

Article

Sequential Deep Learning for Predicting Shareholder Value Creation: Evidence from the Moroccan Stock Market

Youssef Jamil *, Imane El Yamlahi and Nabil Bouayad Amine

Department of Economics and Management, Faculty of Polydisciplinary Studies, Sultan Moulay Slimane University, Khouribga 25000, Morocco; imane.elyamlahi@usms.ac.ma (I.E.Y.); aminenabil.bouayad@usms.ac.ma (N.B.A.)

* Correspondence: youssef.jamil@usms.ac.ma

Abstract

This study investigates whether shareholder value creation, defined as beta-adjusted out-performance relative to a market benchmark, can be effectively predicted in an emerging market using a sequential machine learning framework. While prior research has predominantly focused on profitability forecasting or stock return prediction, the prediction of risk-adjusted shareholder value creation remains relatively underexplored, particularly in emerging economies such as Morocco. To address this gap, the study develops a predictive framework that combines market-based indicators, macroeconomic variables, and accounting fundamentals using only information realistically available to investors at each decision date. These variables are organized into firm-level temporal sequences based on a monthly decision-date panel of non-financial firms listed on the Casablanca Stock Exchange over the period 2010–2024. To capture nonlinear relationships and temporal dependencies in financial data, the empirical analysis compares baseline models with deep learning architectures, including GRU, LSTM, and CNN1D. The results indicate that deep learning models consistently outperform naïve and linear benchmark models, suggesting that shareholder value creation exhibits a measurable degree of predictability. With an AUC of 0.700 and a PR-AUC of 0.727, CNN1D achieves the strongest performance in the final evaluation setting and ranks as the best-performing model according to the primary AUC criterion. The findings also reveal that macroeconomic variables generate the strongest standalone predictive signal, whereas market-based variables exhibit comparatively weaker predictive power when considered in isolation. By extending financial prediction toward a risk-adjusted, benchmark-based, and investor-oriented framework, and by providing new empirical evidence on the value of temporal modeling and multi-source financial information for forecasting shareholder value creation in an emerging market context, this study contributes to the growing literature at the intersection of financial forecasting and artificial intelligence.



Academic Editor: Herbert Kimura

Received: 7 May 2026

Revised: 19 June 2026

Accepted: 21 June 2026

Published: 1 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: shareholder value creation; total shareholder return; deep learning; macroeconomic variables; emerging markets; time-series analysis

1. Introduction

A central objective in corporate finance and investment research is the creation of shareholder value, which reflects a firm's ability to generate returns that exceed market expectations (Jensen, 2001). In this context, Total Shareholder Return (TSR) has emerged as a widely used performance measure because it captures both capital gains and dividend

distributions over a given investment horizon, thereby providing a comprehensive investor-oriented assessment of corporate performance (Boudoukh et al., 2007).

Forecasting shareholder value creation remains a challenging and complex task (Cochrane, 2005; Rapach & Zhou, 2013). Financial markets evolve continuously and are influenced by a wide range of interconnected factors, including firm-specific fundamentals, market-based indicators, and macroeconomic conditions. In this context, operating performance, risk exposure, valuation metrics, and broader economic factors such as inflation, monetary policy, and economic growth are closely associated with firms' future performance (Fama & French, 1993; Campbell et al., 1997). Moreover, financial relationships often exhibit nonlinear dynamics and evolve over time, making it difficult for traditional linear models to adequately capture the complexity of these interactions (Gu et al., 2020).

Recent advances in machine learning have provided new opportunities to address these challenges more effectively. In particular, machine learning methods offer a flexible framework for modeling complex and nonlinear relationships among variables without relying on restrictive parametric assumptions (Hastie et al., 2009; Gu et al., 2020). Beyond capturing static associations, deep learning techniques are particularly well suited to modeling the temporal dynamics inherent in financial data (Fischer & Krauss, 2018). Recurrent architectures, such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU), are specifically designed to capture long-term dependencies in sequential data, whereas one-dimensional convolutional neural networks (CNN1D) are effective in identifying local temporal patterns and short-term dynamics (Hochreiter & Schmidhuber, 1997; Cho et al., 2014; LeCun et al., 2015). Even with these developments, deep learning is not yet widely used to predict shareholder value, especially in countries like Morocco. Rather than directly predicting shareholder value creation, most research focuses on stock return prediction or business profitability (Campbell & Thompson, 2008; Gu et al., 2020; Penman, 2013). Furthermore, many empirical studies rely on static representations of financial variables and often ignore sequential information and time dynamics when explaining company performance (Fischer & Krauss, 2018; Bao et al., 2017).

This paper proposes a predictive approach based on a sequential modeling framework that incorporates macroeconomic variables, market-based indicators, and firm-level accounting variables to deal with these limitations. By converting firm-level observations into rolling time sequences, the models can capture temporal dependencies and dynamic patterns in financial data. The analysis compares baseline models and deep learning architectures, such as CNN1D, GRU, and LSTM, applied to a panel of companies listed between 2010 and 2024 on the Casablanca Stock Exchange.

Instead of aggregating all predictors to one monthly frequency, the design includes lower frequency accounting and macroeconomic data into a monthly decision-data framework because variables are observed at different frequencies (Ghysels et al., 2007; Andreou et al., 2013).

Despite the growing literature, several important limitations remain in machine learning and financial prediction. Furthermore, many empirical methods overlook temporal dynamics and sequential information and instead rely on static representations of financial data (Bao et al., 2017; Fischer & Krauss, 2018). Lastly, although financial markets in emerging economies have specific structural features (Bekaert & Harvey, 2003), the use of deep learning models in financial prediction remains less explored in emerging market contexts, particularly when accounting, market, and macroeconomic data are combined (Sezer et al., 2020).

The Moroccan stock market provides a particularly relevant setting for examining shareholder value prediction because it combines the characteristics of an emerging market with a relatively limited number of listed firms and lower liquidity than major developed

markets. Emerging markets are often characterized by information asymmetries, liquidity constraints, and market frictions that may influence the speed at which new information is incorporated into security prices (Bekaert & Harvey, 2003). Consequently, accounting disclosures, market signals, and macroeconomic developments may affect investor expectations progressively rather than being immediately reflected in stock prices. Such conditions increase the potential importance of temporal dependencies and nonlinear interactions in financial data, making sequential machine learning approaches especially relevant (Fischer & Krauss, 2018; Sezer et al., 2020). Furthermore, while larger and more liquid markets have already been extensively examined in the financial machine learning literature, evidence from emerging markets remains comparatively limited. Studying the Moroccan market therefore provides an opportunity to evaluate whether the predictive advantages of sequential deep learning models remain relevant under different institutional, informational, and liquidity conditions.

This study uses a machine learning approach that incorporates firm-level accounting variables, market-based indicators, and macroeconomic factors into a sequential modeling structure to assess the predictability of shareholder value creation. The objective is to evaluate how effective deep learning models are in capturing nonlinear relationships and temporal dependencies that drive company performance.

This study makes three main contributions. First, it suggests using Total Shareholder Return in comparison to a beta-adjusted market benchmark as a risk-adjusted indicator of shareholder value creation. Second, it makes it possible to explicitly model temporal dynamics by introducing a sequential representation of financial data using rolling firm-specific time periods. Third, it offers fresh data on the predictive performance of deep learning models (GRU, LSTM, and CNN1D) in an emerging market setting by empirically comparing them with baseline classifiers.

Beyond the methodological comparison between deep learning architectures, the originality of this study lies in the prediction of shareholder value creation through a risk-adjusted and investor-oriented framework specifically designed for an emerging market environment. Unlike most previous studies that focus on stock return forecasting or accounting profitability prediction, this research examines whether firms can outperform a beta-adjusted market benchmark using information that would realistically be available to investors at each decision date.

The study is also positioned at the intersection of asset pricing, financial forecasting, and artificial intelligence. By combining accounting fundamentals, market-based indicators, and macroeconomic information within a unified sequential framework, it extends existing research beyond conventional return prediction models and contributes to a more comprehensive understanding of the determinants of shareholder value creation (Kelly et al., 2019; Gu et al., 2020; Jiang et al., 2023).

Furthermore, as in many emerging markets, information may not be incorporated into prices instantaneously. Accounting disclosures, market signals, and macroeconomic developments may influence investor expectations progressively over time rather than through immediate market adjustments. In such an environment, temporal dependencies become economically meaningful, making sequential architectures theoretically relevant for capturing evolving information patterns and dynamic interactions among financial variables (Bekaert & Harvey, 2003; Sezer et al., 2020; Fischer & Krauss, 2018). Consequently, this study contributes not only to the growing literature on artificial intelligence in finance but also to the understanding of how temporal deep learning models can be applied to predict risk-adjusted shareholder value creation in relatively underexplored emerging market settings.

Inspired by the body of the current literature, the following research question is addressed in this study:

Research Question: Can shareholder value creation be predicted using accounting, market-based, and macroeconomic information within a sequential machine learning framework?

To address this question, the study further examines:

- (i) The complementary contribution of accounting, market-based, and macroeconomic information.
- (ii) The role of temporal dependence captured through rolling firm-specific sequences.
- (iii) The comparative performance of sequential deep learning and baseline classification models.

General Hypothesis:

Shareholder value creation is not purely random and can be predicted, to some extent, using firm-level fundamentals, market-based information, and macroeconomic conditions within a temporally consistent machine learning framework.

1st hypothesis: Firm-level accounting variables, market-based indicators, and macroeconomic factors provide predictive information for identifying shareholder value creation.

H1.1. *Accounting variables related to operating performance, profitability, leverage, and firm size provide predictive information about the probability of shareholder value creation.*

H1.2. *Market-based variables capturing valuation, risk exposure, and price dynamics provide additional predictive information about shareholder value creation beyond accounting fundamentals.*

H1.3. *Macroeconomic indicators related to inflation, monetary policy, liquidity, exchange rates, and economic growth influence the probability of shareholder value creation through their impact on discount rates and overall market conditions.*

2nd hypothesis: Sequential machine learning models capture nonlinear and temporal patterns that improve the prediction of shareholder value creation relative to naïve and linear benchmark classifiers.

This second hypothesis is broken down into the following sub-hypotheses.

H2.1. *Rolling firm-specific sequences provide useful temporal information for predicting shareholder value creation.*

H2.2. *Nonlinear architectures capture interactions between financial and macroeconomic variables that are not adequately represented by linear decision boundaries.*

H2.3. *Sequential deep learning models outperform naïve and linear benchmark classifiers, in terms of out-of-sample predictive performance.*

Paper Structure

The remainder of the paper is organized as follows. Section 2 presents the literature review and theoretical foundations. Section 3 describes the methodological approach, including data preparation and the research workflow. Section 4 introduces the evaluation metrics used to assess model performance. Section 5 presents the data and variable construction, while Section 6 details the empirical design and model specification. Section 7 reports the empirical results. Section 8 discusses the findings, and Section 9 concludes by presenting the main limitations and directions for future research.

2. Literature Review and Theoretical Foundations

This section reviews the theoretical and empirical literature relevant to the prediction of shareholder value creation, drawing on insights from financial economics, asset pricing, and machine learning. It first examines the economic mechanisms underlying shareholder value creation and the role of accounting, market-based, and macroeconomic determinants. It then discusses the theoretical foundations of the machine learning approaches employed in this study and their relevance for modeling complex financial relationships and temporal dynamics.

2.1. Economic and Financial Foundations of Shareholder Value Creation

Financial economics has long regarded shareholder value creation as a central objective of corporate performance, commonly assessed through stock returns and value-based performance measures. Traditional frameworks, including market efficiency theories and value-based management approaches, posit that a firm's ability to generate returns exceeding its cost of capital is ultimately reflected in shareholder value creation (Rappaport, 1986; Stewart, 1991; Fama & French, 1993).

In this context, Total Shareholder Return (TSR), which incorporates both capital gains and dividend distributions over a given investment horizon, is widely recognized as a comprehensive measure of shareholder value creation (Boudoukh et al., 2007). However, traditional performance assessments often rely on raw returns and may fail to adequately account for differences in systematic risk across firms. To address this limitation, recent research has emphasized the importance of risk-adjusted performance measures, particularly those based on beta-adjusted returns derived from the Capital Asset Pricing Model (CAPM), which provide a more meaningful evaluation of value creation from an investor's perspective (Sharpe, 1964; Lintner, 1965; Jensen, 1968).

According to this framework, a firm's ability to outperform an appropriate benchmark after accounting for its level of systematic risk is generally regarded as an indicator of shareholder value creation. This perspective is particularly relevant in emerging markets, where structural characteristics, market imperfections, and macroeconomic volatility can significantly influence asset-pricing dynamics (Bekaert & Harvey, 2003).

Building on this risk-adjusted perspective, shareholder value creation can be formulated as a predictive problem by defining a binary target variable that indicates whether a firm outperforms its benchmark. This formulation naturally lends itself to a classification framework, which is widely used in machine learning to predict discrete outcomes and support decision-making under uncertainty (Fawcett, 2006; Provost & Fawcett, 2013).

2.2. Role of Accounting, Market, and Macroeconomic Variables

Predicting shareholder value creation requires the integration of information from multiple sources that capture macroeconomic conditions, market dynamics, and firm-specific characteristics. Accordingly, the explanatory variables considered in this study are organized into three broad categories: accounting variables, market-based indicators, and macroeconomic variables. Each category provides complementary information regarding the determinants of corporate performance and shareholder value creation.

2.2.1. Accounting Variables

Accounting variables provide valuable insights into a firm's operating performance, profitability, and financial structure. In this study, the selected accounting indicators include changes in sales (Δ Sales), changes in EBITDA (Δ EBITDA), earnings per share (EPS), changes in earnings per share (Δ EPS), leverage, and firm size (SIZE). These variables capture key

dimensions of corporate performance and are frequently used in the financial forecasting literature to assess a firm's growth prospects, profitability, and financial stability.

These variables capture essential dimensions of firm fundamentals. In particular, EPS and its variation reflect shareholder-oriented earnings performance, while growth in sales and EBITDA provides information on operating expansion and profitability dynamics (Penman, 2013). Firm size is commonly associated with scale effects, information availability, and market positioning (Fama & French, 1993), whereas leverage reflects financing decisions and the level of financial risk borne by the firm (Bhandari, 1988). A substantial body of empirical research has highlighted the importance of accounting information in explaining expected returns, firm performance, and value creation (Sloan, 1996; Fama & French, 2006; Penman, 2013). Consequently, these variables constitute a relevant foundation for predicting shareholder value creation.

2.2.2. Market, Valuation, and Risk Variables

Market-based variables capture investor expectations, valuation dynamics, and perceived risk. The variables used in this study include the price-to-earnings ratio (PER), market-to-book ratio (M/B), stock return volatility (σ_t), firm beta ($\beta_{i,t}$ (12m)), close price, and the MASI monthly closing index.

Two widely used valuation ratios help explain how the market values firms relative to earnings and book value are PER and M/B (Campbell & Shiller, 1988; Fama & French, 1993). While beta quantifies systematic risk in relation to the market portfolio (Sharpe, 1964), volatility captures the fluctuation of stock returns and indicates uncertainty (Ang et al., 2006). While the MASI index represents the general market trend and acts as a stand-in for aggregate market conditions, the inclusion of the close price enables the capture of firm-level price changes and momentum-related phenomena (Jagadeesh & Titman, 1993).

These factors are based on frameworks for market efficiency and asset pricing theory, which highlight how market data reflects and predicts corporate performance (Fama, 1970; Grossman & Stiglitz, 1980).

2.2.3. Macroeconomic Variables

Macroeconomic variables reflect the broader economic environment in which firms operate and play a fundamental role in shaping corporate performance, investment decisions, and asset valuation. The macroeconomic indicators considered in this study include inflation, the policy interest rate, the yield curve slope (SLOPE), money supply growth ($\Delta M2$), exchange rate variation (ΔFX), and economic growth (ΔGDP).

These indicators influence discount rates, financing conditions, market liquidity, and overall economic activity. For instance, inflation and interest rates directly affect the cost of capital, whereas changes in money supply and exchange rates can influence financial market conditions and investment decisions. The yield curve slope is particularly informative, as it reflects market expectations regarding future economic activity and monetary policy developments (Estrella & Hardouvelis, 1991). These macroeconomic variables reflect broader economic conditions that influence both firm-level cash flows and discount rates and play an important role in asset pricing and financial performance (Cochrane, 2005). Macroeconomic factors have a strong impact on stock returns and firm performance, according to empirical data, especially in emerging economies that tend to be more sensitive to economic shocks (Fama & Schwert, 1977; N.-F. Chen et al., 1986; Bekaert & Harvey, 2003; Rapach et al., 2013; Neely et al., 2014).

2.2.4. Complementarity and Interaction Effects

Accounting fundamentals, market-based indicators, and macroeconomic variables should be viewed as complementary sources of information rather than as independent

determinants of shareholder value creation. These dimensions interact in complex ways, jointly influencing firms' performance, market expectations, and valuation outcomes (Rapach et al., 2013; Kelly et al., 2019). Because such relationships are often nonlinear and evolve over time, they are difficult to capture using conventional linear modeling approaches (Hastie et al., 2009; Gu et al., 2020).

This perspective is particularly relevant in emerging markets, where financial performance is shaped by the interaction between firm-specific characteristics and broader economic conditions (Bekaert & Harvey, 2003). Integrating multiple sources of information within a unified modeling framework enables a more comprehensive assessment of the factors driving shareholder value creation and enhances the ability to predict its evolution over time.

2.3. Mixed-Frequency Information and Information Timing

A common challenge in financial forecasting arises from the heterogeneous frequencies at which different types of information become available. Macroeconomic indicators are generally observed at monthly or lower frequencies, whereas accounting information is typically disclosed on an annual or semiannual basis. In contrast, market-based variables, such as stock prices, returns, volatility, and beta, are available at much higher frequencies and can be aggregated to monthly intervals when required.

The main methodological difficulty in this situation is to make sure that each predictor is included in a way that is consistent with its actual informational availability rather than arbitrarily imposing a uniform frequency across all variables. As long as temporal alignment respects the information set accessible at each decision point, lower-frequency variables can be incorporated into higher-frequency prediction frameworks, according to the mixed-frequency econometrics literature (Fama, 1970; Ghysels et al., 2004; Andreou et al., 2010; Schorfheide & Song, 2015).

In financial markets, the informational content of accounting variables is closely linked to the timing of their disclosure. Investors typically revise their expectations when new financial statements are released rather than continuously updating their assessments based on accounting information. Nevertheless, between reporting dates, investment decisions continue to rely on the most recently available financial information. From this perspective, accounting variables should not be viewed as continuously updated monthly series but rather as lower-frequency firm fundamentals whose informational content remains relevant until new disclosures become available (Ball & Brown, 1968; Beaver, 1968; Chambers & Penman, 1984).

Similarly, in a higher-frequency predictive environment, macroeconomic variables like GDP, which are not naturally seen at a monthly frequency, can be included as low-frequency signals (Marcellino et al., 2006). In this instance, their function is to record more general economic conditions that change more slowly over time rather than to accurately depict intra-period dynamics.

Building on these considerations, this study adopts a mixed-frequency framework that integrates macroeconomic, market-based, and accounting information within a unified monthly decision-date structure. This design enables the models to capture both short-term market dynamics and the more gradual effects of firm fundamentals and macroeconomic conditions. By aligning variables according to their effective availability while preserving their original informational content, the framework maintains temporal consistency and enhances the economic interpretability of the predictive process.

2.4. Machine Learning and Baseline Approaches in Financial Prediction

Machine learning approaches can model complex interactions in high-dimensional data without relying on strict parametric assumptions and have become increasingly important in finance research (Hastie et al., 2009; Varian, 2014; Gu et al., 2020). In particular, classification models are often used to predict financial outcomes by estimating the likelihood of specific events, such as shareholder value creation (Fawcett, 2006).

Recent empirical evidence suggests that machine learning techniques can substantially improve predictive performance in financial applications, including asset pricing, stock return forecasting, and corporate performance prediction, by capturing complex nonlinear relationships and interactions among financial variables (Feng et al., 2020; Kelly et al., 2019; Gu et al., 2020). Logistic regression is used in this study as a standard machine learning model. Logistic regression is a probabilistic linear classifier that uses a sigmoid function to estimate the probability of a binary outcome (Hosmer et al., 2013). Because of its interpretability and solid statistical underpinnings, it is still commonly employed in financial research despite its simplicity, making it a useful benchmark for performance comparison.

In addition to logistic regression, a majority-class classifier is employed as a naïve benchmark. This model simply predicts the most frequently observed class in the training sample and therefore does not involve any learning process. Such baseline models are commonly used in classification tasks to establish a lower-bound performance benchmark and to verify that more sophisticated approaches generate predictive improvements beyond naïve decision rules (Provost & Fawcett, 2013).

While these baseline methods provide useful reference points, their predictive performance may be constrained in complex financial environments characterized by nonlinear relationships and temporal dependencies. This limitation stems from their reliance on restrictive assumptions, including linearity and the independence of observations, which may not adequately reflect the dynamic nature of financial markets.

2.5. Sequential Modeling and Deep Learning Architectures

Recent research has increasingly focused on sequential modeling approaches that incorporate time dependencies to address the limitations of static models (Goodfellow et al., 2016). Long short-term memory (LSTM) and Gated Recurrent Units (GRU) are two examples of recurrent neural networks (RNNs) that are designed to capture long-term relationships in sequential data (Hochreiter & Schmidhuber, 1997; Cho et al., 2014; Goodfellow et al., 2016; Fischer & Krauss, 2018).

Convolutional neural networks (CNNs), which were first created for image processing, have also been adapted for time-series analysis. From sequential data, one-dimensional CNNs (CNN1D) can extract hierarchical features and local temporal patterns (Kiranyaz et al., 2015; LeCun et al., 2015; Zhang et al., 2017; Bai et al., 2018).

These architectures are capable of learning hierarchical representations of financial data, enabling them to capture complex nonlinear relationships and temporal dynamics. Their strengths are complementary: convolutional networks are particularly effective at detecting local temporal patterns, whereas recurrent architectures are specifically designed to model long-term dependencies in sequential data. Examining the performance of these models therefore provides valuable insights into the underlying data-generating process and the relative importance of different temporal structures. Their application to financial forecasting has yielded promising results, particularly in environments characterized by noisy, high-dimensional, and complex datasets (Nelson et al., 2017; Heaton et al., 2017; Fischer & Krauss, 2018; Sezer et al., 2020).

2.6. Theoretical Foundations of Model Selection

The selection of predictive models for financial applications should account for several important characteristics of financial data, including temporal dependencies, nonlinear relationships, and interactions among variables (Campbell et al., 1997; Cochrane, 2005). Conventional approaches, such as logistic regression, may be less effective in complex financial environments because they rely on restrictive assumptions, including linear relationships and the independence of observations (Hastie et al., 2009; Gu et al., 2020).

In contrast, deep learning models are better suited to modeling complex nonlinear relationships and dynamic patterns because of their flexibility in approximating highly intricate functions (Goodfellow et al., 2016; Heaton et al., 2017). Their capacity to learn evolving financial structures is further strengthened by the use of rolling windows and firm-specific temporal sequences, which enable the models to exploit both cross-sectional and temporal information embedded in the data (Bao et al., 2017; Fischer & Krauss, 2018).

This approach provides a way to evaluate the added value of nonlinear and sequential model techniques by comparing baseline models with deep learning architectures (Varian, 2014). By comparing different model designs within a single empirical framework, our study builds on this perspective and adds to our understanding of model selection in financial prediction tasks.

3. Methodology

This section outlines the empirical framework adopted in the study. It describes the preparation and temporal alignment of the data, the specification of the predictive models, and the procedures used to evaluate their performance. Particular attention is devoted to the monthly decision-date structure, the prevention of information leakage, and the implementation of a temporally consistent forecasting framework.

3.1. Data Preparation and Preliminary Diagnostics

This study is based on a firm-level panel structured around monthly decision dates for companies listed on the Casablanca Stock Exchange over the period 2010–2024, combining variables observed at different frequencies.

As illustrated in Figure 1, financial institutions, leasing companies, investment firms, and holding companies were excluded because of their distinct regulatory environments, accounting frameworks, and reporting practices. Additional selection criteria related to data availability, first-year stability, listing continuity, and fiscal-year comparability were subsequently applied. The final sample comprises 30 non-financial firms, representing 5400 firm-month observations.

After preprocessing and aligning the data over time, the sample is reduced to 5040 usable observations. This reduction mainly comes from the construction of the 12-month forward TSR used for target definition and by the rolling sequence structure ($L = 12$), which mechanically truncates observations at the beginning and end of each firm's time series.

The relatively small cross-sectional dimension of the sample should be explicitly acknowledged. Although the panel contains 5040 usable firm-month observations, these observations are generated from only 30 non-financial listed firms. This feature reflects the structure of the Moroccan stock market and the strict sample selection criteria adopted in this study, including sectoral exclusion, listing continuity, data availability, and fiscal-year comparability. Consequently, the empirical design relies more on the temporal dimension of the panel than on a large cross-sectional universe of firms.

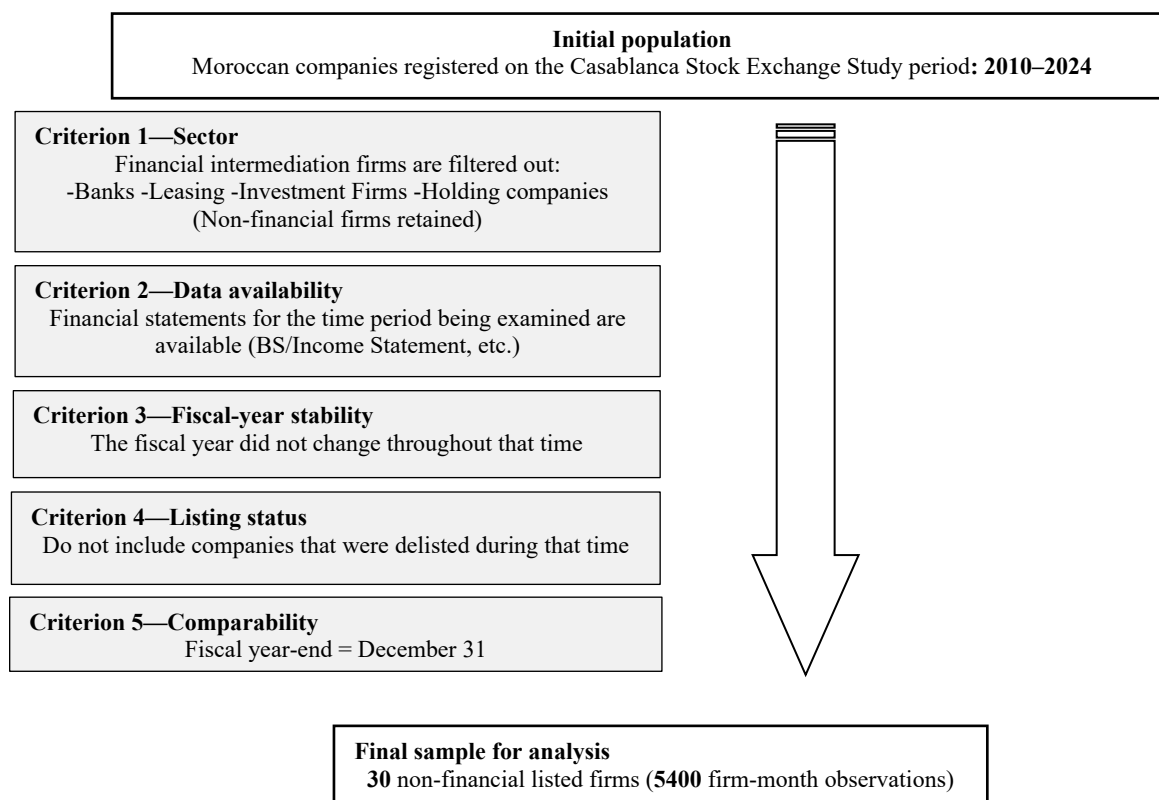


Figure 1. The process of selecting analytical samples for non-financial companies registered on the Casablanca stock exchange. Source: Author’s own elaboration.

This limitation is particularly relevant for deep learning models, which are generally data-intensive and may be sensitive to overfitting when applied to small samples. To mitigate this risk, the architectures used in this study were intentionally kept relatively parsimonious, and several regularization and validation procedures were implemented, including dropout, chronological train–validation–test splitting, early stopping based on validation AUC, restoration of the best validation weights, and block-bootstrap confidence intervals. Therefore, the results should be interpreted as evidence of predictive patterns within the Moroccan listed-firm universe rather than as unrestricted general evidence applicable to all emerging or developed markets.

The dataset is built inside a rigorous decision date structure to reduce any look-ahead bias resulting from data delivery delays. Only data that would have been accessible to the public at or before each monthly decision point (t) is utilized for forecasting.

To preserve the temporal integrity of the dataset, accounting variables are aligned according to their effective public disclosure dates, thereby incorporating the reporting lag between the end of a fiscal period and the publication of financial statements. Similarly, macroeconomic variables are synchronized with their respective release dates to ensure that only information available at each decision date is included in the predictive framework. This procedure prevents the inadvertent introduction of forward-looking information and reduces the risk of information leakage.

This temporal alignment helps mitigate the risk of information leakage by preserving the chronological integrity of the dataset and providing a realistic approximation of information availability at each decision date, even though exact release dates are not explicitly modeled. More specifically, accounting variables are incorporated into the feature set with an explicit reporting lag of approximately six months following the fiscal year-end, reflecting the typical delay between the preparation of financial statements and

their public disclosure. Likewise, macroeconomic variables are aligned according to their official publication schedules, which generally imply a delay of approximately one month relative to the observation period. As a result, the predictive framework relies exclusively on information that would have been available to investors at the time the prediction was generated. Furthermore, no future market information is included in the target design since the rolling beta ($\beta_{i,t}$) utilized for risk adjustment is only calculated using historical return data up to time t . In addition to removing any potential look-ahead bias in feature development and target definition, this explicit lag structure strengthens the dataset's temporal consistency.

In addition to temporal alignment, several preprocessing procedures were implemented to enhance data quality and ensure the robustness of the empirical framework. Extreme observations, which are common in financial datasets, were mitigated through winsorization at the 1st and 99th percentiles. Missing values were handled using a firm-level forward-fill procedure when necessary, ensuring that only information available prior to each decision date was propagated through time.

Furthermore, all input variables were standardized using z-score normalization based exclusively on the training sample. The estimated transformation parameters were subsequently applied to the validation and test sets, thereby preserving consistency across datasets while preventing information leakage. Together, these preprocessing procedures improve model stability, enhance the comparability of variables measured on different scales, and maintain the temporal integrity of the forecasting framework.

3.2. Research Workflow

The research workflow follows a structured sequence designed to ensure the rigorous development and evaluation of the predictive models. The process begins with the collection of accounting, market-based, and macroeconomic data covering the period 2010–2024. Market variables and most macroeconomic indicators are observed at a monthly frequency, whereas accounting variables and GDP are available at lower frequencies. Rather than artificially transforming all variables into monthly series, the proposed framework aligns them to monthly decision dates in accordance with their effective availability. This approach preserves the economic meaning of each variable while ensuring that the information set remains consistent with what would have been available to investors prior to preprocessing and sequence construction. To further reduce the risk of information leakage, all predictors were aligned according to their effective availability date. Market variables were computed exclusively from historical observations available up to the decision date, while accounting variables were incorporated with an appropriate reporting lag reflecting the delay between fiscal year-end and public disclosure. Missing observations were handled using a firm-level chronological forward-fill procedure. This approach is consistent with the low-frequency nature of accounting and certain macroeconomic variables, as the most recently disclosed information remains available to investors until a subsequent update is released. Importantly, only past observations were propagated forward, ensuring that no future information entered the predictive modeling process. Next, firm-level fundamentals, market indicators and macroeconomic signals were organized within a monthly decision framework and combined with lagged variables to capture temporal dynamics. Next, using company TSR relative to the market benchmark, the target variable was defined based on shareholder value creation. Panel data were converted into temporal sequences using a 12-month rolling window to enable the use of sequential deep learning models. The dataset was then split chronologically into training, validation, and test sets. The modeling strategy was carried out in three stages: stage 1 consists of an initial comparison between baseline and deep learning models under a preliminary specification; stage 2

focuses on comparing deep learning architectures (GRU, LSTM, and CNN1D) under a stricter classification setting; and Stage 3 involves final validation including comparison with baseline models, along with diagnostic analyses, and robustness tests to assess the reliability of the selected model. The overall workflow is summarized in Figure 2:

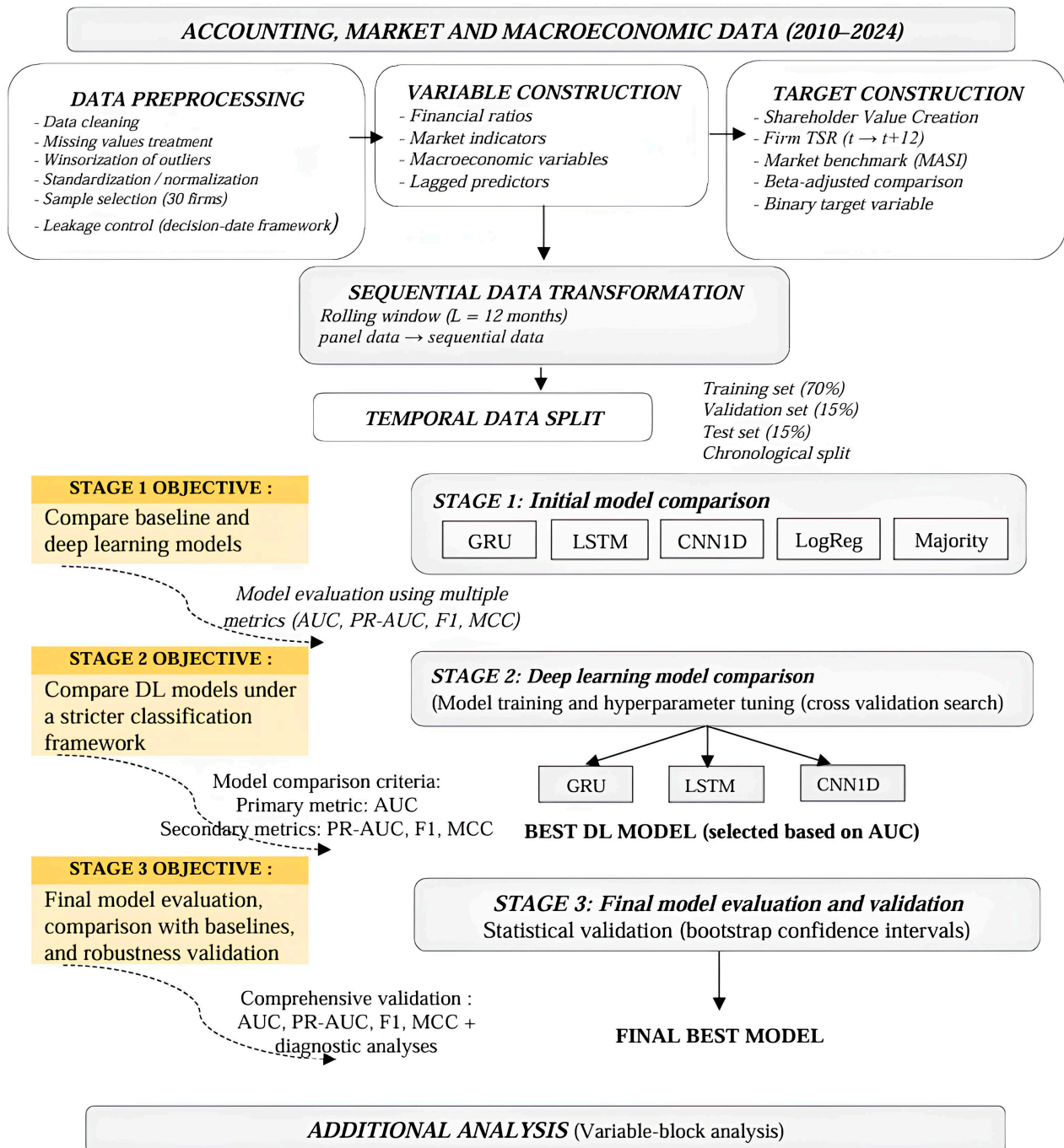


Figure 2. Methodological workflow of the study. Source: Author’s elaboration based on accounting, market and macroeconomic data (2010–2024).

4. Evaluation Metrics

We use a set of complementary criteria frequently employed in binary classification issues to assess the prediction performance of the suggested models (Fawcett, 2006; Saito & Rehmsmeier, 2015; Chicco & Jurman, 2020; Provost & Fawcett, 2013). Let $Y \in \{0, 1\}$ denote

the true class label and $\hat{Y} \in \{0, 1\}$ the predicted class. In addition, let $\hat{P}(X) = P(Y = 1|X)$ denote the predicted probability that an observation belongs to the positive class.

Unlike traditional studies that model shareholder value creation as a continuous outcome, this study adopts a binary classification framework. The target variable is constructed from the beta-adjusted Total Shareholder Return (TSR) relative to the market benchmark. A positive class ($Y = 1$) indicates that a firm generated positive shareholder value creation by outperforming its risk-adjusted benchmark over the evaluation horizon, whereas a negative class ($Y = 0$) indicates that the firm failed to outperform the benchmark.

This formulation is motivated by the practical decision-making context faced by investors. In many investment applications, the primary objective is not necessarily to predict the exact magnitude of future returns, but rather to identify firms that are more likely to create shareholder value. Consequently, the classification framework transforms shareholder value prediction into a decision-oriented problem that is directly aligned with portfolio screening and investment selection processes.

Based on the confusion matrix, we define the following quantities:

$$TP = \sum_{i=1}^n 1(Y_i = 1, \hat{Y}_i = 1) \quad (1)$$

$$TN = \sum_{i=1}^n 1(Y_i = 0, \hat{Y}_i = 0)$$

$$FP = \sum_{i=1}^n 1(Y_i = 0, \hat{Y}_i = 1)$$

$$FN = \sum_{i=1}^n 1(Y_i = 1, \hat{Y}_i = 0)$$

where $1(\cdot)$ is the indicator function.

Accuracy measures the probability of correct classification:

$$\text{Accuracy} = P(\hat{Y} = Y) = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

4.1. Balanced Accuracy

By averaging class-wise conditional probability, Balanced Accuracy corrects for class imbalance:

$$\begin{aligned} \text{Balanced Accuracy} &= \frac{1}{2} [P(\hat{Y} = 1 | Y = 1) + P(\hat{Y} = 0 | Y = 0)] \\ &= \frac{1}{2} \left[\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right] \end{aligned} \quad (3)$$

4.2. Precision

The percentage of accurately anticipated positive observations among all predicted positives is known as precision.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

4.3. Recall

Recall (sensitivity) measures the probability of correctly detecting positive cases:

$$\text{Recall} = P(\hat{Y} = 1 | Y = 1) = \frac{TP}{TP + FN} \quad (5)$$

4.4. F1-Score

The harmonic mean of precision and recall yields the F1-score:

$$F1 = 2 \cdot \frac{P(Y = 1 | \hat{Y} = 1) \cdot P(\hat{Y} = 1 | Y = 1)}{P(Y = 1 | \hat{Y} = 1) + P(\hat{Y} = 1 | Y = 1)} \quad (6)$$

4.5. Matthews Correlation Coefficient (MCC)

A correlation coefficient between observed and anticipated classes is called the MCC.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (7)$$

4.6. Area Under the ROC Curve (AUC)

The probability that a randomly selected positive observation would obtain a higher projected score than a randomly selected negative observation is assessed by the AUC:

$$AUC = P(\hat{p}(X^+) > \hat{p}(X^-)) \quad (8)$$

where X^+ and X^- denote positive and negative observations, respectively.

4.7. Precision–Recall Area Under Curve (PR-AUC)

The relationship between precision and recall over categorization thresholds is summed up by the PR-AUC:

$$PR-AUC = \int_0^1 \text{Precision}(r) dr \quad (9)$$

where r = Recall.

When these measures are used together, a thorough and reliable assessment of categorization performance is produced. While Precision, Recall, and F1-score evaluate the quality of positive class recognition, Accuracy and Balanced Accuracy reflect total predictive correctness. Furthermore, MCC offers a fair and trustworthy indicator of classification performance, especially when there is class imbalance, and AUC and PR-AUC offer threshold-independent measures of discrimination.

The Area Under the ROC Curve (AUC), which provides a threshold-independent measure of overall discriminative ability, is used as the main criterion for model selection in this study. A more thorough evaluation of sensitivity to class imbalance, error trade-offs, and the quality of positive class predictions is supported by additional metrics used as supplementary diagnostic indicators.

5. Data and Variable Construction

Figure 1 shows the sequential filtering process used to define the final analytical sample of non-financial firms listed on the Casablanca Stock Exchange over the 2010–2024 period. While the initial sample includes 5400 firm-month observations, the number of usable observations is reduced to 5040 due to the construction of forward TSR and the use of rolling sequences ($L = 12$), which remove observations at the beginning and end of the sample.

5.1. Variable Definition and Construction

5.1.1. Explanatory Variables

This study employs a comprehensive set of explanatory variables capturing firm fundamentals, market information, risk characteristics, and macroeconomic conditions. The

selection of these variables is grounded in the asset-pricing, corporate finance, and shareholder value creation literature, which highlights the interconnected roles of accounting performance, market expectations, and macroeconomic factors in shaping firms' ability to generate shareholder value.

The explanatory variables are divided into three groups to make understanding easier: macroeconomic variables, market/valuation/risk variables, and accounting variables. All of the variables, together with their construction formulas and brief definitions, are summarized in Table 1.

Table 1. Definition and construction of explanatory variables.

Variable	Formula	Definition
Accounting variables		
ΔSales	$(\text{Sales}_{(t)} - \text{Sales}_{(t-1)}) / \text{Sales}_{(t-1)}$	Sales growth rate
ΔEBITDA	$(\text{EBITDA}_{(t)} - \text{EBITDA}_{(t-1)}) / \text{EBITDA}_{(t-1)}$	Operating profitability growth.
EPS	$\text{Net Income}_{(t)} / \text{Shares Outstanding}_{(t)}$	Earnings per share.
ΔEPS	$(\text{EPS}_{(t)} - \text{EPS}_{(t-1)}) / \text{EPS}_{(t-1)}$	Earnings per share growth.
Leverage	$\text{Total Debt}_{(t)} / \text{Total Assets}_{(t)}$	Financial leverage ratio.
Size	$\text{Ln}(\text{Total Assets}_{(t)})$	Firm size measure.
Market/Valuation/Risk Variables		
PER	$\text{Price}_{(t)} / \text{EPS}_{(t)}$	Price-to-earnings ratio.
M/B	$\text{Market Value}_{(t)} / \text{Book Value}_{(t)}$	Market-to-book ratio.
Volatility σ_t	$\text{Std}(\text{R}_{i,t-11:t})$	Stock return volatility.
$\beta_{i,t} (12m)$	$\text{Cov}(\text{R}_{i,t-11:t}, \text{R}_{m,t-11:t}) / \text{Var}(\text{R}_{m,t-11:t})$	Systematic market risk.
Close Price	$P_{i,t}$	Monthly stock closing price.
MA SI monthly close	$I_{(t)}^{\text{MA SI}}$	Market index level.
Macroeconomic Variables		
Inflation (t)	$\Delta \text{CPI}_{(t)}$	Monthly inflation rate.
Policy Rate	$r_{(t)}$	Central bank policy rate.
SLOPE (t)	$\text{Long Rate}_{(t)} - \text{Short Rate}_{(t)}$	Yield curve slope.
ΔM2	$(\text{M2}_{(t)} - \text{M2}_{(t-1)}) / \text{M2}_{(t-1)}$	Money supply growth.
$\Delta \text{FX}_{(t)}$	$(\text{FX}_{(t)} - \text{FX}_{(t-1)}) / \text{FX}_{(t-1)}$	Exchange rate variation.
ΔGDP	$(\text{GDP}_{(t)} - \text{GDP}_{(t-1)}) / \text{GDP}_{(t-1)}$	Economic growth rate.

Source: Individual creation by the author.

Market variables are observed monthly. Most macroeconomic variables are monthly indicators. GDP is incorporated as a lower-frequency macroeconomic signal embedded in the monthly decision framework. Accounting variables are not treated as naturally monthly observations; instead, they are treated as lower-frequency firm fundamentals aligned to monthly decision dates based on the latest publicly available information.

5.1.2. Shareholder Value Creation Measures

Using a beta-adjusted benchmarking approach, firm performance is evaluated relative to a risk-adjusted market return. A binary classification target is then defined to indicate whether a firm creates shareholder value relative to its benchmark.

Shareholder value creation is measured using Total Shareholder Return (TSR), capturing both capital gains and dividend income over a given horizon H . For firm I , TSR is defined as:

$$TSR_{(i,t \rightarrow t+H)} = (P_{(i,t+H)} - P_{(i,t)} + D_{(i,t \rightarrow t+H)}) / P_{(i,t)} \quad (10)$$

A benchmark return is calculated as follows to account for market-wide fluctuations:

$$TSR_{(m,t \rightarrow t+H)} = (I_{(t+H)} - I_t) / I_t \quad (11)$$

Shareholder value creation adjusted for systematic risk is defined as:

$$TSR_{i,t \rightarrow t+H}^{\beta} = TSR_{i,t \rightarrow t+H} - \beta_{i,t} \cdot TSR_{m,t \rightarrow t+H} \quad (12)$$

A binary target variable is then defined:

$$y_{(i,t)} = 1 \left(TSR_{(i,t \rightarrow t+H)} > \beta_{(i,t)} \cdot TSR_{(m,t \rightarrow t+H)} \right) \quad (13)$$

The use of a beta-adjusted TSR benchmark is motivated by the objective of evaluating shareholder value creation relative to the systematic market risk borne by investors. Rather than measuring raw stock performance, the proposed target variable captures whether a firm generates returns exceeding those expected given its exposure to market risk. The CAPM-based adjustment was selected because it provides a parsimonious, transparent, and widely used benchmark in the asset-pricing literature (Sharpe, 1964; Lintner, 1965). In addition, the implementation of multifactor models such as the Fama–French three-factor model (Fama & French, 1993) or the Carhart four-factor model (Carhart, 1997) requires the construction of reliable size, value, and momentum factor portfolios over long periods. Given the relatively small number of listed non-financial firms available in the Moroccan stock market, the estimation of such factors may be affected by substantial sampling noise and limited portfolio diversification. Consequently, the CAPM-adjusted benchmark was retained as the main specification. Nevertheless, future research may extend the analysis by considering multifactor abnormal return measures derived from Fama–French or Carhart frameworks when sufficiently stable factor portfolios become available.

From a practical perspective, even a modest positive beta-adjusted return may be informative because it indicates that a firm has outperformed the return expected given its systematic risk exposure. Consequently, the proposed classification rule is intended to provide a risk-adjusted screening signal for identifying firms more likely to create shareholder value rather than to measure the exact magnitude of abnormal performance. Consistent with decision-oriented approaches in finance, this binary formulation casts shareholder value forecasting as a classification problem. From an investment perspective, the primary objective is often not to predict exact return magnitudes but rather to distinguish between firms that outperform their benchmark and those that do not. This formulation also facilitates the application of deep learning and classification-based machine learning techniques, which are particularly effective in modeling complex interactions, nonlinear relationships, and heterogeneous patterns in financial data.

This formulation also reduces sensitivity to extreme values and noise in financial returns, which are often highly volatile.

To model this classification problem, the input feature space and its temporal structure need to be defined.

6. Empirical Design and Model Specification

6.1. Data Processing Environment

The empirical analysis was carried out using Python (version 3.13.2) mainly because of its flexibility and the wide range of available libraries for data processing, statistical analysis, and machine learning tasks, and deep learning. Python was used throughout the research workflow, including data preprocessing, variable construction, sequence preparation, model training, and evaluation. This setup also allowed the implementation and comparison of deep learning models (GRU, LSTM, CNN1D) together with benchmark classifiers, ensuring efficient analysis and making the results reproducible.

6.2. Feature Representation and Sequential Structure

$$x_{i,t} = [\Delta Sales_{i,t}, \Delta EBITDA_{i,t}, PER_{i,t}, \sigma_{i,t}, LEV_{i,t}, EPS_{i,t}, \Delta EPS_{i,t}, Size_{i,t}, MB_{i,t}, \beta_{i,t}, z_t]^\top$$

with:

$$z_t = [Infl_t, PolicyRate_t, Slope_t, \Delta M2_t, \Delta FX_t, \Delta GDP_t]^\top \quad (14)$$

To explicitly account for temporal dependencies and dynamic patterns, the feature vectors are organized into rolling firm-specific sequences of fixed length L . This representation allows the models to exploit historical information and capture the temporal structure inherent in financial data.

$$X_{(i,t)}^{(L)} = [x_{(i,t-L+1)}, \dots, x_{(i,t)}] \in \mathbb{R}^{(L \times p)} \quad (15)$$

6.3. Baseline Models

To provide reference points for evaluating predictive performance, two baseline models are considered.

The majority classifier assigns all observations to the most frequent class in the training sample:

$$\hat{y}_{i,t}^{(MC)} = \arg \max_{c \in \{0,1\}} \Pr(Y = c) \quad (16)$$

This naïve model provides a lower bound for predictive performance.

The logistic regression model estimates the probability of shareholder value creation as:

$$p_{i,t} = \sigma(\alpha + w^\top x_{i,t}) \quad (17)$$

where $\sigma(u) = \frac{1}{1+e^{-u}}$.

6.4. Sequential Deep Learning Models

To capture temporal dynamics and non-linear interactions, deep learning models are applied to the sequential input $X_{i,t}^{(L)}$.

Recurrent models (GRU and LSTM)

Recurrent architectures transform the input sequence into a latent representation:

$$h_{i,t} = f_\theta^{(RNN)}(X_{i,t}^{(L)}) \quad (18)$$

where $f_\theta(\cdot)$ denotes either a GRU or LSTM model.

This formulation allows the model to capture temporal dependencies and long-term dynamics in firm-level data.

Convolutional model (CNN1D)

Alternatively, convolutional filters extract local temporal patterns:

$$g_{i,t} = \text{Conv1D}(X_{i,t}^{(L)}) \quad (19)$$

Unlike recurrent models, CNN1D focuses on extracting local temporal patterns and short-term structures within the input sequence.

The architecture and training configuration of the evaluated models are summarized in Table 2.

Table 2. The prediction model’s architecture and training setup.

Model	Key Hyperparameters	Values Used	Training Configuration
Logistic regression (baseline)	Regularization parameter	Default sklearn configuration	Baseline comparison
Majority classifier	-	No parameters	Baseline comparison
GRU	GRU units	64–32	Batch size = 64
	Dropout units	0.2	
	Dense layer units	32	
	Output activation	Sigmoid	
	Training epochs	200	
LSTM	LSTM units	64–32	Batch size = 64
	Dropout rate	0.2	
	Dense layer units	32	
	Output activation	Sigmoid	
	Training epochs	200	
CNN1D	Filters	64–32	Batch size = 64
	Kernel size	3	
	Dropout rate	0.2	
	Dense layer units	64	
	Output activation	Sigmoid	
	Training epochs	200	

Source: Author’s own elaboration.

6.5. Output Layer and Decision Rule

The representations learned by the models are then mapped to a probability of shareholder value creation through a sigmoid output layer.

$$p_{i,t} = \sigma(w_o^\top h_{i,t} + b_o) \quad (20)$$

6.6. Training, Validation and Model Selection Procedure

To enhance transparency and address model tuning concerns, the training setup and hyperparameter values used for the deep learning models are reported in Table 3 to improve transparency and solve model tuning issues. The optimization procedure, convergence behavior, and regularization approach used during model training are all governed by these requirements.

Table 3. Training and hyperparameter configuration of deep learning models.

Element	Specification
Optimizer	Adam
Learning rate	0.001
Loss function	Binary cross-entropy
Evaluation metric during training	Validation AUC and accuracy
Maximum epochs	200
Batch size	64
Early stopping	Yes
Early stopping monitor	Validation AUC
Early stopping patience	12 epochs
Restore best weights	Yes
Learning rate scheduler	Reduce LR On Plateau
LR reduction monitor	Validation AUC
LR reduction factor	0.5
LR reduction patience	6 epochs
Minimum learning rate	1×10^{-6}
Model checkpoint	Best validation AUC
Model selection criterion	Best validation AUC
Validation strategy	Chronological validation set
Final evaluation	Held-out chronological test set

All deep learning models were trained using the Adam optimizer with binary cross-entropy loss and a learning rate of 0.001, as indicated in Table 3. With a batch size of 64, training was done for up to 200 epochs. Early halting was used to reduce overfitting by reinstating the top-performing weights while monitoring validation AUC with a patience of 12 epochs. Furthermore, when validation performance reached a plateau, a learning rate reduction method was implemented. A tightly held-out test set was used to evaluate final performance, and validation AUC within a temporal validation framework guided model selection.

7. Empirical Results

7.1. Preliminary Empirical Evidence

7.1.1. Descriptive Statistics

The descriptive statistics of the macroeconomic and firm-level variables utilized in the empirical research are in Table 4. The table displays the number of observations, mean, standard deviation, minimum, median, maximum, skewness, and kurtosis for each variable.

Table 4. Descriptive statistics of financial and macroeconomic variables for Moroccan listed firms (2010–2024).

Variable	Obs.	Mean	Std. Dev.	Min	Median	Max	Skewness	Kurtosis
Δ sales	5040	0.004	0.063	−0.287	0.001	0.349	1.067	15.739
Δ EBITDA	5040	−0.000	0.076	−0.419	0.002	0.382	−0.511	16.241
PER	5040	1.123	3.615	−0.120	0.017	17.814	3.403	10.265
Volatility σ_t	5040	0.075	0.049	0.022	0.064	0.355	3.148	13.164
Leverage	5040	0.903	0.869	0.005	0.610	4.710	1.796	3.186
EPS	5040	5.956	7.167	−8.933	4.402	32.721	0.842	0.794
Δ EPS	5040	0.008	3.216	−14.297	0.016	11.448	−0.840	6.992
Size	5040	20.818	2.069	16.267	21.032	25.218	−0.012	−0.734
M/B	5040	2.305	3.709	0.014	0.743	25.320	2.922	9.653
$\beta_{i,t}$ (12m)	5040	0.351	0.699	−1.414	0.138	2.957	1.536	3.788
Close-Price	5040	786.95	953.72	9.30	340	5615	1.86	3.68
MASI monthly close	5040	11,158.92	1415.32	8413.72	11,350.73	14,837.22	0.13	−0.65
Inflation (t)	5040	0.001	0.002	−0.004	0.000	0.016	2.462	10.030
Policy-Rate	5040	0.025	0.005	0.015	0.025	0.032	−0.400	−0.636
SLOPE (t)	5040	0.007	0.003	−0.002	0.007	0.018	−0.305	1.069
Δ M2	5040	0.005	0.010	−0.022	0.003	0.035	0.374	0.130
Δ FX (t)	5040	0.001	0.019	−0.057	0.001	0.065	0.523	1.284
Δ GDP	5040	0.185	0.086	0.036	0.186	0.318	−0.050	−1.249

Source: Authors' calculations based on financial statements of companies listed on the Casablanca Stock Exchange (2010–2024).

Due to forward TSR building and sequence window limits, the original dataset of 5400 observations is reduced to 5040 usable observations. Significant variation among firm-level, market-based, and macroeconomic variables is revealed by the descriptive statistics. A number of firm-level metrics deviate significantly from normality, especially Δ sales and Δ EBITDA, which show high kurtosis (15.739 and 16.241, respectively), suggesting heavy-tailed distributions and the presence of extreme observations. Strong positive skewness and elevated kurtosis are also shown in PER, volatility, and M/B, indicating that a small number of enterprises have unusually high risk or value levels.

Some variables, however, appear to be more stable. Size is nearly symmetric, but there is little dispersion in the policy rate, SLOPE(t), and M2. Inflation is still leptokurtic and positively skewed, with sporadic spikes throughout the data period. Overall, the distributional characteristics in Table 4 suggest non-normality, asymmetry, and the presence

of outliers in several predictors, which supports the use of careful preprocessing and more flexible modeling strategies that can handle these features better than strictly linear models.

7.1.2. Correlation Analysis

The pairwise correlation matrix for the macroeconomic, market-based, and firm-level variables used in the analysis is shown in Figure 3. Before model estimation, this matrix helps identify potential multicollinearity issues and provides an initial relationship between variables.

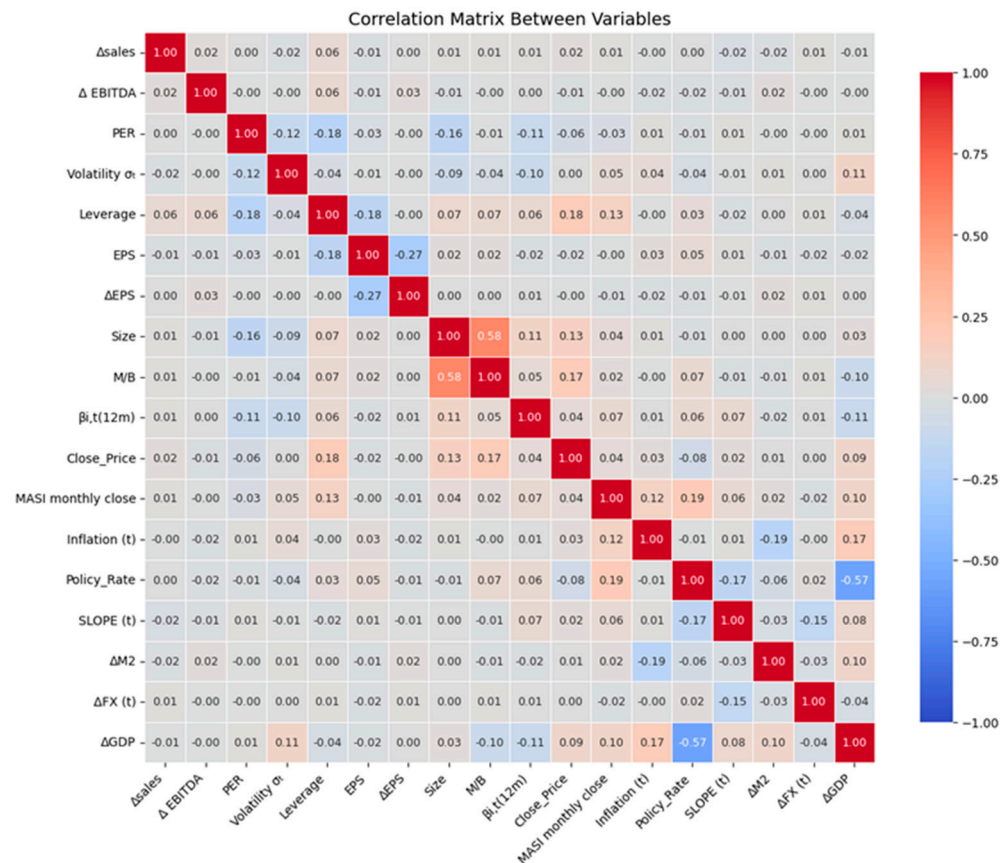


Figure 3. Correlation matrix of explanatory variables. Source: Author's computation using the research dataset.

The correlation matrix reveals that the majority of pairwise correlations are still weak, with absolute correlation coefficients typically falling below 0.20, suggesting that there is little linear dependency between the explanatory variables. Correlations between Δsales and ΔEBITDA at the firm level are nearly nil ($p = 0.02$), indicating that short-term variations in sales do not automatically translate into changes in operating profitability. The market-to-book ratio and firm size show a moderately positive connection ($p = 0.58$), indicating that larger firms typically have higher market valuations. As predicted, stock prices have a positive correlation with the market index ($p \approx 0.58$), indicating that larger firms typically have higher market valuations. As predicted, stock prices have a positive correlation with the market index ($p \approx 0.19$ with MASI), but their correlation with firm-specific risk metrics is still modest ($|p| < 0.10$).

The policy rate and GDP growth have the greatest correlation among macroeconomic variables ($p = -0.57$), suggesting a counter-cyclical monetary policy stance. Other macroeconomic correlations are still weak, such as the one between monetary aggregates and inflation ($|p| < 0.20$). The combined inclusion of firm-level, market-based, and macroe-

conomic variables in the empirical models is supported by the overall magnitude of the correlations, which indicates a minimal probability of multicollinearity.

Furthermore, individual correlations are still weak, with absolute values often below 0.15, according to a univariate examination of the link between each explanatory variable and the target variable. Although their magnitudes are still constrained, macroeconomic factors like the policy rate and the MASI index show the strongest correlations with the goal. Firm-specific variables, on the other hand, show fewer correlations, indicating that no single predictor drives the generation of shareholder value. The use of flexible modeling techniques that may capture nonlinear interactions between variables is further supported by these findings.

In general, the adoption of flexible nonlinear and sequential learning models which are better suited to represent complex and dynamic linkages in shareholder value creation is motivated by the weak pairwise correlations and the existence of non-normal and heavy-tailed distributions.

7.2. Predictive Performance of Models

This section reports the main results of the empirical analysis. To improve clarity and facilitate interpretation, the empirical evaluation is organized into three complementary stages. Table 5 summarizes the role and objectives of each stage and highlights how the different components of the evaluation framework contribute to the overall assessment of model performance and robustness.

Table 5. Summary of the empirical design across the three stages.

Component	Stage 1	Stage 2	Stage 3
Objective	Initial benchmark comparison	Deep learning architecture selection	Robustness and final validation
Models evaluated	Majority Class Baseline, Logistic Regression, GRU, LSTM, CNN1D	GRU, LSTM, CNN1D	Selected deep learning model and benchmark models
Main focus	Preliminary performance assessment	Identification of the best-performing deep learning architecture	Statistical reliability, temporal stability, and robustness assessment
Additional procedures	Standard evaluation metrics	Architecture comparison under identical settings	Bootstrap confidence intervals, temporal stability analysis, alternative beta estimation, variable block analysis, and continuous-target robustness analysis
Main outcome	Baseline comparison	Selection of the best-performing architecture	Validation of the robustness of the main conclusions
Validation design	Chronological train–validation–test split	Same validation framework	Bootstrap confidence intervals and temporal stability analyses
Model-selection criterion	Overall predictive performance metrics	Best-performing deep learning architecture based on AUC and complementary metrics	Robustness and consistency of the selected model

7.2.1. Stage 1: Initial Comparison Between Baseline and Deep Learning Models

This stage's goal is to present a preliminary comparison between deep learning models and baseline classifiers under a preliminary prediction specification. Establishing a reference framework and identifying potential variations in prediction behavior among model classes are the goals of this first stage.

Baseline Models

In order to create a baseline for further comparison, we start by assessing the prediction performance of benchmark models. A common linear probabilistic model in empirical finance is logistic regression, whereas the majority class classifier is a naïve reference that reflects the unconditional class distribution.

The majority class baseline produces an AUC near 0.50, as predicted, suggesting little discriminatory power. With an AUC marginally higher than random classification and minor improvements in balanced accuracy and F1-score, logistic regression outperforms this benchmark. These findings suggest that linear models might only be able to represent a small percentage of the underlying relationships, which encourages the investigation of more adaptable modeling techniques.

Deep Learning Models

We then assess deep learning architectures, such as GRU, LSTM, and CNN1D, to investigate their capacity to capture possible non-linear and temporal dependencies in the data, building on the benchmark results presented in Section Baseline Models.

The additional metrics used to assess model performance include AUC, precision–recall AUC, accuracy, balanced accuracy, F1-score, precision, recall, and the Matthews correlation coefficient (MCC). This allows for a more comprehensive comparison across different aspects of model performance.

The out-of-sample predictive performance of all assessed models, including benchmark classifiers (LogReg and Majority) and deep learning architectures (GRU, LSTM, CNN1D), on the test set is shown in Table 6.

Table 6. Predictive performance of deep learning models and benchmark classifiers.

Model	AUC	PR-AUC	Accuracy	Balanced Acc.	F1-Score	Precision	Recall	MCC
GRU	0.669	0.645	0.641	0.593	0.368	0.766	0.243	0.268
LSTM	0.593	0.561	0.583	0.570	0.496	0.519	0.475	0.142
CNN1D	0.570	0.539	0.588	0.543	0.313	0.557	0.218	0.114
LogReg (baseline)	0.538	0.484	0.571	0.516	0.193	0.511	0.119	0.053
Majority (baseline)	0.500	0.432	0.568	0.500	0.000	0.000	0.000	0.000

Source: Author's computation using the research dataset.

The predictive performance of deep learning models in comparison to benchmark classifiers is shown in Table 6. With an AUC of 0.669 and a PR-AUC of 0.645, GRU outperforms CNN1D (AUC = 0.570) and LSTM (AUC = 0.593) among the assessed models.

GRU also obtains the best accuracy (0.641) and balanced accuracy (0.593) in terms of total classification performance, indicating a more advantageous trade-off between sensitivity and specificity under this initial specification.

Despite having a better F1-score (0.496) and recall (0.475) than GRU (F1 = 0.368; Recall = 0.243), the LSTM model's lower AUC suggests that it has an inferior ranking ability. CNN1D performs mediocly on most evaluation metrics, but it still lags behind GRU in this initial comparison.

In general, deep learning models perform better than benchmark classifiers. The majority classifier produces AUC of 0.500, showing near-random discrimination, whereas logistic

regression achieves an AUC of 0.538. By contrast, GRU's discriminative performance clearly outperforms these baselines.

These preliminary findings imply that using temporal structures and non-linear modeling could improve predictive accuracy in comparison to linear techniques, but more validation is needed in later phases.

Although the outcomes presented in Table 6 offer a preliminary evaluation of out-of-sample prediction performance, they do not take into consideration possible statistical uncertainty resulting from temporal dependency in the data. Specifically, rolling input sequences and overlapping 12-month forward TSR horizons may cause serial correlation across data. Table 7 presents 95% block bootstrap confidence intervals for both AUC and PR-AUC, calculated using 12-month blocks, in order to solve this problem and guarantee a more stringent evaluation framework. This method offers reliable uncertainty estimates for model performance while maintaining the temporal structure of the data.

Table 7. Stage 1. Out-of-sample predictive performance with 95% Block Bootstrap Confidence intervals.

Model	AUC	AUC (95% CI)	PR-AUC	PR-AUC (95% CI)
GRU	0.669	[0.545; 0.775]	0.645	[0.482; 0.789]
LSTM	0.593	[0.454; 0.718]	0.561	[0.393; 0.720]
CNN1D	0.570	[0.448; 0.683]	0.539	[0.380; 0.692]
LogReg	0.538	[0.408; 0.650]	0.484	[0.360; 0.622]

Source: Author's computation using the research dataset.

In terms of key performance indicators, the findings verify that deep learning models perform better than the logistic regression baseline. With an AUC of 0.669 and a confidence interval of [0.545; 0.775], the GRU model in particular performs the best and most consistently. With confidence intervals that partially overlap the random benchmark, the logistic regression baseline, on the other hand, has poorer and less trustworthy prediction power. Although deep learning models perform better overall, the provided intervals show that there is no negligible uncertainty.

Robustness and Stability Analysis

To assess the robustness of the projected results, a number of additional analyses are performed. Classification performance is initially evaluated across different probability thresholds to demonstrate that the observed patterns are not solely impacted by a certain cutoff selection. Second, permutation testing shows that the reported AUC values are unlikely to have happened by chance. Finally, multi-seed research demonstrates that deep learning models function quite consistently across a range of random initializations.

Together, these results provide more proof of the consistency of the findings, but further validation is carried out in subsequent stages.

The permutation test results for the top-performing deep learning model (GRU) are shown in Figure 4. While the empirical null distribution derived from 300 random label permutations is centered around 0.50, with the majority of simulated AUC values falling between around 0.45 and 0.56, the observed test AUC exceeds 0.669.

With an empirical p -value below the 5% significance level, the observed AUC is located in the extreme right tail of the null distribution. This finding implies that random chance is unlikely to be the only factor influencing the model's predictive performance.

These results provide additional evidence that the model's discriminative power is resilient under the original specification.

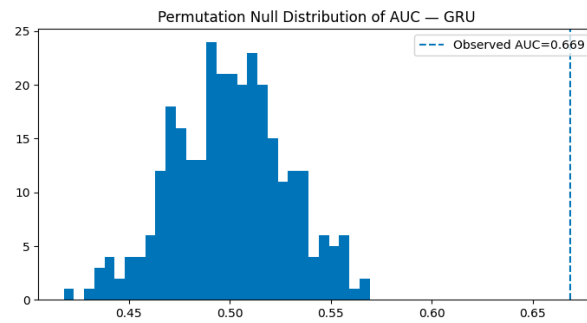


Figure 4. Permutation test for the AUC of the best-performing deep learning model.

The stability of the GRU model under different random initialization is shown in Figure 5. The test AUC has an average value of about 0.583 and ranges from 0.536 to 0.613. The spread across seeds is about 0.077 AUC points, with seed 7 showing the best performance (AUC = 0.613) and seed 77 showing the lowest (AUC = 0.536).

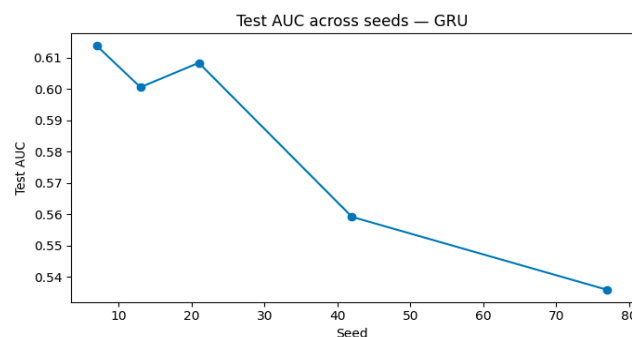


Figure 5. Stability of test AUC across random initializations.

Despite this variation, all AUC values remain above the random benchmark of 0.50, suggesting that the predictive signal is not driven by a specific initialization. The GRU model shows a respectable level of stability and reproducibility under this initial setup, as reflected in the relatively small variation between seeds.

The (ROC) curves for each model on the test sample are shown in Figure 6. With an AUC of 0.669, the GRU model outperforms CNN1D (0.570), LSTM (0.593), and the logistic regression baseline (0.538) in terms of overall discrimination.

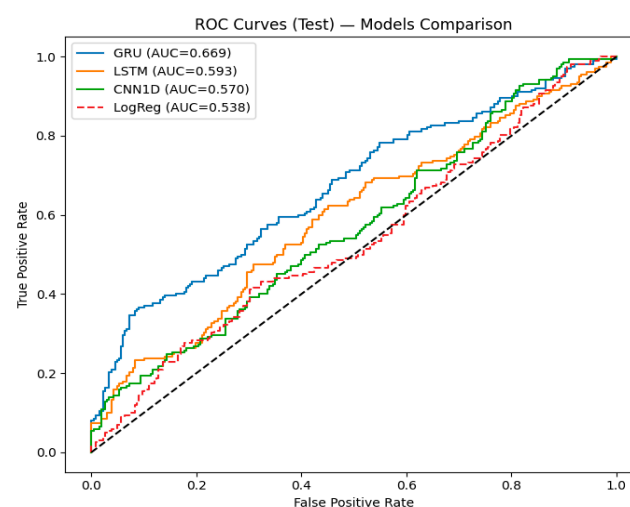


Figure 6. ROC curves of deep learning models and baseline classifiers on the test sample. The dashed diagonal line represents the performance of a random classifier (AUC = 0.5).

The difference in performance is notable: GRU shows an AUC gain of 0.076 over LSTM and 0.131 over the logistic regression benchmark. Although CNN1D and LSTM also outperform the random benchmark ($AUC = 0.500$), their gains remain modest.

The GRU curve generally lies above the other models across most false positive rate levels, indicating a higher true positive rate at similar false positive levels. These results suggest that, under this initial specification, sequential architectures particularly GRU may provide better discriminative performance in predicting beta-adjusted shareholder value.

To complement the ROC-based evaluation, Figure 7 presents the confusion matrices of the deep learning models and the logistic regression benchmark on the test sample.

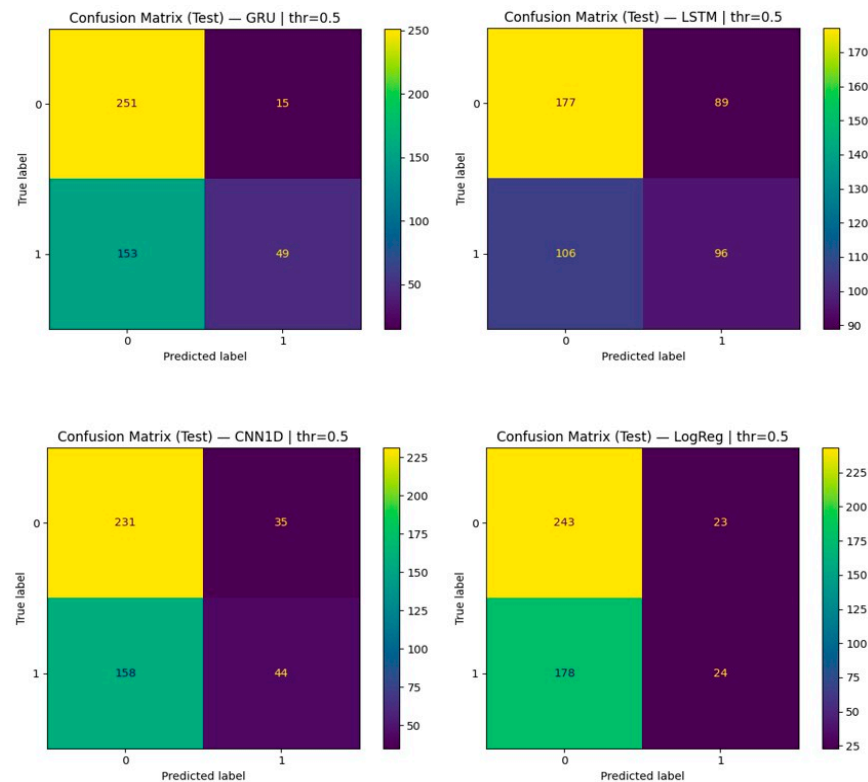


Figure 7. Confusion matrix of the best-performing deep learning model on the test sample. Source: Author's calculations based on data from Moroccan listed firms (2010–2024).

The model produces 15 false positives and 153 false negatives, while correctly identifying 251 true negatives and 49 true positives. This corresponds to a recall of about 24.2% ($49 / (49 + 153)$) and a specificity of 94.4% ($251 / (251 + 15)$), suggesting a strong ability to identify non-performing firms but a more limited ability to detect outperforming firms.

With a recall of 47.5% ($96 / (96 + 106)$) and a specificity of 66.6% ($177 / (177 + 89)$), the LSTM model shows better detection of positive cases, recognizing 96 true positives, but with 89 false positives. This may explain its more balanced F1-score compared to logistic regression and CNN1D.

With a recall of 21.8% and a specificity of 86.8%, CNN1D finds 44 true positives and 231 true negatives, along with 35 false positives and 158 false negatives.

Overall, these findings show that different models under this initial specification have different trade-offs between sensitivity and specificity.

The GRU model's F1-score progression over various probability thresholds is shown in Figure 8. At a threshold of 0.10, the F1-score reaches its maximum value of roughly 0.56, suggesting that, under this initial specification, a relatively low cutoff produces the best balance between precision and recall.

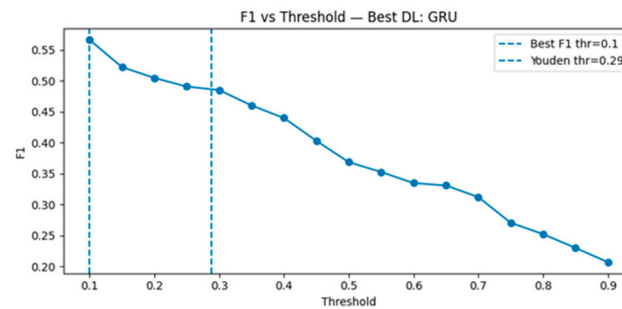


Figure 8. F1-score sensitivity to classification thresholds for the best-performing deep learning model.

The F1-score drops to about 0.37 at the traditional threshold of 0.50, which is a fall of almost 34% from its peak level. The dashed vertical lines in Figure 8 indicate the threshold maximizing the F1-score (0.10) and the threshold selected using the Youden criterion (0.29), respectively. The F1-score stays relatively high at about 0.49 when the Youden threshold (0.29) is applied, indicating that minor threshold tweaks do not significantly affect model performance.

The F1-score decreases as the threshold rises above 0.70, reaching about 0.21 at 0.90, which is consistent with a clear decrease in recall.

Overall, performance remains relatively stable within the 0.10–0.40 threshold range.

The GRU model's precision–recall curve on the test sample is shown in Figure 9, with an average precision (AP) of 0.645. This result indicates strong predictive ability on the positive class and is above the no-skill baseline, which reflects the prevalence of the positive class.

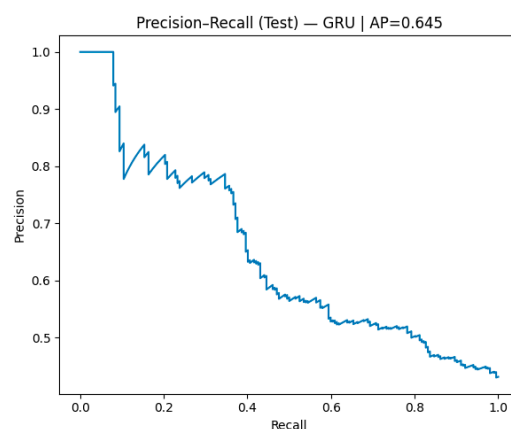


Figure 9. Precision–recall curve of the GRU model on the test set.

Precision stays near 1.00 for low recall levels (below 0.10), indicating that the model's most confident positive predictions are highly accurate. Precision stays between 0.63 and 0.65 when recall rises to roughly 0.40, suggesting that a large share of outperforming firms is correctly identified. The model maintains relatively strong precision in identifying positive cases, as evidenced by precision remaining over 0.52 at recall levels close to 0.60.

Precision progressively drops toward 0.44–0.45 as recall gets closer to 1.00, illustrating the usual trade-off between higher recall and increased false positives. Overall, the AP of 0.645 suggests that the GRU model maintains strong precision–recall performance, particularly when precision remains above 0.60.

Overall, this stage suggests that GRU provides the best overall performance under the initial specification and highlights the usefulness of sequential modeling for this task.

7.2.2. Stage 2: Deep Learning Comparison Under Stricter Classification

At this stage, a more stringent panel-safe classification approach is applied to analyze just deep learning architectures (GRU, LSTM, and CNN1D). The modeling setup becomes more constrained, avoiding potential knowledge leakage and ensuring a more realistic forecasting setting, even though the notion of shareholder value generation is still based on a beta-adjusted market benchmark.

Because the models must capture more nuanced and economically meaningful patterns related to abnormal performance rather than raw returns, the prediction task becomes more difficult. This improved methodology allows for a more thorough evaluation of the relative performance of deep learning models.

The out-of-sample performance of deep learning models under the more stringent classification framework is reported in Table 8. CNN1D shows the strongest discriminative performance in predicting future shareholder value creation, with the highest AUC (0.635) among the evaluated models.

Table 8. Predictive performance of deep learning models.

Model	AUC	Accuracy	F1-Score	Precision	Recall
CNN1D	0.635	0.600	0.267	0.642	0.168
GRU	0.605	0.579	0.273	0.536	0.183
LSTM	0.597	0.600	0.344	0.590	0.243

Source: Author's computation using the research dataset.

Nonetheless, LSTM achieves the highest recall (0.243) and F1-score (0.344), suggesting a better ability to identify positive cases. This highlights a trade-off between minority class detection (F1/Recall) and overall ranking performance (AUC).

GRU shows moderate performance across most metrics, indicating performance that is relatively stable but not leading.

The current specification relies on a more constrained modeling setup and a more stringent panel-safe temporal design than the first stage. These changes make the classification task more challenging and may explain the observed changes in model ranking.

This methodological change makes the classification task more challenging, as reflected in the following:

Lower recall levels (varying from 0.168 to 0.243), indicating more cautious detection of positive events, a relative compression of AUC values across models (CNN1D = 0.635, GRU = 0.605, LSTM = 0.597).

A shift in the model ranking, with CNN1D emerging as the top-performing architecture in terms of AUC, while recurrent models were more competitive in the earlier phase.

Table 9 presents 95% block bootstrap confidence interval for AUC and PR-AUC using 12-month blocks to account for possible statistical uncertainty resulting from temporal dependence.

Table 9. Stage 2. Out-of-sample predictive performance with 95% Block Bootstrap Confidence intervals.

Model	AUC	AUC (95% CI)	PR-AUC	PR-AUC (95% CI)
CNN1D	0.635	[0.568; 0.759]	0.632	[0.483; 0.767]
GRU	0.605	[0.544; 0.769]	0.639	[0.479; 0.777]
LSTM	0.597	[0.495; 0.718]	0.561	[0.410; 0.711]

Source: Author's computation using the research dataset.

The findings show that all models have moderate but consistent predictive performance, with a confidence interval that is completely above the random benchmark. CNN1D achieves the highest AUC (0.635), indicating comparatively strong discriminative ability. A

confidence interval that overlaps the 0.5 threshold indicates that LSTM produces weaker and less stable results, whereas GRU performs similarly but slightly worse. The reported intervals show a non-negligible degree of uncertainty, underscoring the difficult nature of the prediction task under the more stringent modeling framework, even though deep learning models generally outperform random classification.

The three deep learning architectures' out-of-sample ROC curves under the more stringent classification framework are shown in Figure 10. The models with the largest area under the curve (AUC = 0.635) are CNN1D, GRU (AUC = 0.605), and LSTM (AUC = 0.597). These findings imply that CNN1D has the best overall discriminative performance in identifying beta-adjusted shareholder value creation events, as model comparison is mostly based on AUC.

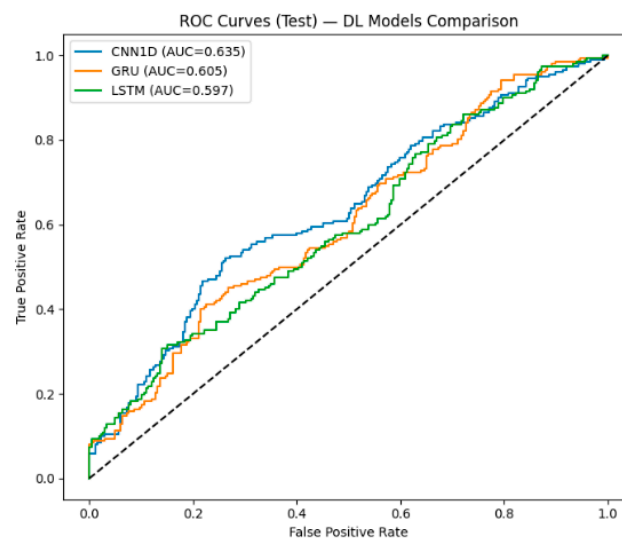


Figure 10. ROC curves (test set) deep learning models comparison. The dashed diagonal line represents the performance of a random classifier (AUC = 0.5).

Additionally, the ROC curves also show that CNN1D generally outperforms the other models across intermediate ranges (FPR between 0.2 and 0.6, where classification trade-offs are most informative). This trend implies that CNN1D would be more appropriate than recurrent architectures for capturing pertinent temporal patterns under this specification, even though the performance differences are still modest.

To complement the ROC-based evaluation, Figure 11 presents the confusion matrix of the best-performing CNN1D model on the test set.

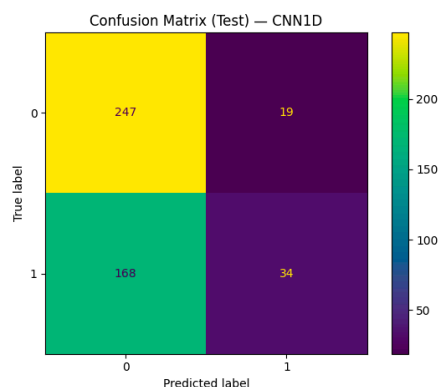


Figure 11. Confusion matrix analysis—CNN1D (test set, threshold = 0.5).

A more detailed understanding of classification performance is offered by the CNN1D model confusion matrix. The model accurately identifies 34 value-creation events (true

positives) and 247 non-value-creation situations (true negatives) out of 468 out-of-sample observations. Nevertheless, it wrongly identifies 19 negative cases as positive (false positives) and misclassifies 168 real positive cases as negatives (false negatives).

These numbers correspond to a precision of 0.642, meaning that 64.2% of anticipated value-creation signals are accurate. Only 16.8% of actual value creation events are effectively identified, according to the recall, which is still restricted at 0.168. The model prioritizes avoiding false alarms (low FP = 19) at the risk of missing a significant number of real positive opportunities (FN = 168). This asymmetry represents a conservative classification behavior. This imbalance between recall and precision is confirmed by the resulting F1-score of 0.267. Under the set threshold of 0.5, the model's sensitivity to actual shareholder value creation events remains relatively moderate, despite its great reliability when predicting outperformance.

The three deep learning architectures' validation AUC progression throughout training epochs is shown in Figure 12. During the first few epochs, CNN1D's performance increases quickly, going from roughly 0.57 to nearly 0.65 by epoch 5, following which it stabilizes at this level.

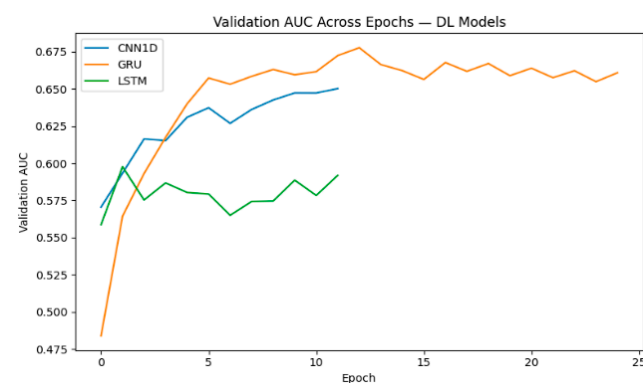


Figure 12. Validation AUC across epochs—DL models.

GRU, by contrast, shows more variation over time, with validation AUC ranging between about 0.60 and 0.67. LSTM shows more stable performance at a lower level, typically remaining below 0.60 during training.

For the selected CNN1D model, Figure 13 shows how the ROC curve (AUC) evolves during epochs for both the training and validation sets.

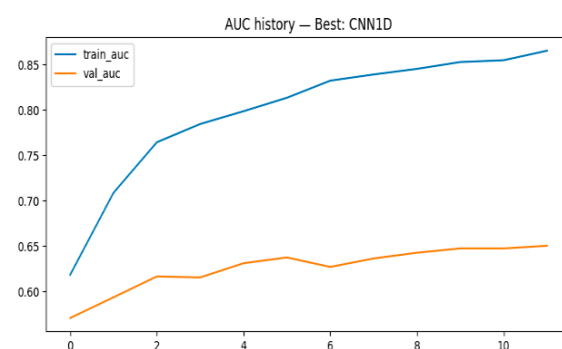


Figure 13. AUC history (best model).

Increasing from about 0.62 in the first epoch to nearly 0.86 in the final epoch, the training AUC shows a steady upward trend. This pattern suggests that the panel-structured financial data helps the model capture relevant temporal patterns.

The validation AUC, by contrast, increases more gradually and stabilizes between 0.63 and 0.65 after the initial epochs. This suggests a moderate level of overfitting, which is com-

mon in deep learning model applications involving noisy financial data, may be indicated by the difference between training and validation AUC. Nonetheless, the validation AUC consistently stays above 0.60, indicating that the model still has significant discriminative power outside of the sample.

The benefit of early stopping in limiting excessive divergence between training and validation performance is consistent with the validation AUC stabilizing after the mid-training epochs.

The training and validation loss development for the chosen CNN1D model over epochs is shown in Figure 14. Improved in-sample fit during training is indicated by the training loss, which steadily decreases from roughly 0.30 in the first epoch to roughly 0.23 in the last epoch.

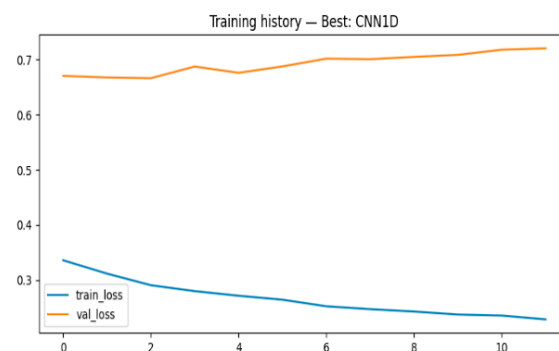


Figure 14. Training loss evolution best model (CNN1D).

The validation loss, on the other hand, varies within a comparatively small range, rising marginally from around 0.67 to roughly 0.72 in the last epochs. Since the model keeps optimizing the training objective while validation performance stabilizes, the ensuring divergence between training and validation loss may be a sign of moderate overfitting.

The validation loss shows no signs of sudden instability, indicating relatively stable training dynamics. This is consistent with the stabilization observed in the validation AUC, suggesting that despite some overfitting, the model still maintains reasonable generalization performance.

Overall, these findings provide more evidence for the dependability of the chosen model under the more stringent categorization framework by indicating that the learning process is steady under the chosen training configuration.

7.2.3. Stage 3: Final Model Evaluation and Validation

This stage reintroduces benchmark models and expands the study to a wider range of performance diagnostics and validation processes, in contrast to stage 2, which solely concentrates on relative performance among deep learning architectures. It seeks to verify the chosen architecture's performance in comparison to both deep learning and baseline models while assessing its resilience, stability, and usefulness. The chosen model configuration is used in this step, which offers a more thorough and accurate evaluation of out-of-sample prediction performance.

Final Comparative Performance

Using fixed classification threshold of 0.5, Table 10 presents the predictive performance of the benchmark models (logistic regression and the majority classifier) and deep learning architectures (CNN1D, GRU, and LSTM) on the test set.

Table 10. Predictive performance of deep learning and benchmark models (test set, threshold = 0.5).

Model	AUC	PR-AUC	ACC	BACC	F1	Precision	Recall	MCC
CNN1D	0.700	0.727	0.647	0.595	0.337	0.894	0.208	0.312
GRU	0.638	0.537	0.643	0.608	0.460	0.664	0.351	0.255
LSTM	0.607	0.585	0.650	0.613	0.457	0.690	0.342	0.272
LogReg (baseline)	0.538	0.484	0.571	0.516	0.193	0.511	0.119	0.053
Majority (baseline)	0.500	0.432	0.568	0.500	0.000	0.000	0.000	0.000

Source: Author's computation using the research dataset.

With an AUC of 0.700, the results show that CNN1D performs the best in terms of discrimination under the major AUC criterion. This result is much higher than the majority classifier (AUC = 0.500) and logistic regression (AUC = 0.538), indicating that deep learning architectures in this context capture more relevant temporal patterns than linear and naïve baselines.

CNN1D's highest PR-AUC (0.727) and MCC (0.312), which indicate better overall classification performance, further support its selection as the best performing model. Despite achieving slightly higher balanced accuracy (0.608 and 0.613, respectively) and accuracy (0.643 and 0.650, respectively), GRU and LSTM's lower AUC values indicate relatively poorer global ranking ability.

LSTM outperforms baseline models but falls short of CNN1D and GRU with a decent discriminative performance (AUC = 0.607). In contrast, logistic regression only slightly surpasses random classification, suggesting that only a small portion of the underlying predictive structure is captured by linear decision boundaries.

CNN1D attains the maximum accuracy (0.894) in terms of classification trade-offs, but GRU and LSTM show greater recall values (0.351 and 0.342, respectively), emphasizing a trade-off between sensitivity to positive instances and prediction purity. All things considered, our findings confirm that CNN1D is the best model in terms of global discriminative performance.

Table 11 presents 95% block bootstrap confidence intervals for AUC and PR-AUC using 12-month blocks to account for statistical uncertainty caused by temporal dependency.

Table 11. Final stage. Out-of-sample predictive performance with 95% Block Bootstrap Confidence intervals.

Model	AUC	AUC (95% CI)	PR-AUC	PR-AUC (95% CI)
CNN1D	0.700	[0.585; 0.803]	0.727	[0.467; 0.768]
GRU	0.638	[0.538; 0.734]	0.537	[0.413; 0.687]
LSTM	0.607	[0.494; 0.711]	0.585	[0.422; 0.726]
LogReg (baseline)	0.538	[0.408; 0.650]	0.484	[0.360; 0.622]

Source: Author's computation using the research dataset.

With an AUC of 0.700 and a 95% confidence interval of [0.585;0.803], the findings verify that CNN1D delivers the best discriminative performance. Although their intervals show more modest predictive stability, GRU and LSTM likewise perform better than the logistic regression baseline. Overall, the confidence intervals show a non-negligible level of uncertainty in the final prediction framework while supporting the superiority of deep learning models over the linear baseline.

To formally assess the significance of performance differences across models, Table 12 presents pairwise bootstrap comparisons of AUC values.

Table 12. Pairwise Bootstrap Comparison of AUC Differences Between Competing Models.

Comparison	Δ AUC	95% Confidence Interval	Bootstrap p -Value
CNN1D – GRU	−0.023	[−0.093; 0.043]	0.584
CNN1D – LSTM	0.038	[−0.044; 0.127]	0.430
CNN1D – Log Reg	0.132	[0.059; 0.214]	0.002
GRU – LogReg	0.154	[0.062; 0.258]	0.002
LSTM – LogReg	0.094	[0.001; 0.196]	0.050
GRU – LSTM	0.061	[−0.045; 0.168]	0.296

Source: Author's computation using the research dataset.

The pairwise bootstrap results provide additional insight into the statistical significance of performance differences across models. Both CNN1D and GRU significantly outperform the logistic regression baseline, with AUC improvements of +0.132 and +0.154, respectively ($p = 0.002$). LSTM also shows a statistically significant improvement over logistic regression (Δ AUC = +0.094; $p = 0.050$). In contrast, the differences among the deep learning architectures themselves remain statistically uncertain. The AUC difference between CNN1D and GRU is not significant (Δ AUC = −0.023; $p = 0.584$), nor are the differences between CNN1D and LSTM (Δ AUC = +0.038; $p = 0.430$) and between GRU and LSTM (Δ AUC = +0.061; $p = 0.296$). These findings indicate that the superiority of deep learning models over the linear benchmark is statistically supported, whereas the ranking among deep learning architectures should be interpreted with caution despite CNN1D achieving the highest overall AUC in the final evaluation stage.

ROC-Based Validation

The ROC curves of the benchmark logistic regression model and the deep learning models (CNN1D, GRU, and LSTM) that were assessed on the set for the classification of beta-adjusted stock outperformance are shown in Figure 15.

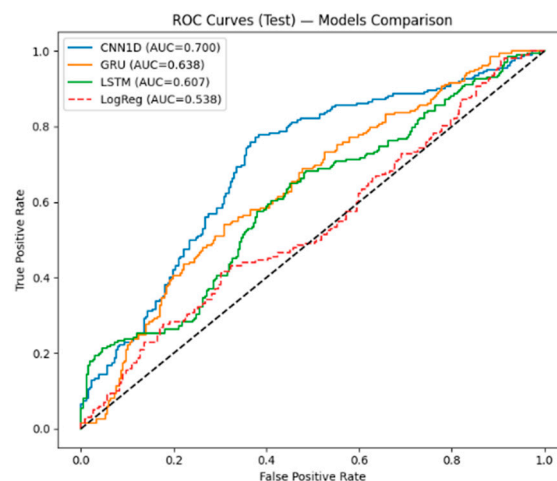


Figure 15. Comparative ROC curves of CNN1D, GRU, LSTM, and logistic regression models on the test set. The dashed diagonal line represents the performance of a random classifier (AUC = 0.5).

The findings demonstrate that the CNN1D model obtains the greatest discriminative performance with an AUC of 0.700, which is obviously higher than the random benchmark (AUC = 0.5). The GRU (AUC = 0.638) and LSTM (AUC = 0.607) models are ranked next, while the logistic regression baseline (AUC = 0.538) stays around the diagonal reference line, showing low predictive power.

Over the majority of the false positive rate range, the CNN1D curve often lies above the other models, suggesting a better ability to distinguish between organizations that are

operating well and those that are not. Even if the performance differences are still modest, the consistent model differences are still modest, the consistent model ranking validates the results shown in Table 10.

These findings imply that deep learning architectures are better at capturing temporal connections in financial data than linear models. Specifically, CNN1D's relative performance within the chosen sequential framework could be a reflection of its capacity to identify pertinent local temporal patterns.

Confusion Matrix Diagnostics

The final confusion matrices used for model evaluation are presented in Figure 16.

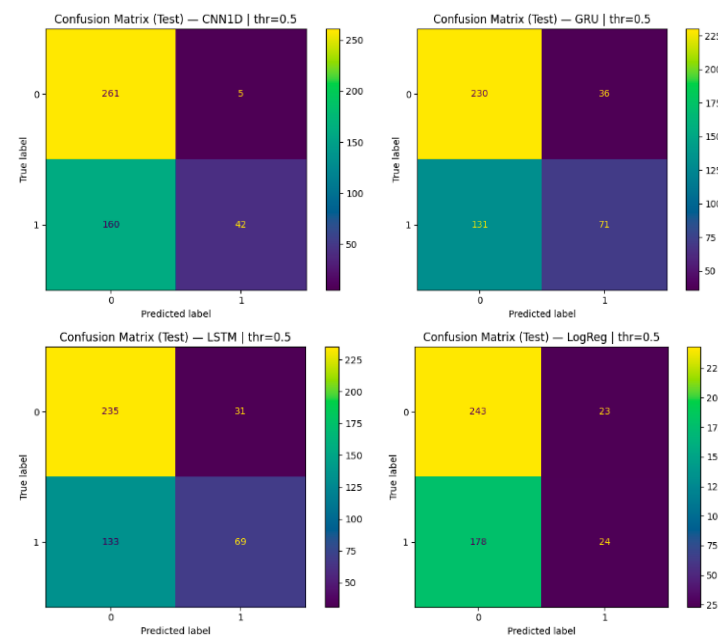


Figure 16. Confusion matrices of CNN1D, GRU, LSTM, and logistic regression (test set, threshold = 0.5).

Beyond overall performance indicators, the confusion matrices provide a thorough understanding of each model's categorization behavior. CNN1D shows a precision-oriented classification pattern among the assessed models. It achieves a high precision of 0.894, indicating that most predicted outperformance cases are correct, with 261 true negatives and only 5 false positives. Nevertheless, its recall is still low at 0.208, indicating a conservative identification of positive cases.

With 71 true positives properly identified and a relatively high number of true negatives (230), the GRU model shows a more balanced classification performance. In comparison to CNN1D, this yields a higher recall (0.351), but at the expense of more false positives (36), which leads to lower precision (0.664). This reflects a trade-off between sensitivity and precision, which is reflected in this pattern.

With 69 true positives and 133 false negatives, the LSTM model shows intermediate performance, producing a recall of 0.342 and an accuracy of 0.690. Its performance reflects a trade-off between classification reliability and detection ability, without outperforming other models in every dimension.

Logistic regression, on the other hand, performs poorly in identifying high-performing firms. Although it correctly identifies many negative cases (243 true negatives), it achieves a low recall (0.119) with only 24 true positives and 178 false negatives. This suggests that linear models may be less effective at capturing complex patterns in the data.

Overall, these findings show that different models have different classification profiles, with CNN1D emphasizing better sensitivity in identifying favorable results.

Threshold Optimization and Decision Rule Diagnostics

For the top-performing deep learning model (CNN1D), Figure 17 shows how sensitive the F1-score is to changes in the classification threshold. As the barrier climbs from 0.10 to roughly 0.30, the F1-score grows gradually and reaches its maximum value (≈ 0.69) at a threshold of 0.30. The best trade-off between recall and precision is found at this moment.

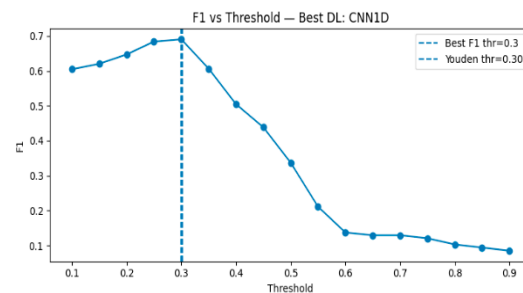


Figure 17. F1-score as a function of the classification threshold for the best deep learning model (CNN1D).

Higher thresholds decrease the model’s capacity to accurately detect positive cases (outperformance events), which lowers recall and weakens the balance conveyed by the F1 metric. This is demonstrated by the F1-score’s sharp reduction beyond this level. The fixed-threshold performance metrics shown in Table 10, which are calculated using the traditional classification threshold of 0.5, are not directly comparable to the F1-score values reported in Figure 17, which are produced under threshold optimization

This contrast emphasizes how crucial threshold selection is in classification problems because the decision rule selected can have a substantial impact on model performance.

Crucially, the Youden J threshold and the ideal F1 threshold (≈ 0.30) overlap, indicating a consistent decision boundary from both angles:

- The F1-max criterion, which balanced recall and precision;
- Youden’s J statistic (maximization of specificity + sensitivity – 1).

These findings imply that this classification problem may not be best served by the traditional criterion of 0.5. Precision threshold is increased to about 0.30, especially when identifying companies that outperform the beta-adjusted market benchmark.

Precision–Recall Validation

The CNN1D model’s precision–recall (PR) curve, as assessed on the out-of-sample test set, is shown in Figure 18. The PR curve average precision, AP of 0.727, indicates solid performance in identifying positive cases, with firms outperforming the beta-adjusted market benchmark.

Precision remains very high (above 0.9) at low recall levels (around 0.1–0.2), suggesting that high confidence positive predictions are generally reliable. Precision decreases as recall increases, reflecting the usual trade-off between maintaining precision and identifying more true positives.

Precision is comparatively stable (around 0.65–0.70) at intermediate recall levels (about 0.5–0.6), suggesting that the model retains acceptable classification quality despite recognizing a significant percentage of outperforming enterprises.

Because the classification problem in this study is somewhat unbalanced ($\approx 54\%$ class 0 vs. 46% class 1), the PR curve is particularly instructive. In these situations, compared to

ROC-AUC, the PR-AUC (or AP) offers a supplementary and more insightful assessment of performance on the positive class.

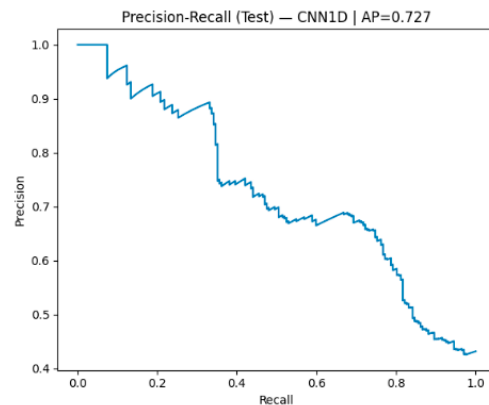


Figure 18. Precision–recall curve of the CNN1D model on the test set (AP = 0.727).

Overall, the AP value of 0.727 supports the findings from the ROC study by showing that the CNN1D model has good precision–recall performance in identifying businesses that provide anomalous performance in comparison to their beta-adjusted benchmark.

Statistical Significance and Stability

The empirical null distribution of the AUC derived from 500 random permutations of the target variable is shown in Figure 19. The distribution of AUC value under the null hypothesis that there is no predictive association between the explanatory variables and the target (random labeling) is shown by the histogram. The CNN1D model’s observed test AUC (AUC = 0.700) is shown by the dashed vertical line.

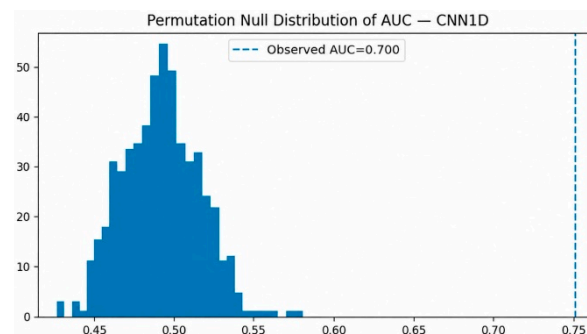


Figure 19. Permutation null distribution of AUC—CNN1D.

The performance of a random classifier is represented by the null distribution, which is centered at roughly 0.50. The observed AUC is obviously outside the range of permuted AUC values and is located far in the distribution’s right tail. This distinction offers compelling evidence that random labeling is unlikely to account for CNN1D’s predictive effectiveness.

The measured AUC is statistically significant at conventional levels, according to the empirical p -value ($p < 0.01$). When combined, these findings offer compelling evidence in favor of the theory that the CNN1D model has significant discriminative power when it comes to forecasting firm-level outperformance in comparison to the beta-adjusted market benchmark.

Out-of-Sample Temporal Stability Analysis

To further assess the robustness of the predictive results, the out-of-sample test set was divided into three consecutive chronological sub-periods (P1, P2, and P3). The pre-

dictive performance of CNN1D, GRU, LSTM, and Logistic Regression was then evaluated separately within each sub-period using AUC and PR-AUC.

The predictive performance of CNN1D, GRU, LSTM, and Logistic Regression was then evaluated separately within each sub-period using AUC and PR-AUC. The results are summarized in Table 13.

Table 13. Temporal Stability of Predictive Performance Across Consecutive Out-of-Sample Test Sub-Periods.

Test Sub-Period	Model	AUC	PR-AUC
Period 1	CNN1D	0.754	0.774
	GRU	0.820	0.817
	LSTM	0.794	0.784
	LogReg	0.618	0.537
Period 2	CNN1D	0.765	0.758
	GRU	0.535	0.448
	LSTM	0.692	0.726
	LogReg	0.435	0.498
Period 3	CNN1D	0.782	0.754
	GRU	0.624	0.598
	LSTM	0.566	0.515
	LogReg	0.655	0.656

Source: Author's computation using the research dataset.

The results reveal notable differences in temporal stability across the competing models. Although GRU achieves the highest AUC in P1 (0.820), its performance declines markedly in P2 (0.535) and remains relatively weak in P3 (0.624). LSTM exhibits a similar pattern, with AUC values decreasing from 0.794 in P1 to 0.566 in P3. In contrast, CNN1D maintains the most stable performance across all sub-periods, with AUC values ranging from 0.754 to 0.782 and PR-AUC values from 0.754 to 0.774. Logistic Regression also shows considerable fluctuations, with AUC values varying between 0.435 and 0.655. Overall, these findings indicate that the superior performance of CNN1D is not driven by a single favorable interval but reflects a more persistent and temporally stable predictive signal throughout the out-of-sample period.

Continuous Regression Robustness Analysis

To examine whether the binary formulation of shareholder value creation influences the results, an additional robustness analysis was conducted using abnormal TSR as a continuous target variable.

The continuous regression robustness analysis reveals limited predictive performance across all specifications. Linear Regression and Ridge Regression produce very poor out-of-sample results, with R^2 values of -7.982 and -7.977 , respectively, and RMSE values exceeding 1.34. Although ensemble methods perform relatively better, their predictive power remains weak. Random Forest achieves an RMSE of 0.935 and an R^2 of -3.353 , while Gradient Boosting delivers the best performance with an RMSE of 0.587, an MAE of 0.336, and an R^2 of -0.714 . Nevertheless, all models generate negative R^2 values, indicating that they fail to predict the exact magnitude of abnormal shareholder returns more accurately than a simple mean-based benchmark. Directional accuracy remains modest, ranging from 43.8% for linear specifications to 55.3% for Random Forest. These findings suggest that predicting the continuous abnormal TSR is particularly challenging in the Moroccan market context. Consequently, the binary classification framework adopted in the main analysis appears more suitable for distinguishing shareholder value creation from value destruction and yields more robust predictive performance, as evidenced by the substantially higher classification results reported for the deep learning models (AUC up to 0.692).

The performance of the regression models under the continuous robustness analysis is summarized in Table 14.

Table 14. Continuous Regression Robustness Analysis.

Model	RMSE	MAE	R ²
Linear Regression	1.343	1.092	−7.982
Ridge Regression	1.342	1.092	−7.977
Random Forest Regressor	0.935	0.444	−3.353
Gradient Boosting Regressor	0.587	0.336	−0.714

Source: Author's computation using the research dataset.

7.2.4. Robustness to Alternative Beta Estimation

To evaluate the sensitivity of the results, an alternative 24-month rolling beta is used in a robustness study. The objective is to assess how sensitive the results are to the selection of the beta estimation window. This alternative specification is used to reevaluate the predictive performance. The predictive performance under this alternative specification is summarized in Table 15.

Table 15. Robustness analysis. Out-of-sample predictive performance with 95% block bootstrap confidence intervals ($\beta = 24$ months).

Model	AUC	AUC (95% CI)	PR-AUC	PR-AUC (95% CI)
CNN1D	0.682	[0.567; 0.782]	0.740	[0.598; 0.853]
LSTM	0.678	[0.573; 0.780]	0.758	[0.625; 0.858]
GRU	0.642	[0.517; 0.754]	0.695	[0.549; 0.822]
LogReg (baseline)	0.643	[0.512; 0.765]	0.699	[0.559; 0.827]

Source: Author's computation using the research dataset.

All models continue to outperform the random benchmark in predicting performance, according to the robustness study based on a 24-month rolling beta. Although the performance disparities amongst models become more compressed when compared to the baseline specification, CNN1D earns the greatest AUC (0.682).

In terms of PR-AUC (0.758), LSTM performs competitively, indicating a significant capacity to detect positive occurrences under the alternative risk adjustment. All models maintain significant predictive power, as seen by confidence intervals that are significantly higher than the random benchmark, even if GRU and logistic regression exhibit very similar performance levels.

Overall, these results show that, despite a little increase in uncertainty, the primary conclusions are not influenced by the short window beta assumption and are resilient to other beta estimation specifications.

Across specifications, the variations in AUC are still rather small. While LSTM improves from 0.607 to 0.678 ($\Delta = +0.071$) and GRU stays steady, slightly rising from 0.638 to 0.642 ($\Delta = +0.004$), CNN1D exhibits a tiny reduction from 0.700 to 0.682 ($\Delta = -0.018$). These differences show that the primary results are unaffected by the selection of the beta estimation window and that model performance is generally steady. The comparative results under the two beta estimation windows are summarized in Table 16.

Table 16. Direct comparison of predictive performance under alternative beta specification (β 12m vs. β 24m).

Model	AUC (β 12m)	AUC (β 24m)	Δ AUC
CNN1D	0.700	0.682	−0.018
LSTM	0.607	0.678	+0.071
GRU	0.638	0.642	+0.004

Source: Author's computation using the research dataset.

7.2.5. Variable Block Analysis

This section looks at model results across various groupings of explanatory variables to further explore the role of various information sources on predicted performance. In particular, accounting variables, market-based indicators, macroeconomic parameters, and their different combinations are used to re-estimate models.

The predictive performance of all assessed models across these variable blocks is shown in Table 17.

Table 17. Predictive performance by variable blocks.

Variable Block	Model	Number of Features	AUC	MCC
Accounting variables only	LSTM	6	0.580	0.092
	GRU	6	0.525	−0.024
	CNN1D	6	0.518	−0.010
	LogReg	6	0.516	0.070
	Majority	6	0.500	0.000
All variables	CNN1D	18	0.700	0.312
	GRU	18	0.638	0.255
	LSTM	18	0.607	0.272
	LogReg	18	0.538	0.053
	Majority	18	0.500	0.000
All variables excluding accounting	CNN1D	12	0.763	0.293
	LSTM	12	0.692	0.362
	GRU	12	0.690	0.161
	LogReg	12	0.531	0.014
	Majority	12	0.500	0.000
All variables excluding macroeconomic variables	CNN1D	12	0.544	0.101
	GRU	12	0.542	0.077
	LSTM	12	0.500	0.000
	LogReg	12	0.490	−0.035
	Majority	12	0.457	0.000
All variables excluding market variables	GRU	12	0.728	0.328
	CNN1D	12	0.702	0.283
	LSTM	12	0.672	0.266
	LogReg	12	0.583	0.119
	Majority	12	0.500	0.000
Macroeconomic variables only	GRU	6	0.723	0.322
	CNN1D	6	0.717	0.277
	LSTM	6	0.717	0.264
	LogReg	6	0.575	0.089
	Majority	6	0.500	0.000
Market variables only	CNN1D	6	0.531	0.035
	GRU	6	0.506	0.013
	Majority	6	0.500	0.000
	LSTM	6	0.494	−0.065
	LogReg	6	0.494	−0.004

Source: Author's computation using the research dataset.

Based on the findings presented in Table 17, Table 18 presents the top-performing model for each variable block to aid in interpretation.

While MCC is reported as a complementary metric, model comparison in this study mainly relies on AUC, as it provides a threshold-independent view of discriminative performance.

The results show clear differences across variable blocks. Interestingly, the full specification does not produce the highest AUC. The best result is instead obtained when accounting variables are excluded, with CNN1D reaching an AUC of 0.763. This suggests

that adding all available variables does not always improve predictive performance, and that some inputs may introduce noise or redundant information.

Table 18. Best-performing model by variable block.

Variable Block	Best Model	Number of Features	AUC	MCC
Accounting variables only	LSTM	6	0.580	0.092
All variables	CNN1D	18	0.700	0.312
All variables excluding accounting	CNN1D	12	0.763	0.293
All variables excluding macroeconomic variables	CNN1D	12	0.544	0.101
All variables excluding market variables	GRU	12	0.728	0.328
Macroeconomic variables only	GRU	6	0.723	0.322
Market variables only	CNN1D	6	0.531	0.035

Macroeconomic variables, taken on their own, already provide a strong signal, with GRU reaching an AUC of 0.723. In contrast, market variables alone lead to much weaker results. The role of macroeconomic information becomes even clearer when it is removed, as performance drops sharply, with the best AUC falling to 0.544.

Overall, these results point to the central role of macroeconomic variables in the predictive setup and highlight the importance of selecting relevant inputs rather than simply increasing their number.

8. Analysis of Results and Discussion

8.1. Predictive Performance Across Modeling Stages

Based on organized financial and macroeconomic data, the empirical findings consistently demonstrate that the development of shareholder value is predictable. Non-naïve models support the general predictability hypothesis by achieving AUC values above the random benchmark at all stages.

GRU outperforms LSTM (0.593), CNN1D (0.570), and benchmark models in the first stage (Table 6) with the greatest AUC (0.669). This outcome supports Hypotheses H2.2 and H2.3, while H2.2 (temporal dependence) is indirectly supported because sequence-based designs' performance indicates that temporal structure makes a significant contribution to prediction.

CNN1D has the greatest AUC (0.635) in the second stage (Table 8), followed by GRU (0.605) and LSTM (0.597). The more stringent classification framework and panel-safe temporal design, which raise task difficulty and alter model ranking, are reflected in this modification.

The strongest evidence is found in the third stage (Table 10). CNN1D outperforms GRU, LSTM, and benchmark models with the highest AUC (0.700). PR-AUC (0.727), and MCC (0.312). ROC and PR analysis, threshold optimization (optimal threshold ≈ 0.30), and permutation testing ($p < 0.01$) all support these findings. According to the key AUC criterion, CNN1D is the best model overall in terms of threshold-independent discriminative performance.

8.2. Variable Block Analysis and Predictive Evidence

The block-wise analysis provides predictive evidence consistent with H1 and its sub-hypotheses (Tables 17 and 18). Under the “All variables” specification, CNN1D achieves the best performance (AUC = 0.700), indicating that combining accounting, market-based, and macroeconomic information improves prediction. Macroeconomic variables exhibit the strongest standalone predictive power (AUC = 0.723), followed by accounting variables (AUC = 0.580), while market variables show the lowest standalone performance

(AUC = 0.531). The exclusion analysis further shows a substantial decline in predictive performance when macroeconomic variables are removed (AUC = 0.544). These findings highlight the predictive relevance of different variable groups for shareholder value forecasting without implying direct causal relationships.

8.3. Interpretation of Model Performance and Implications

The findings show distinct variations in the behavior of the models. While GRU and LSTM produce stronger recall in some configurations, CNN1D offers better ranking performance (AUC and PR-AUC). This suggests that the goal, either ranking accuracy or the identification of favorable results, determines which model is chosen.

From a methodological standpoint, the findings demonstrate that deep learning models perform better than linear and naïve baselines, demonstrating the value of sequential and nonlinear modeling in financial prediction. The differences in model ranking between stages imply that distinct architectures capture distinct facets of temporal financial dynamics.

From an economic standpoint, the significant contribution of macroeconomic variables suggests that both firm-level trends and general economic conditions have an impact on the development of shareholder value. Although accounting and market factors offer complementary information, when taken separately, their predictive ability is still lower.

From an investment perspective, the proposed framework may be viewed as a screening tool rather than a complete trading strategy. By identifying firms that are more likely to outperform their risk-adjusted benchmark, the model can assist investors in narrowing the investment universe and prioritizing firms for further analysis. However, the study does not evaluate portfolio construction rules, transaction costs, turnover, or realized investment performance. Consequently, the results should be interpreted as evidence of predictive usefulness rather than proof of a profitable trading strategy.

8.4. Summary of Evidence Related to the Hypotheses

In general, there is substantial evidence for hypothesis H2. More precisely, H2.1 (temporal dependence) is indirectly supported because sequence-based structures perform well, indicating that temporal organization significantly influences prediction. However, a direct comparison with static representations is outside the purview of this work. The results are consistent with H2.2 and H2.3.

Since the greatest results are achieved when all variable groups are joined, the findings are consistent with H1 in an integrated sense. The findings are more consistent with H1.3, moderately consistent with H1.1, and less consistent with H1.2 among the sub-hypotheses.

In conclusion, the results verify that machine learning models may forecast the production of shareholder value, with performance contingent upon both model architecture and input variable structure.

A summary of the predictive performance across the three modeling stages is presented in Table 19.

Table 19. Summary of predictive performance across modeling stages.

Stage	Best Model	AUC	PR-AUC	Accuracy	Balanced Accuracy	Key Interpretation
Stage 1	GRU	0.669	0.645	0.641	0.593	First indications that sequential models perform better in terms of discrimination when using the AUC criteria
Stage 2	CNN1D	0.635	-	0.600	-	Stricter panel-safe categorization framework; deep learning model comparison; increased of local temporal patterns
Stage 3	CNN1D	0.700	0.727	0.647	0.595	Strong final model with statistical validation and the best discriminative performance under the primary AUC criterion

9. Conclusions

The purpose of this study was to determine whether firm-level, market-based, and macroeconomic data could be used in a sequential machine learning framework to predict shareholder value creation, which is defined as beta-adjusted outperformance in comparison to a market benchmark. Using structured financial and macro-financial data, the empirical analysis of a panel of Moroccan listed companies from 2010 to 2024 suggests that shareholder value creation is not purely random and appears to be partially predictable.

The findings suggest the presence of nonlinear and time-dependent patterns in shareholder value generation as predictive performance tends to improve when moving from simple linear baseline models to more advanced sequential deep learning architectures. Overall, the results provide predictive evidence consistent with the second main hypothesis (H2), whereas H2.1 (temporal dependence) is indirectly supported because sequence-based model performance indicates that temporal structure significantly influences prediction.

In terms of global discriminative performance, CNN1D is the best-performing model among the assessed architectures under the primary AUC criterion, demonstrating the importance of convolutional structures in capturing local temporal patterns. Nevertheless, the findings also show that model performance is still dependent on the predictive framework's structure, indicating that various designs capture complementary parts of financial dynamics.

The block-wise study further demonstrates that combining information from several sources maximizes predictive performance. Although they still play a complementary function to accounting and market-based data, macroeconomic variables offer the largest independent predictive contribution.

Overall, this work adds to the body of literature by demonstrating that the creation of shareholder value can be represented as a dynamic, data-driven process influenced by the interplay of macroeconomic variables, market conditions, and business fundamentals. The study expands previous research beyond conventional return prediction framework and offers fresh empirical evidence from an emerging market environment by combining a risk-adjusted concept of value creation with a sequential modeling technique.

Given the relatively limited number of listed firms and the moderate length of firm-level time series, the findings should be interpreted with caution. The results provide evidence of predictive patterns within the Moroccan listed non-financial sector, but they should not be viewed as definitive or universally generalizable conclusions. Instead, the study should be understood as an emerging-market empirical contribution showing that shareholder value creation may be partially predictable when accounting, market-based, and macroeconomic information are structured in a temporally consistent framework.

Limitations of the Study

Despite its contributions, this study has several limitations that should be considered when interpreting the findings.

First, the empirical analysis is based on a relatively limited number of listed firms. Although the final dataset contains 5040 usable firm-month observations, these observations originate from only 30 non-financial companies listed on the Casablanca Stock Exchange. This characteristic reflects both the size of the Moroccan equity market and the strict sample selection criteria adopted in this study. While the temporal dimension of the panel provides a substantial number of observations, the limited cross-sectional dimension may restrict the generalizability of the findings to larger, more liquid, or more diversified financial markets.

Accordingly, the findings should be interpreted as context-specific evidence from the Moroccan listed non-financial sector rather than as universal conclusions applicable to all emerging markets or developed financial systems.

This issue is particularly important because deep learning models are generally data-intensive and may be more sensitive to overfitting when applied to relatively small samples. To mitigate this risk, several safeguards were implemented, including parsimonious model architectures, dropout regularization, chronological train–validation–test splitting, early stopping procedures, and block-bootstrap confidence intervals. Nevertheless, the limited number of firms remains an inherent constraint of the study and suggests caution when extrapolating the results beyond the Moroccan market context.

Second, the target variable is constructed as a binary classification of shareholder value creation based on beta-adjusted outperformance relative to a market benchmark. Although this formulation is consistent with an investor-oriented perspective and facilitates the implementation of classification-based machine learning models, it inevitably transforms continuous performance differences into discrete outcomes. Consequently, some of the information regarding the magnitude of shareholder value creation may be lost.

In addition, the benchmark relies on a CAPM-based risk adjustment rather than multifactor abnormal return measures. Although the CAPM provides a transparent and parsimonious framework, alternative specifications based on the Fama–French or Carhart models may capture additional dimensions of systematic risk and could potentially refine the measurement of shareholder value creation.

Third, despite their superior predictive performance, deep learning models remain less interpretable than conventional econometric approaches. The black-box nature of GRU, LSTM, and CNN1D architectures makes it difficult to identify the precise contribution of individual predictors and limits the possibility of drawing causal inferences. Accordingly, the results should be interpreted as evidence of predictive relationships rather than causal mechanisms.

Fourth, although the study compares several benchmark and deep learning models, the empirical analysis does not include a comprehensive set of modern machine learning algorithms such as Random Forest, XGBoost, LightGBM, CatBoost, or Support Vector Machines. The objective of the study was primarily to evaluate whether sequential deep learning architectures provide additional predictive value in a temporally structured forecasting setting rather than to conduct an exhaustive comparison of all available machine learning approaches. Nevertheless, the absence of these additional benchmarks limits the breadth of the comparative assessment and should be considered when interpreting the relative performance of the proposed models.

Fifth, while the reported predictive results indicate statistically meaningful classification performance, the study does not directly evaluate whether these predictions can be translated into economically profitable investment strategies after accounting for transaction costs, market frictions, liquidity constraints, or portfolio rebalancing effects. Consequently, the findings should be interpreted primarily as evidence of predictive capability rather than direct proof of economically exploitable trading opportunities.

Future research could extend this framework in several directions. One promising avenue would be to investigate hybrid approaches combining deep learning and traditional econometric methods, or to explore ensemble architectures integrating recurrent and convolutional networks. Another extension would involve incorporating additional sources of information such as textual disclosures, news sentiment, ESG indicators, or higher-frequency financial data. Expanding the analysis to larger datasets, longer time horizons, or multi-country emerging market samples would also provide an opportunity to further assess the robustness and external validity of the findings. Finally, future studies could evaluate the economic usefulness of the proposed models through portfolio construction, trading strategies, or risk-management applications in order to assess their practical value for investors and financial decision-makers.

Overall, this study highlights the potential of deep learning techniques for predicting shareholder value creation in an emerging market environment while emphasizing the importance of combining accounting, market-based, and macroeconomic information within a temporally consistent predictive framework.

A comparative overview of related studies and their relationship to the present research is provided in Table 20.

Table 20. Comparative overview of related studies in financial prediction using machine learning.

Study	Data Type	Methodology	Dependent Variable	Frequency	Key Contribution	Similarity with Our Study
Gu et al. (2020)	Firm characteristics (including-based indicators) + macro variables	Machine learning (RF, NN, LASSO)	Stock returns	Monthly	Demonstrates superior predictive power of ML models in asset pricing	Very high similarity (ML + macro variables)
L. Chen et al. (2021)	Firm-level data + latent factors	Deep learning	Stock returns	Monthly	Captures nonlinear relationship in asset pricing using deep learning evaluates relevance of numerous financial predictors	Very high similarity (Deep Learning + nonlinear modeling)
Feng et al. (2020)	Large set of firm characteristics	Machine Learning factor selection	Stock returns	Monthly	Evaluates relevance of numerous financial predictors	High similarity (multi-variable financial modeling)
Kelly et al. (2019)	Firm characteristic (accounting and market-based)	Econometric + ML approach	Stock returns	Monthly	Links firm characteristics to asset pricing through covariance structure	High similarity (firm-level predictor)
Rapach et al. (2013)	Macroeconomic variables + market data	Econometric forecasting models	Excess returns	Monthly	Highlights predictive role of macroeconomic variables	(macro-driven prediction)
Neely et al. (2014)	Macroeconomic indicators	Machine Learning & forecasting	Stock returns	Monthly	Shows macro variables improve return predictability	Moderate similarity (macro + prediction)
Krauss et al. (2017)	Price-based financial data	Deep Neural Networks	Returns direction	Daily	Applies deep learning to trading strategies	Moderate similarity (DL + prediction)
Sirignano and Cont (2019)	High-frequency market data	Deep Learning	Price movements	High frequency	Model's financial dynamics using DL	Low to moderate similarity (methodological)
Bekaert and Harvey (2003)	Emerging market data (macro + financial)	Empirical analysis	Returns	Mixed	Analyzes characteristics of emerging markets	High relevance (context justification)
Harvey (1995)	Macro + market data (emerging markets)	Econometric models	Returns	Monthly	Shows predictability in emerging markets	High relevance (predictability in EM context)

Author Contributions: Conceptualization, Y.J.; methodology, Y.J.; software, Y.J.; validation, Y.J., I.E.Y. and N.B.A.; formal analysis, Y.J.; investigation, Y.J.; data curation, Y.J.; writing original draft preparation, Y.J.; writing review and editing, Y.J., I.E.Y. and N.B.A.; visualization, Y.J.; supervision, I.E.Y. and N.B.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are available from publicly accessible financial and macroeconomic sources.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Andreou, E., Ghysels, E., & Kourtellis, A. (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics*, 158(2), 246–261. [\[CrossRef\]](#)
- Andreou, E., Ghysels, E., & Kourtellis, A. (2013). Should macroeconomic forecasters use daily financial data and how? *Journal of Business & Economic Statistics*, 31(2), 240–251. [\[CrossRef\]](#)
- Ang, A., Hodrick, R. J., Xing, Y., & Zhang, X. (2006). The cross-section of volatility and expected returns. *The Journal of Finance*, 61(1), 259–299.
- Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv*, arXiv:1803.01271.
- Ball, R., & Brown, P. (1968). An empirical evaluation of accounting income numbers. *Journal of Accounting Research*, 6(2), 159–178. [\[CrossRef\]](#)
- Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long short-term memory. *PLoS ONE*, 12(7), e0180944. [\[PubMed\]](#)
- Beaver, W. H. (1968). The information content of annual earnings announcements. *Journal of Accounting Research*, 6, 67–92. [\[CrossRef\]](#)
- Bekaert, G., & Harvey, C. R. (2003). Emerging markets finance. *Journal of Empirical Finance*, 10(1–2), 3–55. [\[CrossRef\]](#)
- Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The Journal of Finance*, 43(2), 507–528. [\[CrossRef\]](#)
- Boudoukh, J., Michaely, R., Richardson, M., & Roberts, M. R. (2007). On the importance of measuring payout yield: Implications for empirical asset pricing. *The Journal of Finance*, 62(2), 877–915. [\[CrossRef\]](#)
- Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). *The econometrics of financial markets*. Princeton University Press.
- Campbell, J. Y., & Shiller, R. J. (1988). Stock prices, earnings, and expected dividends. *The Journal of Finance*, 43(3), 661–676. [\[CrossRef\]](#)
- Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4), 1509–1531.
- Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, 52(1), 57–82. [\[CrossRef\]](#)
- Chambers, A. E., & Penman, S. H. (1984). Timeliness of reporting and the stock price reaction to earnings announcements. *Journal of Accounting Research*, 22(1), 21–47. [\[CrossRef\]](#)
- Chen, L., Pelger, M., & Zhu, J. (2021). Deep learning in asset pricing. *Management Science*, 67(2), 714–729.
- Chen, N.-F., Roll, R., & Ross, S. A. (1986). Economic forces and the stock market. *The Journal of Business*, 59(3), 383–403. [\[CrossRef\]](#)
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 6. [\[CrossRef\]](#) [\[PubMed\]](#)
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1724–1734). Association for Computational Linguistics.
- Cochrane, J. H. (2005). *Asset pricing* (Rev. ed.). Princeton University Press.
- Estrella, A., & Hardouvelis, G. A. (1991). The term structure as a predictor of real economic activity. *The Journal of Finance*, 46(2), 555–576. [\[CrossRef\]](#)
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417. [\[CrossRef\]](#)
- Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56. [\[CrossRef\]](#)
- Fama, E. F., & French, K. R. (2006). Profitability, investment, and average returns. *Journal of Financial Economics*, 82(3), 491–518. [\[CrossRef\]](#)
- Fama, E. F., & Schwert, G. W. (1977). Asset returns and inflation. *Journal of Financial Economics*, 5(2), 115–146. [\[CrossRef\]](#)
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. [\[CrossRef\]](#)
- Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3), 1327–1370. [\[CrossRef\]](#)
- Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669. [\[CrossRef\]](#)
- Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). The MIDAS touch: Mixed data sampling regression models. In *CIRANO working paper*. UCLA.
- Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, 26(1), 53–90. [\[CrossRef\]](#)
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3), 393–408.
- Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273. [\[CrossRef\]](#)

- Harvey, C. R. (1995). Predictable risk and returns in emerging markets. *The Review of Financial Studies*, 8(3), 773–816. [\[CrossRef\]](#)
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning in finance. *Annual Review of Financial Economics*, 9, 145–181.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.
- Jagadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1), 65–91. [\[CrossRef\]](#)
- Jensen, M. C. (1968). The performance of mutual funds in the period 1945–1964. *The Journal of Finance*, 23(2), 389–416.
- Jensen, M. C. (2001). Value maximization, stakeholder theory, and the corporate objective function. *Journal of Applied Corporate Finance*, 14(3), 8–21. [\[CrossRef\]](#)
- Jiang, J., Kelly, B., & Xiu, D. (2023). (Re-)Imag(in)ing price trends. *The Journal of Finance*, 78(6), 3193–3249.
- Kelly, B., Pruitt, S., & Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3), 501–524. [\[CrossRef\]](#)
- Kiranyaz, S., Ince, T., Abdeljaber, O., Avci, O., & Gabbouj, M. (2015). 1-D convolutional neural networks for signal processing applications. In *2015 international conference on industrial informatics (INDIN)* (pp. 836–841). IEEE.
- Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2), 689–702.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, 47(1), 13–37. [\[CrossRef\]](#)
- Marcellino, M., Stock, J. H., & Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1–2), 499–526. [\[CrossRef\]](#)
- Neely, C. J., Rapach, D. E., Tu, J., & Zhou, G. (2014). Forecasting the equity risk premium: The role of technical indicators. *Management Science*, 60(7), 1772–1791. [\[CrossRef\]](#)
- Nelson, D. M. Q., Pereira, A. C. M., & de Oliveira, R. A. (2017). Stock market's price movement prediction with long short-term memory neural networks. In *Proceedings of the international joint conference on neural networks (IJCNN)* (pp. 1419–1426). IEEE.
- Penman, S. H. (2013). *Financial statement analysis and security valuation* (5th ed.). McGraw-Hill.
- Provost, F., & Fawcett, T. (2013). *Data science for business*. O'Reilly Media.
- Rapach, D. E., Strauss, J. K., & Zhou, G. (2013). International stock return predictability: What is the role of the United States? *The Journal of Finance*, 68(4), 1633–1662. [\[CrossRef\]](#)
- Rapach, D. E., & Zhou, G. (2013). Forecasting stock returns. In G. Elliott, & A. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 2A, pp. 328–383). Elsevier.
- Rappaport, A. (1986). *Creating shareholder value: The new standard for business performance*. Free Press.
- Saito, T., & Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), e0118432. [\[CrossRef\]](#) [\[PubMed\]](#)
- Schorfheide, F., & Song, D. (2015). Real-time forecasting with a mixed-frequency VAR. *Journal of Business & Economic Statistics*, 33(3), 366–380.
- Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review. *Applied Soft Computing*, 90, 106181. [\[CrossRef\]](#)
- Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3), 425–442. [\[CrossRef\]](#)
- Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets: Perspectives from deep learning. *Quantitative Finance*, 19(9), 1449–1459. [\[CrossRef\]](#)
- Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, 71(3), 289–315. [\[CrossRef\]](#)
- Stewart, G. B. (1991). *The quest for value: A guide for senior managers*. HarperBusiness.
- Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3–28. [\[CrossRef\]](#)
- Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2017). A survey on deep learning for big data. *Information Fusion*, 42, 146–157.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.