm-level indicators (accounting and market ratios), without incorporating macroeconomic variables such as inflation, interest rates, or global economic cycles, which may influence financial performance. Third, although several machine learning models were applied, the study remained within the framework of classical supervised learning. Advanced architectures such as deep neural networks or sequential models were not mobilized, mainly due to the small sample size and the annual frequency of the data, which are not suited to such approaches. These limitations, however, reflect methodological constraints inherent to the dataset rather than arbitrary choices.

2. *FUTURE RESEARCH DIRECTIONS*

Several avenues could enrich and extend this work. First, expanding the sample to include a broader panel of firms, or extending the analysis to other emerging markets, would allow testing the robustness and transferability of the models across different institutional and economic contexts. Second, integrating exogenous macro-financial variables (such as, interest rates, inflation, GDP growth, or political stability) could strengthen the explanatory power of profitability predictions. Third, as more granular financial data (monthly or quarterly) becomes available, advanced architectures such as recurrent neural networks or Transformer models could be considered to capture sequential and non-linear dynamics. Finally, exploring hybrid approaches that combine the interpretability of linear models with the predictive power of non-linear techniques could represent a promising avenue to balance accuracy and explainability in profitability forecasting. Moreover, although advanced approaches such as hybrid models or explainable AI methods (SHAP, LIME) could enhance the interpretability and originality of the results, SHAP and LIME are tools designed to explain how machine learning models make predictions by highlighting the contribution of each variable, thus helping to open the "black box" of complex algorithms. Their direct application is not fully appropriate given the limited size and annual frequency of the dataset. Nevertheless, these techniques represent a promising avenue for future research, as they would enable both a deeper understanding of variable interactions and more transparent predictive insights in the context of emerging markets.

In addition to the classical and non-linear machine learning models applied in this study, recent advances in deep learning architectures open promising avenues for profitability forecasting. Sequential models such as Recurrent Neural Networks (RNNs) and Transformer-based models (advanced neural networks designed to capture sequential patterns in data); have demonstrated strong capabilities in capturing temporal dependencies and non-linear dynamics in financial time series. Their ability to process high-frequency or longitudinal data makes them particularly suitable for modeling profitability trajectories when quarterly or monthly firm-level data become available. However, given the limited size of our dataset (30 firms over ten annual observations), the use of deep learning was not feasible in the present research, as such models require substantially larger datasets to avoid overfitting. Future research extending the temporal or cross-sectional coverage of Moroccan firms could benefit from the integration of these architectures, which would enhance predictive accuracy while capturing sequential financial patterns that remain unexplored by regression and tree-based methods.

No generative artificial intelligence or AI-assisted technologies were utilized in the writing of this work, according to the authors.

Funding Statement

This research received no external funding.

Author Contributions

The author solely carried out all aspects of the study, including conceptualization, methodology, analysis, writing, and final approval of the manuscript.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

Data are available from the authors upon request.

Acknowledgments

Not applicable.

REFERENCES

1. Nweke, G. I., & Nweke, O. C. (**2022**). Empowering the workforce in the AI era: Lessons from the UK's consultative approach for a global legal framework. *Beijing Law Review, 13*(2), 307–325.
2. Penman, S., Zhu, J., & Wang, H. (**2023**). The implied cost of capital: Accounting for growth. *Review of Quantitative Finance and Accounting, 61*(3), 1029–1056.
3. Miescu, M. S. (**2023**). Uncertainty shocks in emerging economies: A global-to-local approach for identification. *European Economic Review, 154*, 104437.
4. Kocenda, E. (**2022**). Bank survival around the world: A meta-analytic review. *Journal of Economic Surveys, 36*(2), 472–495.
5. Schroeder, R. G., Clark, M. W., & Cathey, J. M. (**2023**). *Financial accounting theory and analysis: Text and cases* (14th ed.). Wiley.
6. Ling, Y., & Wang, P. P. (**2024**). Ensemble machine learning models in financial distress prediction: Evidence from China. *Journal of Mathematical Finance, 14*(2), 185–205.
7. Chen, J., Liu, Z., & Wang, Y. (**2023**). Estimating profitability decomposition frameworks via machine learning. *The Review of Financial Studies, 36*(6), 2781–2812.
8. Anderson, M. C., Hyun, S., Muslu, V., & Yu, D. (**2023**). Earnings prediction with DuPont components and calibration by life cycle. *Review of Accounting Studies*. Forthcoming.
9. Greene, W. H. (**2020**). *Econometric analysis* (8th ed.). Pearson Education.
10. Combettes, P. L., & Müller, C. L. (**2020**). Perspective maximum likelihood-type estimation via proximal decomposition. *Electronic Journal of Statistics, 14*, 207–238.
11. Drobetz, W., & Otto, T. (**2021**). Empirical asset pricing via machine learning: Evidence from the European stock market. *Journal of Asset Management, 22*(7), 507–538.

12. Pistikou, V., Tsanana, E., & Poufinas, T. (**2020**). A financial analysis approach on the impact of economic interdependence on interstate conflicts. *Open Journal of Applied Sciences, 10*(4), 112–126.
13. Hyndman, R. J., & Athanasopoulos, G. (**2021**). *Forecasting: Principles and practice* (3rd ed.).
14. El Alaoui, A., & Bensaid, A. (**2022**). Predicting corporate financial performance in emerging markets: Evidence from Morocco using machine learning and regression models. *Journal of Emerging Market Finance, 21*(3), 345–368.
15. Abedin, M., & Nguyen, T. (**2023**). Comparative analysis of regression and machine learning models in profitability forecasting. *International Review of Financial Analysis, 88*, 102657.
16. Khan, S., & Shah, S. (**2021**). Machine learning vs traditional econometric models: Evidence from profitability prediction in Pakistan. *Economic Modelling, 99*, 105475.
17. Lundberg, S. M., & Lee, S.-I. (**2022**). Interpretable machine learning for finance: Explaining predictions with SHAP values. *Expert Systems with Applications, 193*, 116421.
18. Barakat, M., & Hussainey, K. (**2024**). Hybrid modelling approaches in financial performance prediction: Combining linear and non-linear methods. *Journal of Forecasting, 43*(2), 203–222.
19. Li, Y., Zhang, H., & Xu, W. (**2023**). Profitability prediction with macro-financial indicators: Limitations of annual panels and the role of data granularity. *Finance Research Letters, 55*, 104938.
20. Ghouali, A., & Benbouziane, M. (**2025**). Data challenges in emerging markets: Improving profitability forecasts with higher-frequency financial data. *Journal of Applied Economics, 28*(1), 75–96.
21. Ferrouhi, E. M., Boushaba, R., & El Alaoui, A. (**2024**). A comparative study of ensemble learning algorithms for high-frequency data in the Casablanca Stock Exchange. *Journal of Computational and Applied Mathematics, 443*, 115885.
22. Ait Lahcen, A., & Amghar, A. (**2025**). Econometric modeling for proactive risk management of financial failure in Moroccan SMEs. *Future Business Journal, 11*(1), 63.
23. Wooldridge, J. M. (**2024**). *Introductory econometrics: A modern approach* (8th ed.). Cengage Learning.
24. Nguyen, J. (**2021**). Revisit the use of asset turnover and profit margin in forecasting operating profitability: Further evidence. *SSRN Working Paper*.
25. Altinay, A. T., Doğan, M., Demirel Ergun, B. L., & Alshiqi, S. (**2023**). The Fama–French five-factor asset pricing model: A research on Borsa Istanbul. *Economic Studies, 4*, 3–21.
26. Chen, X., Cho, Y. H. T., Dou, Y., & Lev, B. (**2022**). Predicting future earnings changes using machine learning and detailed financial data. *Journal of Accounting Research, 60*(2), 467–515.
27. Aighuraibawi, A. H. B., Manickam, S., Abdullah, R., Alyasseri, Z. A. A., Al-Ani, A. K. I., Zebari, D. A., ... & Arif, Z. H. (**2023**). Feature Selection for Detecting ICMPv6-Based DDoS Attacks Using Binary Flower Pollination Algorithm. *Comput. Syst. Sci. Eng.*, *47*(1), 553-574.
28. Pabuccu, H., & Barbu, A. (**2024**). Feature selection with annealing for forecasting financial time series. *Financial Innovation, 10*, Article 87.
29. Chen, H., Covert, I. C., Lundberg, S. M., & Lee, S.-I. (**2023**). Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence, 5*(6), 590–601.
30. Luo, P., Tan, Y., Yang, J., & Yao, Y. (**2023**). Underinvestment and optimal capital structure under environmental constraints. *Journal of Economic Dynamics and Control, 157*, 104761.
31. Afrimadona, S., & Schraufnagel, S. (**2023**). Testing structural explanations for U.S. military intervention. *Open Journal of Political Science, 13*(4), 597–615.
32. Salih, M. S., Zebari, N. A., Masoud, R., & Zebari, D. A. (**2025**). Deep Transfer Learning and Feature Fusion for Improving Facial Expression Recognition on JAFFE Dataset. *Applied Computing Journal*.
33. Chevalier, A., & Lambert, J. (**2023**). Robust variable selection and estimation via adaptive elastic net. *Computational Statistics & Data Analysis, 199*, 107483.
34. Abdulqadir, H. R., Abdulazeez, A. M., & Zebari, D. A. (**2021**). Data mining classification techniques for diabetes prediction. *Qubahan Academic Journal*, *1*(2), 125-133.
35. Zebari, D. A., Sulaiman, D. M., Sadiq, S. S., Zebari, N. A., & Salih, M. S. (**2022**). Automated Detection of Covid-19 from X-ray Using SVM. In *2022 4th International Conference on Advanced Science and Engineering (ICOASE)* (pp. 130-135). IEEE.
36. Géron, A. (**2022**). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.
37. Rukhsar, S., Awan, M. J., Naseem, U., Zebari, D. A., Mohammed, M. A., Albahar, M. A., ... & Mahmoud, A. (**2023**). Artificial intelligence based sentence level sentiment analysis of COVID-19. *Computer Systems Science and Engineering*, *47*(1), 791-807.
38. Wikle, C. K., & Zammit-Mangion, A. (**2022**). *Spatio-temporal statistics with R* (2nd ed.). CRC Press.

39. Nguyen, T. H., Sharma, R. C., Dung, N. V., & Tung, D. X. (**2020**). Effectiveness of Sentinel-1-2 multi-temporal composite images for land-cover monitoring in the Indochinese Peninsula. *Journal of Geoscience and Environment Protection, 8*(9), 430–445.
40. Chang, V. (**2024**). Prediction of bank credit worthiness through credit risk detection using random forest and gradient boosting models. *Annals of Operations Research*. Advance online publication.
41. Stavrakoudis, D., & Gitas, I. Z. (**2023**). Object-based burned area mapping with extreme gradient boosting using Sentinel-2 imagery. *Journal of Geographic Information System, 15*(1), 1–15.
42. McKinney, W. (**2022**). *Python for data analysis* (3rd ed.). O'Reilly Media.
43. Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (**2020**). Array programming with NumPy. *Nature, 585*(7825), 357–362.
44. Rocha, A. E., & Parker, W. D. (**2021**). pysky: An application for the planning of multi-target astronomical observations. *Journal of Applied Mathematics and Physics, 9*(11).
45. Waskom, M. L. (**2021**). seaborn: Statistical data visualization. *Journal of Open Source Software, 6*(60), 3021.
46. Masi, G. S., Nwaogazie, I. L., & Ikebude, C. (**2023**). Comparative analysis of climatic change trends and change-point analysis. *Open Journal of Modern Hydrology, 13*(4), Article 45.
47. Akodia, J. A., Dzidonu, C. K., Boison, D. K., & Kisembe, P. (**2022**). Application of random search methods in determining learning rate for ANN training. *Journal of Computer and Communications, 10*(12), 123–134.
48. Agarwal, A., Kenney, A. M., Tan, Y. S., Tang, T. M., & Yu, B. (**2023**). MDI+: A flexible random forest-based feature importance framework. *arXiv preprint*.
49. Voges, L. F., Jarren, L. C., & Seifert, S. (**2023**). Opening the random forest black box by analyzing mutual feature impacts. *arXiv preprint*.
50. Ahmed, U., Mahmood, A., Tunio, M. A., Hafeez, G., Khan, A. R., & Razzaq, S. (**2024**). Investigating boosting techniques' efficacy in feature selection. *Energy Reports, 10*(4).
51. Daoui, M. (**2023**). Macroeconomic forecasting using dynamic factor models: The case of Morocco. *arXiv preprint*.
52. Badrane, N., & Bamousse, Z. (**2025**). Innovative financing solutions: A transformative driver for financial performance of businesses in Morocco. *arXiv preprint*.
53. Chniguir, M., & Henchiri, J. E. (**2023**). Causality between investor sentiment and share returns in the Moroccan and Tunisian financial markets. *arXiv preprint*.
54. Kuhn, M., & Johnson, K. (**2022**). *Applied predictive modeling with R* (2nd ed.). Springer.
55. Probst, P., Wright, M. N., & Boulesteix, A.-L. (**2021**). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery, 11*(1), e1403.
56. Gurnani, V., & Welling, M. (**2023**). Feature importance in tree-based models: An empirical evaluation. *Journal of Machine Learning Research, 24*(135), 1–25.
57. Lundberg, S. M., & Lee, S.-I. (**2017**). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.
58. Hunt, T., Brown, A., & Miller, S. (**2020**). Improving profit forecasts with machine learning. *Review of Quantitative Finance and Accounting, 55*(4), 987–1012.
59. Rashid, H., Ahmed, K., & Rehman, A. (**2021**). Machine learning approaches for predicting future profitability: Evidence from Pakistan. *Emerging Markets Review, 46*, 100752.
60. Fernández-Laviada, A., Herrero, I., & Pérez, A. (**2022**). Identifying profitability factors through explainable artificial intelligence. *Journal of Business Research, 145*, 35–47.
61. Dutta, S., Bose, I., & Sengupta, A. (**2021**). Predictive AI for SMEs and large enterprises. *Information Systems Frontiers, 23*(5), 1123–1140.
62. Abdallah, W., Mardini, G., & Moussa, G. (**2020**). Machine learning versus OLS for profitability prediction in GCC firms. *International Review of Economics & Finance, 69*, 750–764.
63. Nguyen, T., & Tran, M. (**2021**). Machine learning models for profitability prediction: Evidence from Vietnam. *Asia-Pacific Journal of Accounting & Economics, 28*(3), 367–389.
64. Silva, R., Santos, J., & Oliveira, P. (**2022**). Machine learning methods for profitability forecasting of Brazilian SMEs. *Journal of Applied Economics, 25*(1), 110–128.
65. Krauss, C., & Rösch, D. (**2023**). Hybrid econometric–machine learning models for financial performance prediction in European firms. *European Journal of Operational Research, 310*(2), 642–655.