International Journal of
Financial Studies

MDPI

Article

Predictive Power of ESG Factors for DAX ESG 50 Index Forecasting Using Multivariate LSTM

Manuel Rosinus [1,*] **and Jan Lansky** [2]

[1] Department of Finance, Faculty of Economic Studies, University of Finance and Administration, 101 00 Prague, Czech Republic
[2] Department of Computer Science and Mathematics, Faculty of Economic Studies, University of Finance and Administration, 101 00 Prague, Czech Republic; lansky@mail.vsfs.cz
* Correspondence: 39384@mail.vsfs.cz

Abstract

As investors increasingly use Environmental, Social, and Governance (ESG) criteria, a key challenge remains: ESG data is typically reported annually, while financial markets move much faster. This study investigates whether incorporating annual ESG scores can improve monthly stock return forecasts for German DAX-listed firms. We employ a multivariate long short-term memory (LSTM) network, a machine learning model ideal for time series data, to test this hypothesis over two periods: an 8-year analysis with a full set of ESG scores and a 16-year analysis with a single disclosure score. The evaluation of model performance utilizes standard error metrics and directional accuracy, while statistical significance is assessed through paired statistical tests and the Diebold–Mariano test. Furthermore, we employ SHapley Additive exPlanations (SHAP) to ensure model explainability. We observe no statistically significant indication that incorporating annual ESG data enhances forecast accuracy. The 8-year study indicates that using a comprehensive ESG feature set results in a statistically significant increase in forecast error (RMSE and MAE) compared to a baseline model that utilizes solely historical returns. The ESG-enhanced model demonstrates no significant performance disparity compared to the baseline across the 16-year investigation. Our findings indicate that within the one-month-ahead projection horizon, the informative value of low-frequency ESG data is either fully incorporated into the market or is concealed by the significant forecasting capability of the historical return series. This study's primary contribution is to demonstrate, through out-of-sample testing, that standard annual ESG information holds little practical value for generating predictive alpha, urging investors to seek more timely, alternative data sources.

Keywords: financial markets; econometrics; forecast; machine learning; LSTM; ESG

check for updates

Academic Editors: Brian Bolton and Panagiotis Dimitropoulos

Received: 6 August 2025
Revised: 1 September 2025
Accepted: 2 September 2025
Published: 4 September 2025

Citation: Rosinus, M., & Lansky, J. (2025). Predictive Power of ESG Factors for DAX ESG 50 Index Forecasting Using Multivariate LSTM. *International Journal of Financial Studies*, *13*(3), 167. https://doi.org/10.3390/ijfs13030167

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

ESG (Environmental, Social, and Governance) factors have become increasingly important in influencing stock prices and investor decisions. Strong ESG performance typically leads to higher stock prices and reduced price volatility, which can affect how investors respond to new information. However, the impact can vary depending on market context, investor behavior, and the level of agreement among ESG raters.

The complex and rapidly evolving regulatory landscape in the European Union, particularly for German firms, underscores the importance of ESG factors but also creates challenges in data standardization and analysis. There is no consensus on a universally

accepted framework for ESG assessment in the market. Several indices and standards are available to guide corporate reporting. The most prominent frameworks are the United Nations Principles for Responsible Investment (UNPRI), the United Nations Sustainable Development Goals (UNSDGs), the Sustainability Accounting Standards Board (SASB) Standards, and the Global Reporting Initiative (GRI) Standards. The acceptance of companies varies across frameworks because each one employs a unique scoring methodology and evaluates different dimensions. Additionally, there are ongoing discussions regarding the quantification methods and agendas used by these ESG systems. As no standardized model, algorithm, or heuristic currently exists, this allows for interpretation by human judgment.

One of the most famous German indices, which tries to capture the 50 largest, most liquid German stocks, is the DAX ESG 50[1] index. According to the creator company, STOXX (Germany) of the index, it considers companies that "have passed standardized ESG screens related to Global Standards Screening, as well as the involvement in controversial weapons, tobacco production, thermal coal, nuclear power, and military contracting, and feature comparably good performance based on their Environmental, Social and Governance criteria" 1-DAX 50 ESG (DAXESGK) (DE000A0S3E04) (n.d.).

Reporting under specific frameworks is voluntary for companies; however, legally mandatory reporting requirements also exist. EU directives are translated into national laws for DAX ESG 50 companies, specifically German national law. EU standards are non-legally binding technical specifications or guidelines that are typically used in conjunction with legally binding regulations or national laws. Applicable regulations, rules, and standards are listed in Appendix A, Table A1.

Companies in the EU criticized the complex regulatory obligations and challenging timeline, which forced the European Commission to readjust the scope and timeline. In an EU Omnibus Directive, the EU has made significant adjustments to the sustainability reporting obligations within the European Union, aiming to simplify and ease the administrative burden for companies. Adopted officially in April 2025, this directive revises existing regulations, particularly the Corporate Sustainability Reporting Directive (CSRD) and the associated European Sustainability Reporting Standards (ESRS). The Omnibus package introduces substantial changes to reduce the administrative burden on companies (European Commission, Directorate-General for Financial Stability, Financial Services and Capital Markets Union, 2025). A key change is raising the employee threshold for mandatory ESG reporting from 250 to 1000. This will significantly reduce the number of companies legally required to produce detailed sustainability disclosures. The regulation requires all EU member states, including Germany, to incorporate these changes into their domestic laws by 31 December 2025. The German government is currently implementing this to reduce compliance costs and simplify ESG reporting for German companies.

As it is a challenge to obtain comparable ESG ratings from different companies based on the mentioned reports, several commercial data providers, such as Bloomberg, Sustainalytics (by Morningstar), ISS Global, and MSCI, provide their own scoring.

Previous work compares the forecasting performance for the DAX ESG 50 index using ARIMA and a univariate LSTM model during the 2023–2024 period. Considering only the daily closing prices, the LSTM model provided higher accuracy, as measured by metrics such as RMSE and MAE, compared to classical ARIMA models (Rosinus, 2025). The DAX ESG 50 index was chosen as a representative selection of multi-sector companies, which are highly liquid and have an established reputation in ESG reporting.

Sustainable Finance (or Green Investing/Eco-Investing) is incorporating Environmental, Social, and Governance factors into financial decision-making. Financial institutions and policymakers promote sustainability through strategies, regulations, and initiatives

implemented by banks, asset managers, and institutional investors to include sustainability factors in their investments.

Additionally, an increasing number of products such as green bonds, social impact bonds, sustainable funds, and other sustainable investment vehicles are offered (Dahbi et al., 2024; Maltais & Nykvist, n.d.). This study aims to extend the previous work by testing the predictive value of ESG data within a modern machine learning framework.

For the following three core research questions, answers are sought:

- (RQ1): Does the inclusion of a comprehensive set of ESG scores lead to a statistically significant improvement in the forecast accuracy of monthly stock returns compared to a baseline model?
- (RQ2): Does the predictive value of a single ESG factor (the Disclosure Score) persist over a longer, 16-year time horizon?
- (RQ3): Which specific ESG factors (e.g., Environmental, Social, and Governance), if any, have the most significant impact on the model's predictions?

To answer these questions, the earlier analysis based on a univariate LSTM model is expanded by developing a multivariate version. This extended model incorporates not only the monthly return but also historical ESG data at the company level. The dataset includes firms that were part of the DAX ESG 50 index between March 2020 and July 2025, representing the biggest and most liquid companies in Germany. As ESG data is not available for all companies in the index, only tickers for which data was used and had sufficient data available are included. LSTM models are extraordinarily good at capturing temporal relationships seen in financial time series data because of their internal memory architectures. Having the capabilities to combine multiple variables into one sequential deep neural network is the ideal method to combine ESG scores and time series data.

Integrating monthly index data with annual ESG data provides the network with a sufficient amount of historical context, allowing for a balanced consideration that does not overly prioritize closing prices relative to the yearly data frequency.

The rest of the paper is organized as follows Section 2 provides a Literature Review, Section 3 covers the data sources, the steps taken to clean and prepare the data, and the structure of both the univariate and multivariate LSTM models. Section 4 presents the main results and provides an interpretation of the outcomes. Section 5 discussed the outcomes, and Section 6 concludes with a summary that highlights current limitations.

This work offers the following main contributions:

- **Methodological Contribution:** We provide a pragmatic and direct application of a multivariate LSTM network to systematically test the predictive value of annual ESG data on monthly returns. Our contribution is not the development of a new hybrid model, but rather the design of a clean, transparent, and easily replicable out-of-sample experiment that directly simulates a real-world investment process.
- **Empirical Finding:** We provide robust evidence of a null result, demonstrating that the inclusion of different annual ESG scores does not lead to a statistically significant improvement in forecast accuracy for the DAX ESG 50 index. In our 8-year analysis, adding ESG data significantly increased forecast error, suggesting it acted as noise rather than a valuable signal.
- **Practical Implications:** Our findings have a direct impact on quantitative investors and portfolio managers. The findings show that typical ESG data cannot easily be leveraged to generate alpha since its informational content appears to have been either priced in by the market or exceeded by the forecasting potential of past returns. This emphasizes the importance of more immediate, forward-looking ESG signals for accurate financial forecasts.

2. Literature Review

Traditional time series forecasting models, including classic econometric approaches such as autoregressive integrated moving average (ARIMA) and standard machine learning techniques, often rely solely on historical price data (Hyndman & Athanasopoulos, 2018, p. 221). While models like LSTM networks, a type of RNN, have demonstrated considerable success in capturing complex, non-linear temporal dependencies inherent in financial time series, wind speeds, and COVID-19 trends (ArunKumar et al., 2022; Elsaraiti & Merabet, 2021; Fischer & Krauss, 2018; Latif et al., 2023; Siami-Namini & Namin, 2018), classical ARIMA models can still outperform LSTM models when strong linear trends, limited data, or single features (e.g., historical prices) are used (Gu et al., 2019; Kobiela et al., 2022; Ning et al., 2022; Zhang et al., 2022). Our central research question involves testing the predictive power of several ESG scores alongside historical returns. LSTMs are naturally suited to handling multiple input variables simultaneously, allowing the model to learn the complex, time-varying interactions between them (Gao et al., 2023). Disadvantages of LSTM compared to ARIMA are the complexity and interpretability. LSTMs are "black box" models with their internal decision-making not transparent, unlike ARIMA models, which have statistically interpretable components.

The integration of ESG factors into investment decisions has led to an essential body of academic research. Still, the relationship between ESG performance and corporate financial performance remains a source of intense debate. The literature is divided, with findings that often appear to be contradictory. A critical synthesis shows that such different conclusions are due to discrepancies in methodology, the specific ESG elements examined, and the metrics employed to define "performance."

One stream of research suggests a positive link between strong ESG credentials and financial outcomes. These studies often frame ESG as a proxy for good management and risk mitigation. While fulfilling environmental and social responsibilities does not directly improve corporate financial performance, it allows firms to communicate effectively with stakeholders, including the government, regulatory agencies, financial institutions, and investors (Feng et al., 2022). As ESG factors become increasingly important for all stakeholders, specialized market indices such as the DAX ESG 50 have been developed to track companies with strong ESG profiles, providing benchmarks for responsible investment portfolios. Accurate forecasting of such an index is crucial for portfolio management, risk assessment, and derivative pricing.

ESG factors may influence stock performance, affecting both returns and risk. Recent research in this direction has not provided clear evidence and yields different conclusions. However, research on this topic does not lead to one clear conclusion. Some studies suggest that ESG has a positive effect on stock performance. For example, Kulal et al. (2023) found that companies with strong ESG scores often show higher stock prices and better investment results. These companies are viewed as more sustainable and financially sound. Luo et al. (2024) also found that strong ESG practices can reduce the risk of stock price crashes, especially in privately owned companies. This is likely because good ESG performance may limit risky behavior and reduce the need for aggressive earnings management. Another study by Mechrgui and Theiri (2024) reports that higher ESG scores, particularly in the social category, are linked to lower volatility in stock prices.

Other studies show neutral or even negative effects. Gibson Brandon et al. (2021) examined situations where ESG rating agencies disagree. They found that such disagreement, especially about environmental factors, can lead to a risk premium and higher returns. This suggests that uncertainty itself may influence prices. Some research, including work by La Torre et al. (2020) and the article "Analysis of the Impact of ESG on Financial Performance," did not find a significant link between ESG scores and stock returns. These results suggest

that ESG may not be a primary driver of financial performance in all instances. Teja and Liu (2024) found a negative relationship between ESG risk scores and expected returns, indicating that higher ESG risks are associated with lower future returns. Morea et al. (2022) observed mixed results, as while companies with strong ESG profiles may benefit, specific strategies, such as adopting a circular economy model, do not always lead to improved stock performance. ESG information triggers investor reactions, and negative ESG information causes stock prices to decline, while positive ESG information has no impact on stock prices (Serafeim & Yoon, 2023). A foundational study found that "sin stocks" (e.g., alcohol and tobacco) have higher expected returns than comparable companies (Hong & Kacperczyk, 2009). This return premium is a direct result of social norms, which cause certain institutional investors to avoid these stocks, leading to them being underpriced.

In response to the challenges of quantifying ESG factors, researchers are investigating methods to classify and evaluate ESG information in non-standardized texts automatically. Ferjančič et al. (2024) extracted textual evidence for ESG scores, dividing the "S" in ESG into social capital and human capital. Their study revealed a strong correlation between ESG scores and textual evidence, which could be beneficial for the automated evaluation of ESG scores from textual resources.

Traditional machine learning techniques (support vector machines and XGBoost) are currently being compared to state-of-the-art language models (FinBERT-ESG) and fine-tuned LLMs (e.g., Llama 2) in the context of ESG text classification in current research (Chung & Latifi, 2024). In a non-peer-reviewed study (preprint), the authors demonstrated that fine-tuning LLMs using a novel technique, known as Quantized Low-Rank Adaptation (QLRA), yields substantial enhancements in all ESG domains. Research is also in the process of developing domain-specific fine-tuned models, including EnvLlama 2-Qlora, SocLlama 2-Qlora, and GovLlama 2-Qlora, which have demonstrated promising results in classifying ESG text.

From Univariate to Multivariate Forecasting

Building upon the established strength of LSTMs in capturing non-linear dependencies in financial time series, our study extends this approach to a multivariate setting to test whether their predictive power can be augmented by incorporating non-price, fundamental information in the form of ESG scores. Traditional time series models, such as the univariate LSTM used in prior research, are intended to predict future values using a single input variable, often the historical price or return series. In this setup, the model learns patterns and dependencies simply from the previous behavior of that one variable. While this strategy is well-suited for capturing trends inherent in a single data stream, it does not take into account external, potentially predictive information.

- **Process:** Unlike a univariate model that processes one data point at each time step (e.g., last month's return), our multivariate model processes a vector of features simultaneously. At each step, it takes into account not only the historical return but also the associated ESG scores (e.g., *esg_governance_score*, *esg_social_score*, etc.). The LSTM's internal gates then learn the (potential) complex, non-linear relationships between these different input variables over time.
- **Advantages:** The primary advantage of this multivariate approach is its ability to create a better-informed forecasting model. It allows us to test the central hypothesis of this paper: whether ESG factors contain predictive information that is not already captured in the historical return series. By combining multiple variables, the model can identify potential cross-correlations and lead-lag effects that a univariate model is structurally blind to, making it the ideal method for this research question.

While our study employs a direct approach, the field has also seen the development of more complex hybrid architectures, such as LSTM-augmented GARCH-MIDAS models, designed for sophisticated volatility forecasting with mixed-frequency data (Verma, 2021).

3. Materials and Methods

3.1. Data

The Yahoo Finance API provided the monthly closing prices. The last traded closing price of each ticker in scope was used to calculate the monthly returns.

Bloomberg[2] provided the following ESG scores: *esg_disclosure_score*, *esg_environmental_score*, *esg_governance_score*, *esg_score*, and *esg_social_score*. While the *esg_disclosure_score* is available for most companies in scope since 2007, the other scores gained traction only in later years. Therefore, the main analysis (RQ1) considered the period from January 2016 to December 2023. The additional analysis (RQ2) was conducted for the period from January 2008 to December 2023. Only for a few companies were consistent scores for 2024 onwards available; consequently, the following years were omitted. No data cleaning or filling of missing values was required, as only companies with complete available data for the selected period were considered (see Appendix A, Table A2). Out of the whole index composition, 56 tickers had ESG scores available for the observed duration for the primary analysis and 45 tickers for the additional analysis.

Scores are typically published annually at the end of the year. The scores of a year are paired with the monthly returns of the following year. Due to the low frequency of ESG information, this analysis was conducted on monthly returns (as opposed to daily returns).

The standard econometric approach to this problem is the Mixed Data Sampling (MIDAS) regression model, developed by Ghysels, Santa-Clara, and Valkanov (Ghysels et al., 2007). MIDAS models provide a direct and flexible way to incorporate data sampled at different frequencies into the same model, typically by using a weighting function to aggregate the higher-frequency data to match the lower-frequency variable. While sophisticated, the MIDAS approach is not directly applicable here, as our goal is to use a low-frequency variable to improve the prediction of a high-frequency one. Instead of developing a complex, hybrid MIDAS-LSTM model, we adopt a more direct and transparent approach: treating the annual ESG score as a constant feature vector for each of the subsequent 12 months.

We designed our experiment to simulate a realistic investment scenario. The ESG scores published for a given calendar year (Year Y) were used to forecast the monthly returns of the following year (Year Y + 1). For example, the complete set of ESG scores from 2015 was treated as a constant feature vector for all 12 monthly predictions from January to December 2016. This approach ensures that the model only uses information that would have been available to an investor at the start of the forecast period. Alternative methods, such as interpolating between annual scores, would introduce severe lookahead bias by using future information (e.g., the end-of-year score) to make predictions for the current year. This would artificially inflate the model's performance and invalidate the results. Our approach, while highlighting the static nature of the data, accurately reflects the information an investor would have had at the time of the forecast.

The main analysis (RQ1) used scores from 2015 to 2022, applied to prices from January 2016 to December 2023. Additional analysis (RQ2) used scores from 2007 to 2022, applied to prices from January 2008 to December 2023.

Table 1 (main analysis) and Table 2 (additional analysis) provide the descriptive statistics for the key variables used in our analysis. The monthly returns show the typical characteristics of financial time series, with a mean close to 0 and significant volatility. Bloomberg's single ESG scores (*esg_environmental_score*, *esg_social_score*, and *esg_governance_score*), which

are scaled from 0 to 10 (where 10 is best), show variation across the different pillars. Bloomberg's combined *esg_score* and *esg_disclosure_score* (which does not directly measure the company's ESG performance but rather focuses on the quality and comprehensiveness of the disclosures) are scaled from 0 to 100 (where 100 is best).

Table 1. Descriptive statistics (main analysis).

	Mean	Median	Std. Dev.	Skewness	Kurtosis	Min	Max
monthly_return	0.0069	0.0073	0.0875	0.0154	2.2330	−0.5025	0.5338
esg_score	3.7955	3.8700	1.2457	−0.0546	−0.2165	0.7700	7.0600
esg_environmental_score	3.3577	3.3800	2.1724	0.1912	−0.5480	0.0000	9.1000
esg_social_score	3.4126	3.0300	1.9890	0.8769	0.3297	0.1800	9.7500
esg_governance_score	5.5343	5.6800	1.1658	−0.3083	−0.6242	2.0400	8.1200
esg_disclosure_score	51.4454	52.7311	13.6460	−0.3311	−0.1590	14.5559	86.6311

Table 2. Descriptive statistics (additional analysis).

	Mean	Median	Std. Dev.	Skewness	Kurtosis	Min	Max
monthly_return	0.0086	0.0083	0.0932	0.3007	8.5424	−0.5187	1.3161
esg_disclosure_score	43.5219	44.6333	16.5364	−0.1769	−0.6415	5.7037	86.6311

3.2. Model Architecture

Hochreiter and Schmidhuber (1997) introduced the LSTM architecture, a type of RNN, which addresses the vanishing and exploding gradient problem, enabling it to capture long-term dependencies. This allows it to effectively learn long-range dependencies in sequential data, making it particularly well-suited for financial time series analysis.

The core of the LSTM architecture is the memory cell, which includes a cell state (C_t) and three primary "gates" that regulate the flow of information: the forget gate, the input gate, and the output gate (see Figure 1). These gates, controlled by sigmoid activation functions, determine which information is stored, updated, or discarded at each time step.

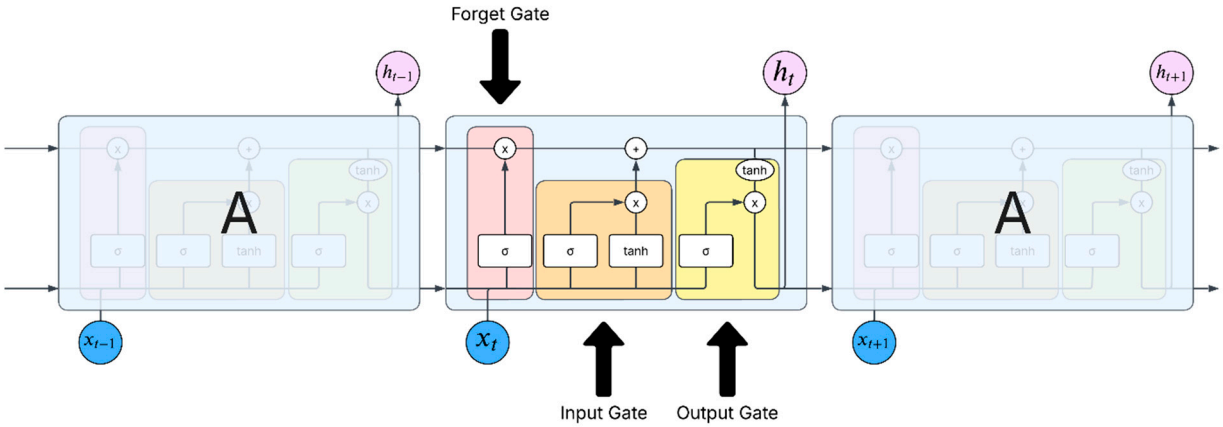


Figure 1. LSTM gates (own illustration based on concepts of Hochreiter & Schmidhuber, 1997).

The forward pass of an LSTM cell incorporating a forget gate can be expressed in the following compact mathematical form:

- **Forget gate** (f_t) : Decides what information is discarded from the memory/previous cell state (C_{t-1}) (using a sigmoid activation function σ).

$$f_t = \sigma_g\Big(W_f x_t + U_f h_{t-1} + b_f\Big) \tag{1}$$

- **Input gate** (i_t) : Decides which new information from the input is considered in the memory (using a sigmoid activation function σ and a tanh layer that creates a vector of new candidate values $\widetilde{C}_t$, which could be added to the state).

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{2}$$

$$\widetilde{C}_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{3}$$

The old cell state (C_{t-1}) is then updated to the new cell state (C_t) by multiplying the old state by the forget gate's output and adding the product of the input gate and the candidate values.

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t \tag{4}$$

- **Output gate** (o_t) : Determines the next hidden state (h_t), which is a filtered version of the cell state. First, a sigmoid layer decides which parts of the cell state will be output. Then, the cell state is passed through a *tanh* function (to push the values to be between -1 and 1) and multiplied by the output of the sigmoid gate.

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{5}$$

$$h_t = o_t \odot \sigma_h(C_t) \tag{6}$$

In these equations, W and b represent the weight matrices and bias vectors for each respective gate, σ is the sigmoid activation function, and $\odot$ denotes element-wise multiplication (Hadamard product).

In our models, the input vector at each time step x_t is expanded from a scalar to a multidimensional vector:

$$x_t = [return_t, esg_score_t, env_score_t, soc_score_t, gov_score_t, disc_score_t]$$

$$x_t = [return_t, disc_score_t]$$

The model is trained using backpropagation through time. The weights are optimized using the Adam optimizer, a widely used adaptive learning rate optimization algorithm. The objective is to minimize the mean squared error (MSE) between the model's predicted returns (y_i) and the actual returns (y_i), which serves as the loss function.

The selection of a 12-month lookback window for the LSTM model was a deliberate choice based on three key factors. First, it aligns with the annual frequency of the ESG data, allowing the model to process a full year of market returns that correspond to the period over which the ESG performance was measured. Second, a 12-month window is well-suited to capture potential annual seasonality in stock returns. Finally, this value was not chosen arbitrarily but was confirmed through a hyperparameter tuning process using a grid search methodology on a validation set, where it demonstrated a strong balance between capturing sufficient historical context and avoiding unnecessary noise from older, less relevant data.

The final architecture and hyperparameters for our LSTM models, detailed in Table 3, were selected through a systematic tuning process. We employed a grid search on a

validation set (a held-out portion of the training data) to evaluate various combinations of layers, units, and learning rates. The chosen configuration represents the architecture that provided the best out-of-sample performance on the validation set, offering a robust balance between model complexity and generalization capability.

Table 3. LSTM model architecture and hyperparameters. This table details the final architecture and key hyperparameters selected for the LSTM models used in this study.

Parameter	Value/Choice	Justification
LSTM Layers	2	Provides sufficient capacity to learn complex patterns without excessive risk of overfitting.
Units per Layer	50	A standard choice offering a good balance between representational power and computational cost.
Activation Function	ReLU	Widely used for its efficiency and effectiveness in preventing vanishing gradients in hidden layers.
Dropout Rate	0.2	A regularization technique to prevent overfitting by randomly dropping 20% of units during training.
Output Layer	Dense (1 unit)	A single output neuron to produce the continuous value forecast for the next month's return.
Optimizer	Adam	An efficient and widely adopted optimization algorithm that adapts the learning rate during training.
Loss Function	Mean Squared Error	Standard loss function for regression tasks, penalizing larger errors more heavily.
Batch Size	32	A common batch size that provides a stable gradient estimate during training.
Epochs	100	An initial upper limit for training, used in conjunction with early stopping to prevent overfitting.

3.3. Feature Importance

To ensure model explainability and to understand the drivers of our forecasts, we chose SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) to measure feature importance. SHAP values indicate how each feature influences the final forecast prediction, the significance of each feature compared to others, and the model's dependence on the interaction between features. While other methods like permutation importance exist, SHAP was selected for its significant theoretical advantages:

1. Grounded in game theory: We chose SHAP because it is not just another ad hoc method; it's built on the Nobel Memorial Prize-winning Shapley value concept from game theory. This provides a solid theoretical basis for fairly distributing a single prediction's outcome among the features that contributed to it.
2. Trustworthy and consistent results: A key advantage of SHAP is its guarantee of consistency. If a feature's actual contribution to the model increases, its measured importance will not decrease. This prevents illogical outcomes found in some other methods and makes the feature importance rankings more intuitive and reliable.
3. Versatility for different questions: SHAP is a versatile tool that can answer two different questions. It can provide a high-level, global view of which features matter most across the entire dataset, which is our focus. It can also zoom in to explain the specific drivers of a single, individual forecast, making it a powerful instrument for any diagnostic work.

3.4. Experimental Design

The following two main analyses were done:

- **Main Analysis (RQ1):**
  January 2016 to December 2023 comparison of a baseline model (using only historical returns) against an ESG model (historical returns and all five ESG scores).
- **Additional Analysis (RQ2):**
  January 2008 to December 2023, comparing a baseline model against an ESG model using only the esg_disclosure_score.
- **Feature Importance in Both Analyses (RQ3):**
  The SHAP values are calculated for the features used in each analysis of RQ1 and RQ2.

Each dataset is split into a 70/30 train–test set and a 10% validation set, which is used for model training. Each ticker was trained and forecasted individually, and both analyses were calculated separately.

3.5. Evaluation Metrics

The forecast errors are measured against commonly used scale-dependent methods: the "mean absolute error" (MAE), the "root mean squared error" (RMSE), and the "mean absolute percentage error" (MAPE). While MAE is defined as the arithmetic mean of the absolute differences between predicted values $\hat{y}_i$ and observed (actual) values $\boldsymbol{y_i}$ RMSE is defined as the square root of the average of the squared differences between predicted and observed values. MAPE provides a scale-independent error as a percentage of actual values.

Formulas:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100 \tag{9}$$

$$MDA = \frac{1}{n}\sum_{t} 1_{sgn(y_i - y_{i-1} = sgn(\hat{y}_i - y_{i-1}))} \tag{10}$$

where n is the number of observations, y_i is the actual value, and $\hat{y}_i$ is the predicted value.

MAE is robust to outliers due to its linear penalty structure, while RMSE is more sensitive to large deviations because of the quadratic penalty. As MAE might be preferable in the presence of outliers or heavy-tailed error distributions, RMSE is beneficial when large errors are undesirable. Therefore, both methods and the percentage value of MAPE are used. Mean directional accuracy (MDA) compares the forecast direction (upward or downward) to the actual realized direction. MDA counts the quantity of how often the forecasted direction is the actual direction.

A Diebold–Mariano test was used to compare the predictive accuracy of the baseline model with that of the corresponding ESG model.

4. Results

This section presents the empirical findings from the two analytical scenarios designed to evaluate the predictive value of ESG data. The results are structured to directly address the research questions concerning model performance and feature importance.

4.1. Main Analysis: 8-Year Period with All ESG Scores

The primary analysis was conducted on a set of 56 tickers spanning the period from January 2016 to December 2023, comparing a baseline LSTM model that utilized only historical returns against an ESG-enhanced model that incorporated all five ESG factors.

As shown in Table 4, a descriptive comparison of the error metrics suggests an advantage for the baseline model, which achieved a lower mean root mean squared error (RMSE) and mean absolute error (MAE). In terms of directional accuracy, both models performed almost identically, correctly predicting the direction of the following month's return in the median approximately 53% of cases.

Table 4. Descriptive statistics of model performance (main analysis).

Metric	Model	Mean	Median	Std. Dev.	Paired Statistical Tests Wilcoxon (*p*-Values)
RMSE	BASELINE	0.093267	0.08972	0.03307	0.0073
RMSE	ESG	0.097681	0.09788	0.03327	
MAE	BASELINE	0.075014	0.07026	0.02581	0.0008
MAE	ESG	0.079447	0.07949	0.02688	
MDA	BASELINE	52.91116	52.9412	11.4258	0.2921
MDA	ESG	51.86074	52.9412	10.8065	

The Diebold–Mariano Test (see Figure 2), which compares forecast accuracy on a ticker-by-ticker basis, showed no significant differences for the majority of cases. Only seven companies had a significantly better baseline model ("afx.de," "boss.de," "cbk.de," "g1a.de," "hot.de," "pah3.de," and "qia.de").

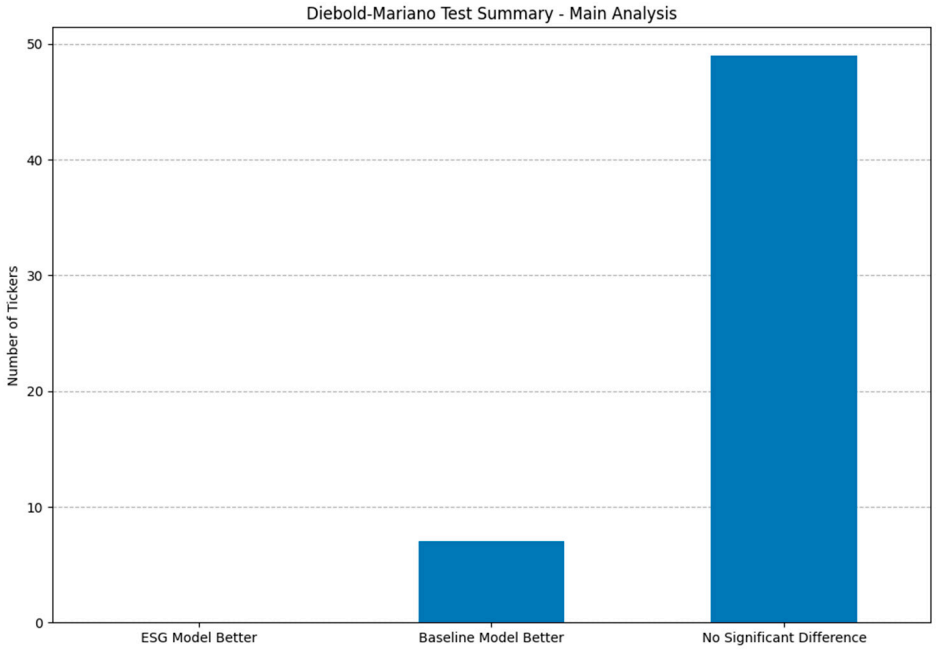


Figure 2. Diebold–Mariano test (main analysis).

4.2. Additional Analysis: 16-Year Period with All ESG Scores

To test the long-term robustness of a single ESG factor, a second analysis was performed on 45 tickers over the period from January 2008 to December 2023. This analysis compared the baseline model to a model enhanced only with the *esg_disclosure_score*. This

is because it was the only ESG parameter with consistent data available throughout the long time frame, back to 2007. The other, broader ESG scores only gained traction and became widely accessible in later years, making a long-term study with the full feature set impossible due to considerable missing data for the first half of the period.

The results in Table 5 show a near-identical performance, confirming that the inclusion of the disclosure score alone also fails to provide a statistically significant predictive advantage over the longer time horizon.

Table 5. Descriptive statistics of model performance (additional analysis).

Metric	Model	Mean	Median	Std. Dev.	Paired Statistical Tests Wilcoxon (*p*-Values)
RMSE	BASELINE	0.10144	0.09867	0.03332	0.9911
RMSE	ESG	0.10283	0.10057	0.03204	
MAE	BASELINE	0.07861	0.07409	0.02461	0.9022
MAE	ESG	0.07994	0.07812	0.02428	
MDA	BASELINE	48.9497	50.0000	6.37896	0.5969
MDA	ESG	49.9417	50.0000	6.31231	

The Diebold–Mariano (see Figure 3) test indicates that the ESG model was significantly better for only five tickers ("bas.de," "bei.de," "cbk.de," "g1a.de," and "wch.de"). In contrast, the baseline model was superior for four tickers ("dhl.de," "pum.de," "qia.de," and "tlx.de"). For the majority of companies (37), no statistically significant difference was found. This corroborates the finding that there is no systematic, positive contribution from the ESG data. The complete test results for each ticker, including the DM statistic and *p*-value, are provided in Appendix A Table A3.

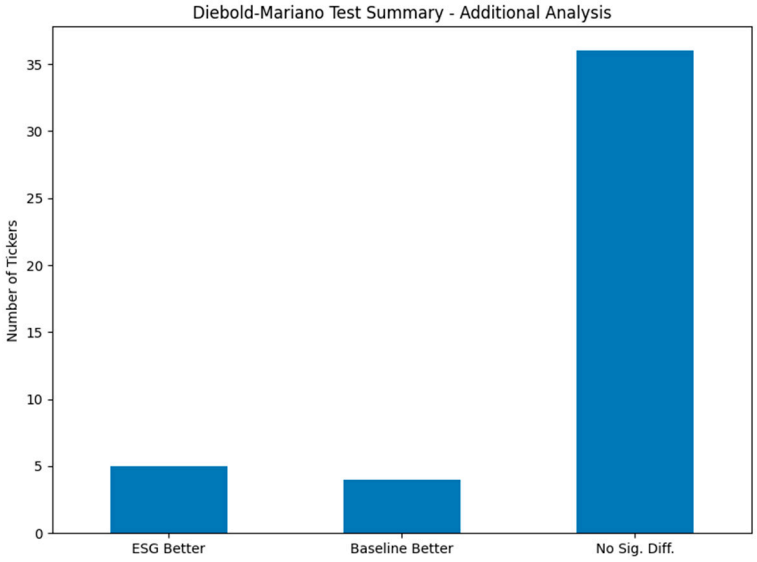


Figure 3. Diebold–Mariano test (additional analysis).

4.3. Feature Importance (SHAP Analysis)

To answer which features were most influential (RQ3), a SHAP analysis was performed on the ESG-enhanced models. To ensure robustness against outliers caused by numerical instability in the SHAP algorithm for certain tickers, the median of the absolute SHAP

values was used as the primary measure for aggregated feature importance. It is worth noting that the absolute SHAP value was used to determine the importance of each feature. It is essential to note that the importance itself does not indicate whether a feature improves forecast accuracy; instead, the question is: "Which feature has the biggest impact on the model's predictions, on average?" It is important to note that the absolute SHAP values reported are on a small scale; this is an expected result, as they explain the model's output, which is the monthly stock return—itself a small numerical value.

4.3.1. Feature Importance in the Main Analysis

The aggregated median SHAP values for the main analysis reveal a notable finding (see Figure 4). Contrary to the expectation that historical returns would dominate, several ESG factors exhibited a higher median importance. Specifically, *esg_governance_score* (Median SHAP: 0.000725) and *esg_social_score* (0.000722) emerged as the most influential features for the typical firm in the dataset. Their importance slightly surpassed that of esg_environmental_score (0.000640) and, critically, the monthly_return itself (0.000619). The overall esg_score and esg_disclosure_score were found to be the least influential factors in this comprehensive model (see Table 6).

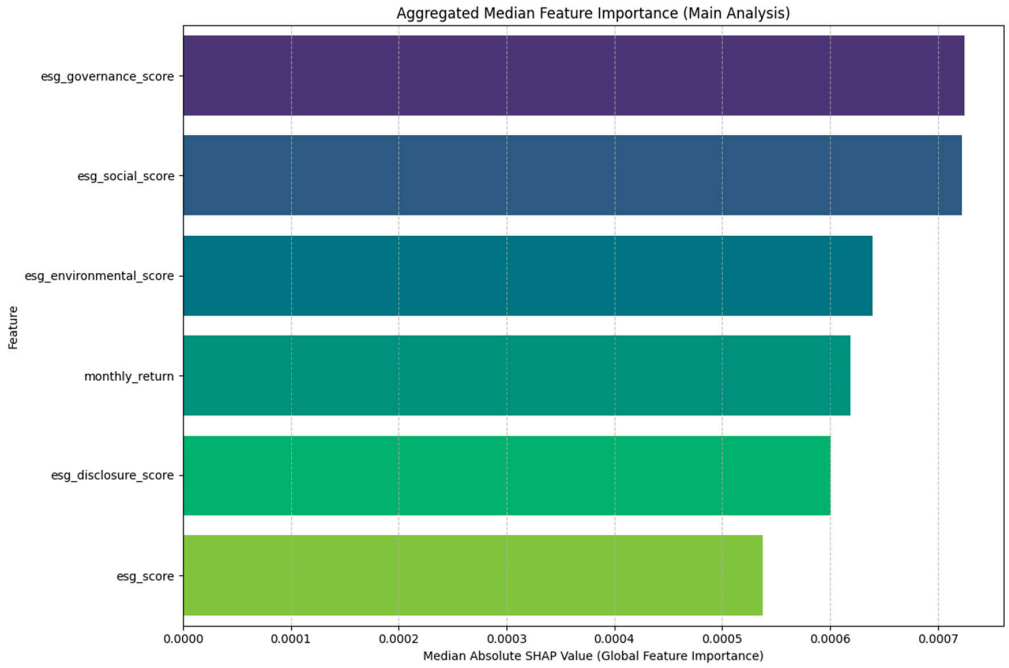


Figure 4. SHAP summary plot (global feature importance). The bar chart shows the median absolute SHAP value for each feature across all tickers, indicating its average impact on the model's output magnitude in the main analysis.

Table 6. Aggregated MEDIAN feature importance (main analysis).

Feature	Median SHAP Value
esg_governance_score	0.000725
esg_social_score	0.000722
esg_environmental_score	0.000640
monthly_return	0.000619
esg_disclosure_score	0.000601
esg_score	0.000537

4.3.2. Feature Importance in the Additional Analysis

In the long-term analysis (18 years), which focused solely on the *esg_disclosure_score*, the feature's importance was even more pronounced. The median absolute SHAP value for the *esg_disclosure_score* (0.001919) was substantially higher than that of the *monthly_return* (0.001082) (see Figure 5 and Table 7). This suggests that over a longer time horizon, the model attributes significant predictive relevance to a company's data transparency, even if this does not ultimately result in a statistically significant improvement in the final performance metrics.

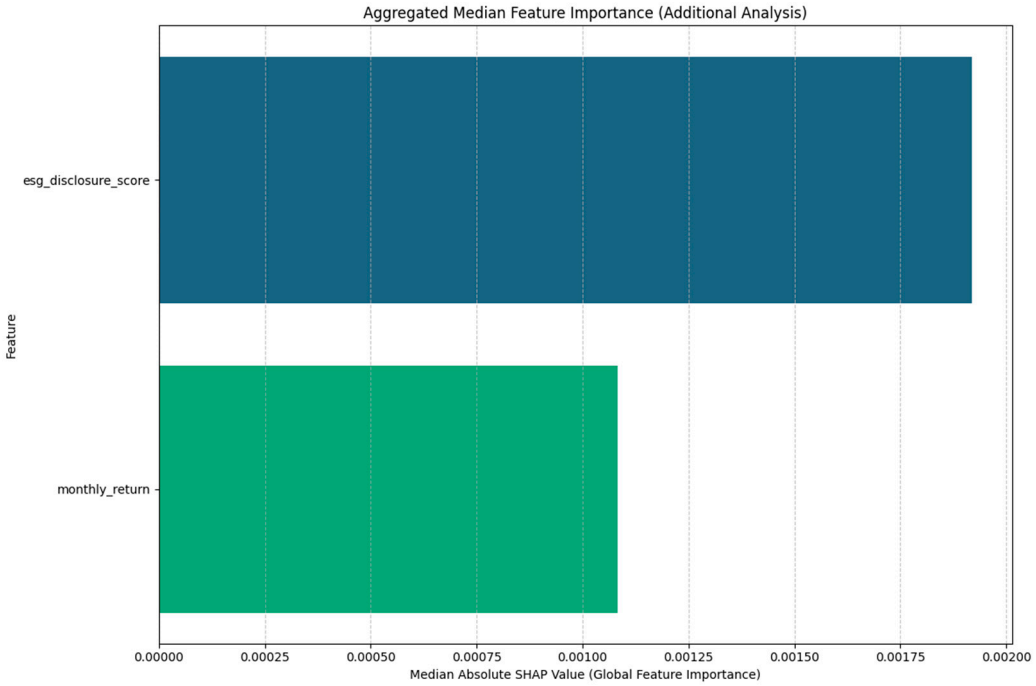


Figure 5. SHAP summary plot (global feature importance). The bar chart shows the median absolute SHAP value for each feature across all tickers, indicating its average impact on the model's output magnitude in the additional analysis.

Table 7. Aggregated MEDIAN feature importance (additional analysis).

Feature	Median SHAP Value
esg_disclosure_score	0.001919
monthly_return	0.001082

While other SHAP visualizations like dependence plots (showing marginal effects) or force plots (explaining single predictions) exist, they were deliberately omitted. Given that our study's core finding is a null result, presenting plots that detail individual predictions could be misleading by overemphasizing the importance of features that are not helpful in aggregate. Therefore, the SHAP summary plot of global feature importance is the most relevant tool for our analysis, as it directly answers which features the model relied on most, on average, across all predictions.

5. Discussion

The empirical results of this study provide a clear and statistically robust answer to the primary research question: for the German market, the inclusion of low-frequency (annual) ESG data into a higher frequency (monthly) LSTM-based forecasting model does not lead to a significant improvement in predictive accuracy.

For the main analysis period, the addition of a comprehensive set of ESG scores resulted in a statistically significant increase in forecast error, suggesting that the data acted as noise rather than a valuable signal. Over a longer time horizon, a single ESG disclosure score showed no evident impact, with the model's performance being statistically indistinguishable from the baseline.

Our findings indicate that forecast errors are substantial for both the baseline model and the ESG-augmented model. In the main analysis (Table 4), the mean root mean squared error (RMSE) of the baseline specification is around 0.093. This corresponds to an average deviation of about 9.3 percentage points in predicted monthly stock returns, which is a considerable error margin that makes reliable return forecasting highly challenging.

The relative difference between the two models is negligible from an economic standpoint. The baseline model has a mean absolute error (MAE) of 0.0750, compared to 0.0794 for the ESG-enhanced version. The resulting gap of 0.0044 (0.44 percentage points) is not statistically significant for the majority of cases and is too small to have practical relevance. Any trading strategy designed to exploit such a marginal difference would almost certainly fail to generate profits once transaction costs, bid-ask spreads, and market impact are taken into account.

These results suggest that our conclusion is sound in both statistical and economic terms. Annual ESG information does not offer a meaningful predictive gain that could be converted into a viable investment or forecasting strategy.

5.1. Context Within Existing Research

Our results align strongly with the strand of literature that has found a neutral or inconclusive link between ESG scores and stock returns. For instance, our findings are consistent with the work of La Torre et al., who did not find a significant connection between ESG indices and stock returns (La Torre et al., 2020). Furthermore, the observation that adding ESG data can be detrimental to model performance resonates with studies like Teja and Liu, who identified a negative relationship between ESG risk scores and expected returns (Teja & Liu, 2024). In contrast, our findings contradict research that suggests a definite positive influence, such as studies that show that high ESG scores lead to higher stock prices or better investment performance. A critical methodological distinction may account for the disparity. Much research looks at the connection between ESG performance and financial returns in real time or within a sample. However, our research analyzes the information's out-of-sample, forward-looking predictive effectiveness.

Reconciling Feature Importance with Predictive Performance: The "Loud But Unhelpful" Signal

A central and seemingly paradoxical finding of this study is that while the SHAP analysis identifies ESG scores as highly influential (Figure 4), their inclusion fails to improve. It even worsens the model's predictive accuracy. This highlights a critical distinction between a feature's influence and its utility.

- **SHAP Measures Influence, Not Correctness:** SHAP values quantify the magnitude of a feature's impact on a model's output. A high SHAP value for *esg_governance_score* means the model learned a relationship and is actively using that score to adjust its forecasts up or down. It tells us the model is listening to that feature.
- **Performance Metrics Measure Correctness:** Metrics like RMSE and MAE, however, measure whether those adjustments were correct.

The results show that our LSTM model did indeed learn patterns from the annual ESG data and gave them significant weight. However, the patterns learned from this

slow-moving, low-frequency data did not generalize effectively to predict the noisy, high-frequency movements of next-month stock returns.

5.2. The Challenge of a Low-Variability Signal

A key factor contributing to our findings is the inherent nature of the ESG data itself. By design, our experiment treated the annual ESG score as a constant for all 12 subsequent months to simulate a realistic investment scenario. A direct consequence of this is that the primary predictive variable (the ESG score) has zero variability within any given year.

While this lack of intra-year variance presents a significant challenge for any forecasting model, it is a true reflection of the low-frequency information available to practitioners. Our study was designed to test whether this static, annual signal contains enough information to predict the direction of the much more volatile monthly returns. The resulting insignificance is, therefore, an important finding in itself: it suggests that the market does not systematically react over the course of a year to the single, static piece of ESG information published for the prior year.

5.3. Limitations

This study is limited to the LSTM machine learning algorithm, which, in general, shows promising results compared to other time series forecasting algorithms. Additionally, lower frequency data, such as yearly ESG information, has limited predictive power, as factors influencing the ESG values in the first place will be reflected in the company's stock value according to the efficient market theory. Due to the low frequency of ESG reporting, there is a delay (lag) until an ESG score is determined for the company. From a forecasting perspective, there is no added value in predicting monthly returns using this modeling approach. Future work could investigate the text-based ESG sentiment of news and social media activity, as well as companies' ESG reports.

A further limitation relates to the data requirements of our chosen model. LSTM networks are known to be data-intensive and typically perform best on very long time series. Our sample sizes, particularly the 8-year period of the main analysis (96 monthly observations), are on the shorter side for such a complex model. While the 16-year period of our additional analysis provides a more substantial dataset (192 observations), even this may not be sufficient for the LSTM to uncover very subtle, long-term patterns. However, this limitation also reinforces our conclusion: if a meaningful predictive relationship between annual ESG scores and monthly returns existed, it should have become detectable over a 16-year period. The model's failure to do so suggests that the signal is not merely subtle, but practically nonexistent.

5.4. Future Work

While our findings proved robust across different time periods, future work could further validate these results by employing alternative models (e.g., GRU or tree-based methods) to confirm that the conclusions are not algorithm-specific.

Due to limited data provider availability to the authors, annual data provided by Bloomberg was used. ESG data is typically reported annually, with reporting timelines varying by company. Some data providers might incorporate other non-standardized ESG data with a higher frequency. Future studies could investigate ESG data, provided by different data providers, e.g., on a monthly basis.

Also, to more effectively capture real-time market reactions, future research may investigate ESG sentiment analysis and the combination of official reports and other relevant information, which is published on a higher frequency (e.g., news releases, blog posts, or social media posts).

6. Conclusions

This study was designed to test a practical and widespread hypothesis: can the standard, annual ESG scores that are widely available to investors be used in a sophisticated machine learning model to forecast monthly stock returns? While the efficient market hypothesis suggests this is unlikely, the pervasive narrative in the practitioner community made this an essential question to test empirically.

Our comprehensive, out-of-sample analysis provides a clear and resounding no. Across two different time horizons and feature sets, the inclusion of low-frequency annual ESG data failed to provide any statistically or economically significant improvement in forecast accuracy. In fact, the addition of a full suite of ESG scores was detrimental to the model's performance, demonstrating that this backward-looking data acted as noise rather than a valuable signal.

The primary contribution of this paper is this robust null result. We did not aim to invent a new hybrid model, but to apply a powerful, standard tool—the multivariate LSTM—to give the practitioner hypothesis its best possible chance of success. The failure of this hypothesis under such conditions is a significant finding, providing strong, empirical evidence that the informational content of annual ESG reports is already fully incorporated into market prices.

Our findings, derived from a time series forecasting perspective, align remarkably well with recent, methodologically different research. In the portfolio construction domain, Bruno et al. (2025) similarly find that while ESG information appears valuable in-sample, these benefits vanish entirely in a more realistic out-of-sample setting. Where their work demonstrates the failure of ESG data to improve out-of-sample Sharpe Ratios, our study shows its failure to improve out-of-sample forecast accuracy. This is further corroborated by studies of investor behavior, such as Hartzmark and Sussman (2019), who show that even salient ESG information does not trigger reactions from institutional investors in a way that would suggest clear alpha generation.

The practical implications for quantitative investors and asset managers are direct. Our work serves as a critical, data-driven word of caution: the search for a predictive edge, or "alpha," in standard, delayed ESG disclosures is likely a fruitless endeavor. The true value of ESG data may lie in risk management and aligning investor preferences, not in outperformance. The future of predictive ESG analysis, as our findings implicitly suggest, must move beyond these standard disclosures and toward the high-frequency, alternative data sources that can provide timely, forward-looking signals.

Finally, while our study is confined to the German market, the findings may have broader relevance. The DAX 50 ESG Index represents a highly liquid segment of a major developed economy, and the Bloomberg ESG factors used are comprehensive and internationally recognized. Therefore, it is plausible that similar challenges in leveraging annual ESG data for monthly forecasting exist in other developed markets. However, further research is needed to confirm these findings in different geographical contexts and with different modeling approaches.

Author Contributions: Conceptualization, M.R.; methodology, M.R.; software, M.R.; validation, M.R.; formal analysis, M.R.; investigation, M.R.; resources, M.R.; data curation, M.R.; writing—original draft preparation, M.R.; writing—review and editing, J.L.; visualization, M.R.; supervision, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript/study, the author(s) utilized GenAI (Google Gemini 2.5) for Python code development for data gathering (Yahoo Finance API), data processing, and running the experiments. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
ARIMA	Autoregressive Integrated Moving Average
CSRD	Corporate Sustainability Reporting Directive
CSR-RUG	CSR-Richtlinie-Umsetzungsgesetz (German national law implementing the EU Non-Financial Reporting Directive)
DAX	Deutscher Aktienindex (German Stock Index)
ESG	Environmental, Social, and Governance
ESRS	European Sustainability Reporting Standards
EU	European Union
FY	Fiscal Year
GRI	Global Reporting Initiative
LkSG	Lieferkettensorgfaltspflichtengesetz
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MDA	Mean Directional Accuracy
MIDAS	Mixed Data Sampling
NFRD	Non-Financial Reporting Directive
QLRA	Quantized Low-Rank Adaptation
Qlora	Quantized Low-Rank Adapter (context: Llama 2-Qlora models)
RNN	Recurrent Neural Network
RMSE	Root Mean Squared Error
RQ	Research Question
SASB	Sustainability Accounting Standards Board
SDGs	Sustainable Development Goals
SFDR	Sustainable Finance Disclosure Regulation
SHAP	SHapley Additive exPlanations
std dev	Standard Deviation
UNPRI	United Nations Principles for Responsible Investment
UNSDGs	United Nations Sustainable Development Goals

Appendix A

Table A1. ESG regulations applicable for DAX ESG 50 companies.

Regulation/Law	Legal Level	Applies To	Effective Since	Applicability to DAX ESG 50 Companies?
Non-Financial Reporting Directive (NFRD)	EU Directive	Large public-interest entities with >500 employees	1 January 2017	Yes; implemented in Germany via the CSR-Richtlinie-Umsetzungsgesetz

Table A1. *Cont.*

Regulation/Law	Legal Level	Applies To	Effective Since	Applicability to DAX ESG 50 Companies?
Corporate Sustainability Reporting Directive (CSRD)	EU Directive	Expands scope to large companies and listed SMEs	5 January 2023 (reporting from FY 2024)	Yes; mandatory reporting from FY 2024 onwards
European Sustainability Reporting Standards (ESRS)	EU Standards	Companies subject to CSRD	Adopted 31 July 2023	Yes; binding with CSRD compliance from FY 2024
CSR-Richtlinie-Umsetzungsgesetz (CSR-RUG)	German National Law	German transposition of NFRD	11 April 2017	Yes; requires non-financial statements
Lieferkettensorgfaltspflichtengesetz (LkSG)	German National Law	Companies with >3000 employees (from 2023); >1000 employees (from 2024)	1 January 2023	Yes; many DAX ESG 50 companies meet the employee threshold
EU Taxonomy Regulation	EU Regulation	Financial market participants and large companies	1 July 2021	Yes; alignment disclosures mandatory
Sustainable Finance Disclosure Regulation (SFDR)	EU Regulation	Financial market participants	10 March 2021	Yes; applicable to DAX ESG 50 companies in the financial sector
Global Reporting Initiative (GRI) Standards	International Standard	Voluntary, but widely adopted	First released in 2000; GRI Standards since 2016	Yes; many DAX ESG 50 companies use GRI for sustainability reporting

Table A2. Companies in the scope of the analysis.

Company Name	Bloomberg Ticker	Yahoo Ticker	Main Analysis	Additional Analysis
Adidas	ADS GY Equity	ads.de		X
Allianz	ALV GY Equity	alv.de		X
Aroundtown	AT1 GY Equity	at1.de		X
BASF	BAS GY Equity	bas.de	X	X
BMW	BMW GY Equity	bmw.de	X	X
Bayer	BAYN GY Equity	bayn.de	X	X
Bechtle AG	BC8 GY Equity	bc8.de	X	
Beiersdorf	BEI GY Equity	bei.de	X	X
Brenntag	BNR GY Equity	bnr.de		X
Carl Zeiss Meditec	AFX GY Equity	afx.de	X	
Commerzbank	CBK GY Equity	cbk.de	X	X
Continental	CON GY Equity	con.de	X	X
Covestro	1COV GY Equity	1cov.de	X	
Daimler (jetzt Mercedes-Benz)	MBG GY Equity	mbg.de	X	X
Deutsche Bank	DBK GY Equity	dbk.de	X	X
Deutsche Börse	DB1 GY Equity	db1.de	X	X
Deutsche Post (now DHL Group)	DHL GY Equity	dhl.de	X	X

Table A2. *Cont.*

Company Name	Bloomberg Ticker	Yahoo Ticker	Main Analysis	Additional Analysis
Deutsche Telekom	DTE GY Equity	dte.de	X	X
Deutsche Wohnen	DWNI GY Equity	dwni.de	X	
E.ON SE	EOAN GY Equity	eoan.de	X	X
Evonik	EVK GY Equity	evk.de	X	X
Fraport	FRA GY Equity	fra.de	X	X
Freenet AG	FNTN GY Equity	fntn.de	X	
Fresenius	FRE GY Equity	fre.de	X	X
Fresenius Medical Care AG	FME GY Equity	fme.de	X	
Fuchs Petrolub SE	FPE3 GY Equity	fpe3.de	X	
GEA Group	G1A GY Equity	g1a.de	X	X
Hannover Rück	HNR1 GY Equity	hnr1.de	X	X
Heidelberg Materials (previously HeidelbergCement)	HEI GY Equity	hei.de	X	X
Henkel	HEN3 GY Equity	hen3.de	X	X
Hochtief	HOT GY Equity	hot.de	X	X
Hugo Boss	BOSS GY Equity	boss.de	X	
Infineon Technologies	IFX GY Equity	ifx.de	X	X
K+S	SDF GY Equity	sdf.de	X	X
KION Group	KGX GY Equity	kgx.de	X	X
Knorr-Bremse	KBX GY Equity	kbx.de	X	X
LEG Immobilien	LEG GY Equity	leg.de	X	X
Lanxess	LXS GY Equity	lxs.de	X	
Lufthansa	LHA GY Equity	lha.de	X	X
Merck	MRK GY Equity	mrk.de	X	X
Münchener Rück	MUV2 GY Equity	muv2.de	X	X
Porsche Automobil Holding SE	PAH3 GY Equity	pah3.de	X	X
ProSiebenSat.1 Media	PSM GY Equity	psm.de	X	
Puma	PUM GY Equity	pum.de	X	X
Qiagen N.V.	QIA GY Equity	qia.de	X	X
Rational AG	RAA GY Equity	raa.de	X	
SAP	SAP GY Equity	sap.de	X	X
Sartorius	SRT3 GY Equity	srt3.de	X	X
Sartorius AG	SRT GY Equity	srt.de	X	X
Scout24	G24 GY Equity	g24.de	X	
Siemens	SIE GY Equity	sie.de	X	X
Symrise	SY1 GY Equity	sy1.de	X	
TAG Immobilien AG	TEG GY Equity	teg.de	X	
Talanx AG	TLX GY Equity	tlx.de	X	X
Thyssenkrupp	TKA GY Equity	tka.de	X	X

Table A2. *Cont.*

Company Name	Bloomberg Ticker	Yahoo Ticker	Main Analysis	Additional Analysis
United Internet AG	UTDI GY Equity	utdi.de	X	X
Volkswagen AG	VOW3 GY Equity	vow3.de	X	X
Vonovia SE	VNA GY Equity	vna.de	X	X
Wacker Chemie AG	WCH GY Equity	wch.de	X	X
Zalando	ZAL GY Equity	zal.de	X	

Table A3. Detailed Diebold–Mariano Test Results. The table shows the Diebold–Mariano (DM) test statistic and *p*-value for each ticker. A negative DM statistic indicates that the baseline model had a lower forecast error, while a positive statistic indicates the ESG-enhanced model performed better. *p*-values below 0.05 are marked with an asterisk (*) to denote statistical significance.

Ticker	Main Analysis DM-Stat	Main Analysis *p*-Value	Additional Analysis DM-Stat	Additional Analysis *p*-Value
1cov.de	0.231	0.8199	N/A	N/A
ads.de	N/A	N/A	−0.908	0.3688
afx.de	−2.675	0.0166 *	N/A	N/A
alv.de	N/A	N/A	1.539	0.1308
at1.de	N/A	N/A	−0.060	0.9532
bas.de	−0.718	0.4830	2.285	0.0271 *
bayn.de	0.889	0.3871	0.589	0.5589
bc8.de	−0.236	0.8164	N/A	N/A
bei.de	−1.130	0.2750	2.129	0.0388 *
bmw.de	−1.047	0.3107	−1.706	0.0949
bnr.de	N/A	N/A	1.493	0.1439
boss.de	−2.886	0.0107 *	N/A	N/A
cbk.de	−2.594	0.0196 *	2.288	0.0269 *
con.de	1.704	0.1077	−1.105	0.2750
db1.de	0.228	0.8229	−0.629	0.5325
dbk.de	0.486	0.6335	−0.225	0.8230
dhl.de	−1.265	0.2240	−2.256	0.0290 *
dte.de	−1.931	0.0715	−1.523	0.1347
dwni.de	−1.468	0.1616	N/A	N/A
eoan.de	1.458	0.1641	1.082	0.2849
evk.de	0.802	0.4340	0.339	0.7376
fme.de	0.824	0.4223	N/A	N/A
fntn.de	1.305	0.2104	N/A	N/A
fpe3.de	0.400	0.6948	N/A	N/A
fra.de	−0.896	0.3838	0.946	0.3491
fre.de	0.701	0.4937	0.124	0.9020
g1a.de	−2.571	0.0205 *	3.200	0.0025 *

Table A3. *Cont.*

Ticker	Main Analysis DM-Stat	Main Analysis *p*-Value	Additional Analysis DM-Stat	Additional Analysis *p*-Value
g24.de	0.297	0.7700	N/A	N/A
hei.de	0.666	0.5148	1.671	0.1018
hen3.de	−0.785	0.4437	−1.791	0.0800
hnr1.de	−0.407	0.6891	−1.443	0.1560
hot.de	−3.733	0.0018 *	−0.856	0.3966
ifx.de	0.547	0.5918	−1.308	0.1975
kbx.de	−0.462	0.6605	−1.257	0.2556
kgx.de	0.088	0.9309	1.417	0.1687
leg.de	−0.121	0.9050	0.439	0.6644
lha.de	0.651	0.5245	−0.642	0.5242
lxs.de	−1.170	0.2590	N/A	N/A
mbg.de	−0.612	0.5493	0.482	0.6324
mrk.de	−0.654	0.5226	0.565	0.5746
muv2.de	0.960	0.3515	0.578	0.5664
pah3.de	−4.413	0.0004 *	0.604	0.5491
psm.de	0.424	0.6775	N/A	N/A
pum.de	−1.804	0.0901	−2.092	0.0421 *
qia.de	−3.778	0.0016 *	−5.376	0.0000 *
raa.de	0.378	0.7105	N/A	N/A
sap.de	−0.722	0.4806	−0.046	0.9639
sdf.de	1.613	0.1263	−1.815	0.0763
sie.de	0.641	0.5305	−0.871	0.3884
srt.de	−1.970	0.0664	0.552	0.5838
srt3.de	−1.022	0.3218	−0.130	0.8972
sy1.de	−1.724	0.1039	N/A	N/A
teg.de	−0.177	0.8615	N/A	N/A
tka.de	1.337	0.1998	−1.302	0.1997
tlx.de	−1.053	0.3079	−4.126	0.0003 *
utdi.de	−0.620	0.5437	0.477	0.6356
vna.de	0.048	0.9620	0.848	0.4046
vow3.de	−0.431	0.6725	1.530	0.1329
wch.de	−0.461	0.6510	2.468	0.0175 *
zal.de	−0.373	0.7139	N/A	N/A

Notes

[1] https://stoxx.com/index/DAXESGK/ (accessed on 1 May 2025).
[2] https://www.bloomberg.com/professional/solutions/sustainable-finance/#overview (accessed on 1 May 2025).

References

1-DAX 50 ESG (DAXESGK) (DE000A0S3E04). (n.d.). *DAX® 50 ESG a new benchmark for German ESG investing*. Available online: https://stoxx.com/index/daxesgk/ (accessed on 30 April 2025).

ArunKumar, K. E., Kalaga, D. V., Kumar, C. M. S., Kawaji, M., & Brenza, T. M. (2022). Comparative analysis of gated recurrent units (GRU), long short-term memory (LSTM) cells, autoregressive integrated moving average (ARIMA), seasonal autoregressive integrated moving average (SARIMA) for forecasting COVID-19 trends. *Alexandria Engineering Journal*, *61*(10), 7585–7603. [CrossRef]

Bruno, G., Goltz, F., & Naly, A. (2025). Does ESG information deliver investment value? A high-dimensional portfolio perspective. *SSRN*. [CrossRef]

Chung, T. Y., & Latifi, M. (2024). Evaluating the performance of state-of-the-art esg Domain-specific pre-trained large language models in text classification against existing models and traditional machine learning techniques. *arXiv*, arXiv:2410.00207.

Dahbi, F., Carrasco, I., & Petracci, B. (2024). A systematic literature review on social impact bonds. *Finance Research Letters*, *62*, 105063. [CrossRef]

Elsaraiti, M., & Merabet, A. (2021). A comparative analysis of the ARIMA and LSTM predictive models and their effectiveness for predicting wind speed. *Energies*, *14*(20), 6782. [CrossRef]

European Commission, Directorate-General for Financial Stability, Financial Services and Capital Markets Union. (2025). *Commission simplifies rules on sustainability and EU investments, delivering over €6 billion in administrative relief*. European Commission, Directorate-General for Financial Stability, Financial Services and Capital Markets Union.

Feng, J., Goodell, J. W., & Shen, D. (2022). ESG rating and stock price crash risk: Evidence from China. *Finance Research Letters*, *46*, 102476. [CrossRef]

Ferjančič, U., Ichev, R., Lončarski, I., Montariol, S., Pelicon, A., Pollak, S., Šuštar, K. S., Toman, A., Valentinčič, A., & Žnidaršič, M. (2024). Textual analysis of corporate sustainability reporting and corporate ESG scores. *International Review of Financial Analysis*, *96*, 103669. [CrossRef]

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654–669. [CrossRef]

Gao, Z., Chen, J., Wang, G., Ren, S., Fang, L., Yinglan, A., & Wang, Q. (2023). A novel multivariate time series prediction of crucial water quality parameters with long short-term memory (LSTM) networks. *Journal of Contaminant Hydrology*, *259*, 104262. [CrossRef] [PubMed]

Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, *26*(1), 53–90. [CrossRef]

Gibson Brandon, R., Krueger, P., & Schmidt, P. S. (2021). ESG rating disagreement and stock returns. *Financial Analysts Journal*, *77*(4), 104–127. [CrossRef]

Gu, S., Kelly, B. T., & Xiu, D. (2019). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273. [CrossRef]

Hartzmark, S. M., & Sussman, A. B. (2019). Do investors value sustainability? A natural experiment examining ranking and fund flows. *The Journal of Finance*, *74*(6), 2789–2837. [CrossRef]

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780. [CrossRef] [PubMed]

Hong, H., & Kacperczyk, M. (2009). The price of sin: The effects of social norms on markets. *Journal of Financial Economics*, *93*(1), 15–36. [CrossRef]

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

Kobiela, D., Krefta, D., Król, W., & Weichbroth, P. (2022). ARIMA vs LSTM on NASDAQ stock exchange data. *Procedia Computer Science*, *207*, 3836–3845. [CrossRef]

Kulal, A., Abhishek, N., Dinesh, S., & Divyashree, M. S. (2023). Impact of environmental, social, and governance (ESG) factors on stock prices and investment performance. *Macro Management & Public Policies*, *5*(2), 14–26. [CrossRef]

Latif, N., Selvam, J. D., Kapse, M., Sharma, V., & Mahajan, V. (2023). Comparative performance of LSTM and ARIMA for the short-term prediction of bitcoin prices. *Australasian Accounting, Business and Finance Journal*, *17*(1), 256–276. [CrossRef]

La Torre, M., Mango, F., Cafaro, A., & Leo, S. (2020). Does the ESG index affect stock return? Evidence from the Eurostoxx50. *Sustainability*, *12*(16), 6387. [CrossRef]

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17* (pp. 4768–4777). Curran Associates Inc.

Luo, W., Tian, Z., Fang, X., & Deng, M. (2024). Can good ESG performance reduce stock price crash risk? Evidence from Chinese listed companies. *Corporate Social Responsibility and Environmental Management*, *31*(3), 1469–1492. [CrossRef]

Maltais, A., & Nykvist, B. (n.d.). Understanding the role of green bonds in advancing sustainability. *Journal of Sustainable Finance & Investment*, 1–20. [CrossRef]

Mechrgui, S., & Theiri, S. (2024). The effect of environmental, social, and governance (ESG) performance on the volatility of stock price returns: The moderating role of tax payment. *Journal of Financial Reporting and Accounting*. *ahead-of-print*. [CrossRef]

Morea, D., Mango, F., Cardi, M., Paccione, C., & Bittucci, L. (2022). Circular economy impact analysis on stock performances: An empirical comparison with the Euro Stoxx 50® ESG index. *Sustainability*, *14*(2), 843. [CrossRef]

Ning, Y., Kazemi, H., & Tahmasebi, P. (2022). A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and Prophet. *Computers & Geosciences*, *164*, 105126. [CrossRef]

Rosinus, M. (2025). Comparison of classical ARIMA forecasting methods to the machine learning LSTM method: A case study on DAX 50 ESG index. *ACTA VŠFS*, *19*(1). [CrossRef]

Serafeim, G., & Yoon, A. (2023). Stock price reactions to ESG news: The role of ESG ratings and disagreement. *Review of Accounting Studies*, *28*(3), 1500–1530. [CrossRef]

Siami-Namini, S., & Namin, A. S. (2018). Forecasting economics and financial time series: ARIMA vs. LSTM. *arXiv*, arXiv:1803.06386. [CrossRef]

Teja, K. R., & Liu, C.-M. (2024). ESG investing: A statistically valid approach to data-driven decision making and the impact of ESG factors on stock returns and risk. *IEEE Access*, *12*, 69434–69444. [CrossRef]

Verma, S. (2021). Forecasting volatility of crude oil futures using a GARCH–RNN hybrid approach. *Intelligent Systems in Accounting, Finance and Management*, *28*(2), 130–142. [CrossRef]

Zhang, R., Song, H., Chen, Q., Wang, Y., Wang, S., & Li, Y. (2022). Comparison of ARIMA and LSTM for prediction of hemorrhagic fever at different time scales in China. *PLoS ONE*, *17*(1), e0262009. [CrossRef] [PubMed]

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