UNIVERSITY OF COPENHAGEN
DEPARTMENT OF ECONOMICS

Master Thesis

Christian Birk Gustafson

Machine Learning for Empirical Asset Pricing

- LSTM's Ability to Predict Excess Stock Returns

Faculty: Faculty of Social Sciences

Submitted on: 30/05/2024

Supervisor: Stefan Voigt

ECTS: 30

Keystrokes: 125,663 (52 standard pages)

Machine Learning for Empirical Asset Pricing

Christian Birk Gustafson (ldg790)

University of Copenhagen

Department of Economics

Spring 2024

Abstract

This thesis investigates the predictive capability of five Long Short-Term Memory (LSTM) models, varying from one to five LSTM layers, in forecasting monthly US excess stock returns. The optimal network identified from this comparison utilises the 20 most important features for out-of-sample predictions. These features are selected from 94 stock characteristics and 13 macroeconomic predictors using the permutation feature importance (PFI) technique. The predicted excess returns are employed to construct a 10-1 hedge portfolio to assess economic gains. The multivariate LSTM model is compared against a traditional feed-forward neural network (FNN), a buy-and-hold strategy, and a univariate LSTM model. The results reveal that the optimal LSTM network comprises two LSTM layers. Yet, its R^2 of -3.547 indicates poor predictive accuracy relative to the FNN. Additionally, the 10-1 hedge portfolio strategy does not outperform a buy-and-hold strategy of the S&P 500 index. However, the multivariate LSTM model achieves a higher Sharpe ratio and generates statistically significant positive excess returns compared to the univariate LSTM model, suggesting that incorporating additional features enhances performance. Among the most crucial features, two pertain to risk measures, three to liquidity, and ten to valuation ratios and fundamental indicators, with no momentum-related features deemed important. Finally, the five most critical macroeconomic predictors are inflation, 3-Month Treasury Bill, long-term yield, term spread, and long-term rate of returns.

Keywords: *Machine Learning, LSTM, Forecast Excess Stock Returns, PFI.*

Table of Contents

1 Introduction

This thesis conducts a comprehensive examination of Long Short-Term Memory (LSTM) models' ability to predict monthly US excess stock returns. The field of empirical asset pricing has traditionally relied on foundational models such as the Capital Asset Pricing Model (CAPM), introduced by Sharpe (1964), and further expanded upon by the Fama-French Three-Factor model (FF3) as detailed in Fama and French (1993). These models utilise fundamental characteristics to predict future excess stock returns. While fundamental, these conventional methodologies have come under increasing scrutiny. Critics argue that their linear and relatively simplistic assumptions fail to adequately capture financial markets' dynamic and intricate behaviour, thereby leading to poor excess stock return predictions.

The advent of machine learning (ML) techniques, as seen in Gu et al. (2020), promises a paradigm shift in this respect, offering sophisticated tools to dissect and predict equity risk premiums with greater precision. LSTM models especially emerge as promising methods among these ML techniques. As illustrated in the studies by Moghar and Hamiche (2020), Ghosh et al. (2022), and Gaur (2023), LSTMs demonstrate an advantageous ability to capture temporal dependencies and predict future excess stock returns.

Building upon the framework established by Gu et al. (2020), this thesis investigates whether LSTM models can effectively predict monthly US excess stock returns and achieve economic gains. Furthermore, it explores whether LSTM models can outperform the traditional feed-forward neural network (FNN) employed by Gu et al. (2020). Unlike FNNs, LSTM networks incorporate a form of memory, allowing them to excel in detecting long-term relationships in time series data, which are pivotal for understanding market dynamics. For a thorough exploration of this topic, the analysis is segmented into the following three hypotheses:

H1 *Employing permutation feature importance in conjunction with an LSTM model uncovers the most critical features for predicting excess stock returns.*

H2 *LSTM models exhibit superior accuracy in predicting excess stock returns compared to traditional feed-forward neural networks.*

H3 *A 10-1 hedge portfolio constructed using LSTM predictions generates significant economic gains, net of transaction costs.*

To test these hypotheses, this thesis examines 94 stock characteristics and 13 macroeconomic predictors, outlined by Green et al. (2017) and Welch and Goyal (2008), respectively. These features are obtained from *The Center for Research in Security Prices (CRSP)* and *Compustat*. Previously analysed by Gu et al. (2020), these features are now reevaluated in this thesis to determine their effectiveness in predicting monthly excess stock returns using an LSTM model.

In addition, five LSTM architectures are explored, ranging from a single LSTM layer to a five-layer structure. These models were trained and optimised using K-fold cross-validation on data from 1960 to 2008. The optimal network structure is then employed for variable selection, utilising permutation feature importance (PFI) to identify the most important predictors for forecasting monthly excess stock returns.

The analysis identifies the optimal LSTM model to contain two LSTM layers, denoted as LSTM2. This model selects the 15 most important stock characteristics and five key macroeconomic predictors, along with past monthly excess stock returns, to generate forecasts for the out-of-sample period from 2009 to 2020. These predictions are used to sort stocks into deciles, representing portfolios. The highest decile (tenth) contains stocks with the highest predicted excess returns, while the lowest decile (first) contains stocks with the lowest predicted excess returns.

Finally, a long-short investment strategy, also referred to as the 10-1 hedge portfolio strategy, is implemented by going long on stocks in the highest decile and shorting stocks in the lowest decile. The effectiveness of the portfolios is evaluated using the Sharpe ratio and the FF3. The analysis closes with an examination of the constructed portfolios to identify the patterns and industries that characterise the stocks selected by the LSTM model.

My contributions to the field of empirical asset pricing are threefold: (i) Building upon the foundational research of Gu et al. (2020), I enhance the analysis by incorporating five new macroeconomic predictors. (ii) I expand the temporal scope of the dataset beyond the period examined by Gu et al. (2020), extending the analysis to include data from 2017 to 2020. (iii) I introduce another model, specifically the LSTM model, into the framework established by Gu et al. (2020).

Accurate forecasts of excess stock returns are crucial for investors and financial institutions, enabling informed decision-making, risk management, and strategic asset allocation. However, it is inherently challenging due to the returns' unpredictable nature, influenced by factors such as unemployment rates, retail sales, investor psychology, and global events.

Empirical asset pricing seeks to address these complexities by analysing historical data to identify patterns and test theories for predicting excess stock returns. Traditional methods, such as conditional portfolio sorting and linear regression models, have played an important role in identifying predictors and understanding asset pricing dynamics.

However, ML offers significant advancements over traditional methods by handling a broader set of predictors and analysing complex, nonlinear relationships between variables. This expands the modelling scope for asset prices and enhances the precision of risk premium measurements. The integration of ML allows researchers to uncover intriguing patterns and insights in financial data, often overlooked by traditional approaches. Furthermore, the ML approach allows for estimating excess stock returns without the constraints of fixed functional form assumptions, eliminating the reliance on preconceived notions about the data's distribution characteristics. This shift offers a more nuanced perspective on the factors driving asset prices, enriching our understanding of financial markets.

The methodology employed in this thesis offers a promising path for institutional investors to overcome some of the traditional challenges associated with excess stock return prediction. The focus on institutional investors, such as banks and large funds, arises from the substantial computational resources and extensive data LSTM models require for peak performance. Despite their high cost, adopting an LSTM method can be particularly beneficial for entities with the requisite computational capacity and data access, offering a potential edge in the market.

This thesis reveals that among the most crucial features for predicting excess stock returns, two pertain to risk measures, three to liquidity, and ten to valuation ratios and fundamental indicators. Additionally, it identifies the five most critical macroeconomic predictors to be inflation (**infl**), 3-Month Treasury Bill (**tbl**), long-term yield (**lty**), term spread (**tms**), and long-term rate of returns (**ltr**). However, the LSTM model did not outperform the traditional FNN employed in the study by Gu et al. (2020). Nevertheless, the implemented

long-short investment strategy based on the multivariate LSTM2 did generate small yet statistically significant positive excess returns. Thus, the multivariate LSTM2 model outperformed the univariate LSTM model, suggesting that incorporating a broader range of features can enhance performance. However, the LSTM2 failed to outperform a passive buy-and-hold strategy based on the market index. Specifically, the multivariate LSTM2 model achieved an annual Sharpe ratio of 0.447, compared to the market index's Sharpe ratio of 0.96.

The remainder of this paper is structured as follows: Section 2 reviews past literature on asset pricing from 1964 to the present, providing a historical and contemporary context for this thesis. Section 3 delves into the theories presented in the literature. Section 4 outlines the methodology employed in this thesis. Section 5 illustrates the empirical results of this thesis, focusing on the analysis of US stocks. Section 6 discusses and critically assesses the methodology and results of this thesis. Additionally, it outlines directions for future research and considerations for real-world application. Finally, Section 7 summarises the main findings and their implications.

2 Literature

This section first explores the fundamental factor models in empirical asset pricing and highlights notable advancements that have influenced the field. It then addresses the limitations inherent in these models and explains how Machine Learning (ML) techniques are employed to mitigate these challenges. The literature extends beyond traditional ML methodologies, such as feed-forward neural networks (FNN), and explores the field of Long Short-Term Memory (LSTM) networks. It clarifies the advantages of LSTM networks, outlines my contributions, and presents the main hypotheses.

Introduced by Sharpe (1964), the Capital Asset Pricing Model (CAPM) posits that an asset's expected excess returns are fully explained by its beta, which measures its correlation with the market portfolio. However, Ross (1976) introduces the Arbitrage Pricing Theory (APT) through a linear factor model, challenging the assumptions of the CAPM. The APT suggests the presence of critical factors not accounted for in the CAPM. Specifically, the APT posits that excess stock returns are influenced not only by their riskiness relative to the market but also by macroeconomic factors, such as inflation or GNP.

Building on the foundational principles established by the CAPM and APT, Banz (1981) in-

troduces the concept of the *size premium*, followed by Rosenberg et al. (1985) proposing the *value premium*. These factors aim to account for a more significant fraction of the fluctuation in excess returns beyond what market risk explains, thereby challenging the foundational assertions of the CAPM. As a result, by incorporating these two additional factors, Fama and French (1993) introduced the Fama-French Three-Factor model (FF3), offering a more nuanced understanding of excess stock returns.

The study by Lewellen (2014) utilises the Fama-MacBeth regression technique, introduced by Fama and MacBeth (1973), which is a two-step regression method to assess expected returns. First, betas for each asset are estimated. Then, these betas are used in a second set of cross-sectional regressions over various dates to examine the relationship between betas and stock returns. Lewellen (2014) simultaneously analyse the predictive capacity of various firm characteristics. The study focuses on 15 firm-specific features characterised by low-frequency data and reports a monthly predicted return of 0.74. This demonstrates that a select number of characteristics significantly impact expected stock returns.

Addressing ongoing critiques and identifying new determinants influencing excess stock returns, Fama and French (2015) introduced the Fama-French Five-Factor model (FF5). This expanded model includes two new factors, profitability and investment, in the foundational framework, aiming for a more comprehensive explanation of excess stock returns. Despite the improvements introduced by the FF5 model, it has limitations in comprehensively understanding the variations of excess stock returns across portfolios. Notably, they acknowledge that the FF5 model does not fully account for the lower average returns of smaller stocks. This is attributed to these smaller stocks exhibiting patterns similar to heavily invested firms despite their lower profitability.

Building upon these findings, Green et al. (2017) expand the scope of the investigation to 94 characteristics, applying the Fama-MacBeth regression method with corrections for biases, including data-snooping and the disproportionate influence of microcap stocks. Their study from 1980 to 2014 reveals that merely 12 characteristics serve as reliable stock return predictors. This confirms the findings of Lewellen (2014), who demonstrated that only a small fraction of investigated features have significant predictive power.

Despite the valuable insights derived from using the Fama-MacBeth regression, this method-

ology has shortcomings. One significant limitation is the method's capacity to analyse only a finite number of characteristics effectively. Moreover, the approach is predicated on the assumption that the relationship between predictors and stock returns is linear. This reliance on linearity is primarily motivated by ease of interpretation and computational efficiency rather than concrete empirical evidence supporting linear relationships in financial markets. Such a simplification overlooks the complex and potentially nonlinear interactions between various firm attributes and their impact on expected stock returns. Consequently, studies started to explore new nonlinear methodologies within empirical asset pricing.

Gu et al. (2020) advocate for the use of various ML methods to measure asset risk premiums. Contrary to the traditional factor models, ML can efficiently process a broad set of predictors and manage complex functional forms. However, the definition of ML remains fluid and tends to be context-dependent. This thesis adopts elements of Gu et al. (2020)'s definition of ML to describe: (i) A narrow spectrum of LSTM models aimed at statistical prediction. (ii) Techniques known as 'regularisation' for selecting models and reducing overfitting. (iii) Advanced algorithms that facilitate the exploration of numerous potential model configurations efficiently.

According to Gu et al. (2020), the growing interest in ML within empirical asset pricing stems from several factors: First, empirical asset pricing primarily focuses on understanding the varying expected returns across different assets and the market's equity risk premium. Determining an asset's risk premium involves predicting its future excess returns. Machine learning excels in predictive analytics, so it is exceptionally well-matched for this task.

Second, the landscape of empirical asset pricing is dotted with numerous stock characteristics and macroeconomic indicators that have been put forth as potential predictors. However, if many of these predictors are highly correlated, traditional prediction techniques face a challenge. This is where ML comes in, simplifying the analysis by focusing on variable selection and reducing data complexity.

Lastly, a significant challenge is the uncertainty about how predictors should influence risk premiums. Questions about modelling these relationships, incorporating nonlinear dynamics, and considering predictor interactions complicate the model-building process. With its range of techniques, from simple linear models to complex structures such as regression trees and

neural networks, ML is well-equipped to navigate these complexities and uncover intricate nonlinear patterns.

Gu et al. (2020) investigates nearly 30,000 individual stocks from 1957 to 2016. Their predictor set includes 94 stock characteristics, interactions of each characteristic with eight macroeconomic variables, and 74 industry sector dummy variables totalling more than 900 baseline signals. The study highlights the superior performance of traditional FNN models compared to linear regressions and tree-based models, particularly in achieving the highest Sharpe ratio. This thesis adopts parts of their framework, specifically using the same eight macroeconomic predictors, 94 stock characteristics, and companies listed on the same US stock exchanges. Moreover, this thesis incorporates aspects of their neural network approach and applies elements of their hyperparameter tuning method. Lastly, this thesis applies the same metrics and tests for economic contribution, such as constructing 10-1 hedge portfolios.

Gu et al. (2020) provide a foundational framework for calculating expected excess returns, establishing a general equation to guide such predictions:

$$r_{i,t+1} = E_t(r_{i,t+1}) + \epsilon_{i,t+1}, \tag{1}$$

where

$$E_t(r_{i,t+1}) = g(z_{i,t}). \tag{2}$$

In this equation, $g(\cdot)$ represents a nonlinear function, with stocks indexed as $i = 1, \ldots, N_t$, months by $t = 1, \ldots, T$, and the predictors noted as the P-dimensional vector $z_{i,t}$. A notable limitation of this approach emerges during out-of-sample forecasting, where $g(\cdot)$ depends neither on i nor t. In addition, $g(\cdot)$ depends on z only through $z_{i,t}$, which implies that the predictions do not leverage information from the history before t or from individual stocks other than the ith.

Given the limitations of FNNs, research by Naik and Mohan (2019) highlight the superior performance of LSTM models in forecasting excess stock returns. This advantage is attributed to the model's unique ability to incorporate an intermediate storage, known as the memory cell, during out-of-sample prediction. Hence, the LSTM model depends on both i and t, implying that its predictions leverage information from the individual stock and its history prior to t.

This is further reinforced by Gaur (2023), who emphasises the significance of historical data in predictive analysis. The LSTM models' ability to retain and utilise extensive sequences of past data allows for a more detailed understanding of individual stock's return dynamics. This feature is particularly valuable in stock market forecasting, where understanding historical trends and patterns can significantly improve the prediction of future excess returns.

However, upon closer examination of the application of LSTM models for excess stock return prediction, a notable disparity in the number of predictive features used becomes evident, especially when compared to Gu et al. (2020). For instance, K. Chen et al. (2015) employ ten features, Hansson (2017) focuses solely on historical returns, Moghar and Hamiche (2020) use only the opening price, Adila et al. (2022) concentrate on the closing price, Mi et al. (2023) incorporates seven features, and Gaur (2023) utilises a mere four features, encompassing opening, closing, highest, and lowest prices. This disparity underscores the significant gap in the complexity and depth of predictive features across different studies, suggesting potential areas for further exploration and refinement in using LSTM models for predicting excess stock returns.

Drawing from the literature and the results presented above, my contribution to the field is threefold: (i) I build on the work of Gu et al. (2020) by integrating five additional macroeconomic predictors. (ii) I extend the dataset period beyond that used by Gu et al. (2020) to include the years 2017 to 2020. (iii) I introduce another model, specifically the LSTM model, into the framework established by Gu et al. (2020). As a result, this approach integrates a total of 107 features, surpassing the number of features used in previous studies that also employ an LSTM model. This expansion not only enriches the model's predictive capacity but also sets a new benchmark in terms of feature complexity.

Based on my contributions, informed by the literature and the findings presented herein, the primary objective of this thesis is to assess the ability of LSTM models to forecast excess stock returns and achieve significant economic gains. Additionally, it aims to demonstrate the superior performance of LSTM models over traditional methods and to determine whether the LSTM outperforms the traditional FNN employed by Gu et al. (2020). To systematically address this question, the analysis is structured around three main hypotheses:

Hypothesis 1 (H1): *Employing permutation feature importance in conjunction with an LSTM model uncovers the most critical features for predicting excess stock returns.*

Hypothesis 2 (H2): *LSTM models exhibit superior accuracy in predicting excess stock returns compared to traditional feed-forward neural networks.*
Hypothesis 3 (H3): *A 10-1 hedge portfolio constructed using LSTM predictions generates significant economic gains, net of transaction costs.*

3 Theory

Section 3.1 establishes a baseline understanding of asset pricing models, such as the CAPM and APT. This provides the theoretical foundation for deriving an FF3 regression model to evaluate the long-short strategy. Subsequently, Section 3.2 explores the differences between FNNs and recurrent neural networks (RNNs), underscoring the enhanced ability of LSTM to manage the complex, nonlinear relationships present in financial datasets. Additionally, Section 3.3 addresses the role of permutation feature importance in identifying key predictors within LSTM models.

3.1 Traditional asset pricing models

The CAPM is an equilibrium model that assumes investors make asset allocation decisions based on a trade-off between expected returns and portfolio variance. This implies that each investor holds a mean-variance efficient portfolio, which optimises expected returns for a given level of risk. Furthermore, the CAPM asserts that the market portfolio, representing the aggregate of all individual portfolios, is also mean-variance efficient, assuming that all investors share identical expectations about the returns and (co)variances of individual assets and that there are no transaction costs, taxes, or trading restrictions (Verbeek, 2017). As a result, a linear relationship can be established between the expected excess returns of individual assets and the expected excess return of the market portfolio. Thus, the following equation holds that:

$$E[R_{i,t} - R_{f,t}] = \beta_i \cdot (E[R_{m,t}] - R_{f,t}), \tag{3}$$

where $E[R_{i,t}]$ signifies the expected return for asset i at time t, $R_{f,t}$ is a risk-free asset, and $E[R_{m,t}] - R_{f,t}$ represents the premium for market risk. The coefficient β_i is defined as

$$\beta_i = \frac{Covariance(R_{i,t}, R_{m,t})}{Variance(R_{m,t})}, \tag{4}$$

in which $R_{i,t}$ is the return on an individual stock, and $R_{m,t}$ is the market's overall return. β_i quantifies the extent to which changes in asset $i's$ returns are correlated with overall market

movements. It is a measurement of systematic risk or market risk. Investors are compensated for bearing this type of risk through a risk premium $E[R_{m,t}]-R_{f,t} > 0$ since it is impossible to eliminate systematic risk through portfolio diversification without reducing expected return (Verbeek, 2017).

Assuming rational expectations only for this derivation, which posits that the expectations of economic agents coincide with mathematical expectations, it is possible to derive a linear regression model from equation (3).

Initially, define the unexpected return of asset i as

$$u_{i,t} = R_{i,t} - E\{R_{i,t}\}.$$

Next, specify the unexpected returns of the market portfolio as

$$u_{m,t} = R_{m,t} - E\{R_{m,t}\}.$$

Equation (3) is now reformulated as

$$R_{i,t} - R_{f,t} = \beta_i(R_{m,t} - R_{f,t}) + \epsilon_{i,t}, \tag{5}$$

where

$$\epsilon_{i,t} = u_{i,t} - \beta_i u_{m,t}.$$

Equation (5) is a regression model without an intercept. The error term $\epsilon_{i,t}$ is a function of unexpected return, and it can be illustrated that it satisfies some minimal requirements for a regression error term. According to Verbeek (2017), following the definitions of $u_{i,t}$ and $u_{m,t}$, which is they both have a mean zero, i.e.,

$$E\{\epsilon_{i,t}\} = E\{u_{i,t}\} - \beta_i E\{u_{m,t}\} = 0. \tag{6}$$

In addition, it is uncorrelated with the regressor $R_{m,t} - R_{f,t}$. This can be inferred from the definition of β_i, which may also be written as

$$\beta_i = \frac{E\{u_{i,t}u_{m,t}\}}{Variance\{u_{m,t}\}},$$

and the outcome of

$$E\{\epsilon_{i,t}(R_{m,t} - R_{f,t})\} = E\{(u_{i,t} - \beta_i u_{m,t})u_{m,t}\} = E\{u_{i,t}u_{m,t}\} - \beta_i E\{u_{m,t}^2\} = 0, \tag{7}$$

where $R_{f,t}$ is not stochastic. This implies that any component of the asset's excess return that is correlated with the excess market return is captured by the term $\beta_i \cdot (R_{m,t} - R_{f,t})$.

Based on the above results, the OLS estimator provides a consistent estimator for β_i. This consistency arises because the error term has a zero mean and is uncorrelated with the regressor $R_{m,t} - R_{f,t}$. These conditions ensure no systematic bias in the estimates, implying that, on average, the estimates are accurate. Furthermore, assuming that $\epsilon_{i,t}$ does not exhibit autocorrelation or heteroskedasticity, the OLS estimates, standard errors, and tests are considered valid. This validity is attributed to the asymptotic result and the approximate distributional result provided by Verbeek (2017).

According to the APT, a stock's expected return, $r_{t,n}$, can be written as a linear combination of underlying factors, $f_{t,k}$:

$$r_{t,n} = \alpha_n + \sum_{k=1}^{K} \beta_{n,k} f_{t,k} + \epsilon_{t,n}, \tag{8}$$

where α_n is the constant to stock n, the error term related to the stock is $\epsilon_{t,n}$ at time t, and the sensitivity of stock n to factor k is shown by $\beta_{n,k}$.

The main difference between CAPM and APT is the introduction of multiple factors in the APT model that could influence an asset's returns beyond the market risk factor accounted for in the CAPM.

Expanding upon the CAPM's foundational principles and incorporating insights from the APT, the FF3 model offers a more nuanced approach to assessing excess asset returns. This model evolves from the singular focus on market risk in the CAPM to include two additional risk factors: size and value. The FF3 model posits that firms' size and book-to-market ratios significantly influence an asset's expected excess returns, thus challenging CAPM's assumption that market beta is the sole determinant of those excess returns.

The derivation of the CAPM regression, as shown in equation (5), can be extended to the FF3 model by incorporating the two extra risk factors. Adopting the notation from Verbeek

(2017), the FF3 model is expressed as follows:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_i(R_{m,t} - R_{f,t}) + b_s \cdot \text{SMB}_t + b_v \cdot \text{HML}_t + \epsilon_{i,t}. \tag{9}$$

Equation (9) is a regression model and is extended to include an intercept term, denoted as α_i. The market risk premium, $R_{m,t} - R_{f,t}$, is also referred to as *MKT*. The difference between small and big market capitalisation is denoted by *SMB*, and the difference between high and low book-to-market ratios is represented by *HML*. Moreover, b_s and b_v denote the sensitivity of the asset's excess return to the SMB and HML factors, respectively. A positive b_s suggests that the asset is weighted towards small-cap stocks, whereas a negative b_s indicates it is weighted towards large-cap stocks. Similarly, a positive b_v indicates that the asset is weighted towards value stocks, whereas a negative b_v suggests it is weighted towards growth stocks.

The intercept term, α_i, captures the portion of the asset's excess return not explained by its exposure to the three risk factors. The FF3 model is utilised in Section 5.4.3 to evaluate the long-short portfolio. If the portfolio generates excess returns above those predicted by the FF3 model, the long-short strategy has generated a positive α_i. Conversely, a negative α_i suggests that the long-short strategy has underperformed relative to the FF3 model.

3.2 From FNN to RNN: Theoretical justification for using LSTM

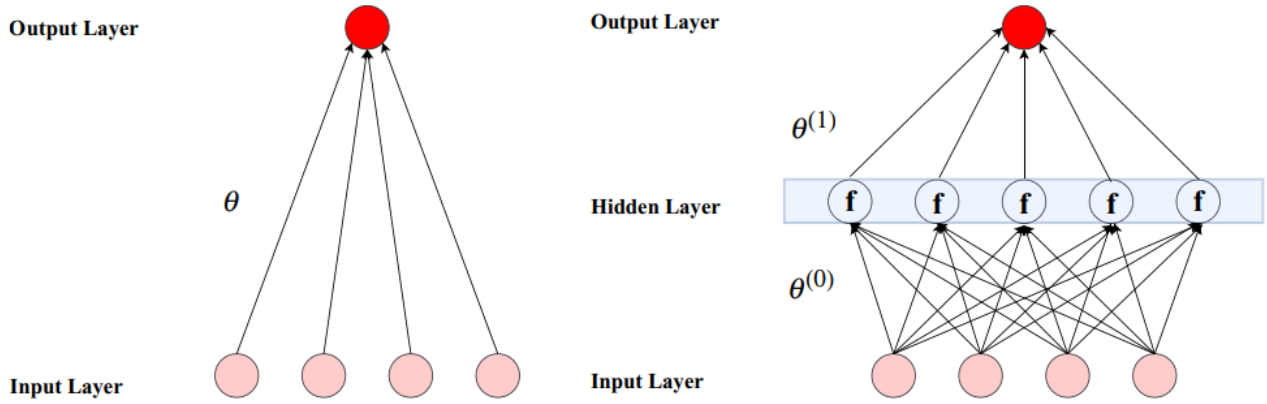
A traditional FNN is an artificial neural network in which connections between the nodes do not form a cycle. This is in contrast to an RNN, which enables information to circulate within the network. The FNN only moves in one direction: from the input layer, through any hidden layers, to the output layer. There are no backward connections from a node to any preceding nodes, preventing the formation of cycles. The output layer typically consists of one neuron for regression or several neurons for classification, while the number of units in the input layer equals the dimension of the predictors. As the fitted value in a regression analysis, the output layer is intended to forecast future values. However, due to their complexity, neural networks are among the most highly parameterised, least transparent, and challenging to interpret machine learning techniques (Scheuch et al., 2023).

Within FNNs, each neuron transmits its output to the subsequent layer following the application of a nonlinear activation function, denoted as f to the aggregated signal:

$$x_k^l = f\left(\theta_0^k + \sum_{j=1}^{N^l} z_j \theta_{l,j}^k\right).$$

Within this equation, z_j represents the input variables, which can either be the raw data or, in the case of several chained layers, the result from a previous layer $z_j = x_k - 1$. N^l denotes the number of units (a hyperparameter to tune). The parameters to fit are denoted by θ. Sigmoid or ReLu, defined in Section 4.2.1, are typical activation functions, although the simplest instance, $f(x) = \alpha + \beta x$, is a linear regression (Scheuch et al., 2023). Figure 1 illustrates two FNN models.

Figure 1: Neural network with or without a hidden layer and one output layer positioned to the right and left, respectively.



Source: Gu et al. (2020).
Note: The input layer is shown by pink circles, and the output layer is indicated by dark red circles. Every arrow has a weight parameter attached to it. A nonlinear activation function f in the network with a hidden layer modifies the inputs before sending them to the output.

The left panel presents the simplest neural network, which contains no hidden layers. The 5-dimensional parameter vector θ, comprising an intercept and one weight parameter per predictor, determines how each predictor signal is amplified or muted. The right panel depicts an example with one hidden layer that includes five units (Gu et al., 2020).

As visualised in Figure 1, the versatility of neural networks is derived from the chaining of multiple layers. This configuration introduces considerable freedom into the network's architecture, where clear theoretical guidance is yet to be established. The construction of a neural network requires, at a minimum, the specification of the number of units, the number of hidden layers, the connection structure of the units (whether dense or sparse), and

the application of regularisation techniques to mitigate the risk of overfitting. Finally, the *learning* process entails optimising the network parameters through numerical optimisation, a task often requiring careful calibration of the learning rate (Scheuch et al., 2023).

This thesis, however, proposes an alternative to the FNN approach employed by Gu et al. (2020). Instead, it advocates using an RNN, leveraging its capability for feedback loops within its architecture. Supporting this suggestion, Karmiani et al. (2019) finds that the LSTM outperforms the FNN in achieving higher accuracy and lower variance when predicting stock prices. However, Karmiani et al. (2019) notes that the LSTM requires more computational resources than the FNN. In contrast, Naeini et al. (2010) compared stock value predictions using both an RNN and an FNN model. They found that the FNN is more promising in predicting stock value changes than the RNN. However, the RNN is better at predicting the direction of the changes in the stock value.

The choice of using RNN models for excess stock return prediction in this thesis is underpinned by their inherent structural benefits for sequential data analysis, as highlighted by GfG (2023). FNNs process data in a unidirectional flow, from input to output, without feedback loops. On the other hand, RNNs have loops, allowing the output from a layer to be fed back into the input, enabling them to maintain a form of memory. However, FNNs' architecture is simpler and has a lower computational complexity. Therefore, it tends to operate faster.

Despite the increase in computational cost and the findings by Naeini et al. (2010), RNNs might offer advantages in capturing the temporal dependencies in stock data, which could be critical for precise excess stock return forecasts. Additionally, the selection of LSTM as the preferred RNN model is primarily due to its ability to address the issues commonly associated with traditional RNNs, notably the vanishing and exploding gradients problems that can hinder the learning process.

The vanishing gradients problem is characterised by the derivatives of the weights with respect to the loss function approaching zero. This poses a significant challenge in training RNNs, as vanishing gradients lead to extremely small weight updates during backpropagation through time, an algorithm used to update the weights and optimise RNNs. Backpropagation through time is similar to backpropagation in an FNN; however, due to the time dependency

in RNNs and LSTMs, it is important to unroll the network through time and apply back-propagation with time dependency (Hanu, 2021).

Conversely, the exploding gradients problem occurs when the derivatives of the weights with respect to the loss function become excessively large. As a result, substantial weight updates during the backpropagation through time occur, causing the network to fail to converge or even diverge. Hence, this leads to highly unstable training of the RNN. For an in-depth understanding of the vanishing and exploding gradients problem, please refer to the work of Bengio et al. (1994).

The architecture of LSTM models consists of input, forget, and output gates designed to capture and retain long-term dependencies in datasets. Based on the literature stating that stock returns demonstrate identifiable patterns and can be explained by stock characteristics and macroeconomic factors, this thesis suggests that the LSTM model is more effective than FNN for predicting excess stock returns.

The LSTM architecture applied in this thesis closely follows the framework established by Hochreiter and Schmidhuber (1997) and has proven effective in a wide range of LSTM applications, as demonstrated by Russakovsky et al. (2015), Silver et al. (2017), Wu et al. (2016), and M. X. Chen et al. (2018).

Figure 2 provides a detailed illustration of an LSTM cell. This cell operates on a batch of input vectors, denoted as $\mathbf{x}_t$, with dimensions $B \times I$, where B signifies the batch size and I represents the number of input features at each discrete time step $t = 0, 1, 2, ...$ Central to the LSTM cell's functionality are its input ($i_t \in \mathbb{R}^{B \times H}$), output ($o_t \in \mathbb{R}^{B \times H}$), and forget gates ($f_t \in \mathbb{R}^{B \times H}$), with H being the hidden units. These gates utilise both feed-forward and recurrent connections to modulate the flow of information within the network. The LSTM cell maintains two crucial memory components: The hidden state $\mathbf{h}_t$, of dimension $B \times H$, representing the short-term memory, and the internal cell state $\mathbf{c}_t$ of the same dimension, representing the long-term memory. These gates take in the current input $\mathbf{x}_t$ and the previous

hidden state $\mathbf{h}_{t-1}$ to compute:

$$i_t = \sigma(\mathbf{x}_t W_{ix} + \mathbf{h}_{t-1} W_{ih} + \mathbf{b}_i) \tag{10}$$

$$o_t = \sigma(\mathbf{x}_t W_{ox} + \mathbf{h}_{t-1} W_{oh} + \mathbf{b}_o) \tag{11}$$

$$f_t = \sigma(\mathbf{x}_t W_{fx} + \mathbf{h}_{t-1} W_{fh} + \mathbf{b}_f). \tag{12}$$

Here, σ denotes the sigmoid activation function applied element-wise, $\mathbf{b}_i$, $\mathbf{b}_o$, and $\mathbf{b}_f$ are bias vectors of size $B \times H$, and $W_{.x}$ and $W_{.h}$ are weight matrices of dimensions $I \times H$ and $H \times H$, respectively. These gate activations facilitate the update of the hidden and memory states according to:

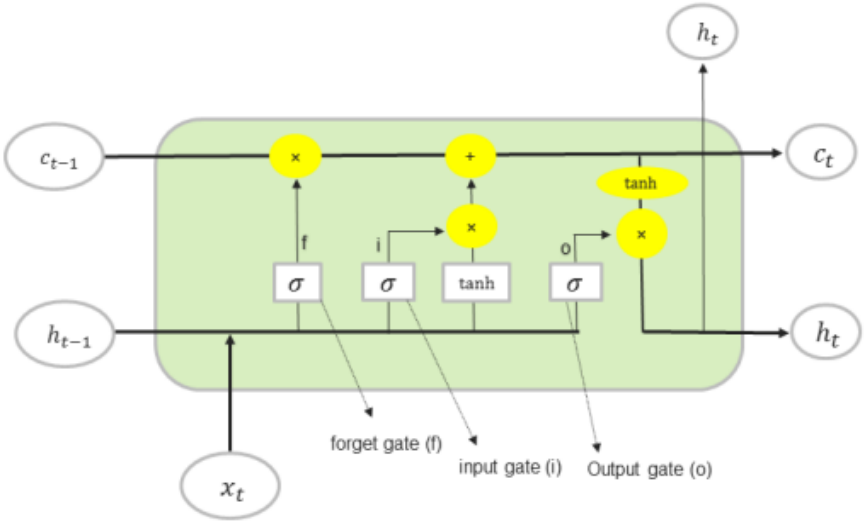
$$\hat{\mathbf{h}}_t = \mathbf{x}_t W_{hx} + \mathbf{h}_{t-1} W_{hh} + \mathbf{b}_{\hat{h}} \tag{13}$$

$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \odot \tanh(\hat{\mathbf{h}}_t) \tag{14}$$

$$\mathbf{h}_t = \tanh(\mathbf{c}_t) \odot o_t. \tag{15}$$

$\hat{\mathbf{h}}_t \in \mathbb{R}^{B \times H}$ denotes the input node, and $\mathbf{b}_{\hat{h}} \in \mathbb{R}^{1 \times H}$ is a bias parameter. The input gate i_t determines the extent to which new data are considered via $\hat{\mathbf{h}}_t$, while the forget gate f_t decides the proportion of the previous cell's internal state $\mathbf{c}_{t-1} \in \mathbb{R}^{B \times H}$ that is preserved. $\odot$ represents the Hadamard (elementwise) multiplication (Chalvatzis & Hristu-Varsakelis, 2019).

Figure 2: Detailed diagram of an LSTM cell



Source: Chalvatzis and Hristu-Varsakelis (2019).

Suppose the forget gate consistently remains at one and the input gate consistently at 0. In

that case, the internal state of the memory cell will perpetually remain the same, transferring unaltered to every following time step. However, the input and forget gates allow the model to learn when to maintain this value steadily and adjust it in reaction to subsequent inputs. This design, in practical terms, mitigates the issue of the vanishing and exploding gradients, leading to models that are simpler to train, mainly when dealing with datasets with extensive sequence lengths (A. Zhang et al., 2023).

Finally, the hidden state $\mathbf{h}_t \in \mathbb{R}^{B \times H}$ is computed by applying the *tanh* activation function to the internal state of the memory cell and subsequently performing elementwise multiplication with the output gate. The *tanh* activation function, defined as:

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{16}$$

ensures that the values of $\mathbf{h}_t$ lie within the range of (-1, 1). While the output gate approaches 1, the internal state of the memory cell can unrestrictedly affect the following layers. Conversely, when the output gate approaches 0, the current memory is effectively isolated from other network layers. This gating mechanism enables a memory cell to store information over extended periods without immediately affecting the network's overall state (as long as the output gate remains close to 0). However, if the output gate shifts from accepting values near 0 to accepting values near 1, the stored memory can abruptly impact the network at a later time step (A. Zhang et al., 2023). Note that this thesis utilises the ReLu activation function, detailed in Section 4.2.1, instead of the tanh activation function. This modification shifts the range of $\mathbf{h}_t$ values to $[0, \infty)$, while the remainder of the operations stays consistent.

In the case of deep networks composed of multiple LSTM layers, the hidden states from each layer serve as the input for the subsequent layer. This process is articulated through a subsequent application of equation (10) to (15), wherein the input variable $\mathbf{x}_t$ within Equations (10)-(13) is substituted by the hidden states produced by the preceding layer (Chalvatzis & Hristu-Varsakelis, 2019).

This thesis employs an LSTM framework for one-step-ahead predictions, implying that the model predicts a single value for each input sequence. The sequences are constructed from sliding windows, where each window ends with the target value that immediately follows the inputs. Thus, the LSTM processes a sequence of numerical vectors denoted as $x_{t-T+1}, ..., x_t$, where T denotes a fixed 'time window' size of three months. Specifically, the model leverages

data from the past three months to predict the fourth-month excess stock return.

The final LSTM layer outputs the hidden state $\mathbf{h}_t$, which is an x-dimensional vector where x depends on the number of neurons in that layer. This vector is fed through a linear fully-connected layer to produce a single prediction. The relationship can be expressed as:

$$\hat{y}_{t+1} = \mathbf{h}_t W_d + b_d,$$

where $\hat{y}_{t+1}$ is the predicted excess stock return for the following month, W_d is a weight matrix with dimensions $[x, 1]$, where x represents the number of neurons in the last LSTM layer and 1 being the number of output units in the dense layer. Lastly, b_d is a scalar bias term.

The sliding window approach maximises the use of all available data for training by continuously shifting the window across the dataset and using each segment of data multiple times in slightly different contexts. Furthermore, this helps maintain the temporal order of the data points, which is crucial for time series forecasting.

Within this thesis, the input feature vectors, $\mathbf{x}_t$, consist of monthly excess stock returns, 94 distinct stock characteristics, and 13 macroeconomic predictors. This set of inputs is utilised to determine the optimal LSTM architecture, ranging from LSTM1, with a single LSTM layer, to LSTM5, which contains five LSTM layers. Upon establishing the optimal number of layers, this thesis employs permutation feature importance to identify the top 20 most influential features. This analysis aims to streamline the input features, reducing them to monthly excess stock returns, 15 selected stock characteristics, and 5 macroeconomic predictors. Subsequently, the optimal model, leveraging these 20 most important features, is fitted and deployed for out-of-sample predictions. The methodologies underpinning this approach, including the model fitting and prediction processes, are elaborated in Section 4.

3.3 Permutation feature importance

This thesis employs the permutation feature importance (PFI) technique, originally introduced by Breiman (2001), to pinpoint the top 20 predictive features vital for forecasting excess stock returns. This method is further endorsed by Molnar (2020) and Deotte (2021). Moreover, Dami and Esterabi (2021) finds that combining PFI with LSTM forms a robust tool for assessing feature importance in financial time series predictions.

PFI determines the importance of individual features by quantifying their impact on the model's MSPE, as defined in Section 4.2.4 equation (17). It operates under the principle that the importance of a feature can be measured by calculating the increase in the MSPE, after permuting the feature. A feature is considered 'important' if shuffling its values increases the MSPE, indicating that the model relies on this feature for more accurate predictions. Conversely, a negligible change in the MSPE following a permutation suggests the feature is relatively unimportant. This thesis employs an algorithm, detailed in Algorithm 1, to estimate feature importance using the optimal LSTM model, trained exclusively on the training data.

Algorithm 1 Permutation feature importance

Initialisation:
Load the training dataset and divide into X_{train} (training features) and Y_{train} (targets).
The feature set X consists of 107 features, and Y represents the target variable, excess stock return.
Initialise parameters for K-fold cross-validation.
Model Selection:
Set n_{stocks} to 0.
while $n_{\text{stocks}} <= 500$ **do**
    Load a fitted model from the optimal LSTM.
    Increment n_{stocks} by 1.
    For each split generated by cross-validation:
        Divide data into $X_{\text{train}}, X_{\text{valid}}$ and $y_{\text{train}}, y_{\text{valid}}$.

    Feature Importance Computation:
    for each feature k in the feature set **do**
        Calculate the mean of the k-th feature across X_{train}.
        Replace the k-th feature in X_{valid} with its calculated mean.
        Predict outcomes using the modified X_{valid} and compute MSPE.
        Accumulate the MSPE for the k-th feature across all stocks and splits.
        Restore the original k-th feature values in X_{valid}.
    end for
end while

Source: Adapted from Deotte (2021).

The average increase in MSPE is computed for each feature in the training set across 500 stocks using K-fold cross-validation with the optimal LSTM structure. Subsequently, the

features are ranked in descending order of their importance to readily identify the most predictive features.

4 Methodology

This section presents the methodology employed in this thesis to conduct empirical asset pricing using LSTM models. Section 4.1 begins with a description of the data, including its sources and the methodology for preprocessing, which involves transformation and division into training and test sets. Subsequently, Section 4.2 details the selection process for the optimal LSTM network and its hyperparameters, explaining the rationale behind using K-fold cross-validation to ensure the statistical reliability and validity of the model's predictive accuracy. Section 4.3 outlines the 10-1 portfolios and the long-short investment strategy as a methodological approach to empirically assess the performance of the LSTM model and the selected features. Finally, Section 4.4 depicts the framework for including transaction costs to provide a comprehensive evaluation of the practical utility of the LSTM model.

4.1 Data description and source

This thesis examines monthly excess stock returns for all companies listed on the NYSE, AMEX, and NASDAQ, using data from the *Center for Research in Security Prices* (CRSP, 2024).[1] The dataset spans from 1960 to 2020, a total of 60 years, and includes approximately 3.2 million observations for excess stock returns before any transformations. The sample comprises 24,851 stocks, with an average of 4,407 stocks per month. The dataset's time-frame aligns with that of Gu et al. (2020) but extends to 2020, four years beyond their study.

The data filters applied in this thesis follow the methodology of Gu et al. (2020), which includes: (i) retaining data exclusively from the specified periods of interest, (ii) limiting the dataset to US-listed stocks, denoted by share codes 10 and 11, and (iii) including data only from months within the unique start and end dates.

Additionally, this thesis includes stocks priced below 5 USD, ensuring that the research findings are applicable across a wider spectrum of securities. This enhances the thesis's relevance to both low-priced and higher-priced stocks. Moreover, excluding stocks based on their price could introduce sample selection bias, potentially skewing the results and omitting significant

[1] Table 10 in the Appendix shows that the normalised ret_excess variable is stationary.

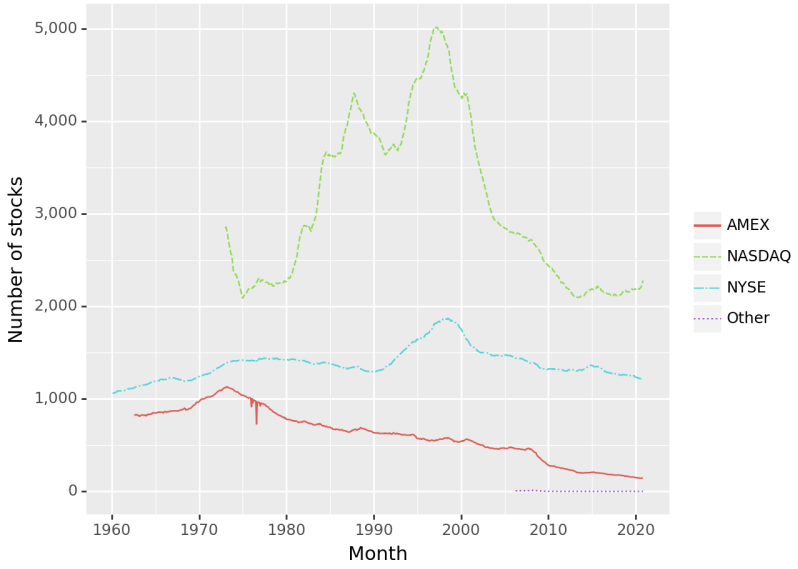
market behaviours.

This thesis utilises the one-month Treasury bill rate from Prof. Kenneth French's finance data library, Fama/French Factors French (2024), as a proxy for the risk-free rate. This rate is used to calculate the individual excess returns. The dataset also includes the market's excess return time series, denoted as $R_{m,t} - R_{f,t}$, representing the value-weighted return of all CRSP firms incorporated in the US and listed on the NYSE, AMEX, or NASDAQ. The Fama/French factors are constructed using six value-weighted portfolios formed on size (SMB), defined as the average return on the three small portfolios minus the average return on the three big portfolios, and book-to-market (HML), defined as the average return on the two value portfolios minus the average return on the two growth portfolios. These factors are used to examine excess portfolio returns in Section 5.4.3.

Figure 3 illustrates the monthly number of securities by listing exchange from 1960 to 2020. The NYSE listing shows a relatively stable trend with a slight increase until the mid-1990s, followed by a modest decline into the 2020s. In contrast, the NASDAQ exhibits more volatility, with a significant surge in the number of securities in the late 1990s, peaking around 2000 before a steep decline. The AMEX shows a moderate decrease from the mid-1970s to 2020. By the end of 2020, the NYSE consisted of 2,281 stocks, the NASDAQ 1,237, and the AMEX 145, with only one stock belonging to the Other category.

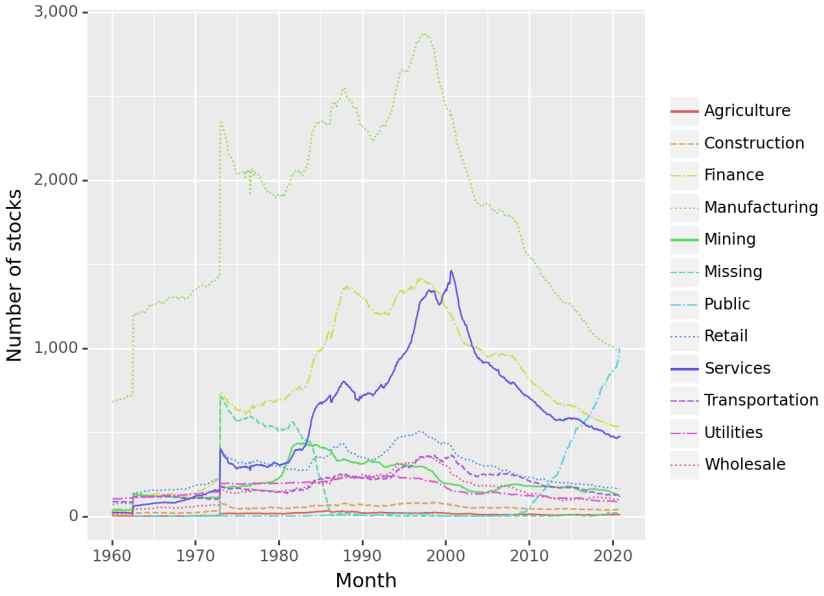
Figure 4 presents the distribution of stocks across various industries within the sample. This is achieved by translating industry codes into descriptive industry names, following the methodology proposed by Bali et al. (2016). The manufacturing industry emerged as the dominant industry for most of the examined period, peaking in the late 1990s with nearly 3,000 stocks. However, a significant downturn was observed after this period, resulting in approximately 1,000 stocks by 2020, mirroring the size of the public industry. The finance and services industries were the second and third largest, respectively, during most of the period. However, both industries experienced a decrease in the number of stocks after the 2000s and were smaller than the public industry by the end of 2020.

Figure 3: Monthly number of stocks by listing exchange from 1960 to 2020



Source: CRSP (2024).

Figure 4: Monthly number of stocks by industry from 1960 to 2020



Source: CRSP (2024).

Following the methodology of Gu et al. (2020), this thesis employs the same 94 stock characteristics derived from Green et al. (2017). These characteristics are extracted from the *Compustat* and CRSP (2024) databases. In addition, 13 macroeconomic predictors are utilised. Welch and Goyal (2008) provide a comprehensive reassessment of the macroeconomic indi-

cators suggested by academic research as reliable predictors of the equity premium. The data are hosted on Goyal (2023)'s website for reference. These features are discussed in more detail in Section 4.1.2.

4.1.1 Transformation and split

The dataset, spanning 60 years, is divided into two subsets. A training subset covering 48 years (1960−2008) and a subsequent 12-year period (2009−2020) designated for out-of-sample testing. This division corresponds to an 80/20 split.

After splitting the dataset, it is crucial for the LSTM model that each stock has enough training data and a sufficient span for out-of-sample testing. Therefore, stocks with less than ten years of training data and fewer than three months of testing data are excluded. However, this may raise concerns of look-ahead bias, as discussed in Section 6.1.1. This filtration process reduces the dataset from 24,851 stocks to 2,552, leading to a decrease in the number of excess stock return observations from approximately 3.2 million to 974,207. Despite the reduction in observations, the results retain their generalised value, ensuring the findings' applicability across a broad spectrum of stocks.

Excluding stocks with less than three months of testing data is necessary due to the input requirements of the LSTM models. The input data for the LSTM model consists of two parts: X and Y. Specifically, X comprises the excess stock returns, stock characteristics, and macroeconomic predictors from the previous three months, whereas Y denotes the excess stock return for the following month. The models are rigorously trained on a dataset covering the past 48 years, which includes numerous instances of these X-Y pairs. This extensive training enables the LSTM to effectively discern patterns and trends in the fluctuations of excess stock returns over three-month intervals. Consequently, after being fed data from the preceding three months, the model is adept at forecasting the excess return for the following month.

4.1.2 Stock characteristics and macroeconomic predictors

This thesis examines 94 stock characteristics, detailed in Appendix Table 8 and categorised into four groups following the classification by Gu et al. (2020). The first group includes variables related to recent price movements, such as short-term reversal (**mom1m**), stock momentum (**mom12m**), change in momentum (**chmom**), industry momentum (**indmom**), highest recent return (**maxret**), and long-term reversal (**mom36m**). The second group

comprises liquidity-related variables, including turnover and its volatility (**turn**, **SD_turn**), logarithm of market equity (**mvel1**), dollar trading volume (**dolvol**), and Amihud's measure of illiquidity (**ill**). The third group encompasses risk measures, such as overall and specific return volatility (**retvol**, **idiovol**), market beta (**beta**), and the square of beta (**betasq**). The final group is dedicated to valuation ratios and fundamental indicators, featuring earnings-to-price (**ep**), sales-to-price (**sp**), asset growth (**agr**), and the count of recent earnings increases (**nincr**).

This thesis employs the same normalisation technique as Gu et al. (2020), normalising the characteristics to a (-1,1) interval. To avoid look-ahead bias, it is crucial to obtain the mean and standard deviation from the training set for normalising the values in the testing set. Furthermore, this thesis follows the same assumptions as Gu et al. (2020), that monthly stock characteristics are delayed by at most one month, quarterly with at least four months lag, and annual with at least six months lag. To predict excess stock returns for the period $t+1$, the LSTM models utilise the most recent monthly characteristics at the end of month t, the most recent quarterly data by the end of month $t-4$, and the most recent annual data by the end of month $t-6$, as well as the two preceding periods for each characteristic. Another challenge is missing characteristics, which this thesis addresses by replacing them with the cross-sectional median for each month and stock, respectively.

The 13 macroeconomic predictors and the excess stock returns used in this thesis are normalised to a median value of zero, following the methodology of Gu et al. (2020). The macroeconomic predictors include:

- **Dividend Price Ratio (dp):** Measures the difference between the logs of dividends and prices, where prices are monthly averages of daily closing prices and dividends are 12-month moving sums of dividends paid on the S&P 500 index (Campbell & Shiller, 1988; Campbell & Yogo, 2006).
- **Dividend Yield (dy):** Represents the difference between the logs of dividends and lagged prices (Ball, 1978).
- **Earnings Price Ratio (ep):** Calculated as the difference between the logs of earnings and prices, where earnings are 12-month moving sums of earnings on the S&P 500 index (Campbell & Shiller, 1988).
- **Dividend Payout Ratio (de):** Defined as the difference between the logs of dividends and earnings (Lamont, 1998).
- **Stock Variance (svar):** The sum of squared daily returns on the S&P 500 (Guo, 2006).
- **Book to Market Ratio (bm):** The ratio of book value to market value for the Dow Jones Industrial Average (Kothari & Shanken, 1997).

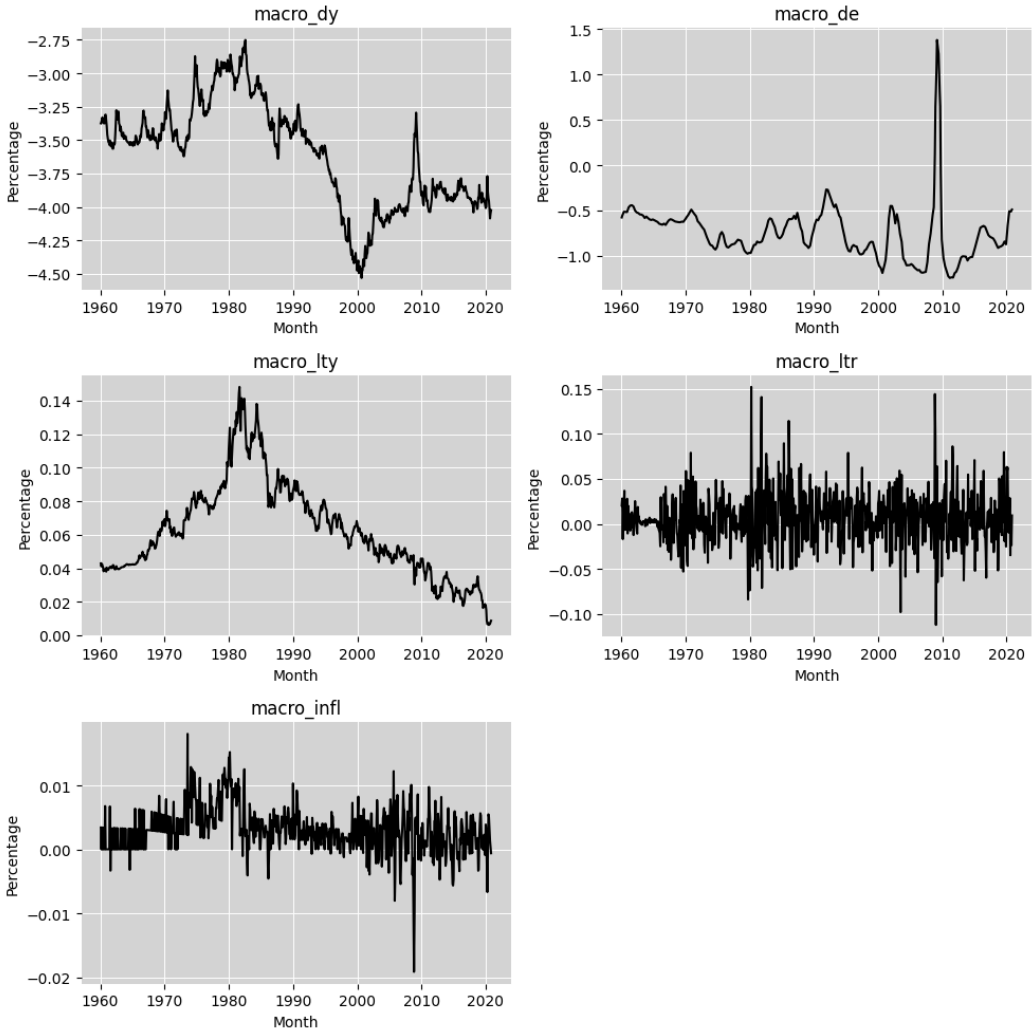
- **Net Equity Expansion (ntis):** Reflects market breadth, calculated from 12-month moving sums of net issues by NYSE listed stocks divided by the total end-of-year market capitalisation of NYSE stocks (Campbell et al., 2008).
- **3-Month Treasury Bill (tbl):** Secondary Market Rate from the economic research database at the Federal Reserve Bank at St. Louis (Campbell, 1987).
- **Long Term Yield (lty):** Yield on long-term government bond based on Ibbotson's Stocks, Bonds, Bills, and Inflation Yearbook (Welch & Goyal, 2008).
- **Long Term Rate of Returns (ltr):** Returns on long-term government bond based on Ibbotson's Stocks, Bonds, Bills, and Inflation Yearbook (Welch & Goyal, 2008).
- **Term Spread (tms):** Defined as the difference between the long-term yield on government bonds and the T-bill rate (Campbell, 1987).
- **Default Yield Spread (dfy):** Represents the difference between yields on BAA- and AAA-rated corporate bond yields (Fama & French, 1989).
- **Inflation (infl):** Denotes the Consumer Price Index (All Urban Consumers) based on the Bureau of Labor Statistics (Campbell & Vuolteenaho, 2004).

This thesis expands upon the macroeconomic predictors used by Gu et al. (2020), incorporating dividend yield (**dy**), dividend payout ratio (**de**), long-term yield (**lty**), long-term rate of returns (**ltr**), and inflation (**infl**). Figure 5 visualises these predictors.

The first plot, 'macro_dy', displays the dividend yield over time, showing some volatility and a general downward trend from 1983 to 2020. As Ang and Bekaert (2006) suggest, dividend yield is positively correlated with one-month ahead excess returns, a relationship that weakens or reverses over longer periods. This correlation is particularly relevant to this thesis, focusing on short-term, one-month-ahead predictions. The second plot, 'macro_de', represents the dividend payout ratio. This series appears to be more stable. However, it shows a significant spike from 2008 to 2009, potentially indicating economic events that substantially impacted earnings yield, such as the 2008 financial crisis. Research by Da et al. (2014) suggests combining dividend yield with earnings yield can enhance stock return predictions. The third plot, 'macro_lty', depicts the long-term yield, which peaked around the 1980s before a pronounced long-term decline. Bianchi et al. (2021) indicate that neural networks can effectively harness macroeconomic factors such as long-term yields to predict bond returns. The 'macro_ltr' plot illustrates the long-term rate of returns, characterised by high volatility and no clear trend over the entire period. Keim and Stambaugh (1986) posits that variables such as long-term bond returns can forecast returns across various asset types, suggesting a robust relationship between bond and stock returns. The final plot, 'macro_infl', represents inflation, showing periods of high volatility, particularly noticeable during the 2008 financial crisis. M. Z. Zhang (2021) demonstrates an inverse relationship between real stock

returns and inflation, indicating that inflation can help explain the fluctuations in stock returns.

Figure 5: Monthly macroeconomic predictors from 1960 to 2020



Source: Welch and Goyal (2008).
Note: The figure present the dividend yield (**dy**), dividend payout ratio (**de**), long-term yield (**lty**), long-term rate of returns (**ltr**), and inflation (**infl**).

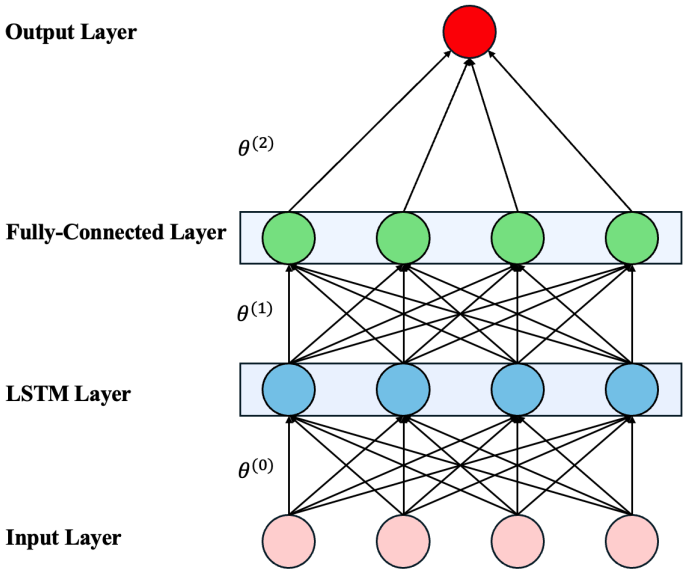
4.2 Model selection: LSTM hyperparameters and validation

According to Gu et al. (2020), identifying an optimal network architecture through cross-validation poses significant challenges due to the impracticality of examining an infinite number of architectures. This thesis predefines a selection of network architectures to address this issue, and it assesses each one within a subset of 500 stocks from a total of 2,552.

The objective is to establish a performance benchmark for five LSTM models of varying complexity, subsequently selecting the most suitable model for making predictions across all 2,552 stocks. The LSTM models considered range from one to five LSTM layers. The simplest model, LSTM1, consists of one LSTM layer with 32 neurons. Progressively, LSTM2 incorporates two LSTM layers with 32 and 16 neurons, LSTM3 includes three layers with 32, 16, and 8 neurons, LSTM4 with four layers containing 32, 16, 8, and 4 neurons, and the most complex, LSTM5, features five layers with 32, 16, 8, 4, and 2 neurons, respectively. This approach adheres to the geometric pyramid rule outlined by Masters (1993) and is also adopted by Gu et al. (2020), who advocate for a pyramidal configuration of neurons from the input layer to the output. This configuration facilitates effective information compression and pattern abstraction without an excessive number of parameters.

Figure 6 depicts the simplified proposed model structure containing 13 units. First, the input data are fed into the first LSTM layer. The LSTM units selectively process the data within the LSTM layer, determining which information is important. The model structure can consist of multiple LSTM layers, where a linear fully-connected layer is set atop the last LSTM layer, ensuring inputs from all preceding layer units. Consistent with the FNN, θ denotes the parameters to fit, where each arrow is associated with a weight parameter.

Figure 6: Proposed LSTM network structure with one LSTM layer



Source: Own visualisation.
Note: Pink circles show the input layer, and the dark red circle indicates the output layer. The blue circles represent the LSTM unit depicted in Figure 2, which processes the inputs before sending them to the fully-connected layer represented by the green circles.

By analysing the performance differences between LSTM1 and LSTM5, this thesis aims to shed light on how varying network depths influence the accuracy of excess return forecasts. All estimates in this thesis share the same objective of minimising the MSPE, as defined in equation (17).

4.2.1 Activation function

In the LSTM cell depicted in Figure 2, sigmoid activation functions are employed within the input, forget, and output gates to modulate the flow of information, while the tanh activation function normalises the cell state updates and computes the output from the cell state in conjunction with the output gate. However, this thesis proposes a modification, substituting tanh with the Rectified Linear Unit (ReLU) activation function, which is expressed as:

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{otherwise.} \end{cases}$$

As suggested by Gu et al. (2020), this modification aims to promote neuron activation sparsity and expedite the computation of derivatives. Characterised by a linear response for positive inputs and zero for negative inputs, the ReLU gradient remains constant, facilitating quicker learning processes (K. & K., 2022). Conversely, the gradients of the sigmoid activation function diminish as the input values increase. The sigmoid function is defined as:

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

It smoothly transitions between 0 and 1 in an S-shaped curve. According to Keras (2021), sigmoid outputs are approximately zero for inputs less than -5 and approach one for inputs greater than 5.

The decision to pair ReLU with sigmoid in this thesis is further supported by K. and K. (2022), who indicate negligible performance disparity between the conventional tanh-sigmoid pairing and the alternative ReLU-sigmoid pairing. Therefore, given ReLU's computational efficiency, this thesis leans towards the ReLU-sigmoid configuration.

4.2.2 Optimiser

Given that the LSTM networks in this thesis contain a large number of parameters, finding a closed-form solution for these parameters would be impractical and computationally

intensive. Therefore, this thesis adopts the Adaptive Moment Estimation (Adam) optimiser presented by Kingma and Ba (2014) to approximate a solution. Adam is an algorithm employed to adjust the weights and biases of the network to minimise the loss function (MSPE in this thesis). By modifying the parameters to minimise the MSPE, the optimiser enhances the network's ability to learn from the training data and improves its out-of-sample prediction accuracy. According to the authors, the optimiser is computationally efficient, requires little memory, and is well suited for large datasets and numerous parameters.

The Adam optimiser combines the advantages of two other extensions of stochastic gradient descent, the Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp). AdaGrad and RMSProp maintain a learning rate for each parameter and improve performance on problems with sparse gradients, and nonstationary problems (e.g., noise), respectively. Adam works by computing individual learning rates for different parameters based on estimates of the first and second moments of the gradients (Kingma & Ba, 2014). The first moment (the mean), averages out the gradients to smooth the update directions. The second moment (uncentered variance) tracks the squared gradients and helps modify the learning rate for each parameter based on how much variability in gradients the parameter experiences. Initially, the first and second moments of the gradients are vectors of zeros, which implies that they are biased towards zero. Adam corrects this by calculating bias-corrected first and second-moment estimates, which are then used to update the parameters. As a result, the Adam optimiser includes the following configuration parameters: The learning rate α, where tuning α involves a trade-off between the accuracy of finding the minimum and computational demand. Lower learning rates lead to slower convergence towards the minimum, while larger learning rates result in faster initial learning but increase the risk of overshooting the minimum. The parameters β_1 and β_2 are the exponential decay rates for the first and second-moment estimates, respectively. ϵ is a small number to prevent any division by zero in the implementation. As recommended by Kingma and Ba (2014), good default settings for the configuration parameters are $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$.

Another significant optimiser, Stochastic Gradient Descent (SGD), operates on the principle of gradient descent. It updates model weights using selected data subsets, thereby enhancing computational efficiency for large datasets, as employed by Gu et al. (2020). Nonetheless, utilising the Adam optimiser in the LSTM model's training phase has resulted in superior predictive accuracy compared to SGD. Thus, the Adam optimiser is the preferred choice.

4.2.3 Regularisation

This thesis employs early stopping and L1 regularisation techniques to mitigate overfitting, which enhances the model's predictive performance on unseen data. Early stopping initiates the optimisation process with a large number of training epochs and ceases training once the model's performance no longer improves. This is achieved by iteratively updating the number of epochs to minimise the MSPE within the training dataset. Simultaneously, the model's performance is evaluated using K-fold cross-validation. The optimisation process is halted once errors on the cross-validation sets begin to increase, typically before the training set's prediction errors reach their minimum. This approach effectively determines the optimal number of iterations for training the model. Early stopping serves as a computationally efficient alternative to L2 regularisation, also known as Ridge Regression in the elastic net, by terminating the optimisation process early. This avoids the exhaustive computation associated with full optimisation under each tuning parameter guess, as suggested by (Gu et al., 2020). Early stopping can be utilised as a standalone regularisation method or in conjunction with L1 regularisation. In this thesis, it is used alongside L1 regularisation to achieve a balanced regularisation effect with reduced computational demands.

The L1 regularisation, also known as lasso regression in the elastic net, is a technique that adds a penalty equal to the absolute value of the magnitude of coefficients to the loss function. The objective is to minimise the combined loss function, which comprises the original loss (the MSPE in this thesis), plus the L1 penalty term. This introduces a new variation of the loss function that now includes the regularisation term:

$$L(\theta) = MSPE(\theta) + \lambda \sum_{i=1}^{n} |\theta_i|,$$

where L is the loss function, λ is the regularisation parameter, $\sum_{i=1}^{n} |\theta_i|$ represents the sum of the absolute values of the parameters θ (or weights), and n is the number of parameters (Moon, 2018). A higher value of λ applies more penalty, pushing more coefficients to become zero, thus leading to a sparser model. A sparse model is one in which only a few features have non-zero coefficients. If $\lambda = 0$, the regularisation term has no effect, and the L1 regularisation reverts to the original loss function.

4.2.4 K-fold cross-validation and hyperparameter tuning

Cross-validation is a statistical method employed in this thesis to estimate the performance of machine learning models. It divides the original dataset into a training set for model training and a validation set for model assessment. This aids the machine learning models in maintaining their predictive accuracy on new, unseen data, a crucial process for dealing with hyperparameter tuning.

Hyperparameter tuning is the process of identifying the optimal values of given parameters for a learning algorithm. An example is the selection of the optimal L1 penalty parameter, as detailed in Table 1. The table presents the range and values of the hyperparameters adjusted during the tuning of the LSTM models. The strength of L1 regularisation, denoted as

Table 1: Hyperparameters for LSTM1-LTM5

Parameter	LSTM1-LSTM5
L1 penalty	$\lambda 1 \in (10^{-8}, 10^{-3})$
Batch Size	8, 16, 32, 64, 128
Epochs	200
Patience	15
Adam Optimiser	Default

Source: Own setup.

$\lambda 1$, spans from 10^{-8} to 10^{-3}. The batch size, which reflects the number of training samples processed in one iteration, varies across the values 8, 16, 32, 64, and 128. This variation helps assess the model's impact on both the efficiency of the learning process and the computational demands placed on the system. The initial number of epochs is set to 200, where one epoch refers to one complete pass through the entire training data. A patience level of 15 epochs is set in conjunction with early stopping. Lastly, the Adam optimiser is employed in its default setting.

Given the use of time series data in this thesis, preserving the temporal order of the observations is essential. Therefore, K-fold cross-validation is utilised, which approximates the true Mean Squared Prediction Error (MSPE) by creating predictions for K new samples of the data, none of which are used to train the algorithm:

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{T}\sum_{t=1}^{T}\left(y_t^k - \hat{y}_t^k\right)^2. \tag{17}$$

In practical terms, the dataset is initially divided into K folds of approximately equal sizes.

Subsequently, K rounds of training and validation are carried out. In each round, a different fold of the data is set aside for validation, while the remaining $K-1$ folds are utilised for training. Then, the MSPE derived from the validation set serves as an indicator of the model's predictive performance. To identify the best hyperparameter settings, the training data are subdivided into multiple parts. For each set of potential hyperparameters (e.g., $\lambda 1$), the model is trained, and its predictive accuracy is assessed using a separate sample (Scheuch et al., 2023).

The procedure unfolds as follows: First, a grid of hyperparameters, as seen in Table 1, is specified. Second, predictors $\hat{y}_i \lambda 1$ are collected for the used parameters $\lambda 1$. Third, the following is computed:

$$\text{MSPE}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{T} \sum_{t=1}^{T} \left(y_t^k - \hat{y}_t^k(\lambda 1) \right)^2 .$$

In K-fold cross-validation, this calculation is repeated K times. The validation set, consisting of $M = T/K$ observations, is randomly selected, with the random samples being $y_1^k, \ldots, y_T^k$, with $k = 1$ (Scheuch et al., 2023).

The value of K depends on the given problem one is trying to solve. Preferably, higher values of K are chosen to ensure the training data better represents the original dataset. However, this benefit comes at the cost of substantially increased computational time. This thesis considers tuning the models and using PFI with a training period of five years and a validation period of three years. The number of splits differs among the stocks, as the quantity of observations varies for each stock.

Each of the five LSTM models, from LSTM1 to LSTM5, is evaluated using identical K-fold cross-validation settings. The model that exhibits the lowest MSPE is preferred as the optimal model for out-of-sample prediction. However, visualisations of the models fitted to the training sample are also provided to enhance the robustness of the model selection process.

4.3 10-1 PORTFOLIOS

Out-of-sample predictions are performed upon selecting the optimal LSTM network with the top 20 features. These predicted excess stock returns are used to sort stocks into monthly

deciles representing portfolios. Hence, stocks with the lowest expected excess returns are assigned to the first portfolio and those with the highest to the tenth portfolio. The value-weighted portfolio returns are computed using lagged market capitalisation as a criterion for weight allocation among the stocks within each portfolio. This thesis believes that employing market capitalisation weighting provides a more realistic perspective in an investment context. This method ensures that stocks with larger market capitalisation receive a higher weighting in the portfolios, reflecting their significant market presence and stability as revenue generators. However, a limitation is the diminished capacity of smaller firms to mitigate the impact of underperformance by larger companies due to their lower weight allocation.

A long-short strategy is implemented to evaluate portfolio performance. This involves purchasing stocks in the tenth portfolio while shorting those in the first. The realised monthly average excess return and standard error are calculated alongside the Sharpe ratio and the application of the FF3 model. These metrics provide insights into the strategy's potential to generate economic gains. In addition, this analysis includes turnover and transaction costs to assess the trading frequency and the financial implications of the strategy.

4.4 Transaction costs

Portfolio rebalancing entails significant costs, necessitating informed decisions based on an investor's existing holdings. The trade-off between the benefits of wealth reallocation and the expenses from portfolio turnover becomes crucial when transaction costs are non-negligible. Inspired by Hautsch and Voigt (2019), this thesis incorporates transaction costs using the following approach to strengthen the evidence supporting the potential gains derived from the long-short strategy.

The initial phase involves monitoring the weight of each holding within the portfolio before rebalancing, denoted as ω_t. If a holding is absent in the current assessment, ω_t is assigned a value of 0, signifying its exclusion from the portfolio at that moment. Subsequently, the change in the weight of each holding is determined by calculating the squared difference between the forthcoming allocation ω_{t+1} and the current holdings ω_t:

$$\nu\left(\omega_{t+1}, \omega_t, \beta\right) = \frac{\beta}{2}\left(\omega_{t+1} - \omega_t\right)^2. \tag{18}$$

$\beta > 0$ represents a predetermined cost parameter. β is measured in basis points (bp), and

according to Hautsch and Voigt (2019), $\beta < 100$ can be associated with small transaction costs. In the literature, a typical value for β is 50 bp, see, for instance, DeMiguel et al. (2009), and Olivares-Nadal and DeMiguel (2018). The division by 2 accommodates the bidirectional nature of transaction costs, capturing expenses from both buying and selling. Transaction costs penalise portfolio performance during the transition from existing holdings ω_t to a new allocation ω_{t+1}. In this framework, transaction costs escalate in a nonlinear manner. Specifically, substantial rebalancing efforts incur greater penalties than minimal adjustments, a concept accentuated by squaring the difference in allocations. This thesis adjusts the realised excess returns for transaction costs by subtracting equation (18) from the realised excess returns.

5 An empirical study of US stocks

This section begins with a comprehensive evaluation of the most effective model architecture, as outlined in Section 5.1. It involves conducting a performance comparison of the LSTM1 through LSTM5. Additionally, a visual representation of the predictive outcomes is provided to facilitate a more informed selection of the optimal network structure. Subsequently, Section 5.2 analyses all 107 predictive factors to identify the 20 most critical features that impact the prediction of excess stock returns. Following this, Section 5.3 applies the optimal LSTM structure to evaluate its out-of-sample performance, comparing it to a baseline univariate LSTM with an identical structure. This comparison seeks to determine to what extent the inclusion of additional features enhances the model's performance and improves out-of-sample prediction accuracy. Section 5.4 utilises these out-of-sample predictions to construct 10-1 portfolios, adopting a long-short strategy. Lastly, Section 5.5 provides an economic perspective on the rationale behind the long-short strategy and the LSTM model's performance, considering both the portfolio patterns and industry characteristics.

5.1 Optimal model selection

This thesis assesses the most effective model architecture among LSTM1, LSTM2, LSTM3, LSTM4, and LSTM5. Initially, this involves an examination of the MSPE across all models, utilising K-fold cross-validation on a representative subset of 500 stocks. The selection of this subset is necessitated by the substantial computational resources required for a full-scale analysis. Nonetheless, this subset is deemed sufficiently representative to accurately identify the superior LSTM model. The MSPE scores for each model are summarised in Table 2.

Table 2: MSPE for the five LSTM models during the training period

	LSTM1	LSTM2	LSTM3	LSTM4	LSTM5
MSPE	0.865	0.870	0.774	0.660	0.491

Source: Own calculations.

The table shows that LSTM1 has an MSPE of 0.865, LSTM2 with 0.870, LSTM3 scoring 0.774, LSTM4 at 0.660, and LSTM5 with the lowest MSPE of 0.491. A lower MSPE score indicates a model with a better fit to the training data, thus suggesting that LSTM5 outperforms the other models in terms of predictive accuracy.

To further comprehend how the five LSTM models have adapted to the training data, visualisations for each model are provided in Figure 7. This helps determine whether the models possess the necessary capacity to reflect the complexity of the training data. If a model's predictions are overly simplistic or fail to capture the data's variability, this could be a sign of underfitting, indicating a model that is too simplistic. Conversely, if the model captures noise more than the actual trends, this might suggest overfitting and a need for regularisation to correct this imbalance.

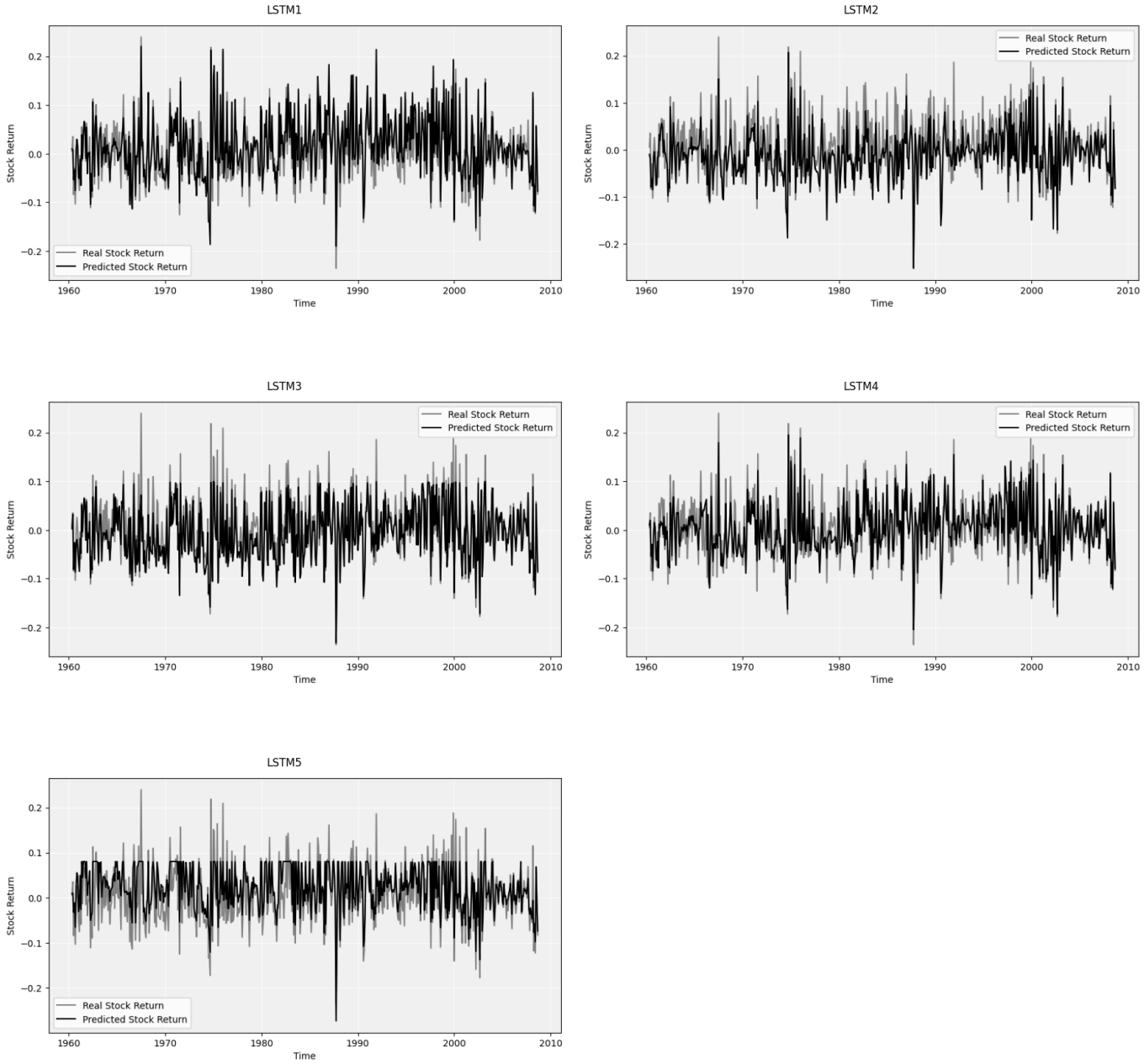
The initial plot in Figure 7 illustrates LSTM1's predicted (black) and realised excess stock returns (grey), adeptly capturing significant fluctuations in the stock return series with minimal periods of deviation. This model captures the data's variability, suggesting a nuanced understanding rather than a simplistic interpretation of the trends. However, an underlying concern regarding its precision arises; the model's close adherence to the data may indicate overfitting, as it could be assimilating noise in addition to underlying patterns.

In contrast, LSTM2 demonstrates less alignment, particularly noticeable during the peaks of the stock's performance, indicating a somewhat superficial understanding of these dynamics. However, the model's deviation from complete alignment with the realised excess returns may indicate a more generalised LSTM model, potentially avoiding the pitfalls of overfitting.

LSTM3 exhibits a marginally improved alignment over LSTM2, yet it does not fully capture the stock's peak behaviours. Additionally, there is a tendency to capture noise, particularly at the lower end of the returns, suggesting a susceptibility to overfitting similar to LSTM1.

The predictions of LSTM4 closely mirror those of LSTM2, suggesting a more realistic modelling of realised excess returns. While certain periods display poor alignment, LSTM4 appears to avoid the excessive noise characteristic of LSTM1 and LSTM3, indicating a more balanced fit to the data.

Figure 7: Monthly LSTM training results depicting one stock



Source: CRSP (2024), and own predictions.

Lastly, LSTM5 is characterised by its diminished performance, primarily due to its inability to accurately interpret patterns during significant spikes in the excess stock returns. This underscores the model's limitations in adapting to complex market dynamics and illustrates the principle that excessive complexity within a model can negatively impact its performance.

Although LSTM5, with the lowest MSPE of 0.491, suggests high accuracy, it does not automatically render it the best choice. Despite its low MSPE, LSTM5 struggles with complex market dynamics, highlighting the significance of model complexity and underscoring the principle that an excessive number of layers can negatively impact the model's performance.

In comparison, LSTM2, with an MSPE of 0.870, strikes a commendable balance by fitting the data reasonably well without capturing excessive noise. The moderate alignment between the realised and predicted excess returns of LSTM2 suggests a more generalised approach, making it preferable for accurately capturing market trends without overfitting. Meanwhile, LSTM1, LSTM3, and LSTM4, despite certain strengths, indicate overfitting risks due to the close alignment between their realised and predicted excess returns.

Consequently, LSTM2 emerges as the optimal model. Hence, a shallow LSTM network indicates better performance than a deep LSTM network, which aligns with the findings of Gu et al. (2020).

5.2 Which features matter

This section proceeds to assess the significance of 13 macroeconomic predictors and 94 stock characteristics to identify the most crucial features for integrating into the optimal LSTM2 model. Using the methodology of PFI, Figure 8 presents a visualisation of the aggregated feature importance derived from the LSTM2 model. Note that the variable importance is normalised to sum to one, facilitating a more intuitive interpretation of the results.

The top 20 most important features, according to Figure 8, are as follows:

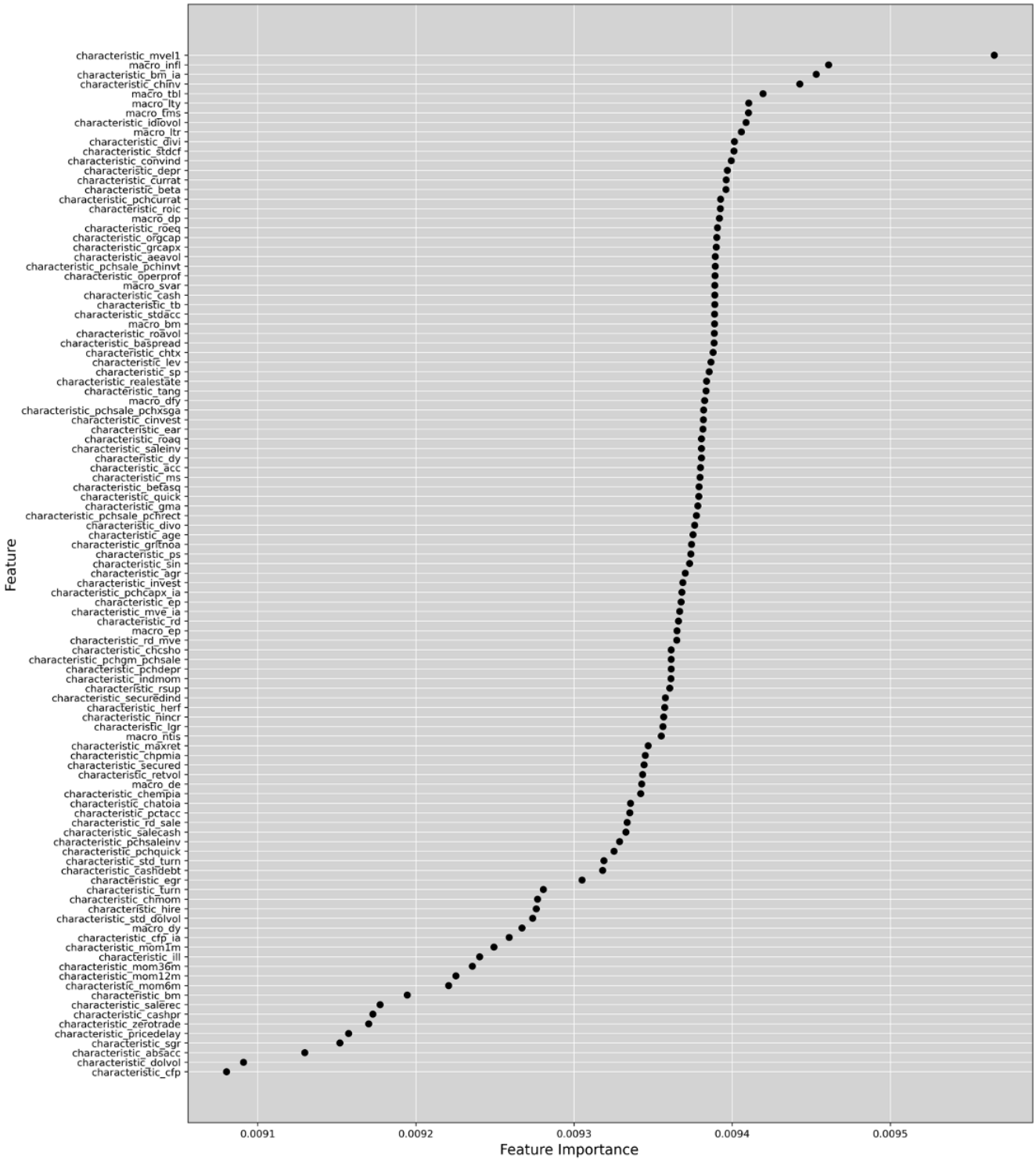
- After selecting the top 15 most important stock characteristics, zero of those are related to the category recent price movements.
- Subsequently, the most important stock characteristics related to liquidity are:
  - The log market equity (**mvel1**).
  - Current assets divided by current liabilities (**currat**).

  - Per cent change in currat (**pchcurrat**).

- Thirdly, features related to risk measures are:
  - The standard deviation of residuals of weekly returns on weekly equal weighted market returns for 3 years prior to month end (**idiovol**).
  - Estimated market beta from weekly returns and equal weighted market returns for 3 years ending month $t-1$ with at least 52 weeks of returns (**beta**).
- Lastly, features related to valuations ratio and fundamental indicators are:
  - The Change in inventory (inv) scaled by average total assets (at) (**chinv**).
  - Industry adjusted book-to-market ratio (**bm_ia**).
  - An indicator variable equal to 1 if the company pays dividends but did not in the prior year (**divi**).
  - Cash flow volatility, (**stdcf**).
  - An indicator equal to 1 if the company has convertible debt obligations (**convind**).
  - Depreciation divided by PP&E (**depr**).
  - Annual earnings before interest and taxes (ebit) minus non-operating income (nopi) divided by non-cash enterprise value (ceq+lt-che) (**roic**).
  - Earnings before extraordinary items divided by lagged common shareholders' equity (**roeq**).
  - Capitalized SG&A expenses (**orgcap**).
  - Per cent change in capital expenditures from year $t-2$ to year t (**grcapx**).
- The most important macroeconomic features are:
  - Inflation (**infl**).
  - The 3-Month Treasury Bill (**tbl**).
  - The long-term yield (**lty**).
  - The term spread (**tms**).
  - The long-term rate of returns (**ltr**).

The only important predictor variables identical to the findings by Gu et al. (2020) are (**beta**), (**idiovol**), and (**mvel1**). Moreover, the LSTM's disregard for recent price movement indicators such as (**mom1m**), (**mom12m**), and (**chmom**) can be attributed to its ability to analyse temporal sequences and recognise long-term dependencies in stock market data. The LSTM model prioritises features related to liquidity, risk measures, and valuation ratios and fundamental indicators over more transient price movements, indicating a model that seeks stable predictors of stock performance rather than reacting to recent trends.

Figure 8: LSTM2 permutation feature importance



Source: Own calculations.

Note: Variable importance is an average of the MSPE for all data splits using K-fold CV, employing 500 stocks with LSTM2. Variable importance is normalised to sum to one. Table 9 in the Appendix provides a summary statistic of the 15 most important stock characteristics.

In the study conducted by Gu et al. (2020), the authors found the most crucial macroeconomic predictors for FNNs to be (**bm**), (**ntis**), (**tbl**), (**tms**), and (**dp**). This thesis agrees on the importance of (**tbl**) and (**tms**) but diverges by highlighting the newly considered macroeconomic predictors: (**lty**), (**ltr**), and (**infl**), underscoring their importance in LSTM models.

Consequently, this section demonstrates that using PFI in conjunction with an LSTM model uncovers the most critical features, thereby affirming hypothesis **(H1)**.

5.3 LSTM out-of-sample performance

Having identified the top 20 most important features, this thesis conducts an out-of-sample performance comparison between the LSTM2 and the univariate LSTM model. It aims to discern the advantage of incorporating a broader range of features by comparing the LSTM2, a multivariate LSTM model that integrates the top 20 most significant features with the univariate model, which relies solely on past excess stock returns as its predictive factor.

Table 3 offers a comparative analysis of performance metrics for LSTM2 and the univariate LSTM. The metrics used for this comparison include the coefficient of determination (R^2), MSPE, and the mean absolute error (MAE). The following equations define the R^2 value and MAE:

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

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$

where n is the number of observations, y_i is the realised excess return of the i^{th} observation, $\hat{y}_i$ is the predicted excess return of the i^{th} observation, and $\bar{y}$ is the mean of the realised excess returns.

Table 3: Monthly out-of-sample model performance

Model	R^2	MSPE	MAE
LSTM2	−3.547	0.103	0.194
LSTM univariate	−0.954	0.043	0.109

Source: Own calculations

According to the table, the univariate LSTM model outperforms LSTM2 across all listed criteria: It has a higher (less negative) R^2 value of -0.954 compared to -3.547 for the LSTM2,

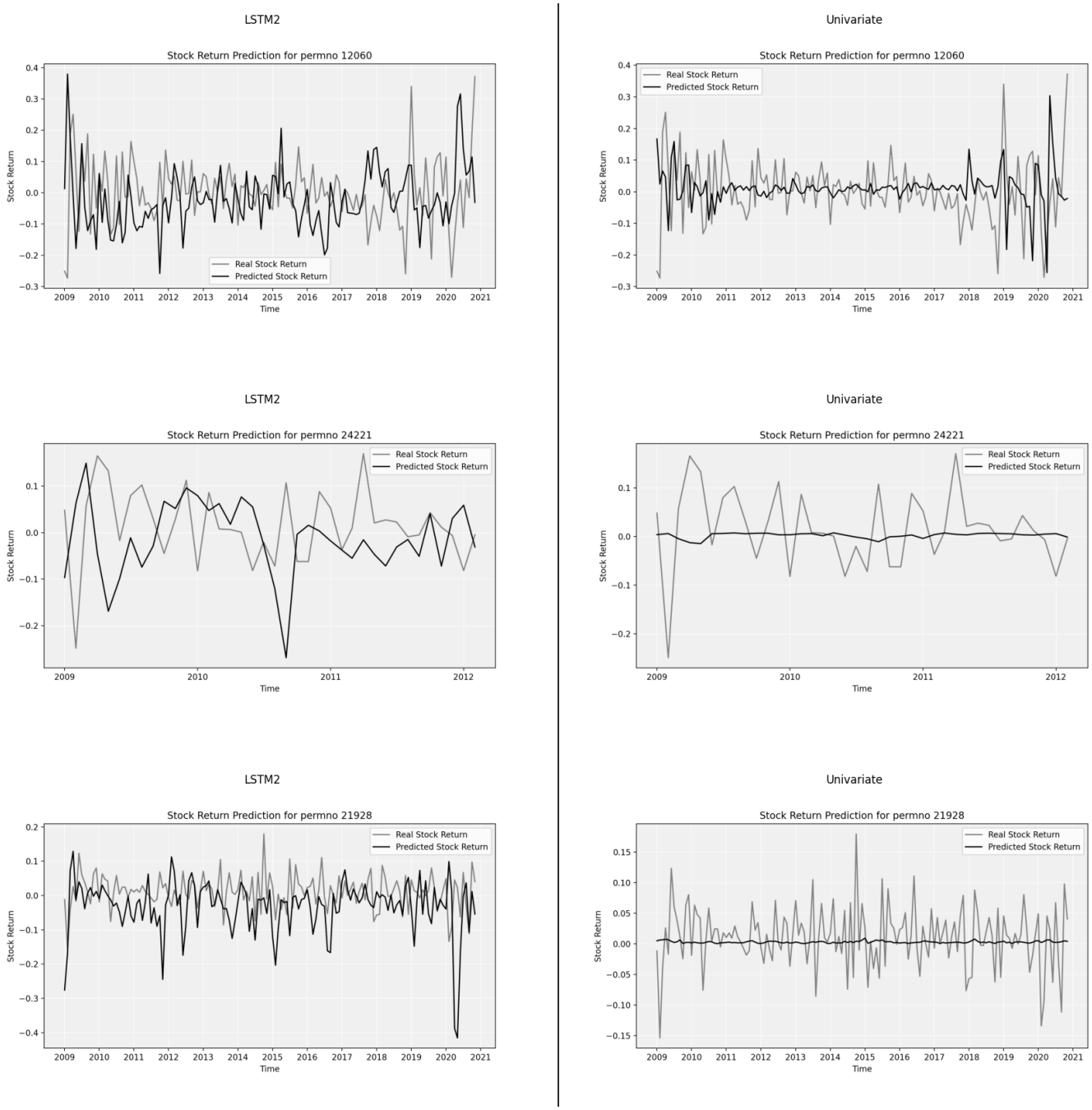
indicating a stronger correlation between the predicted values and the realised values. It also scores lower MSPE and MAE values (0.043, and 0.109, respectively), suggesting it generates predictions closer to the realised excess returns with less error. However, it is crucial to note that both models exhibit a negative R^2, which implies that the models fit the data worse than a simple horizontal line representing the mean of the excess stock returns. This could be attributed to poor model fit, implying that the models do not capture the underlying trend of the data well. Alternatively, the models may be too simplistic for the given dataset. Lastly, the top 20 features might not be sufficiently relevant or informative to predict excess stock returns. Nonetheless, Figure 9 provides additional insights into the performance of the two models.

The Figure presents the LSTM2 excess stock return predictions (black) against the realised excess returns (grey) on the left-hand side and the univariate LSTM predictions on the right-hand side. Comparing the initial excess stock return predictions for permno 12060, the LSTM2 model appears to follow the volatility of the realised excess returns with a certain degree of accuracy, although with a noticeable lag and diminished amplitude in the predictions. The univariate model for the same permno, while appearing to capture the direction of the trends, exhibits a clear divergence in magnitude, failing to capture the peaks and troughs of the realised excess returns.

The observations for permno 24221 reveal that the LSTM2 captures the volatility and the spikes more effectively than the univariate LSTM. However, the LSTM2 demonstrates a noticeable lag in its predictions. The univariate model, although smoother, does not accurately follow the realised stock return pattern, indicating that both models fail to capture the temporal trends.

Lastly, for permno 21928, the LSTM2 predictions track the realised excess returns more closely despite underestimating extreme values. In contrast, the univariate model's predictions are consistently smoother and fail to recognise the finer fluctuations in the stock's performance.

Figure 9: LSTM2 (left) vs univariate LSTM (right) monthly out-of-sample excess stock return predictions



Source: CRSP (2024), and own predictions.

Note: Permno 12060 represents General Electric Co. Permno 24221 denotes Bridge SaaS Ltd, and Permno 21928 is IDACORP Inc.

While the data in Table 3 suggest that the univariate LSTM model outperforms LSTM2 based on numerical metrics, a more detailed examination of the charts uncovers the inherent

challenge of accurately predicting excess stock returns. Both models have difficulty fully capturing the complex behaviour of excess stock returns. However, the LSTM2 model appears to discern specific trends and sudden shifts in the data and reflects the general volatility more accurately. In contrast, the univariate LSTM model seems to disregard these trends and the volatility, resulting in more simplistic predictions and a lack of variability. Therefore, despite the implications of the quantitative metrics, the LSTM2 model might offer better excess stock return predictions, suggesting that the inclusion of additional features provides value.

5.4 Portfolio evaluation

Analysing expected excess returns at the portfolio level is beneficial, given that the LSTM models are designed for individual stock predictions. Aggregating and forecasting at the portfolio level provide an additional indirect assessment of the LSTMs' effectiveness. Furthermore, aggregated portfolios indicate the broader economy, mirroring investment vehicles in risky assets favoured by a large segment of investors, such as mutual or exchange-traded funds. The distribution of portfolio returns is heavily influenced by the correlation among stock returns, suggesting that an LSTM's accuracy at the stock level does not necessarily translate to precision at the portfolio level. Portfolio construction offers a means to test the LSTM's capacity to extend its predictions from individual stocks to more comprehensive investment scenarios. The final benefit of examining portfolio accuracy lies in evaluating the economic gain of the LSTM by analysing its contribution to a risk-adjusted portfolio return.

5.4.1 Sharpe ratio

The Sharpe ratio, defined by Sharpe (1966), is a widely recognised metric to assess the effectiveness of an investment. It proposes the term *reward-to-variability*, which offers investors a way to evaluate the trade-off between additional returns over the risk-free rate and the risk taken to achieve these returns. The formula for the Sharpe ratio is given by the difference between the return of a portfolio R_a and the risk-free rate, R_f, divided by the portfolio's standard deviation of excess return, σ_a:

$$\text{Sharpe Ratio} = \frac{E[R_a - R_f]}{\sigma_a}.$$

This formula denotes the extra return an investor gains for each unit of risk taken. In line with the principles of modern portfolio theory, investors typically aim to optimise their portfolios for the highest mean-variance, thus maximising their Sharpe ratio. Historical data

indicates that the annualised Sharpe ratio for the S&P 500 from 2009 to 2020 is 0.96.[2] To annualise the Sharpe ratio, it is standard practice to assume that returns follow a log-normal distribution, which allows for the multiplication by the square root of 12 to adjust for the annual period.

5.4.2 Long-short strategy

Table 4 compares the out-of-sample performance of the LSTM2 and univariate LSTM model, in managing investment portfolios from 2009 to 2020. Portfolio 1 contains the lowest predicted excess returns, portfolio 10 includes the highest, and the 10-1 portfolio represents the long-short strategy. For each portfolio, the table includes several key metrics: monthly realised excess returns, t-statistics to quantify the significance of these excess returns, the standard deviation (SD) indicating the volatility of these excess returns, both monthly and annualised Sharpe ratios to assess the risk-adjusted excess returns, transaction costs (TC), and turnover rates.

Table 4: Comparison of monthly out-of-sample LSTM2 and univariate LSTM portfolio performance

	Portfolio	1	2	3	4	5	6	7	8	9	10	10-1
LSTM2	**Rea. exc. return**	0.011	0.012	0.010	0.010	0.009	0.008	0.011	0.010	0.012	0.015	0.004
	t-Stat	2.451	3.049	2.796	2.832	2.628	2.284	2.780	2.517	2.735	3.107	1.658
	SD	0.054	0.048	0.043	0.041	0.041	0.043	0.049	0.049	0.051	0.059	0.033
	Sharpe	0.205	0.255	0.234	0.237	0.220	0.191	0.233	0.210	0.229	0.260	0.129
	Ann. Sharpe	0.710	0.883	0.810	0.820	0.761	0.662	0.805	0.729	0.792	0.900	0.447
	TC	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.001
	Turnover	0.541	0.735	0.800	0.792	0.773	0.777	0.806	0.785	0.757	0.618	1.159
Univariate	**Rea. exc. return**	0.011	0.010	0.010	0.010	0.010	0.009	0.011	0.014	0.012	0.016	0.004
	t-Stat	1.961	2.242	2.739	2.844	3.188	2.456	3.177	3.327	2.419	2.474	1.296
	SD	0.070	0.054	0.045	0.041	0.036	0.042	0.042	0.051	0.058	0.075	0.038
	Sharpe	0.164	0.187	0.229	0.238	0.267	0.205	0.266	0.278	0.202	0.207	0.109
	Ann. Sharpe	0.568	0.649	0.793	0.824	0.923	0.712	0.920	0.964	0.701	0.717	0.377
	TC	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.003
	Turnover	0.850	0.874	0.871	0.786	0.675	0.732	0.795	0.781	0.852	0.812	1.662

Source: Own calculations.

Note: *Rea. exc. return* is the monthly realised excess returns including transaction costs, *SD* is their standard deviations, *Sharpe* is the monthly Sharpe ratio, *Ann. Sharpe* is the annualised Sharpe ratio, *TC* is the monthly transaction cost, and *Turnover* is the monthly turnover.

For the LSTM2 model, excess returns demonstrate a nonlinear trend, with portfolio 6 registering the lowest excess return and portfolio 10 achieving the highest. Contrary to expectations, portfolio 1, despite containing stocks with the lowest predicted excess returns, does not exhibit negative excess returns. As a result, the 10-1 portfolio produces a monthly excess return

[2] Own calculations

of 0.004.

The Sharpe ratio peaks in portfolio 10 and reaches its lowest in portfolio 6, with a value of 0.191, indicating that the highest risk-adjusted returns are not uniformly distributed across the deciles. The 10-1 portfolio's Sharpe ratio stands at 0.129, with its annualised counterpart at 0.447. Compared to the S&P 500 index's annualised Sharpe ratio of 0.96, this highlights the LSTM2 long-short strategy's relatively lower excess return per unit of risk over the 2009-2020 period.

Moreover, despite LSTM2's higher excess returns of portfolios 9 and 10, the elevated transaction costs, recorded at 0.002 for both portfolios, underline the cost of achieving these returns. This is primarily due to the need for frequent portfolio reallocations. In contrast, portfolio 1 has the lowest transaction cost at 0.001 and a turnover rate of 0.541, suggesting minimal trading activity. This reduced turnover implies lower associated trading costs, potentially indicative of a strategy favouring long-term market trends over immediate gains. Thus, the transaction cost for the 10-1 portfolio is 0.001.

The univariate LSTM model exhibits a similar pattern in excess returns, with the highest return of 0.016 in portfolio 10 and the lowest return of 0.009 in portfolio 6, while portfolio 1 achieves 0.011. Consequently, the 10-1 portfolio using the univariate LSTM produces an identical monthly excess return of 0.004. However, the t-statistic in portfolio 1 is 1.961 and 2.474 in portfolio 10, leading to a t-statistic of 1.302 for the 10-1 portfolio, indicating a performance close to zero. Additionally, the standard deviation of excess returns for the univariate LSTM is generally higher across all portfolios compared to LSTM2, suggesting greater volatility. The standard deviation for the 10-1 portfolio is 0.038, higher than LSTM2's 0.033. The Sharpe ratios are lower for the univariate model across most portfolios, indicating that, for the level of risk taken, the univariate LSTM's excess returns are not as substantial as those for LSTM2. The 10-1 portfolio also presents a lower annualised Sharpe ratio of 0.377 compared to the 0.447 achieved by LSTM2. Furthermore, the univariate LSTM model incurs slightly higher transaction costs, with significantly greater turnover observed in the 10-1 portfolio. This indicates that operating the univariate model is more costly.

To further assess the performance of the two LSTM models, Figure 10 illustrates the realised annual excess portfolio returns based on the LSTM2 model on the left side and those of the

univariate LSTM model on the right side.

Figure 10: Out-of-sample realised annual excess portfolio returns of LSTM2 (left) and univariate LSTM (right) with TC

Source: Own predictions.
Note: The red figure indicates portfolio 10, the blue figure indicates portfolio 1, and the green figure indicates the long-short strategy (10-1).

Both figures reveal the absence of consistent, prominent patterns over recent years, with each portfolio experiencing periods of positive and negative annual excess returns. Conventionally, one might expect portfolio 1 to consistently exhibit negative excess returns and portfolio 10 to consistently register positive gains. Despite this expectation, portfolio 10 shows positive excess returns for most of the observed years. However, the long-short strategy does not perform as anticipated because portfolio 1 often yields positive excess returns. Nonetheless, portfolio 1 still records more years with negative excess returns compared to portfolio 10, which may suggest some degree of predictive accuracy from both LSTM models.

The final assessment of the long-short strategy is presented in Table 5. This table displays the regression results from analysing the out-of-sample performance of the LSTM2 and the univariate LSTM model from 2009 to 2020. These results focus on the generated excess returns, providing insights into the models' average performance over the specified period when all other variables are held constant.

For the LSTM2 model, the intercept is 0.004, with a standard error of 0.003. The t-statistic is 1.658, leading to a p-value of 0.097. The positive intercept suggests that the LSTM2 model generates an average monthly excess return of 0.004. The p-value of 0.097 indicates

Table 5: Monthly out-of-sample LSTM2 and univariate long-short regression results

		Estimate	Std. Error	t-Statistic	p-value
LSTM2	Intercept	0.004	0.003	1.658	0.097
Univariate	Intercept	0.004	0.003	1.296	0.195

Source: Own calculations.

that the null hypothesis of average excess returns being equal to zero can be rejected at a 10% significance level. This implies that the result is marginally significant, revealing a slight tendency towards generating positive excess returns. However, it should be interpreted cautiously because the p-value is close to the threshold.

For the univariate LSTM model, the intercept is also 0.004, with the same standard error of 0.003. However, the t-statistic is slightly lower at 1.296, resulting in a higher p-value of 0.195. Hence, the null hypothesis of average excess returns being equal to zero cannot be rejected at a 10% significance level. Thus, on average, the univariate LSTM model does not provide statistically significant positive excess returns.

Both models exhibit an average positive performance across the out-of-sample period, as their positive intercepts indicate. However, the statistical evidence supporting positive excess returns is stronger for the LSTM2 model than for the univariate model. Neither result is robustly significant at a 5% significance level, suggesting caution in interpreting these findings. As a result, hypotheses **(H2)** and **(H3)** is rejected. First, the LSTM did not exhibit superior performance compared to the traditional FNN employed by Gu et al. (2020), which found an annualised Sharpe ratio of 1.35 utilising a similar 10-1 hedge portfolio strategy. Second, the LSTM2 model generated small but statistically significant positive excess returns at a 10% significance level, even after accounting for transaction costs. However, since the LSTM did not outperform a passive buy-and-hold strategy, it did not provide substantial economic gains, as the excess returns achieved by the LSTM are too small to be deemed beneficial in a practical trading context.

5.4.3 Comparison with Fama-French model

This section uses the FF3 model to analyse excess portfolio returns. The objective is to determine whether the well-known factors adequately account for the alpha observed in the 10-1 portfolios or if additional characteristics provide incremental predictive capability for the cross-section of expected excess returns.

Table 6 presents regression results from applying the FF3 model to analyse the excess returns of the 10-1 portfolios for both the LSTM2 and the univariate LSTM models.

Table 6: FF3 monthly out-of-sample long-short regression results of LSTM2 and univariate LSTM

Model	Coefficients	Estimate	Std. Error	t-Statistic	p-value
LSTM2	α	0.002	0.003	0.641	0.522
	MKT_excess	0.089	0.068	1.301	0.195
	SMB	-0.004	0.120	-0.033	0.974
	HML	-0.305	0.097	-3.143	0.002
Univariate LSTM	α	0.002	0.003	0.719	0.474
	MKT_excess	0.191	0.079	2.407	0.017
	SMB	-0.375	0.140	-2.683	0.008
	HML	0.056	0.113	0.497	0.620

Source: Own calculations.

The F-statistics for the two regressions are 3.339 and 3.527, respectively. The corresponding p-values are 0.021 and 0.017, which are less than 0.05. This indicates that at least one of the predictors in each test is statistically significant at the 5% level. However, the low R^2 values of 0.067 and 0.071 suggest that the predictors in the FF3 model can only explain a relatively small proportion of the variability in the long-short portfolio.

For the LSTM2 model, the analysis reveals that the alpha coefficient, at 0.002 with a standard error of 0.003, yields a t-statistic of 0.641 and a p-value of 0.522. This indicates a lack of statistical significance, suggesting that the model's returns do not significantly deviate from zero after accounting for risk factors. Thus, including additional features does not result in further economic gains. The market excess return coefficient, observed at 0.089 with a p-value of 0.195, also does not significantly contribute to explaining the model's returns. Additionally, the size premium is virtually negligible, with a coefficient of -0.004 and a p-value of 0.974. Thus, the 10-1 portfolio is not weighted towards either large-cap or small-cap stocks. Conversely, the value premium shows a significant negative influence on the model's excess returns, with a coefficient of -0.305 and a p-value of 0.002, suggesting that the 10-1 portfolio tends to perform better in environments favouring growth stocks over value stocks. Specifically, the -0.305 estimate implies that as the return differential between value and growth stocks increases (i.e., value stocks outperform growth stocks), the 10-1 portfolio's returns are expected to decrease. Conversely, when growth stocks outperform value stocks, the 10-1 portfolio's performance is expected to increase.

The univariate LSTM model's alpha is also 0.002 and lacks significance, with a p-value of 0.474. However, the market excess return plays a significant role in this model's performance, marked by a coefficient of 0.191, a t-statistic of 2.407, and a p-value of 0.017. The significant positive coefficient for market excess return suggests that the model's performance is correlated with overall market performance. A coefficient of 0.191 means that for every 1% increase in market excess return, the portfolio's return is expected to increase by 0.191%. The size premium exhibits a significant negative effect, with a coefficient of -0.375 and a p-value of 0.008, indicating that the 10-1 portfolio is weighted towards large-cap stocks. Meanwhile, the value premium does not significantly affect the model, as evidenced by a coefficient of 0.056 and a p-value of 0.620.

The similarity in risk-adjusted performance is interesting, especially since the LSTM2 model uses the top twenty predictor features yet achieves similar performance to the univariate LSTM model. This raises interesting questions about the patterns of the 10-1 portfolios for both models and whether significant differences exist between them.

5.5 Portfolio patterns

This section explores the underlying patterns of the 10-1 portfolios. Specifically, it analyses the industry sectors, liquidity, risk metrics, dividends, and the industry-adjusted book-to-market ratio to support previous findings. This approach also helps validate the effectiveness of the top twenty features identified through PFI.

Table 7 provides an analysis of the features driving performance for LSTM2's portfolios, including the market value of equity (**mvel1**), market beta (**beta**), industry-adjusted book-to-market ratio (**bm_ia**), and dividend indicator (**divi**). The normalisation of these features complicates direct comparison.

Table 7: Out-of-sample LSTM2 portfolio characteristics

	mvel1	bm_ia	divi	beta
Portfolio 1	-0.005	0.021	-0.003	0.146
Portfolio 10	-0.129	-0.019	-0.005	0.140
All Portfolios	0.156	-0.057	-0.006	0.037

Source: Own calculations.

The mvel1 for portfolio 1 is -0.005, indicating a modest preference for small-cap stocks in short positions. Portfolio 10's mvel1 (-0.129) reveals a definitive bias towards small-cap stocks

for long positions, diverging from the overall portfolio average of 0.156. This trend differs from the FF3 regression analysis, which did not significantly lean towards small-cap stocks, as evidenced by the insignificant SMB coefficient.

The industry-adjusted book-to-market ratio (**bm_ia**) highlights a strategic coherence across the value-growth spectrum. Portfolio 10's bm_ia of -0.019 suggests a lean towards growth stocks in long positions, whereas portfolio 1's bm_ia of 0.021 shows a slight preference for shorting value stocks. LSTM2's negative HML coefficient of -0.305 in Table 6, combined with the bm_ia characteristics of portfolio 10 and portfolio 1, implies a coherent strategy orientation. When growth stocks are expected to outperform value stocks, portfolio 10 stands to gain, and when value stocks underperform, the short position in portfolio 1 is expected to profit.

The dividend variable (**divi**), showing negative values for both portfolio 1 and portfolio 10 of -0.003 and -0.005, respectively, with an overall average of -0.006, suggests a tendency to underweight dividend-paying stocks in the LSTM2 strategy. This tendency is consistent with a preference for growth stocks, which are more likely to reinvest earnings than pay dividends.

The beta values of 0.146 for portfolio 1 and 0.140 for portfolio 10, with a mean of 0.037 for all portfolios, suggest moderate to low market risk relative to the benchmark. The FF3 regression results show a positive but insignificant impact coefficient for MKT_excess (0.089 with a p-value of 0.195), suggesting that while the market's movements influence the LSTM2 model's excess returns, this influence is weak. The moderate beta values reinforce this interpretation, indicating that while the LSTM2 model is responsive to market trends, it is not overly exposed to market volatility. This balanced exposure could explain why the MKT_excess factor's impact on the model's excess returns is positive but insignificant.

Figures 12 and 13 in the Appendix illustrate the industry distribution within the portfolios managed by LSTM2 and the univariate LSTM, highlighting a significant presence of finance, manufacturing, and service sectors, thereby detailing the relative differences across these sectors.

In Figure 12, the LSTM2 favours short positions in the manufacturing sector, representing over 40% of its portfolio, while the service and finance sectors account for about 15%

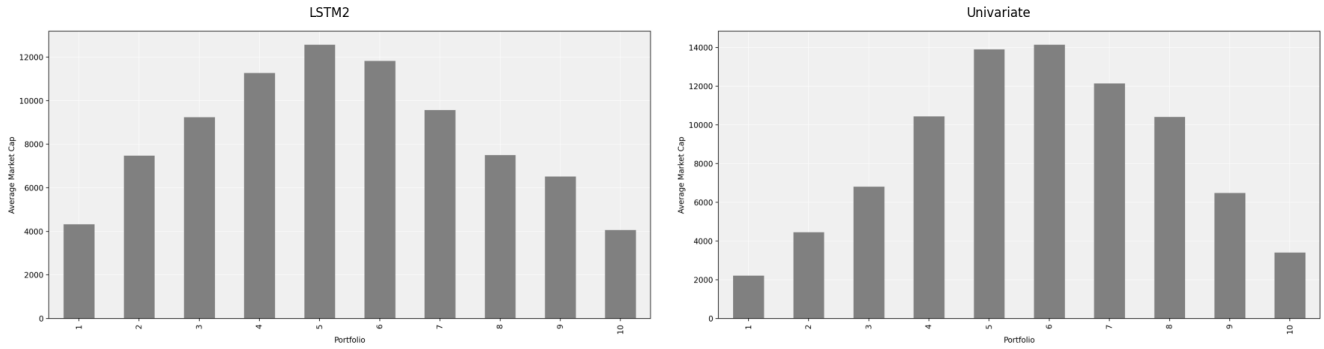
and 14%, respectively. In contrast, the strategy prefers stocks in the finance sector for long positions, making up approximately 19% of its long portfolio, with the service sector at around 13%. Similar to its approach in short positions, LSTM2 invests more than 40% of its long positions in the manufacturing sector, indicating a significant allocation to this industry.

The inclination of LSTM2 to favour finance stocks for long positions during the period 2009-2020 may be attributed to several factors. The aftermath of the 2008 financial crisis saw substantial regulatory reforms and stabilisation measures, which contributed to a robust recovery in the financial sector. This recovery might have presented attractive investment opportunities that LSTM2 identified as favourable for long positions. Despite the diversity in portfolio configurations, all portfolios formulated under the LSTM2 strategy exhibit a consistent pattern. This recurring pattern may largely stem from the dataset's disproportionate representation of the three industries (manufacturing, services, and finance). Consequently, the selection bias towards stocks from these sectors may not necessarily reflect a strategic preference but rather a limitation imposed by the dataset's composition.

Figure 13 displays a uniform industry pattern across all portfolio allocations within the univariate LSTM strategy. This approach predominantly favours short positions in nearly 50% of stocks from the manufacturing sector, coupled with around 15% in the services sector and 14% in the finance sector. Notably, this allocation pattern is also consistently reflected in the strategy's long positions. This identical pattern between the long and short positions may primarily arise from two factors. First, the observed uniformity in sector-specific allocations could directly result from the over-representation of the manufacturing, services, and finance sectors. Second, the univariate LSTM model operates with only one predictor feature compared to the multivariate LSTM2. This limitation restricts the model's capability to accurately forecast excess stock returns across different industries. Consequently, the univariate LSTM tends to generate more homogeneous industry patterns in its portfolio allocations, reflecting its constrained predictive capacity.

Figure 11 presents the average market capitalisation for the portfolios constructed by the LSTM2 and the univariate LSTM. Both strategies exhibit a notably similar pattern, with portfolios 1 and 10 predominantly composed of firms with lower market capitalisation, indicating a strategic preference for small-cap stocks. This observation contrasts with the FF3 results, where the SMB factor for the LSTM2 is insignificant, and the negative, significant SMB factor for the univariate LSTM suggests a weighting towards large-cap stocks.

Figure 11: Market cap by portfolio 1 to 10 for LSTM2 (left) and univariate LSTM (right) during the out-of-sample period



Source: Own calculations.

Further examination reveals a differentiation between portfolios 1 and 10, with the remaining portfolios demonstrating a preference for higher market capitalisation stocks, suggesting a more conservative investment stance compared to the extremes. Hence, portfolios 2 through 9 are characterised by larger average market capitalisation.

To summarise, the performance of the LSTM2, as indicated by the FF3 results in Table 6, presents a significant negative value premium. This suggests that the LSTM2 strategy excels in environments favouring growth stocks. The pattern is consistent with the portfolio characteristics displayed in Table 7, which affirm LSTM2's preference for growth stocks. Conversely, the univariate LSTM strategy, according to its FF3 model results, displays a significant negative size premium, implying a general avoidance of small-cap stocks. This finding contradicts the results in Figure 11, which show that the univariate LSTM favours small-cap stocks.

6 Discussion

Section 6.1 begins with a discussion of the RNN framework, contrasting the LSTM model with the FNN approach. It explores the issue of look-ahead bias in the LSTM method and discusses the use of PFI for variable selection. Following this, Section 6.2 discusses the difficulties associated with identifying the optimal LSTM structure, emphasising the complexity of this task. Additionally, Section 6.3 scrutinises the issues of publication bias and p-hacking, discussing their impact on the validity of research findings. In conclusion, Section 6.4 assesses the practical application of this LSTM model in real-world contexts.

6.1 Methodological approach: FNN vs RNN

Adopting the LSTM methodology over the FNN approach reduced the number of stocks analysed to 2,552. While the original dataset included over 20,000 stocks, the LSTM model in this study is designed to predict the performance of one stock at a time. This design choice is due to the LSTM's dependence on historical data for forecasting excess returns one month ahead, necessitating the use of data exclusively from the same stock for its subsequent month's return prediction. Using historical data from different stocks would compromise the model's forecast accuracy. As a result, each stock needed sufficient historical data prior to 2009, along with a minimum of three months used for out-of-sample testing, thus narrowing the number of stocks. In contrast, employing an FNN approach would have allowed the inclusion of all 20,000 plus stocks, as FNNs rely solely on information at time t to predict the next month's excess returns.

This thesis reveals that leveraging historical data with an LSTM model did not yield superior accuracy compared to the findings of Gu et al. (2020) and failed to outperform a passive buy-and-hold strategy based on the S&P 500 index. This outcome may be attributable to the limited number of features used; the optimal LSTM2 model utilised only 20 features, compared to over 900 features in Gu et al. (2020), primarily due to computational constraints. Future studies could explore the inclusion of additional features, such as a cross-section between the stock characteristics and macroeconomic predictors for a more comprehensive comparison with Gu et al. (2020). This investigation also highlights the significant computational demands of using an LSTM compared to an FNN, given its design to predict the performance of individual stocks sequentially.

Furthermore, LSTMs may be more suitable for predicting daily excess returns of individual stocks, as supported by Krauss et al. (2017), Fischer and Krauss (2018), and Ghosh et al. (2022). Notably, Ghosh et al. (2022) reveals that using a multi-feature LSTM approach can increase daily excess returns to 0.64%, outperforming single-feature models and highlighting the LSTM's capability to enhance excess stock return predictions for day trading. Additionally, future research could investigate the combination of a random forest (RF) with the LSTM. The study by Ma et al. (2019) illustrates the efficacy of utilising the RF to select the most crucial features for the LSTM to use in daily stock price predictions. This model combination outperformed a single LSTM model strategy and a buy-and-hold strategy.

6.1.1 Look-ahead bias

To ensure robustness in the analysis, it is crucial to select stocks with adequate data for both the training and testing phases. Nonetheless, this requirement introduces the potential for look-ahead bias, particularly when identifying stocks that remain active during the out-of-sample period. To address this issue, this thesis employs a three-month waiting period before the forecast point at $t + 1$, explicitly excluding this period from model training and forecasting. In other words, the first three months after the training period are used as a standby period for the LSTM to utilise for the fourth month's prediction. Therefore, stocks with less than three months of data available in the standby period are excluded.

The data split conducted in this thesis, which separates the training and testing data at the onset of the 2008-2009 financial crisis, could also introduce bias into the LSTM model. This period was a significant event that drastically affected stock returns (Hamdaoui et al., 2022). The LSTM models were trained on data up until 2008, just as the financial crisis began. Subsequently, they were tested on data from 2009, during the crisis. This implies that the models were trained in a market environment that is fundamentally different from the one they were tested on. The financial crisis brought about extreme volatility and uncertainty in the market, conditions for which the models were not fully trained to handle. This could be one reason for the poor model performance. However, financial crises are complex events influenced by a multitude of factors, many of which are difficult to quantify or predict. These factors include, but are not limited to, macroeconomic indicators, policy decisions, investor sentiment, and global economic conditions. Thus, it would be difficult for any model to fully capture such events. Future research could attempt to mitigate this bias by incorporating data from the crisis period into the training set. This might allow the models to learn from the crisis and potentially improve their ability to handle similar events in the future.

6.1.2 Feature selection

One of the primary advantages of PFI is its model-agnostic nature, allowing it to be applied across different models without needing adjustments. This flexibility makes PFI a versatile tool for feature importance analysis in diverse machine learning workflows. Subsequently, the rationale behind PFI is straightforward. By measuring the impact of shuffling a feature's values on model performance (e.g., MSPE), it is possible to infer the importance of that feature. This direct approach facilitates clear and intuitive insights into which features drive the model's predictions. Unlike methods that require retraining the model after removing features, PFI saves time by simply permuting feature values, making it a quicker alternative

for assessing feature importance (Molnar, 2020).

However, when features are correlated, permuting one feature can lead to unrealistic data instances, skewing the measurement of importance. This complication can obscure the interpretation of how feature interactions influence model performance, as highlighted by Molnar (2020).

If the emphasis is solely on implementing an LSTM model encompassing numerous features with crucial temporal dynamics and interactions between features, SHAP (SHapley Additive exPlanations) values emerge as a highly effective alternative to PFI. SHAP importance shares traits with variance-based importance assessments, where a significant alteration in the model's output due to a change in a feature underscores that feature's importance. This perspective on importance deviates from the approach taken by PFI, which is predicated on loss. For instance, in the case of model overfitting, where a feature that lacks relevance to the output is utilised, PFI would deem this feature non-contributory, since it does not aid in accurate predictions. In contrast, variance-based importance assessments, such as SHAP, could ascribe significant importance to such a feature, acknowledging that changes to the feature may result in notable prediction variations, as detailed by Molnar (2020).

While SHAP offers compelling advantages, it is important to note that computing SHAP values, especially for models with a large number of features or complex temporal dependencies, can be computationally intensive. Additionally, SHAP might not be as easy to interpret as PFI. Therefore, this thesis finds PFI the preferred choice due to the lack of computational power and focus on interpretability.

6.2 Challenges of model optimisation

Future studies might consider adopting a different LSTM optimisation framework. The LSTM network structure employed in this thesis is based on the geometric pyramid rule adopted by Gu et al. (2020). However, this approach was tested on an FNN and was not applied to an LSTM for predicting excess stock returns. Consequently, it is plausible that a more effective LSTM network structure exists, but discovering such a structure within the given timeframe is a complex task. Moreover, the models showed signs of overfitting despite efforts to mitigate this using early stopping combined with $L1$ regularisation. An alternative strategy could have involved the use of $L2$ regularisation in place of early stopping. If

sufficient computing power were available, $L2$ regularisation might have offered better control over overfitting. However, due to its computational cost, early stopping was the method of choice. The list of potential regularisation and optimisation tools is extensive, and it is believed that investing more time in optimising the LSTM network could result in improved prediction accuracy, thereby leading to higher Sharpe ratios.

6.3 Addressing publication bias and p-hacking

Harvey (2017) addresses the critical issue of publication bias in financial economics, emphasising that, due to journal preferences, researchers often favour studies yielding "significant" results. This leads to the "file drawer effect", where studies with marginal or negative results are not submitted for publication. The bias is further exacerbated by p-hacking, where researchers select only the most important findings for publication. Even if journals were open to publishing less noteworthy findings, the inclination to publish only significant results discourages authors from investing time in potentially valuable studies that may not yield immediately noteworthy results.

Harvey et al. (2016) conducted a meta-analysis of factor studies from 1963 to 2012, revealing a distribution suggestive of publication bias. The analysis showed a nearly equal number of studies reporting t-statistics in the range of 2.0 to 2.57 and 2.57 to 3.14, with relatively few studies with t-statistics less than 2.00 being published. This pattern indicates a preference for publishing findings that meet conventional thresholds of significance, thereby overlooking potentially insightful research that does not meet these criteria.

The discussion by Harvey (2017) concludes that the field of financial economics faces a "complex agency problem", where the drive for significant results often overshadows the pursuit of advancing knowledge. This thesis utilises the PFI method, which assesses a feature's significance through its impact on the MSPE rather than relying on t-statistics. As a result, it is challenging to robustly evaluate the significance of the selected features using PFI alone. For future studies aiming to assess feature significance more rigorously, the Permutation Importance (PIMP) algorithm, introduced by Altmann et al. (2010), offers a viable alternative. A key advantage of the PIMP algorithm is its provision of a p-value for each feature's importance score, enabling a statistical test to assert a feature's importance. However, it is important to acknowledge that the PIMP algorithm significantly increases computational demands compared to PFI.

6.4 Considerations for real-world application

The LSTM method examined in this thesis is designed to illustrate how institutional investors, such as banks and large funds, can utilise machine learning to address the traditional challenges of predicting excess stock returns. However, it is crucial to note that this thesis does not endorse the direct application of this LSTM method, as a passive buy-and-hold strategy for the market index was found to outperform it. Despite this, the analysis revealed that the LSTM strategy generated positive excess returns, even after accounting for transaction costs. However, examining portfolio patterns suggested that the LSTM strategy might not be suitable for generating promising short- or long-term investment results. Moreover, the portfolio patterns indicated that portfolios 1 and 10 consist of almost identical industries and predominantly include growth stocks. This raises two primary concerns: First, the nearly identical stock composition of both portfolios increases risk due to a lack of diversification. Second, focusing primarily on growth stocks may result in substantial transaction costs and slippage, reducing the portfolios' excess returns. Moreover, growth stocks are inherently more volatile and risky than value stocks, offering the potential for substantial returns but also greater exposure to market fluctuations. Consequently, an LSTM strategy heavily emphasising growth stocks will probably experience increased volatility and potential drawdowns. Since this thesis analysed only 2,552 stocks, incorporating a larger number of stocks into the portfolio might lead to increased transaction costs, potentially negatively affecting the observed positive excess returns.

To overcome the challenges of this method, this thesis proposes a different strategy if this method were to be implemented by real-life institutional investors. Instead of focusing on a vast array of individual stocks, the strategy could shift towards sector-based investments. This approach would involve using the LSTM to predict the performance of various sector exchange-traded funds (ETFs) and set up a similar 10-1 portfolio strategy without needing over 20,000 stocks. Sector-based investing can reduce the granularity of stock selection, lowering transaction costs and slippage while providing diversified portfolios (Horst, 2022).

7 Conclusion

This thesis investigates five LSTM network structures, ranging from one to five LSTM layers, to predict monthly US excess stock returns. As a result, the model with two LSTM layers demonstrates the best performance. Subsequently, permutation feature importance is employed to identify the top 20 most significant predictor variables from a pool of 94 stock

characteristics and 13 macroeconomic predictors. The multivariate LSTM, incorporating all 20 features, is then compared to a baseline univariate LSTM, which includes only the monthly excess stock returns. This comparison aims to evaluate the impact of incorporating additional features on predictive performance. The three main hypotheses investigated in this thesis are:

H1 *Employing permutation feature importance in conjunction with an LSTM model uncovers the most critical features for predicting excess stock returns.*

H2 *LSTM models exhibit superior accuracy in predicting excess stock returns compared to traditional feed-forward neural networks.*

H3 *A 10-1 hedge portfolio constructed using LSTM predictions generates significant economic gains, net of transaction costs.*

The first hypothesis (**H1**) is not rejected. While the PFI did not select any features related to recent price movements, it identified two features related to risk measures, three features related to liquidity, and ten features related to valuation ratios and fundamental indicators. Additionally, the five most influential macroeconomic predictors are inflation, 3-Month Treasury Bill, long-term yield, term spread, and long-term rate of returns. The greater performance of the multivariate LSTM model compared to the univariate LSTM model suggests that incorporating these additional features is beneficial.

The second hypothesis (**H2**) is rejected. The LSTM models did not demonstrate superior accuracy in forecasting excess stock returns compared to the traditional feed-forward neural network employed by Gu et al. (2020). This outcome is primarily attributed to challenges related to overfitting and network architecture optimisation within the LSTM models.

The third hypothesis (**H3**) is also rejected. While the 10-1 hedge portfolio strategy based on the multivariate LSTM predictions generated small but statistically significant positive excess returns, it did not outperform a passive buy-and-hold strategy invested in the S&P 500 index. The annualised Sharpe ratio for the multivariate LSTM is 0.447. In contrast, the S&P 500 index achieved a Sharpe ratio of 0.96 during the period from 2009 to 2020.

References

Adila, P. N., Saepudin, D., & Ihsan, A. F. (2022). Prediction of stocks return in the lq45 index with long-short-term-memory (lstm) and its application for portfolio selection. *2022 10th International Conference on Information and Communication Technology (ICoICT)*, 194–199.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347.

Ang, A., & Bekaert, G. (2006). Stock Return Predictability: Is it There? *The Review of Financial Studies*, *20*(3), 651–707. https://doi.org/10.1093/rfs/hhl021

Bali, T. G., Engle, R. F., & Murray, S. (2016). *Empirical asset pricing: The cross section of stock returns*. John Wiley & Sons.

Ball, R. (1978). Anomalies in relationships between securities' yields and yield-surrogates. *Journal of financial economics*, *6*(2-3), 103–126.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of financial economics*, *9*(1), 3–18.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, *5*(2), 157–166.

Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, *34*(2), 1046–1089.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–23.

Campbell, J. Y. (1987). Stock returns and the term structure. *Journal of financial economics*, *18*(2), 373–399.

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of finance*, *63*(6), 2899–2939.

Campbell, J. Y., & Shiller, R. J. (1988). Stock prices, earnings, and expected dividends. *the Journal of Finance*, *43*(3), 661–676.

Campbell, J. Y., & Vuolteenaho, T. (2004). Inflation illusion and stock prices. *American Economic Review*, *94*(2), 19–23.

Campbell, J. Y., & Yogo, M. (2006). Efficient tests of stock return predictability. *Journal of financial economics*, *81*(1), 27–60.

Chalvatzis, C., & Hristu-Varsakelis, D. (2019). High-performance stock index trading: Making effective use of a deep lstm neural network. *arXiv preprint arXiv:1902.03125*.

Chen, K., Zhou, Y., & Dai, F. (2015). A lstm-based method for stock returns prediction: A case study of china stock market. *2015 IEEE international conference on big data (big data)*, 2823–2824.

Chen, M. X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L., Parmar, N., Schuster, M., Chen, Z., et al. (2018). The best of both worlds: Combining recent advances in neural machine translation. *arXiv preprint arXiv:1804.09849*.

CRSP. (2024). Mfe database. https://www.crsp.org

Da, Z., Jagannathan, R., & Shen, J. (2014). *Growth expectations, dividend yields, and future stock returns* (tech. rep.). National Bureau of Economic Research.

Dami, S., & Esterabi, M. (2021). Predicting stock returns of tehran exchange using lstm neural network and feature engineering technique. *Multimedia Tools and Applications*, *80*(13), 19947–19970.

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies*, *22*(5), 1915–1953.

Deotte, C. (2021). Lstm feature importance. https://www.kaggle.com/code/cdeotte/lstm-feature-importance/notebook

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, *74*(366a), 427–431.

Fama, E. F., & French, K. R. (1989). Business conditions and expected returns on stocks and bonds. *Journal of financial economics*, *25*(1), 23–49.

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.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, *116*(1), 1–22.

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, *81*(3), 607–636.

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.

French, K. R. (2024). Data library. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/f-f_factors.html

Gaur, Y. (2023). Stock market price prediction using lstm. *International Journal for Research in Applied Science and Engineering Technology*, *11*, 1881–1887. https://doi.org/10.22214/ijraset.2023.57673

GfG. (2023, January). Difference between feed-forward neural networks and recurrent neural networks. https://www.geeksforgeeks.org/difference-between-feed-forward-neural-networks-and-recurrent-neural-networks

Ghosh, P., Neufeld, A., & Sahoo, J. K. (2022). Forecasting directional movements of stock prices for intraday trading using lstm and random forests. *Finance Research Letters*, *46*, 102280.

Goyal, A. (2023). Ankit goyal [Accessed: 2024-01-16].

Green, J., Hand, J. R., & Zhang, X. F. (2017). The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies*, *30*(12), 4389–4436.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Guo, H. (2006). On the out-of-sample predictability of stock market returns. *The Journal of Business*, *79*(2), 645–670.

Hamdaoui, M., Ayouni, S., & Maktouf, S. (2022). Financial crises: Explanation, prediction, and interdependence. *SN Business & Economics*, *2*(8), 88.

Hansson, M. (2017). On stock return prediction with lstm networks.

Hanu, S. (2021). Lstm - derivation of back propagation through time [Accessed: 30 May 2024]. https://www.geeksforgeeks.org/lstm-derivation-of-back-propagation-through-time/

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *The Journal of Finance*, *72*(4), 1399–1440.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5–68.

Hautsch, N., & Voigt, S. (2019). Large-scale portfolio allocation under transaction costs and model uncertainty. *Journal of Econometrics*, *212*(1), 221–240.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Horst, J. (2022). *Machine learning for exchange traded fund price predictions* [B.S. thesis]. University of Twente.

K., V., & K., S. (2022). Towards activation function search for long short-term model network: A differential evolution based approach. *Journal of King Saud University - Computer and Information Sciences*, *34*(6, Part A), 2637–2650. https://doi.org/https://doi.org/10.1016/j.jksuci.2020.04.015

Karmiani, D., Kazi, R., Nambisan, A., Shah, A., & Kamble, V. (2019). Comparison of predictive algorithms: Backpropagation, svm, lstm and kalman filter for stock market. *2019 amity international conference on artificial intelligence (AICAI)*, 228–234.

Keim, D. B., & Stambaugh, R. F. (1986). Predicting returns in the stock and bond markets. *Journal of financial Economics*, *17*(2), 357–390.

Keras. (2021). Keras documentation: Layer activation functions [Accessed: 2024-02-20]. https://keras.io/api/layers/activations/

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kothari, S. P., & Shanken, J. (1997). Book-to-market, dividend yield, and expected market returns: A time-series analysis. *Journal of Financial economics*, *44*(2), 169–203.

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.

Lamont, O. (1998). Earnings and expected returns. *The journal of Finance*, *53*(5), 1563–1587.

Lewellen, J. (2014). The cross section of expected stock returns. *Forthcoming in Critical Finance Review, Tuck School of Business Working Paper*, (2511246).

Ma, Y., Han, R., & Fu, X. (2019). Stock prediction based on random forest and lstm neural network. *2019 19th International Conference on Control, Automation and Systems (ICCAS)*, 126–130.

Masters, T. (1993). *Practical neural network recipes in c++*. Morgan Kaufmann.

Mi, Y., Xu, D., & Gao, T. (2023). Application of pca-lstm algorithm for financial market stock return prediction and optimization model. *International Journal for Simulation and Multidisciplinary Design Optimization*, *14*, 8.

Moghar, A., & Hamiche, M. (2020). Stock market prediction using lstm recurrent neural network. *Procedia Computer Science*, *170*, 1168–1173.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Moon, C. (2018). L1 and l2 as regularization for a linear model [Accessed on 2024-02-21].

Naeini, M. P., Taremian, H., & Hashemi, H. B. (2010). Stock market value prediction using neural networks. *2010 international conference on computer information systems and industrial management applications (CISIM)*, 132–136.

Naik, N., & Mohan, B. R. (2019). Study of stock return predictions using recurrent neural networks with lstm. *Engineering Applications of Neural Networks: 20th International Conference, EANN 2019, Xersonisos, Crete, Greece, May 24-26, 2019, Proceedings 20*, 453–459.

Olivares-Nadal, A. V., & DeMiguel, V. (2018). A robust perspective on transaction costs in portfolio optimization. *Operations Research*, *66*(3), 733–739.

Rosenberg, B., Reid, K., & Lanstein, R. (1985). Persuasive evidence of market inefficiency (spring 1985). In *Streetwise* (pp. 48–55). Princeton University Press.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, *13*(3), 341–360. https://doi.org/https://doi.org/10.1016/0022-0531(76)90046-6

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*, 211–252.

Scheuch, C., Voigt, S., & Weiss, P. (2023). *Tidy finance with r* (1st). Chapman; Hall/CRC. https://doi.org/https://doi.org/10.1201/b23237

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, *19*(3), 425–442.

Sharpe, W. F. (1966). Mutual fund performance. *The Journal of business*, *39*(1), 119–138.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *nature*, *550*(7676), 354–359.

Verbeek, M. (2017). *A guide to modern econometrics*. John Wiley & Sons.

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455–1508.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). *Dive into deep learning*. Cambridge University Press.

Zhang, M. Z. (2021). *Stock returns and inflation redux: An explanation from monetary policy in advanced and emerging markets*. International Monetary Fund.

Appendix

Table 8: Summary of stock characteristics

No.	Acronym	Definition of the characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
1	absacc	Absolute accruals	Bandyopadhyay, Huang & Wirjanto	2010, WP	Compustat	Annual
2	acc	Working capital accruals	Sloan	1996, TAR	Compustat	Annual
3	aeavol	Abnormal earnings announcement volume	Lerman, Livnat & Mendenhall	2007, WP	Compustat+CRSP	Quarterly
4	age	# years since first Compustat coverage	Jiang, Lee & Zhang	2005, RAS	Compustat	Annual
5	agr	Asset growth	Cooper, Gulen & Schill	2008, JF	Compustat	Annual
6	baspread	Bid-ask spread	Amihud & Mendelson	1989, JF	CRSP	Monthly
7	beta	Beta	Fama & MacBeth	1973, JPE	CRSP	Monthly
8	betasq	Beta squared	Fama & MacBeth	1973, JPE	CRSP	Monthly
9	bm	Book-to-market	Rosenberg, Reid & Lanstein	1985, JPM	Compustat+CRSP	Annual
10	bm ia	Industry-adjusted book to market	Asness, Porter & Stevens	2000, WP	Compustat+CRSP	Annual
11	cash	Cash holdings	Palazzo	2012, JFE	Compustat	Quarterly
12	cashdebt	Cash flow to debt	Ou & Penman	1989, JAE	Compustat	Annual
13	cashpr	Cash productivity	Chandrashekar & Rao	2009, WP	Compustat	Annual
14	cfp	Cash flow to price ratio	Desai, Rajgopal & Venkatachalam	2004, TAR	Compustat	Annual
15	cfp ia	Industry-adjusted cash flow to price ratio	Asness, Porter & Stevens	2000, WP	Compustat	Annual
16	chatoia	Industry-adjusted change in asset turnover	Soliman	2008, TAR	Compustat Annual	Annual
17	chcsho	Change in shares outstanding	Pontiff & Woodgate	2008, JF	Compustat Annual	Annual

Continued on next page

Table 8 – continued from previous page

No.	Acronym	Definition of the characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
18	chempia	Industry-adjusted change in employees	Asness, Porter & Stevens	1994, WP	Compustat Annual	Annual
19	chinv	Change in inventory	Thomas & Zhang	2002, RAS	Compustat Annual	Annual
20	chmom	Change in 6-month momentum	Gettleman & Marks	2006, WP	CRSP Monthly	Monthly
21	chpmia	Industry-adjusted change in profit margin	Soliman	2008, TAR	Compustat Annual	Annual
22	chtx	Change in tax expense	Thomas & Zhang	2011, JAR	Compustat Quarterly	Quarterly
23	cinvest	Corporate investment	Titman, Wei & Xie	2004, JFQA	Compustat Quarterly	Quarterly
24	convind	Convertible debt indicator	Valta	2016, JFQA	Compustat Annual	Annual
25	currat	Current ratio	Ou & Penman	1989, JAE	Compustat Annual	Annual
26	depr	Depreciation / PP&E	Holthausen & Larcker	1992, JAE	Compustat Annual	Annual
27	divi	Dividend initiation	Michaely, Thaler & Womack	1995, JF	Compustat Annual	Annual
28	divo	Dividend omission	Michaely, Thaler & Womack	1995, JF	Compustat Annual	Annual
29	dolvol	Dollar trading volume	Chordia, Subrahmanyam & Anshuman	2001, JFE	CRSP Monthly	Monthly
30	dy	Dividend to price	Litzenberger & Ramaswamy	1982, JF	Compustat Annual	Annual
31	ear	Earnings announcement return	Kishore, Brandt, Santa-Clara & Venkatachalam	2008, WP	Compustat+CRSP Quarterly	Quarterly
32	egr	Growth in common shareholder equity	Richardson, Sloan, Soliman & Tuna	2005, JAE	Compustat	Annual
33	ep	Earnings to price	Basu	1977, JF	Compustat	Annual
34	gma	Gross profitability	Novy-Marx	2013, JFE	Compustat	Annual

Continued on next page

Table 8 – continued from previous page

No.	Acronym	Definition of the characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
35	grCAPX	Growth in capital expenditures	Anderson & Garcia-Feijoo	2006, JF	Compustat	Annual
36	grltnoa	Growth in long term net operating assets	Fairfield, Whisenant & Yohn	2003, TAR	Compustat	Annual
37	herf	Industry sales concentration	Hou & Robinson	2006, JF	Compustat	Annual
38	hire	Employee growth rate	Bazdresch, Belo & Lin	2014, JPE	Compustat	Annual
39	idiovol	Idiosyncratic return volatility	Ali, Hwang & Trombley	2003, JFE	CRSP	Monthly
40	ill	Illiquidity	Amihud	2002, JFM	CRSP	Monthly
41	indmom	Industry momentum	Moskowitz & Grinblatt	1999, JF	CRSP	Monthly
42	invest	Capital expenditures and inventory	Chen & Zhang	2010, JF	Compustat	Annual
43	lev	Leverage	Bhandari	1988, JF	Compustat	Annual
44	lgr	Growth in long-term debt	Richardson, Sloan, Soliman & Tuna	2005, JAE	Compustat	Annual
45	maxret	Maximum daily return	Bali, Cakici & Whitelaw	2011, JFE	CRSP	Monthly
46	mom12m	12-month momentum	Jegadeesh	1990, JF	CRSP	Monthly
47	mom1m	1-month momentum	Jegadeesh & Titman	1993, JF	CRSP	Monthly
48	mom36m	36-month momentum	Jegadeesh & Titman	1993, JF	CRSP	Monthly
49	mom6m	6-month momentum	Jegadeesh & Titman	1993, JF	CRSP	Monthly
50	ms	Financial statement score	Mohanram	2005, RAS	Compustat	Quarterly
51	mvel1	Size	Banz	1981, JFE	CRSP	Monthly
52	mve ia	Industry-adjusted size	Asness, Porter & Stevens	2000, WP	Compustat	Annual
53	nincr	Number of earnings increases	Barth, Elliott & Finn	1999, JAR	Compustat	Quarterly
54	operprof	Operating profitability	Fama & French	2015, JFE	Compustat	Annual

Continued on next page

Table 8 – continued from previous page

No.	Acronym	Definition of the characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
55	orgcap	Organizational capital	Eisfeldt & Papanikolaou	2013, JF	Compustat	Annual
56	pchcapx ia	Industry adjusted % change in capital expenditures	Abarbanell & Bushee	1998, TAR	Compustat	Annual
57	pchcurrat	% change in current ratio	Ou & Penman	1989, JAE	Compustat	Annual
58	pchdepr	% change in depreciation	Holthausen & Larcker	1992, JAE	Compustat	Annual
59	pchgm pchsale	% change in gross margin - % change in sales	Abarbanell & Bushee	1998, TAR	Compustat	Annual
60	pchquick	% change in quick ratio	Ou & Penman	1989, JAE	Compustat	Annual
61	pchsale pchinvt	% change in sales - % change in inventory	Abarbanell & Bushee	1998, TAR	Compustat	Annual
62	pchsale pchrect	% change in sales - % change in A/R	Abarbanell & Bushee	1998, TAR	Compustat	Annual
63	pchsale pchxsga	% change in sales - % change in SG&A	Abarbanell & Bushee	1998, TAR	Compustat	Annual
64	pchsaleinv	% change sales-to-inventory	Ou & Penman	1989, JAE	Compustat	Annual
65	pctacc	Per cent accruals	Hafzalla, Lundholm & Van Winkle	2011, TAR	Compustat	Annual
66	pricedelay	Price delay	Hou & Moskowitz	2005, RFS	CRSP	Monthly
67	ps	Financial statements score	Piotroski	2000, JAR	Compustat	Annual
68	quick	Quick ratio	Ou & Penman	1989, JAE	Compustat	Annual
69	rd	R&D increase	Eberhart, Maxwell & Siddique	2004, JF	Compustat	Annual
70	rd mve	R&D to market capitalization	Guo, Lev & Shi	2006, JBFA	Compustat	Annual
71	rd sale	R&D to sales	Guo, Lev & Shi	2006, JBFA	Compustat	Annual

Continued on next page

Table 8 – continued from previous page

No.	Acronym	Definition of the characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
72	realestate	Real estate holdings	Tuzel	2010, RFS	Compustat	Annual
73	retvol	Return volatility	Ang, Hodrick, Xing & Zhang	2006, JF	CRSP	Monthly
74	roaq	Return on assets	Balakrishnan, Bartov & Faurel	2010, JAE	Compustat	Quarterly
75	roavol	Earnings volatility	Francis, LaFond, Olsson & Schipper	2004, TAR	Compustat	Quarterly
76	roeq	Return on equity	Hou, Xue & Zhang	2015, RFS	Compustat	Quarterly
77	roic	Return on invested capital	Brown & Rowe	2007, WP	Compustat	Annual
78	rsup	Revenue surprise	Kama	2009, JBFA	Compustat	Quarterly
79	salecash	Sales to cash	Ou & Penman	1989, JAE	Compustat	Annual
80	saleinv	Sales to inventory	Ou & Penman	1989, JAE	Compustat	Annual
81	salerec	Sales to receivables	Ou & Penman	1989, JAE	Compustat	Annual
82	secured	Secured debt	Valta	2016, JFQA	Compustat	Annual
83	securedind	Secured debt indicator	Valta	2016, JFQA	Compustat	Annual
84	sgr	Sales growth	Lakonishok, Shleifer & Vishny	1994, JF	Compustat	Annual
85	sin	Sin stocks	Hong & Kacperczyk	2009, JFE	Compustat	Annual
86	sp	Sales to price	Barbee, Mukherji & Raines	1996, FAJ	Compustat	Annual
87	std dolvol	Volatility of liquidity (dollar trading volume)	Chordia, Subrahmanyam & Anshuman	2001, JFE	CRSP	Monthly
88	std turn	Volatility of liquidity (share turnover)	Chordia, Subrahmanyam, & Anshuman	2001, JFE	CRSP	Monthly
89	stdacc	Accrual volatility	Bandyopadhyay, Huang & Wirjanto	2010, WP	Compustat	Quarterly
90	stdcf	Cash flow volatility	Huang	2009, JEF	Compustat	Quarterly

Continued on next page

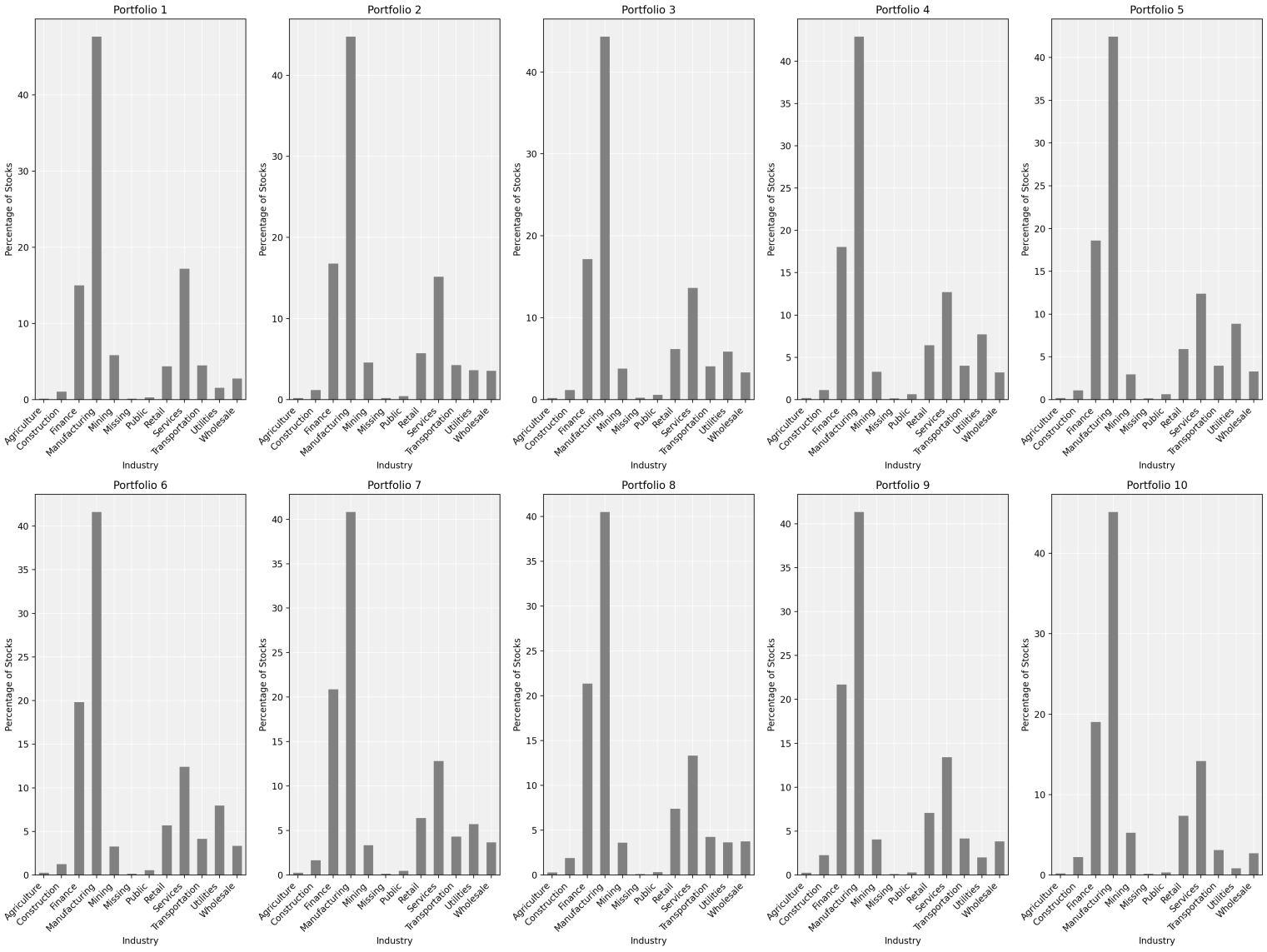
Table 8 – continued from previous page

No.	Acronym	Definition of the characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
91	tang	Debt capacity/firm tangibility	Almeida & Campello	2007, RFS	Compustat	Annual
92	tb	Tax income to book income	Lev & Nissim	2004, TAR	Compustat	Annual
93	turn	Share turnover	Datar, Naik & Radcliffe	1998, JFM	CRSP	Monthly
94	zerotrade	Zero trading days	Liu	2006, JFE	CRSP	Monthly

Source: Green et al. (2017).

Note The list contains all the stock characteristics applied in this thesis.

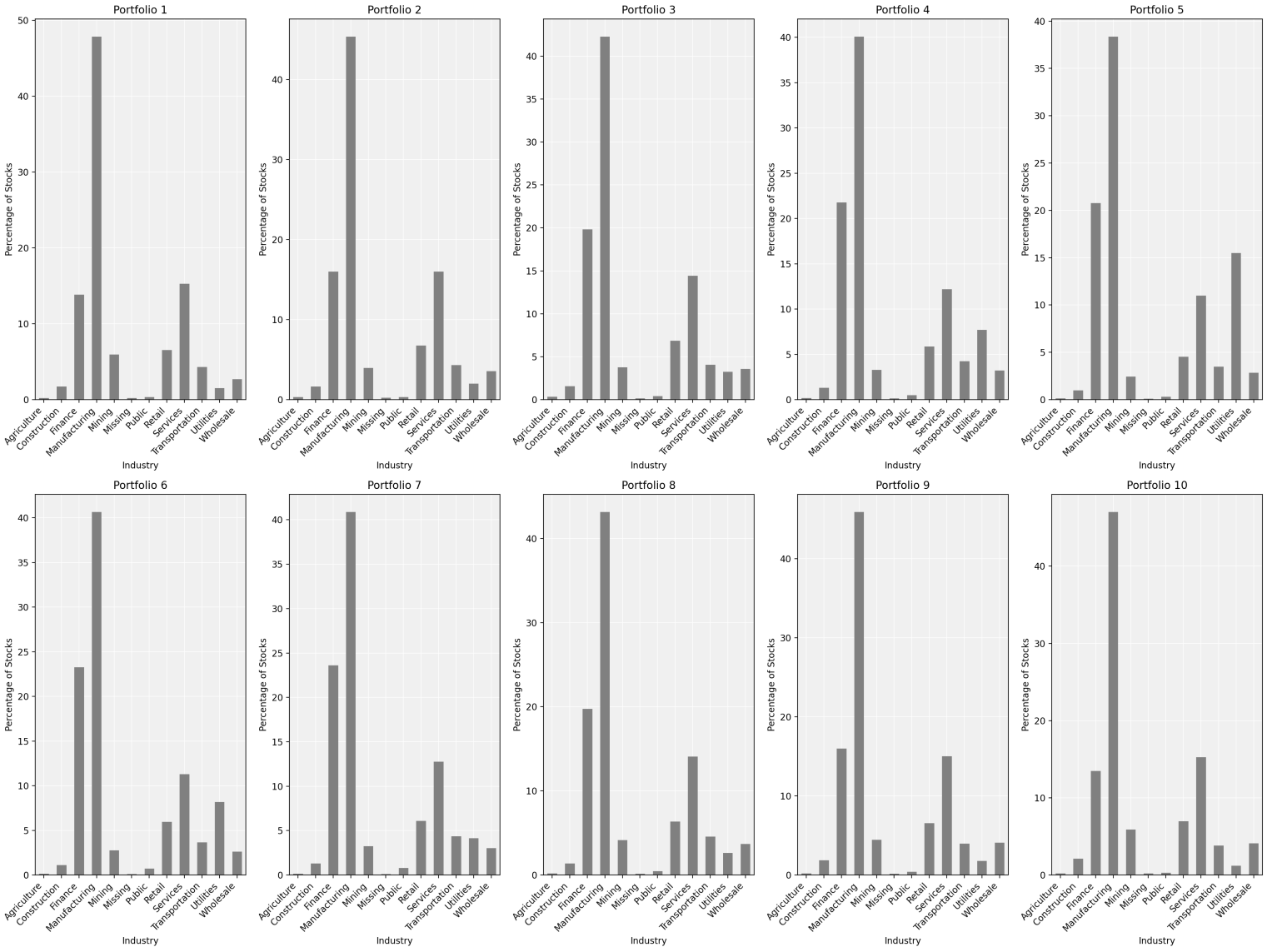
Figure 12: Industry by portfolio 1 to 10 of LSTM2 during the out-of-sample period 2009-2020



Source: Own calculations.

Note: The industry classifications follow Bali et al. (2016) procedure.

Figure 13: Industry by portfolio 1 to 10 of univariate LSTM during the out-of-sample period 2009-2020



Source: Own calculations.

Note: The industry classifications follow Bali et al. (2016) procedure.

Table 9: Summary statistics of the 15 most important stock characteristics and excess return from 1960 to 2020

	mean	std	min	25%	50%	75%	max
mvel1	0.2	0.58	-1	-0.281	0.297	0.72	1
bm_ia	-0.066	0.532	-0.99	-0.51	-0.049	0.363	1
chinv	0.028	0.553	-0.99	-0.442	0.03	0.492	1
idiovol	-0.066	0.516	-1	-0.502	-0.084	0.332	0.991
divi	-0.007	0.151	-0.051	-0.037	-0.031	-0.025	1
stdcf	-0.088	0.449	-1	-0.374	-0.022	0.052	0.991
convind	-0.014	0.316	-0.304	-0.134	-0.112	-0.092	1
depr	-0.038	0.518	-1	-0.448	0.003	0.348	1
currat	-0.002	0.531	-0.99	-0.431	0.01	0.419	1
beta	0.025	0.533	-0.99	-0.411	0.038	0.461	0.991
pchcurrat	0.024	0.504	-0.99	-0.356	0.007	0.416	1
roic	0.106	0.528	-0.99	-0.279	0.077	0.557	1
roeq	0.085	0.499	-0.99	-0.238	0.019	0.487	1
orgcap	0.05	0.457	-1	-0.174	0.046	0.318	1
grcapx	0.016	0.503	-0.99	-0.35	0.004	0.395	1
ret_excess	0.011	0.152	-0.995	-0.057	0.002	0.066	10.34

Source: Own calculations.

Note: The 15 most important features have been normalised with a min-max scalar from -1 and 1. The variable ret_excess in this table shows the raw data, but is later being normalised to have a mean zero.

Upon normalising the dependent variable, the Augmented Dickey-Fuller (ADF) test, developed by Dickey and Fuller (1979), is utilised to ascertain whether the variable excess return is stationary. The test yields a Dickey-Fuller statistic of -24.325. This statistic is derived using a lag order of 110, as determined by the Akaike Information Criterion (AIC), and is accompanied by a p-value of 0.000. The null hypothesis of the ADF test posits that the variable ret_excess is non-stationary. In contrast, the alternative hypothesis contends that the data are stationary. Given the exceedingly low p-value of 0.000, there is compelling evidence to reject the null hypothesis, suggesting that the variable is stationary.

Table 10: Augmented Dickey-Fuller Test

Data	Dickey-Fuller	Lag order	P-value
ret_excess	-24.325	110	0.000

Source: Own calculations.