

Article

Information-Theoretic ESG Index Direction Forecasting: A Complexity-Aware Framework

Kadriye Nurdanay Öztürk ^{1,2}  and Öyküm Esra Yiğit ^{3,*} 

¹ PhD Program in Statistics, Department of Statistics, Graduate School of Science and Engineering, Yildiz Technical University, 34000 Istanbul, Türkiye; kadriye.ozturk@bilecik.edu.tr

² Department of Statistics and Computer Sciences, Faculty of Science, Bilecik Seyh Edebali University, 11100 Bilecik, Türkiye

³ Department of Statistics, Faculty of Arts and Sciences, Yildiz Technical University, 34220 Istanbul, Türkiye

* Correspondence: oeyigit@yildiz.edu.tr

Abstract

Sustainable finance exhibits non-linear dynamics, regime shifts, and distributional drift that challenge conventional forecasting, particularly in volatile emerging markets. Conventional models, which often overlook this structural complexity, can struggle to produce stable or reliable probabilistic forecasts. To address this challenge, this study introduces a complexity-aware forecasting framework that operationalizes information-theoretic meta features, Shannon entropy (SE), permutation entropy (PE) and Kullback–Leibler (KL) divergence to make Environmental, Social, and Governance (ESG) index forecasting more stable, probabilistically accurate, and operationally reliable. Applied in an emerging-market setting using Türkiye’s ESG index as a natural stress test, the framework was benchmarked against a macro-technical baseline with a calibrated XGBoost classifier under a strictly chronological, leakage-controlled nested cross-validation protocol and evaluated on a strictly held-out test set. In development, the framework achieved statistically significant improvements in both stability and calibration, reducing fold-level dispersion (by 40.4–66.6%) across all metrics and enhancing probability-level alignment with Brier score reduced by 0.0140 and the ECE by 0.0287. Furthermore, a meta-analytic McNemar’s test confirmed a significant reduction in misclassifications across the development folds. On the strictly held-out test set, the framework’s superiority was confirmed by a statistically significant reduction in classification errors (exact McNemar $p < 0.001$), alongside strong gains in imbalance-robust metrics such as BAcc (0.618, +12.8%) and the MCC (0.288, +38.5%), achieving an F1-score of 0.719. Overall, the findings of the complexity-aware framework indicate that explicitly representing the market’s informational state and transitions yields more stable, well-calibrated, and operationally reliable forecasts in regime-shifting financial environments, supporting enhanced robustness and practical deployability.

Keywords: ESG forecasting; sustainable finance; information theory; probability calibration; forecast stability



Academic Editor: Boris Ryabko

Received: 1 October 2025

Revised: 12 November 2025

Accepted: 15 November 2025

Published: 17 November 2025

Citation: Öztürk, K.N.; Yiğit, Ö.E. Information-Theoretic ESG Index Direction Forecasting: A Complexity-Aware Framework. *Entropy* **2025**, *27*, 1164. <https://doi.org/10.3390/e27111164>

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

Financial markets are complex adaptive systems shaped by information flows and characterized by nonlinear dynamics such as abrupt regime shifts and volatility clustering [1–4]. The global ascendancy of sustainable finance, with sustainable investment assets under management reaching USD 30.3 trillion as of early 2022 [5], introduces an additional

layer of complexity, with Environmental, Social, and Governance (ESG) criteria increasingly influencing capital allocation decisions [6]. Yet capital flows into ESG assets remain heterogeneous across regions and over time [7], a pattern that is especially pronounced in emerging markets where data limitations, policy uncertainty, and liquidity frictions coexist. Evidence from MSCI [8] suggests that firms managing social and governance risks more effectively tend to exhibit superior financial performance, reframing sustainable finance as a problem of uncertainty management under regime variability.

Against this backdrop, prior research has evolved along three broad lines: portfolio construction and optimization integrating ESG criteria [9,10], machine-learning applications for firm- and market-level prediction [11,12], and volatility/risk forecasting with advanced time-series models [13–15]. In practice, ESG index forecasting typically follows two established approaches [16]: fundamental analysis, which exploits macroeconomic and commodity-linked variables to capture structural drivers [17–19], and technical analysis, which leverages historical price patterns, momentum, and volatility as proxies for short-horizon dynamics [13,20]. Hybrid frameworks combine these inputs [15], yet most remain limited to feature aggregation and treat uncertainty and regime dependence implicitly, lacking mechanisms to represent distributional change or dynamical structure.

Information theory provides a principled framework for modeling financial time series as complex adaptive systems. It encompasses families of entropy, divergence, and dynamical complexity descriptors that quantify, respectively, the unpredictability of returns, shifts in the data-generating distribution, and the temporal organization of price dynamics. Implemented in a rolling, time-respecting manner, these descriptors explicitly capture regime dependence and distributional shifts in the return-generating process, complementing conventional fundamental and technical indicators [21,22].

This information-theoretic perspective is particularly well suited to the practical challenges of sustainable finance. Emerging ESG markets are prone to nonlinear dynamics triggered by policy uncertainty, regulatory interventions, and greenwashing controversies, all of which can induce abrupt regime shifts. Conventional models often struggle to identify such transitions. Information-theoretic measures such as Shannon Entropy (SE), Permutation Entropy (PE), and Kullback–Leibler (KL) divergence are explicitly designed to quantify distributional uncertainty, temporal complexity, and structural shifts. Collectively, they provide a principled quantitative framework for capturing informational instability and policy-induced transitions that characterize sustainability-linked financial systems.

Building on this rationale, the need for such explicit representations is heightened in the burgeoning field of sustainable finance: ESG indices exhibit methodology drift and high constituent turnover, while ESG capital flows are regime-sensitive and nonstationary. The challenge is particularly acute in emerging markets, where volatility couples with limited data depth and policy uncertainty. Although the literature on ESG index forecasting is limited, the explicit operationalization of information-theoretic descriptors within predictive pipelines remains scarce, underscoring a clear research gap in applying these tools to one of the most complex and increasingly important domains in modern finance.

This study introduces a complexity-aware forecasting framework by systematically augmenting a hybrid machine-learning pipeline with information-theoretic measures. The framework represents ESG index dynamics along two complementary dimensions: (i) distributional properties, where SE proxies for uncertainty and KL Divergence detects structural change; and (ii) dynamical structure, where PE captures the temporal and ordinal patterns in price dynamics. By integrating these signals alongside conventional macro-technical inputs, the design moves beyond simple feature aggregation to provide an explicit diagnostic of the market's informational state and regime dependence. The central hypothesis is that such augmentation improves not only predictive accuracy but,

more importantly, the stability and calibration of forecasts across heterogeneous regimes. Its practical utility and robustness were further demonstrated by a stress test in Türkiye's highly volatile market, underscoring effectiveness in a challenging real-world setting and broader relevance to other emerging economies.

To rigorously evaluate the incremental value of information-theoretic features, this study implements a multi-phase, leakage-aware, and time-consistent framework. Phase 1 (Model and Data Preparation) defines the next-day direction target, constructs the competing baseline and entropy-augmented feature sets, and specifies a bespoke time-respecting calibrated XGBoost classifier. Phase 2 (Model Validation and Comparison) employs a forward-chaining nested cross-validation to jointly tune XGBoost hyperparameters and the probability-calibration component (Platt or isotonic); and to produce unbiased out-of-sample predictions for statistical testing at two granularities: at the fold level, robustness of performance differences (Wilcoxon signed-rank, Hodges–Lehmann) and stability ($\Delta CV\%$ with BCa CIs and permutation tests); at the prediction level, probabilistic quality (Brier, ECE) and binary-decision agreement (McNemar's test). Phase 3 (Final Hold-Out Evaluation) refits both specifications on the full development span under the selected configurations, fixes the decision threshold ex-ante from development data, and reports generalization on the strictly held-out test set using the pre-specified calibrator and threshold.

The remainder of this paper was organized as follows. Section 2 reviews the relevant literature to formally establish this research gap. Section 3 describes the data and the proposed methodological framework. Section 4 presents the empirical results, and Section 5 concludes the paper.

2. Literature

The ascendance of sustainable development as a global norm [23,24] has transformed financial thinking, prompting a re-evaluation of sustainability through its three core dimensions: economic efficiency, environmental protection, and social equity [23,25]. One of the key mechanisms for operationalizing this perspective is the ESG index, which serves as a benchmark for firm performance at the micro level and as a potential indicator of systemic resilience at the macro level, though this role remains debated [15]. Major providers such as S&P Global, Dow Jones, and MSCI have introduced widely used indices, including the Dow Jones Sustainability World Index and MSCI's World and Emerging Markets ESG Focus series, which underpin investable products like ETFs from iShares, Vanguard, SPDR, and Fidelity [7,8,26]. Alongside these global benchmarks, national adaptations such as Türkiye's BIST Sustainability Index, Brazil's ISE B3, and South Africa's JSE Responsible Investment Index reflect local macroeconomic priorities and regulatory contexts.

Against this backdrop, a vast and methodologically diverse academic literature has emerged to understand and model these complex assets. This section proceeds as follows: it first examines the multi-scale determinants that drive ESG performance, from macro-financial variables to firm-specific characteristics. It then critically evaluates the evolution of forecasting methodologies, from traditional econometrics to modern machine learning, highlighting their capabilities and limitations. Following this, it introduces information-theoretic concepts as a powerful alternative lens for capturing the complexity that other models overlook. Finally, it synthesizes these streams to identify a critical research gap and motivate the complexity-aware modeling framework proposed in this study.

2.1. Key Determinants of Financial Sustainability in ESG Markets: A Multi-Scale Perspective

Extensive research demonstrates that ESG performance is shaped by a multi-scale interplay of market dynamics, institutional structures, and firm-specific strategies. To provide a concise synthesis of this literature, Table 1 organizes the principal determinant

categories into four analytical domains: (1) macro-financial and market, (2) institutional and structural, (3) corporate and firm-level, and (4) social–environmental and behavioral.

Table 1. Summary of multi-scale determinants of ESG performance.

Determinant Category	Key Variables/Examples	Representative Literature
Macro-Financial & Market	Commodity and energy prices (oil, gold), exchange and interest rates, economic policy uncertainty, systemic shocks (e.g., pandemics, crypto spillovers).	[17–19,27–34]
Institutional & Structural	Economic growth, institutional quality, SDG alignment, demographic and public finance indicators, higher education, circular economy, renewable energy systems.	[35–43]
Corporate & Firm-Level	Financial structure (debt, liquidity), income diversification, intellectual capital efficiency, profitability, capital structure, firm performance.	[44–51]
Social, Environmental & Behavioral	Corporate social responsibility, employee engagement, social media activity, environmental disclosure, governance transparency.	[52–55]

Within this framework, ESG outcomes emerge from complex and interdependent relationships across these domains rather than from isolated causal mechanisms. This documented complexity highlights the importance of developing forecasting methodologies capable of capturing non-linear, cross-level, and dynamic dependencies.

2.2. Traditional and Emerging Approaches to ESG/Sustainability Index Forecasting

The analytical challenges outlined above have driven the evolution of methodologies used to analyze and forecast ESG and sustainability indices. Over the past decade, approaches have gradually shifted from conventional econometric frameworks toward data-driven machine learning (ML) and deep learning (DL) paradigms. As researchers sought to capture the complex, non-linear, and interdependent dynamics of sustainability markets, this transition marked a methodological turning point from explanatory, assumption-driven models to flexible algorithms capable of uncovering hidden patterns and adaptive relationships within ESG and financial data.

Early studies predominantly relied on panel data analysis and regression-based methods, including GMM, quantile regression, and Structural Equation Modeling (SEM), among others [36,39,42,43,46,47,53]. While these approaches provided foundational empirical insights, their dependence on linearity and stationarity assumptions limited their ability to represent the non-linear complexity inherent in sustainability indices.

The most significant methodological advance has been the widespread adoption of ML and DL models, which shift from theory-driven to data-driven forecasting [56]. Early applications employed Decision Trees, Artificial Neural Networks, and Support Vector Machines [57,58], while ensemble methods such as Random Forest (RF) demonstrated superior predictive power in financial sustainability contexts [59]. More recent studies have combined ML with efficiency analysis frameworks, notably Data Envelopment Analysis integrated with RF and interpreted using SHAP analysis [60] and have developed specialized DL models for ESG-specific price and volatility prediction [9,20,61,62]. Among these, Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU), and their hybrid variants are the most widely adopted architectures in financial forecasting [63]. Recent applications in ESG domains confirm their effective-

ness, with studies demonstrating strong predictive performance for volatility and risk modeling [9,15,20,61,62].

To provide a synthetic overview of this rapidly expanding literature, Table 2 summarizes recent ESG forecasting frameworks, organizing them by methodological class, feature composition, and forecasting objective. The comparative synthesis reveals three key trends: a shift from single-model to hybrid architectures; an increasing integration of textual and unstructured information; and growing attention to interpretability and calibration through model-agnostic techniques such as SHAP or focal-loss optimization. Despite these notable advances in predictive accuracy and model flexibility, most existing studies still treat heterogeneous ESG features as static aggregates and lack a principled framework to quantify informational uncertainty, temporal complexity, and regime-dependent behavior. This methodological gap motivates the integration of entropy-based measures into forecasting frameworks.

Table 2. Recent ML- and DL-based frameworks for ESG forecasting.

Source	Determinant Type	Key Variables/Drivers	ESG Forecasting Models
[9]	ESG-focused Portfolios	Stock Returns, ESG Ratings, Portfolio Weights	DRIP with Multivariate Bidirectional LSTM
[15]	Fundamental, Technical, and Macroeconomic Drivers of ESG Index Volatility	Cboe Volatility Index, Interest Rate, Civilian Unemployment Rate, Consumer Sentiment Index, US Dollar Index, Technical Indicators	LSTM, GRU, CNN
[20]	ESG Newsflow-Driven Volatility Determinants	ESG-Related Financial News, Textual Features Extracted from Newsflow, Transformer-Based Language Representations	ESG2Risk Deep Learning Pipeline
[56]	Technical Indicators-Based Market Drivers	Technical Indicators	Decision Tree, Random Forest, AdaBoost, XGBoost, SVC, Naïve Bayes, KNN, Logistic Regression, ANN, RNN, LSTM
[57]	Financial Ratios & Industry-Specific Drivers	Profitability Ratios, Liquidity Ratios, Leverage Ratios, Management Efficiency Ratios, Fraud Checks, Industry Code, Company Size (96 Financial And Industry-Related Indicators)	Decision Tree, Naïve Bayes, ANN
[58]	Urban Sustainability Indicators	Environmental, Social, and Economic Indicators	Decision Tree, ANN, SVM
[59]	Financial Sustainability of Microfinance Institutions	Operational, Financial, and Institutional Variables Of Microfinance Institutions	Random Forest, Quantile Random Forest, Linear Regression, Partial Least Squares, Stepwise Linear Regression, Elastic Net, Bayesian Ridge Regression, KNN, SVR
[60]	Financial Sustainability of Banks	Loans and Leases, Interest Income, Total Liabilities, Total Assets, Market Capitalization, Revenue To Assets, Revenue Per Share	Random Forest Classification, SHAP-based Feature Analysis, Three-Stage Network DEA
[61]	ESG Performance and Investment Decisions	ESG Variables	Light Gradient Boosting Machine, Local Outlier Factors, LSTM, GRU
[62]	Corporate ESG Disclosure and Communication	ESG-Related Sentences in Earnings Calls	Neural Language Modeling
[63]	Systemic Banking Risk & ESG Factors	ESG Risk Score, Inflation Rate, Unemployment Rate, House Prices, Current Account Balance/GDP Ratio	Interpretable Multivariate LSTM with Focal Loss

2.3. Information-Theoretic and Entropy-Based Approaches in Financial Time Series Modeling

Conventional econometric and even many machine learning models often fall short when confronted with the strong non-linearities and regime-dependent patterns of financial markets. To address this limitation, information theory provides a principled lens, viewing markets as adaptive systems defined by uncertainty, complexity, and change [64].

Within this framework, SE remains a foundational measure of distributional uncertainty and informational diversity in financial markets [65,66], with recent applications to market efficiency and volatility analysis [67,68]. Moving beyond static uncertainty measures, PE captures the temporal organization of price dynamics through ordinal patterns, quantifying how the order of successive observations evolves over time [69–71].

This property enables PE to detect hidden regimes, local nonlinearities, shifts in market efficiency [72–74], and reveal instability through irreversibility analysis [75].

In contrast, KL divergence quantifies the distance between probability distributions, offering a complementary view of complexity by identifying structural breaks and regime transitions through changes in the overall shape of return distributions [76–80]. While PE is inherently sensitive to temporal order, KL divergence captures broader distributional reconfigurations, making both measures jointly powerful for modeling structural change and instability in financial systems.

Despite their proven value in detecting hidden complexity, distributional change, and systemic stress in traditional financial markets [77,81,82], these entropy-based tools remain underexplored within sustainability forecasting. This underutilization defines a crucial methodological gap that the present study seeks to address.

2.4. Synthesizing the Literature: The Case for a Complexity-Aware Forecasting Framework

The preceding literature review reveals a central paradox in sustainability finance. While the field has made important progress in identifying the multi-scale determinants of ESG performance and has increasingly adopted machine learning for forecasting, these advances have also exposed a deeper methodological gap: a disconnect between the complexity of ESG markets and the ability of existing models to capture their informational dynamics. Even sophisticated models tend to treat heterogeneous features as simple aggregates, without a principled way to quantify the emergent properties (uncertainty, complexity, and structural instability) that characterize sustainability-linked assets.

This study addresses this gap by proposing a complexity-aware forecasting framework. The core innovation lies in shifting from feature aggregation to the explicit modeling of market information. The framework systematically integrates a curated set of information-theoretic metrics: SE (distributional uncertainty), PE (sequential complexity), and KL divergence (structural shifts). Together, these measures allow the model to quantify both the state of the market and its transitions.

The central hypothesis is that this augmentation improves not only predictive accuracy but also, and more importantly, the stability and calibration of forecasts across heterogeneous regimes. Together, stability and calibration enhance a forecast's practical value in volatile ESG markets: stability ensures consistent performance across different regimes, while calibration provides the quantifiable confidence necessary for granular risk management and reliable decision-making under uncertainty.

To demonstrate its external validity, the framework was applied to Türkiye's BIST Sustainability Index, an emerging-market ESG index that serves as a natural stress test. The index operates in an emerging-market setting and is notably exposed to geopolitical shocks, abrupt policy changes, and volatility in cross-border capital flows. The framework's ability to deliver robust gains in this stress-test setting highlights not only its practical relevance for emerging-market contexts but also its potential as a portable and generalizable methodology for embedding informational dynamics into predictive modeling across global sustainability finance.

3. Data and Methodology

This section presents an end-to-end, complexity-aware framework to forecast the next-day directional movement of the BIST ESG index. The study evaluated, under strict time-respecting and leakage-aware conditions, whether augmenting a baseline of macroeconomic and technical indicators with information-theoretic meta-features improved (i) predictive discrimination, (ii) probability calibration, and (iii) temporal stability of performance.

The workflow was organized as follows. Section 3.1 documents the dataset and the multi-stage feature-engineering protocol. Section 3.2 details the modeling pipeline, including the temporal data-partitioning scheme, the baseline versus information-theoretic augmented feature sets, and a calibrated classifier tailored to time-series classification. Section 3.3 presents the evaluation design, comprising nested cross-validation, a suite of statistical comparison procedures, and the final assessment on a strictly held-out test set.

3.1. Data

3.1.1. ESG Index Data

The primary dataset is the BIST Sustainability Index (XUSRD), the first ESG benchmark launched by Borsa İstanbul on 4 November 2014 and widely recognized as Türkiye's national ESG index. The index comprises firms listed on BIST that meet published ESG performance criteria, and its composition is reviewed on a regular schedule (typically annually), so the number of constituents varies over time.

For the empirical analysis, daily index levels were obtained from the Central Bank of the Republic of Turkey (CBRT) electronic data distribution system for the period 4 November 2014 to 5 November 2024. The dataset includes standard OHLC (Open, High, Low, Close) prices, recorded in TRY. These series formed the basis for all return calculations and the construction of the derived features used in the forecasting models.

3.1.2. Complementary Market Variables

To capture the broader market environment in which the ESG index operates, a curated set of complementary variables was incorporated into the analysis. These macro-financial indicators were selected to represent both global market dynamics and local systematic risks. Specifically, three external variables were employed: the USD/TRY exchange rate, commonly regarded as an indicator of financial stability and exposure to external shocks in emerging markets [83]; the XAU/USD gold price, widely recognized as a proxy for global risk aversion and a safe-haven asset during episodes of financial stress [84]; and the Brent crude oil price, which reflects global economic activity and inflationary pressures [85]. Incorporating these variables enabled the framework to account for macroeconomic conditions in parallel with the endogenous dynamics of the ESG index.

3.1.3. Technical Indicators

Technical indicators were incorporated into the feature set as quantitative proxies for market psychology and trading behavior. A substantial body of empirical research shows that such indicators, by capturing non-fundamental aspects of price formation, can enhance forecasting performance across both econometric and machine-learning approaches [86–92]. In this study, a curated set of indicators was employed and grouped into three functional categories: (i) trend-following measures (Exponential Moving Average (EMA) and Parabolic SAR-PSAR) to capture directional momentum; (ii) momentum oscillators (Relative Strength Index (RSI) and Williams %R) to assess the velocity of price movements and identify overbought/oversold conditions; and (iii) a volatility measure (Average True Range (ATR)) to quantify the scale of market fluctuations. The definitions, formulas, parameterizations, and expected roles of these indicators are summarized in Table 3.

Table 3. Definitions and expected roles of technical indicators.

Definition	Formula
Trend-Based Technical Indicators	
EMA is a trend-following indicator that applies exponentially decaying weights to past observations [93,94]. Unlike SMA (equal weights), EMA emphasizes recent data, enhancing responsiveness while smoothing noise. It captures short- to intermediate-term directional momentum.	$EMA_t = \alpha P_t + (1 - \alpha) \times EMA_{t-1}$ where P_t denotes the closing price and $\alpha = 2/(n + 1)$ is the smoothing coefficient with n lookback window.
PSAR captures trend direction and potential reversals; dots below price → uptrend, above → downtrend. Also used as a trailing stop [95–97].	Uptrend: $SAR_t = SAR_{t-1} + \alpha(EP_{t-1} - SAR_{t-1}).$ Downtrend: $SAR_t = SAR_{t-1} - \alpha(SAR_{t-1} - EP_{t-1}).$ where EP is the extreme point and α is the acceleration factor.
Momentum-Based Technical Indicators	
RSI is a momentum oscillator, bounded between 0 and 100, that measures the speed and change of price movements. It is used to identify overbought (>70) and oversold (<30) conditions movements [98].	$RSI = 100 - \left(\frac{100}{1 + RS} \right)$ where RS is the ratio of average gains to average losses over the lookback period [99].
Williams %R is bounded between 0 and −100, that measures the current closing price in relation to the high/low range over a past period n . It is used to identify overbought (>−20) and oversold (<−80) level [94].	$\%R_t = \left(\frac{H_n - P_t}{H_n - L_n} \right) \times (-100)$ where P_t is the closing price at time t, H_n is the highest price over the lookback period n, and L_n is the lowest price over the same period.
Volatility-Based Technical Indicators	
ATR is a measure of market volatility that incorporates price gaps. It quantifies the degree of price movement or variability, rather than the direction. High values indicate high volatility [94,100–102].	$TR_t = \max \{ (H_t - L_t), \text{abs}(H_t - P_{t-1}), \text{abs}(L_t - P_{t-1}) \}$ $ATR_t = \frac{(n-1)ATR_{t-1} + TR_t}{n}$ where H_t is the current high, L_t the current low, P_{t-1} the previous close, TR_t the true range at time t, and n the lookback period.

3.2. Feature Engineering

Feature engineering in this study targets representations that complement the macroeconomic context by encoding endogenous ESG market dynamics. While macro variables summarize exogenous conditions, they are often insufficient to capture investor-driven regime shifts. Classical technical indicators (e.g., EMA, PSAR, RSI, Williams %R, and ATR) quantify trend, momentum, and volatility, but largely remain surface-level statistics of price paths. To address this representational gap, information-theoretic features (comprising market-state and market-transition indicators) were introduced as principled descriptors of systemic uncertainty, structural change, and sequential complexity in financial time series.

The construction of the final feature space followed a two-stage process. First, a leakage-aware time-series cross-validation procedure was used to optimize the lookback parameters of the classical technical indicators (Section 3.2.1). Second, the set of information-theoretic features was computed under the same time-respecting alignment (Section 3.2.2). Together with the block of raw macro-financial variables, these derived feature sets form the complete pool of predictors used in the modeling phase.

3.2.1. Optimization of Technical Indicator Parameters

Rather than relying on default parameters, a data-driven procedure was applied to select lookback windows for indicators with tunable horizons. Grid searches were defined

a priori for $\text{EMA} \in \{10, 20, 30, 40, 50\}$, $\text{RSI} \in \{7, 14, 21\}$, $\text{ATR} \in \{10, 14, 20\}$ and $\text{Williams \%R} \in \{7, 14, 21\}$. PSAR was retained at conventional settings (acceleration factor starting at 0.02, incrementing by 0.02 on each new extreme, and capped at 0.20) to reflect standardized practice and avoid unnecessary overfitting.

The optimization protocol was designed to be model-agnostic and leakage-aware. For each candidate window, the univariate indicator series was evaluated against the binary next-day direction target using a 5-fold TimeSeriesSplit restricted to the training data. Discrimination was quantified by the Area Under the ROC Curve (AUC), which was directionally adjusted using the transformation $\max(\text{AUC}, 1 - \text{AUC})$. This ensures that the optimization captures the strongest predictive signal, regardless of whether the correlation is positive or negative. The window length yielding the highest mean validation AUC was selected as optimal for each indicator.

The resulting set of optimized horizons for EMA, RSI, ATR, and Williams %R with PSAR at conventional settings constituted the finalized technical-indicator configuration used downstream in the modeling pipeline, with all computations performed under time-respecting alignment (no look-ahead).

3.2.2. Information-Theoretic Feature Extraction

To capture non-linear dynamics and informational properties that price-based technical indicators and macro-financial covariates may not fully encode, the feature set was augmented with information-theoretic meta-features. Following the information-theoretic tradition in econometrics and return predictability [64,76], these measures are not trading rules; rather, they serve as a diagnostic layer that quantifies market uncertainty (entropy), dynamical complexity, and susceptibility to structural change and regime shifts, thereby contextualizing both macroeconomic and technical signals [64,76,77].

Let $r_t = (P_t - P_{t-1})/P_{t-1}$ denote the simple return of the ESG index closing price, where P_t is the closing price at time t . Three information-theoretic features were extracted over a 14-day rolling window of r_t to capture distinct aspects of the series' informational properties. The selected features consist of market state indicators and a market transition indicator which are detailed below. For measures requiring discrete probability distributions, a discretization with $K = 10$ bins was employed; to ensure numerical stability, probabilities were offset by $\varepsilon (=10^{-12})$ prior to logarithmic operations. Further operational details for SE/PE (state) and KL (transition) are provided in the subsections below.

Market State Indicators: Entropy Measures

The informational state of the ESG market was evaluated with two complementary entropy measures: SE, which captures distributional uncertainty, and PE, which reflects sequential (ordinal) complexity.

SE, a foundational metric introduced by Shannon (1948) [65], measures the average uncertainty of the return distribution within each rolling window. For a given window of simple returns, a discrete probability distribution $\{p_i\}_{i=1}^K$ was constructed from histogram-binned returns, where π_i is the empirical probability of the i^{th} bin. Conceptually, higher SE reflects a more dispersed and unpredictable market state, typically associated with elevated volatility and diverse investor expectations, whereas lower SE indicates concentrated and highly predictable returns with minimal informational diversity, conditions that often occur during periods of market consensus or one-sided trading phases. The mathematical formulation for SE was provided in Appendix A.

PE, a model-free measure introduced by Bandt and Pompe (2002) [69], quantifies the sequential (or ordinal) complexity encoded in the relative ordering of consecutive values rather than their magnitudes, making it robust to outliers [92] and suitable for noisy, non-stationary financial data (e.g., [81,82,103]). Conceptually, lower PE indicates a loss of

informational diversity and an increase in synchronized trading behavior, conditions that are typically associated with herd dynamics and reduced market efficiency, whereas higher PE reflects more heterogeneous and adaptive market interactions.

To compute PE, the return series within each window, $\{r_t\}$, was mapped into overlapping vectors of length $d = 3$ (embedding dimension) with a time lag of $\tau = 1$ [104]. Each vector was then assigned to one of the $d!$ possible ordinal patterns (permutations). Let $p(\pi_j)$ denote the empirical frequency of a given permutation π_j . The normalized PE, which scales the entropy value to a range between 0 (perfectly ordered) and 1 (completely random), is formally defined in Appendix A.

Market Transition Indicator: Divergence Measure

While SE and PE were employed to quantify the within-window informational state, a between-window measure was required to detect structural changes in the return-generating process. For this purpose, the KL divergence was adopted as the primary market transition indicator. The choice of KL is motivated by several properties: its asymmetry is advantageous for modeling temporal processes; it is a standard metric in the econometrics of regime detection; and it remains computationally efficient in a rolling-window implementation [76,105,106].

For each rolling window, a discrete probability distribution $P = \{p_i\}_{i=1}^K$ was derived from the returns using the histogram method with K bins. The reference distribution $Q = \{q_i\}_{i=1}^K$ was constructed in the same manner from the immediately preceding window (shifted by one day). KL divergence measures the information change from the previous distribution Q to the current distribution P ; higher values indicate potential regime shifts.

Economically, higher KL values indicate pronounced regime transitions or external shocks that disrupt the prevailing market structure and reflect a sharp increase in distributional instability and informational surprise between consecutive time windows. In contrast, lower KL values correspond to stable market conditions where the return-generating process evolves smoothly and remains consistent with recent history. The formal computation for KL divergence is provided in Appendix A.

The complete set of predictors used in the modeling phase is organized into three coherent categories. The macroeconomic variables were incorporated in raw level form without scaling or transformation. All technical indicators were computed directly from the OHLC price series, with PSAR retained at its standard parameterization and the lookback windows for EMA, RSI, ATR, and Williams %R selected through the leakage-free time-series cross-validation procedure. The information-theoretic features (SE, PE, and KL divergence) were derived from the daily simple-return series (pct change), and no normalization or scaling (e.g., z-score or min–max) was applied, in order to preserve the distributional characteristics essential for valid entropy- and divergence-based diagnostics. This finalized, category-structured feature set is summarized in Table 4, which provides a concise overview of all engineered predictors used in the subsequent modeling workflow.

3.3. Modeling Framework

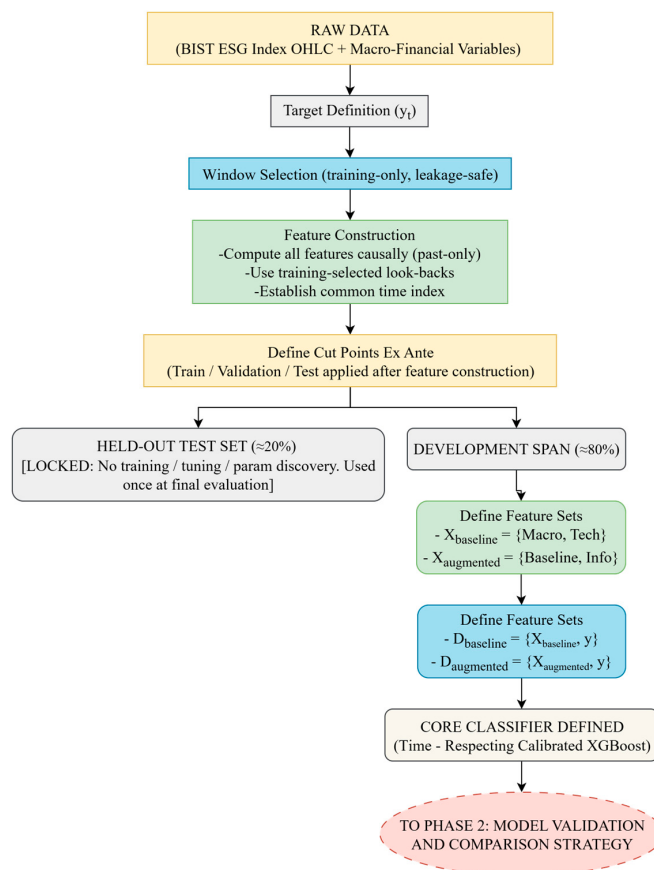
The complexity-aware framework was designed for a fair, leakage-aware comparison between two specifications: a baseline model, comprising macro-financial variables and optimized technical indicators, and an information-theoretic-augmented model, which additionally incorporates SE, PE, and KL divergence features. These information-theoretic signals were treated as diagnostics of market state and transitions and were evaluated for their incremental contribution to probability calibration, fold-to-fold stability, and directional accuracy under a time-consistent design. The workflow proceeded in three main phases: (i) model and data preparation, (ii) validation and comparison strategy, and (iii) final model training and hold-out evaluation.

Table 4. Summary of the final engineered feature set.

Category	Features	Preprocessing Notes
Macroeconomic	exchangerate, gold, oil	Raw levels
Technical	EMA, RSI, ATR, WILLR	Optimized lookback windows
Technical indicator (fixed)	PSAR	Standard configuration
Information-theoretic	SE, PE, KL divergence	Computed on daily returns

3.3.1. Phase 1: Model and Data Preparation

In this phase, the dataset was partitioned chronologically into training, validation, and a strictly held-out test block, and the prediction target was defined as the next-day direction. Two feature specifications were fixed: a baseline set and an information-theoretic-augmented extension adding entropy- and divergence-based measures. All features were computed causally, and the test block remained untouched. The same XGBoost classifier with a time-respecting probability calibration layer was specified for both feature sets. A schematic of this workflow is shown in Figure 1. The diagram illustrates the main data flow and feature construction steps leading from raw inputs to baseline and augmented model definitions under a time-respecting design.

**Figure 1.** Schematic of the Phase 1 data preparation and partitioning workflow.

Temporal Data Partitioning

To prevent look-ahead bias and to obtain an unbiased estimate of generalization performance, the dataset was partitioned chronologically into a development span (≈80%) and a strictly held-out test block (≈20%). Following common practice in financial time-series forecasting, the development span was further divided into a training block (≈60%) and a validation block (≈20%).

The roles of these subsets were strictly defined. The training set was used for initial model fitting and for leakage-aware optimization of technical indicator look-back windows, selected exclusively from training data. The validation set supported model selection and hyperparameter tuning within the development span.

To ensure temporal integrity and consistency, feature engineering was applied prior to the final partitioning. Features were computed causally (using only past information) to establish a common time index, which inherently handled the removal of warm-up periods. The definitive chronological split was then applied to this aligned, feature-rich dataset, ensuring that all subsets fully preserved the causal structure of the data. The test block remained strictly untouched throughout all development phases and was used only once to report the final out-of-sample performance.

Target Construction

The modeling task was formulated as a binary classification problem. The target variable y_t was set to 1 if the next day's closing price P_{t+1} exceeded the current day's closing price P_t , and 0 otherwise. This binary setup formalizes the task as predicting whether the next-day closing price will rise relative to the current day.

Baseline and Augmented Feature Sets

To isolate and evaluate the incremental contribution of the information-theoretic features across multiple dimensions of model performance, two distinct feature sets (design matrices) were constructed for a direct, like-for-like comparison. This evaluation was designed to assess the impact on not only directional accuracy but also on fold-to-fold stability and probability calibration.

X_{baseline} : This specification includes only traditional predictors:

- (i) The block of external macroeconomic variables;
- (ii) The block of empirically optimized technical indicators and the fixed-parameter PSAR. No information-theoretic measures are included.

$X_{\text{augmented}}$: This specification extends X_{baseline} with the full suite of information-theoretic measures, categorized as:

- (i) The block of market state indicators (SE and PE);
- (ii) The market transition measure (KL divergence).

Formally, let X^{macro} , X^{tech} , and X^{info} denote the macroeconomic, technical, and information-theoretic blocks, respectively. The two design matrices are defined as:

$$X_{\text{baseline}} = [X^{\text{macro}}, X^{\text{tech}}], \quad X_{\text{augmented}} = [X^{\text{macro}}, X^{\text{tech}}, X^{\text{info}}].$$

Both specifications were evaluated under an identical protocol on the same temporal partition defined previously ($\approx 60-20-20$). Because the learner is tree-based and invariant to monotonic transformations, no feature scaling was applied; therefore, the shared predictors in X_{baseline} and $X_{\text{augmented}}$ have identical values. Consequently, any observed performance difference is attributable solely to the inclusion of X^{info} . This experimental design yields two parallel datasets for the subsequent modeling phases:

$$D_{\text{baseline}} = \{X_{\text{baseline}}, y\}, \quad D_{\text{augmented}} = \{X_{\text{augmented}}, y\}$$

Core Classifier: Time-Respecting Calibrated XGBoost

The core predictive engine for both datasets defined previously, D_{baseline} and $D_{\text{augmented}}$, was an XGBoost classifier, selected for its state-of-the-art performance on tabular data and built-in regularization against overfitting [107]. However, a known limitation of such classifiers is their tendency to produce miscalibrated probabilities [108]. As this study evaluated the models not only on classification accuracy but also on the quality of

probabilistic forecasts using calibration-sensitive metrics (e.g., Brier score and Expected Calibration Error-ECE), addressing this issue was paramount.

To this end, a time-respecting calibration wrapper was implemented around the XGBoost classifier and applied identically to both D_{baseline} and $D_{\text{augmented}}$. Calibration was performed after the initial model training, not before, to prevent any look-ahead bias, following standard practice in probability calibration (e.g., [108]), where calibration is applied post-training to adjust output probabilities without affecting the base learner. The process was nested within validation: in each outer fold of the nested CV, the calibration split followed the training split in time, and the mapping g was fit only on the scores obtained from this later, non-overlapping data. This protocol enforced causality by ensuring that the calibration step used past-only information relative to its application, thereby eliminating look-ahead and transforming raw scores into well-calibrated probabilities.

The calibration procedure followed a time-respecting, post-training design, ensuring that only past information was used at each step. Specifically, the development span was partitioned into a disjoint, past-only calibration holdout slice: the base learner (f) was trained on the earlier portion of the data, and the mapping function (g) was subsequently fitted on the scores obtained from this later, non-overlapping holdout. This approach enforced causality by calibrating model outputs strictly on future-unseen data within each fold. The formal algorithmic steps and mathematical definitions for this time-respecting calibration procedure are detailed in Appendix B (Time-Respecting Calibration Protocol).

Two well-established alternatives were considered for the mapping function g : Platt Scaling (logistic regression) [109], and isotonic regression (pool-adjacent-violators) [110]. To stabilize calibration, the fraction α was drawn from the range $[0.15, 0.30]$, with a minimum of 100 observations enforced for the calibration split; potential overfitting of isotonic regression was mitigated by applying it only on temporally subsequent holdouts within inner validation. Class imbalance in each training split was handled by setting `scale_pos_weight` dynamically. A temporal gap equal to one feature-window length was enforced between train and validation segments in both inner and outer splits to avoid overlap-induced leakage. The choice of g and its parameters was treated as a tunable option within the development span and applied identically to both feature specifications X_{baseline} and $X_{\text{augmented}}$, providing a consistent foundation for the validation and comparison procedures described in Phase 2.

3.3.2. Phase 2: Model Validation and Comparison Strategy

This phase details the protocol for validating and comparing the baseline and the information-theoretic augmented models on the development span ($\approx 80\%$). The core is a time-respecting NCV in which the Time-Respecting Calibrated XGBoost classifier was tuned and evaluated. Applied identically to both datasets, the NCV used forward-chaining TimeSeriesSplit (3 inner folds, 5 outer folds) with strictly chronological splits (no shuffling, past-only information). The inner loop jointly tuned XGBoost hyperparameters and the probability-calibration component (Platt or isotonic), while the outer loop provided unbiased performance estimates. For each model, the procedure yielded (i) five outer-fold scores and (ii) a pooled set of out-of-fold, per-sample predictions from the calibrated classifier. These outputs underpinned two complementary analyses: fold-level comparisons (stability/robustness) and prediction-level comparisons (probabilistic quality and binary-decision agreement). The held-out test set was not used in this phase. A schematic of this workflow is shown in Figure 2, illustrating the NCV structure used for model tuning and comparison. The diagram highlights the generation of fold-level and prediction-level outputs under identical, time-respecting conditions.

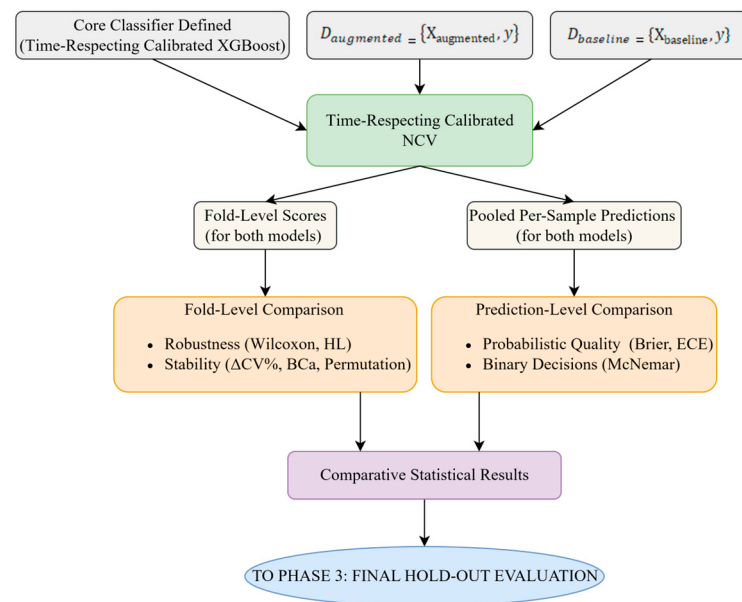


Figure 2. Schematic of the Phase 2 nested cross-validation and model-comparison workflow.

Nested Cross-Validation Protocol for Hyperparameter Tuning and Performance Estimation

To identify the optimal configuration and obtain an unbiased estimate of generalization, a time-respecting NCV procedure was implemented on the development span (the first $\approx 80\%$ of the sample), leaving the remaining $\approx 20\%$ as a strictly held-out test block. The protocol was designed to preserve chronological order and prevent data leakage. Its main steps are summarized below.

1. Fold Structure

A TimeSeriesSplit scheme was employed, which preserved chronological order (no shuffling) with 3 inner folds (for hyperparameter tuning via RandomizedSearchCV) and 5 outer folds (for performance estimation). The choice of a 3×5 -fold structure aimed to balance the bias–variance trade-off, consistent with established recommendations for time-series tasks of this scale [111,112].

2. Inner Loop (Hyperparameter Tuning)

Given the broad hyperparameter space of the XGBoost + calibration wrapper (see Supplementary Table S1), the inner loop used RandomizedSearchCV rather than an exhaustive grid due to the wide parameter ranges and the diminishing returns of exhaustive enumeration. Within this loop, the search treated the choice of calibration method (Platt vs. isotonic) as a tunable hyperparameter, which was optimized jointly with the standard XGBoost hyperparameters and a calibration-holdout fraction drawn from the range $[0.15, 0.30]$, ensuring a minimum of 100 observations in the calibration slice. With a budget of $n_iter = 200$ per inner loop and a 3-fold TimeSeriesSplit, each outer fold evaluates approximately 600 candidate model fits; across 5 outer folds this totals approximately 3000 inner-loop fits per specification ($200 \times 3 \times 5$), plus 5 refits of the selected configurations.

3. Outer Loop (Performance Estimation)

For each outer split, the model was trained on the outer-train slice, calibrated on its past-only calibration slice, and evaluated on the outer-validation slice. The resulting outer-fold scores were then aggregated to obtain an unbiased estimate of generalization performance.

4. Protocol Application and Bias Prevention

The same time-respecting NCV design (chronological splits, no shuffling) and search protocol were applied identically to both the baseline and the augmented specifications. By

strictly insulating hyperparameter (and calibrator) selection from performance estimation, this nested design prevents optimistic bias and yields a robust, unbiased distribution of expected out-of-sample performance under identical experimental conditions.

Performance Evaluation Metrics

Let $y_i \in \{0, 1\}$ denote the true label and $\hat{p}_i \in [0, 1]$ the predicted probability for instance i . Confusion-matrix entries are denoted by TP (true positive), FP (false positive), TN (true negative), FN (false negative).

Directional accuracy metrics quantify thresholded classification performance and are robust to class imbalance. The suite includes: Balanced Accuracy (BACC), which averages sensitivity (recall) and specificity; the F1-Score, the harmonic mean of precision and recall, which emphasizes positive-class detection; and the Matthews Correlation Coefficient (MCC), a robust correlation measure between predicted and observed classifications on a scale of -1 to $+1$. The formal definitions for directional accuracy metrics (BACC, Precision, Recall, F1, MCC) are provided in Appendix C.

Discrimination metric, the receiver operating characteristic area under the curve (ROC AUC) quantifies threshold-independent discrimination: it is the probability that a randomly chosen positive instance receives a higher score $\hat{s}_i$ ($\hat{p}_i$) than a randomly chosen negative one ($1.0 = \text{perfect}$, $0.5 = \text{random}$), i.e., the proportion of correctly ordered positive–negative pairs. Given the mild class imbalance, the F1-score was pre-specified as the primary evaluation metric, with BACC, MCC, and ROC-AUC reported as complementary metrics for directional accuracy and discrimination. The formal definition for AUC is provided in Appendix C.

Probability calibration metrics quantify the alignment between a model’s probabilistic forecasts and the observed outcomes, independently of any decision threshold. The Brier score measures this via the mean squared difference between predicted probabilities and realized labels [113]. The Expected Calibration Error (ECE) summarizes the disparity between a model’s mean confidence and the empirical event rate within probability bins (here $K = 10$, equal width). Let B_k denote bin k with size $n_k = |B_k|$ and $n = \sum_{k=1}^K n_k$ [114]:

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^n (\hat{p}_i - y_i)^2 \quad (1)$$

$$\text{ECE} = \sum_{k=1}^K \frac{n_k}{n} \left| \frac{1}{n_k} \sum_{i \in B_k} y_i - \frac{1}{n_k} \sum_{i \in B_k} \hat{p}_i \right| \quad (2)$$

Model stability metrics quantify the consistency of performance across repeated estimates (e.g., outer folds) by measuring the dispersion of the performance scores, rather than their absolute magnitude. Let $m_1, \dots, m_K$ denote the set of metric values over K repeats, with mean $\bar{m}$ and standard deviation s . Two summary statistics are defined to assess stability: the coefficient of variation (CV%) and a performance-to-stability ratio R ; lower CV% (as it reflects lower dispersion relative to the performance level) and higher R (as it reflects a stronger signal relative to its noise) indicate greater stability.

$$\text{CV\%} = 100 \times \frac{s}{|\bar{m}|}, \quad R = \frac{\bar{m}}{s} \quad (3)$$

Statistical Comparison of Model Performance

This section details the statistical procedures used to test whether the augmented specification provided a significant improvement over the baseline and whether such gains were robust and stable. All comparisons used outer-fold predictions from the time-respecting NCV.

Fold-Level Comparison (Robustness and Stability)

Fold-level comparisons were conducted on the five outer-fold, out-of-sample scores from the calibrated XGBoost to assess both the robustness of performance differences and comparative stability.

First, robustness of performance differences was assessed at the fold level for the probability-calibration metrics (Brier score and ECE). The question was whether one specification exhibited a consistent advantage across the five calibrated NCV outer folds. Let $M_f^{(Aug)}$ and $M_f^{(Base)}$ be the augmented and baseline scores on outer fold f ($= 1, \dots, 5$), respectively, and define the per-fold difference as: $\Delta_f = M_f^{(Aug)} - M_f^{(Base)}$.

Statistical significance of the paired differences $\{\Delta_f\}_{f=1}^5$ was assessed using the Wilcoxon signed-rank (Pratt) test, complemented by the Hodges–Lehmann (HL) median effect [115] with a 90% bias-corrected and accelerated (BCa) bootstrap confidence interval [116], based on $B = 20,000$ bootstrap resamples and a paired sign-flip permutation test obtained by randomly swapping the augmented and baseline labels within folds.

Second, models' stability was assessed to quantify and test differences in performance consistency between the baseline and augmented specifications, using directional-accuracy and discrimination metrics (BAcc, F1, MCC, ROC AUC). The question was whether one specification exhibited a statistically significantly lower variability (i.e., higher stability) across the five calibrated NCV outer folds. For each metric, stability was summarized as CV% computed over the five-fold-level scores for each model, and the stability contrast was defined as: $\Delta CV\% = CV\%^{(Aug)} - CV\%^{(base)}$.

Additionally, to synthesize performance level and stability into a single, intuitive signal-to-dispersion ratio (where higher is better), the R metric was also computed and the reliability contrast between models was then defined as the difference in their respective R values: $\Delta R = R^{(Aug)} - R^{(base)}$.

Uncertainty for both $\Delta CV\%$ and ΔR was quantified with a BCa confidence intervals based on $B = 20,000$ bootstrap resamples.

Prediction-Level Comparison (Paired, Sample-wise)

To complement the fold-level analyses, prediction-level comparisons were conducted on the pooled out-of-sample predictions from the five outer folds of the time-respecting NCV. For each observation, calibrated probabilities and the induced 0/1 decisions from the augmented and baseline specifications were paired, permitting per-sample evaluation of calibration quality and binary decision outcomes.

Out-of-sample predictions from all five outer folds were pooled, and calibrated probabilities from the two specifications were compared on a per-observation basis. For the Brier score, per-sample paired differences were formed as $d_i = (\hat{p}_i^{(Aug)} - y_i)^2 - (\hat{p}_i^{(Base)} - y_i)^2$, where negative values favor the augmented specification. The vector $\{d_i\}$ was evaluated using the Wilcoxon signed-rank test (Pratt variant); the HL median effect was reported. Uncertainty was summarized with a BCa bootstrap confidence interval ($B = 20,000$, index-bootstrap resampling), and a paired sign-flip permutation p -value was also computed for the mean difference.

For calibration error, ECE with $K = 10$ equal-width bins was computed on the same pooled predictions for each model, and the scalar contrast $\Delta ECE = ECE_{K=10}^{(Aug)} - ECE_{K=10}^{(Base)}$ was assessed via BCa confidence intervals ($B = 20,000$, index-bootstrap); statistical significance was concluded when the interval excluded zero.

Finally, to compare paired 0/1 predictions, a fold-stratified McNemar's test was employed. This approach was chosen over a standard pooled test to explicitly assess whether one model exhibited a consistently superior error profile across the different time periods represented by the five NCV outer folds. For fold f with index set D_f , let $\hat{y}_i^{(Aug)}, \hat{y}_i^{(base)} \in \{0, 1\}$. Define discordant counts as:

$$\begin{aligned} n_{01}^{(f)} &= \left| \left\{ i \in D_f : \hat{y}_i^{(Aug)} = 1 \wedge \hat{y}_i^{(base)} = 0 \right\} \right| \\ n_{10}^{(f)} &= \left| \left\{ i \in D_f : \hat{y}_i^{(Aug)} = 0 \wedge \hat{y}_i^{(base)} = 1 \right\} \right| \end{aligned} \quad (4)$$

Exact two-sided McNemar tests were computed per fold to test $H_0 : n_{01}^{(f)} = n_{10}^{(f)}$ (no asymmetry in error rates). Fold-specific p -values were then combined using Stouffer's Z (equal weights) to obtain a single meta-analytic p -value, thereby preserving the NCV block structure and avoiding pooling [117,118].

3.3.3. Phase 3: Final Model Training and Hold-Out Evaluation

This phase evaluated the baseline and the information-theoretic augmented models on the strictly untouched hold-out test block (final $\approx 20\%$). All design and tuning decisions were determined exclusively on the development span (first $\approx 80\%$) using time-respecting procedures.

A schematic of this workflow is shown in Figure 3, which outlines the final training and evaluation stage on the strictly held-out test set, including the calibrated model refit, decision-threshold optimization, and McNemar's statistical comparison of out-of-sample predictions.

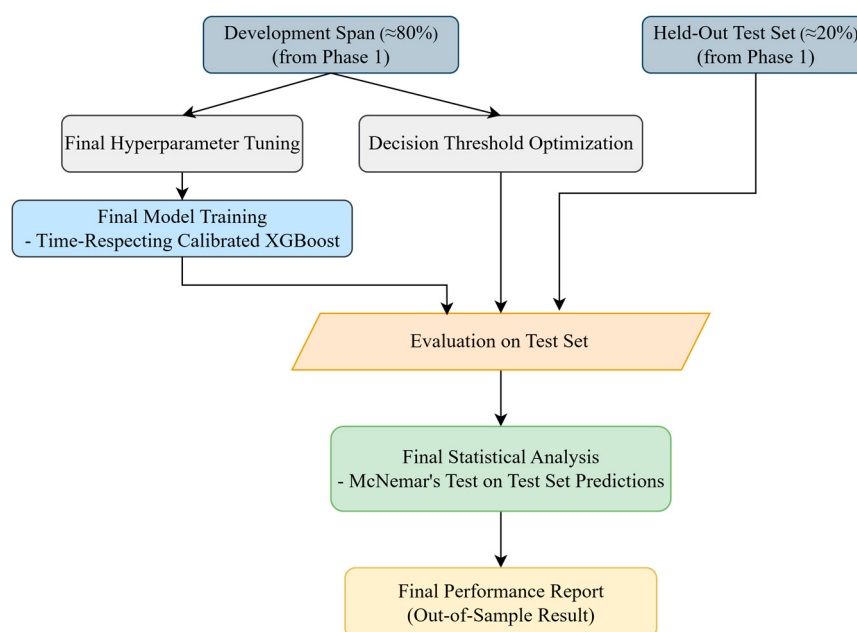


Figure 3. Schematic of the Phase 3 final training and hold-out evaluation workflow.

For each specification, a time-respecting, calibrated XGBoost classifier was re-tuned on the development span with a forward-chaining TimeSeriesSplit (3 inner folds) and RandomizedSearchCV. The hyperparameter search space, CV scheme, scoring, and calibration protocol (isotonic or Platt) were identical to Phase 2.

Using the selected configuration, each model was refit on the full development span with a time-respecting calibration scheme: the base learner was trained on an earlier slice, and a held-out past-only slice within the development span was used to fit the chosen

calibrator. The hold-out test block remained strictly untouched. Decision thresholds were fixed a priori on a per-model basis from out-of-fold calibrated probabilities obtained via forward-chaining CV on the development span, using a dense grid (0.05–0.95; step 0.005) to maximize F1. With calibrated probabilities and the frozen threshold, performance was then computed on the untouched test block.

Paired comparison on the hold-out test set was employed McNemar’s exact test (two-sided, exact binomial). Discordant counts (n_{01} and n_{10}) and the resulting p -value were reported. Thresholds and the calibrator were pre-specified from development data; no test-set tuning was performed.

4. Results

4.1. Experimental Setup and Data Overview

Daily ESG data from 4 November 2014 to 5 November 2024 were analyzed. The raw price series contained no internal missing values and was transformed into daily simple returns, which served as the basis for all entropy and divergence computations. As the return series reflects the natural variation and occasional volatility characteristic of financial markets, no outlier filtering or winsorizing was applied. The only rows removed were the warm-up observations generated by rolling-window operations for entropy and technical indicators. Following these preprocessing steps, the aligned panel comprised 2642 trading days. For out-of-sample assessment, the chronologically last 529 days (~20%) were retained as a strictly held-out test block, while the preceding 2113 days (~80%) constituted the development span. All analyses respected the time order to avoid look-ahead bias (i.e., to ensure a leakage-free evaluation).

The daily simple-return series was examined for stationarity, trend, and serial dependence to characterize the underlying data-generating process. As shown in Table 5, an Augmented Dickey–Fuller test strongly rejected the unit-root hypothesis ($p < 0.001$), indicating stationarity. A Kendall–Tau test detected a mild but statistically significant positive monotonic trend. Ljung–Box statistics showed no joint autocorrelation at short horizons (lag 10), but confirmed significant serial dependence at longer lags (lag 20 and lag 50). The next-day direction target exhibited mild class imbalance (54% up vs. 46% down). Together, these observed behaviors of the ESG return series, namely weak-form stationarity together with a significant monotonic trend and long-horizon dependence, motivate the adoption of a predictive modeling framework.

Table 5. Diagnostic tests for daily ESG index return dynamics (2014–2024).

Test	Statistic/Setting	p -Value	Conclusion ($\alpha = 0.05$)
ADF	ADF = −50.87	<0.001	Stationary; unit root rejected
Kendall-Tau	tau = 0.029	0.024	Upward trend (significant)
Ljung–Box	Lag 10	0.593	No autocorrelation ($\leq$ lag 10)
Ljung–Box	Lag 20	0.017	Serial dependence (lag 20)
Ljung–Box	Lag 50	0.022	Serial dependence (lag 50)

The baseline feature set consisted of 3 macro-financial indicators (USD/TRY, XAU/USD, Brent) and 5 technical indicators (EMA, RSI, ATR, Williams %R, PSAR). The augmented model additionally included 3 information-theoretic features (PE, SE and KL divergence), yielding a total dimensionality of $p = 11$. Leakage-aware, time-respecting optimization selected EMA = 20, RSI = 21, ATR = 14, Williams %R = 21, while PSAR is non-windowed.

Figure 4 provides a diagnostic overview of ESG daily returns and associated information-theoretic signals, each computed on 14-day rolling windows.

- Panel A shows a weakly stationary yet volatility-clustered return process, consistent with the ADF and Ljung–Box test results discussed in Section 4.1.
- Panel B presents the volatility, where shaded regions denote persistent high-volatility regimes, most prominently during the 2020 COVID-19 shock and the 2022–2023 turbulence period. The dashed line marks the high-volatility threshold, defined as the 75th percentile of the rolling volatility distribution.

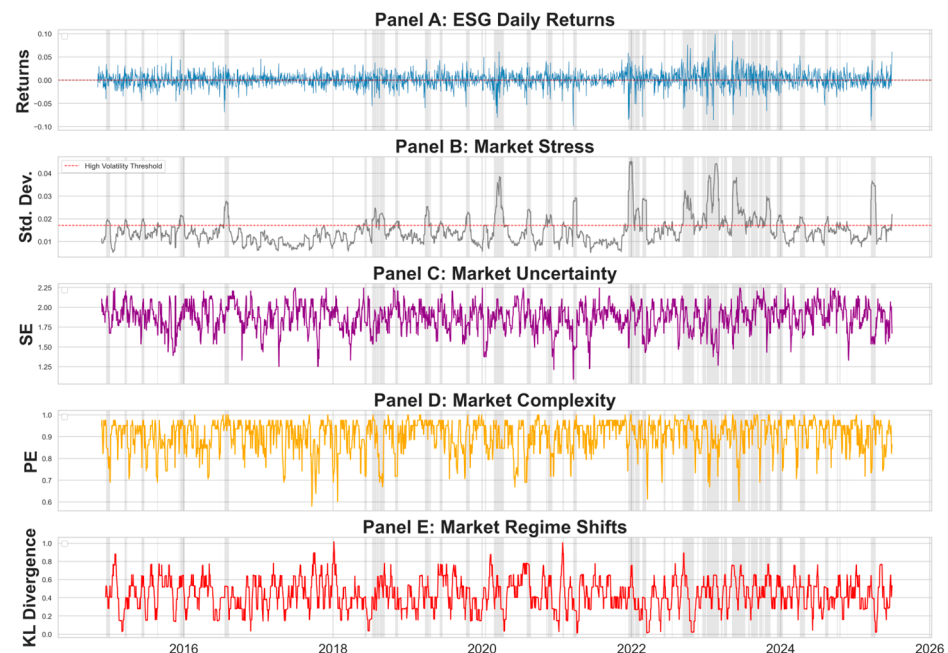


Figure 4. Diagnostic overview of ESG daily returns and information-theoretic signals.

The subsequent panels demonstrate how these information-theoretic signals act as empirical proxies for these market regimes, validating the motivation for their inclusion:

- Panel C reveals that SE tends to decline during and immediately after sharp market drawdowns (e.g., 2020), suggesting a temporary compression of informational diversity and a transition toward more consensus-driven, one-sided trading.
- Panel D shows that PE tends to decline in parallel with SE during high-stress episodes, illustrating its sensitivity to synchronized trading activity. As market stress (Panel B) intensifies, price dynamics appear to simplify and lose ordinal complexity, indicating the emergence of coordinated market movements and herd-driven behavior, conditions that are typically associated with diminished informational diversity and reduced market efficiency.
- Panel E shows that KL values often rise sharply during and immediately after major shocks (e.g., 2020, 2022), indicating its usefulness as a sensitive indicator of market regime transitions. The measure appears to capture both the intensity of structural breaks and the lingering distributional instability that can remain once the market's underlying return-generating structure has been affected by external forces

Taken together, these diagnostics suggest that SE, PE, and KL function not merely as abstract statistical constructs but as empirically grounded indicators of distinct market conditions. SE reflects periods of market consensus or panic, PE captures patterns of herd behavior and reduced efficiency, and KL signals episodes of structural adjustment or external shocks. This interpretation provides a coherent and empirically supported rationale for including these measures in the augmented predictive framework.

4.2. Comparative Performance in Nested Cross-Validation

4.2.1. Overall Performance Summary

The primary results for the baseline and information-theoretic augmented models, obtained under 5-fold time-respecting nested cross-validation with calibrated probabilities, are summarized in Table 6 (mean $\pm$ SD across outer folds). The augmented model attains higher mean BAcc (+0.0175; +2.8% relative), MCC (+0.0196; +7.1% relative), and ROC AUC (+0.0186; +2.7% relative), while the mean F1-score is comparable (0.6648 vs. 0.6646). Importantly, dispersion across folds is lower for the augmented model on every metric, with standard-deviation reductions of approximately 66.6% (F1), 50.1% (BAcc), 57.7% (MCC), and 40.4% (ROC AUC), indicating a more stable out-of-sample profile.

Table 6. NCV results for baseline and augmented specifications with calibrated probabilities.

Metric	Baseline Model	Augmented Model
F1 Score	0.6648 $\pm$ 0.0578	0.6646 $\pm$ 0.0193
BAcc	0.6286 $\pm$ 0.0451	0.6461 $\pm$ 0.0225
MCC	0.2744 $\pm$ 0.1010	0.2940 $\pm$ 0.0427
ROC AUC	0.6957 $\pm$ 0.0574	0.7143 $\pm$ 0.0342

Fold-level distributions (Figure 5) corroborate the tabulated patterns, primarily highlighting the superior stability of the augmented model. Across metrics, the augmented specification yields tighter violins and narrower interquartile ranges, with fewer extreme values. For F1-score, although the means are similar in Table 6 (0.6648 vs. 0.6646), the median is higher for the augmented model (0.6702 vs. 0.6494), and its fold-to-fold spread is markedly smaller. For BAcc and MCC, both the center (e.g., medians around 0.645 vs. 0.631 for BAcc; 0.294 vs. 0.270 for MCC) and the dispersion favor the augmented model. For ROC AUC, the baseline median is slightly higher (0.7266 vs. 0.7167), yet the augmented model achieves a higher mean in Table 6 (0.7143 vs. 0.6957) and substantially lower variability, indicating a more reliable out-of-sample profile. Taken together, Figure 5 and Table 6 indicate that information-theoretic augmentation improves central performance on average and, more importantly, delivers consistently tighter fold-level distributions under identical NCV conditions.

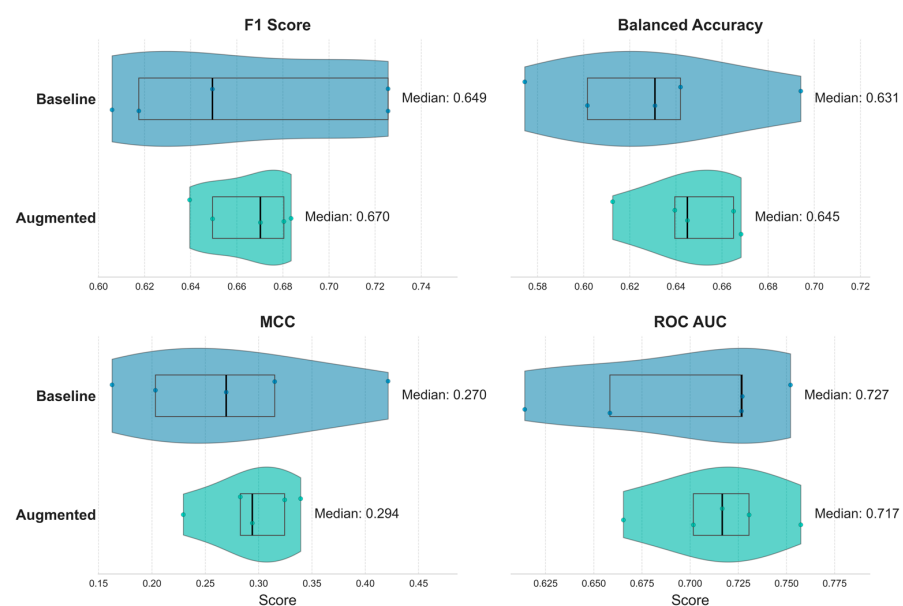


Figure 5. Fold-level raincloud plots of NCV performance metrics.

4.2.2. Statistical Significance of NCV Results

To assess robustness, fold-level differences in calibration quality were examined for Brier score and ECE across the five outer folds (Table 7). For both metrics, the HL estimator of the fold-level difference $\Delta = \text{Aug} - \text{Base}$ is negative, indicating lower error under the augmented specification. The BCa bootstrap confidence intervals for HL do not include zero (Brier: $[-0.02784, -0.00610]$; ECE: $[-0.06678, -0.01868]$), providing evidence consistent with improved calibration. Exact Wilcoxon signed-rank tests (two-sided) yield $p = 0.0625$ for both metrics, offering suggestive evidence at the 10% significance level, which is expected given the small sample (five folds) and the discrete nature of the test statistic. Taken together, these results indicate that, under calibrated predictions within the NCV protocol, the augmented model delivers more reliable probability forecasts across folds, with lower Brier and ECE on average.

Table 7. Fold-level robustness of calibration metrics (NCV, calibrated predictions).

Metric	HL Median Δ (Aug – Base)	90%BCa CI (HL)	Wilcoxon p
Brier	−0.01098	$[-0.02784, -0.00610]$	0.0625
ECE	−0.02797	$[-0.06678, -0.01868]$	0.0625

To assess stability, variability across outer folds was quantified by CV% for each performance metric, and fold-level contrasts ($\Delta\text{CV}\%$) were formed between the augmented and baseline specifications (Table 8). Across all four metrics, $\Delta\text{CV}\%$ was negative with BCa confidence intervals excluding zero, indicating that the augmented model exhibited consistently higher stability (lower fold-to-fold variability).

Table 8. Fold-level stability contrasts ($\Delta\text{CV}\%$) under NCV with calibrated probabilities.

Metric	CV% (Baseline)	CV% (Augmented)	$\Delta\text{CV}\%$	90%BCa CI	Interpretation
F1 Score	8.69	2.91	−5.78	$[-8.22, -4.15]$	Aug more stable
BAcc	7.18	3.48	−3.70	$[-5.46, -0.73]$	Aug more stable
MCC	36.81	14.51	−22.29	$[-31.48, -11.04]$	Aug more stable
ROC AUC	8.24	4.79	−3.45	$[-5.00, -2.22]$	Aug more stable

To complement this variability-focused measure, a performance-to-stability ratio R was also computed (Table 9), capturing performance per unit dispersion. Across all metrics, the augmented specification achieved substantially higher R values, more than doubling relative to the baseline for F1, BAcc, and MCC, confirming that improvements arise not only in average level but also in the reliability of performance across folds.

Table 9. Performance-to-stability ratio R under NCV with calibrated probabilities.

Metric	R (Baseline) [BCa CI]	R (Augmented) 90% [BCa CI]	% Improvement
F1 Score	11.51 [9.97, 12.76]	34.35 [27.26, 40.97]	+198.4%
BAcc	13.93 [9.49, 21.61]	28.72 [21.17, 52.03]	+106.2%
MCC	2.72 [1.88, 3.67]	6.89 [4.80, 13.27]	+153.6%
ROC AUC	12.13 [8.86, 16.27]	20.89 [13.93, 29.64]	+72.2%

Prediction-Level Comparison

To assess calibration at the prediction level, pooled out-of-sample predictions from the five outer folds were analyzed. Using calibrated probabilities paired per observation (within fold), effects were evaluated for both probabilistic accuracies. As shown in Table 10,

the augmented model achieved a lower mean Brier loss ($\Delta = -0.0140$), with the BCa CI $[-0.0199, -0.0084]$ excluding zero; Wilcoxon signed-rank testing ($p = 0.0037$) and a paired sign-flip permutation test ($p < 0.001$) corroborated this improvement. With 10 equal-width bins, ΔECE was -0.0287 and the BCa CI $[-0.0440, -0.0117]$ excluded zero, evidencing a robust reduction in calibration error.

Table 10. Prediction-level calibration contrasts (pooled out-of-sample; calibrated probabilities).

Metric	Mean Δ (Aug – Base)	90%BCa CI	Wilcoxon p	Perm p	Interpretation
Brier	-0.0140	$[-0.0199, -0.0084]$	0.0037	0.0001	Aug better
ECE	-0.0287	$[-0.0440, -0.0117]$	†	†	Aug better

Note: † ECE@10 is a distribution-level scalar that yields a single value per evaluation set; per-observation paired tests (Wilcoxon/sign-flip permutation) are therefore not applicable. Inference relies on the BCa bootstrap confidence interval for the paired difference $\Delta = \text{Aug} - \text{Base}$.

Finally, to compare the final 0/1 predictions, a fold-stratified exact McNemar’s test was conducted. Exact p -values from the five NCV outer folds were combined via Stouffer’s method, yielding a single meta-analytic result $p = 0.0402$. This indicates that the augmented specification achieved a statistically significant and consistent reduction in misclassifications relative to the baseline.

4.3. Definitive Performance on the Held-Out Test Set

This section reports definitive out-of-sample performance on the strictly held-out test set. Both specifications were trained on the full development span with their optimal hyperparameters. Decision thresholds were selected on out-of-fold calibrated probabilities from the development span to maximize F1 and then fixed for the test evaluation (Baseline: 0.070; Augmented: 0.415). The augmented threshold lies closer to a neutral 0.5 cut, consistent with a more balanced operating point under calibrated scores, whereas the baseline required a notably low cut-off to attain its best F1. The final hyperparameters used in the held-out evaluation are reported in Supplementary Table S2.

As shown in Table 11, at these fixed operating points the augmented specification outperformed the baseline on the threshold-dependent classification metrics. The largest gains appeared in imbalance-robust criteria: BAcc improved from 0.5480 to 0.6180 ($\Delta = +0.0700$, +12.8%), and MCC from 0.2080 to 0.2880 ($\Delta = +0.0800$, +38.5%). F1 increased from 0.7060 to 0.7190 ($\Delta = +0.0130$, +1.8%), reflecting a better precision–recall balance at the chosen operating point. ROC AUC, which is threshold-independent, remained essentially unchanged (0.7210 vs. 0.7230; $\Delta = +0.0020$, +0.3%). Calibration measures were broadly comparable across models (Brier ≈ 0.21 ; ECE ≈ 0.07).

Table 11. Definitive Out-of-Sample Performance.

Model	F1	BAcc	ROC-AUC	MCC
XGB-Calib (Baseline)	0.7060	0.5480	0.7210	0.2080
XGB-Calib (Augmented)	0.7190	0.6180	0.7230	0.2880
$\Delta\%$ (Aug – Base)	(+1.8%)	(+12.8%)	(+0.3%)	(+38.5%)

The performance on the held-out test set confirms the patterns observed during cross-validation, supporting the overall robustness of the findings. While calibration quality is broadly comparable across models, the augmented specification demonstrates a clear advantage in discrimination, particularly in imbalance-robust metrics. This performance gain is statistically significant and practically meaningful, as shown by a paired McNemar’s exact test, which considers only discordant days in the out-of-sample period, that is, cases

where the two models disagree, and evaluates whether one is systematically more accurate on the final predictions. The entropy-augmented model was correct on substantially more of these discordant observations ($n_{01} = 58$, $n_{10} = 25$; $p < 0.001$), indicating a higher conditional probability of correctness. Because the test window spans both tranquil and volatile ESG conditions, this asymmetric improvement suggests stronger generalization across changing market regimes rather than gains confined to stable periods.

4.4. Model Interpretability (SHAP Analysis)

To enhance interpretability and assess the directional influence of each predictor on the next-day ESG index movement, a formal interpretability analysis was conducted using SHAP (SHapley Additive exPlanations). This method quantifies the magnitude and direction of each feature's contribution to the final, calibrated model output. Figure 6 presents comparative SHAP summary plots for both the baseline and augmented models, computed on the full development span.

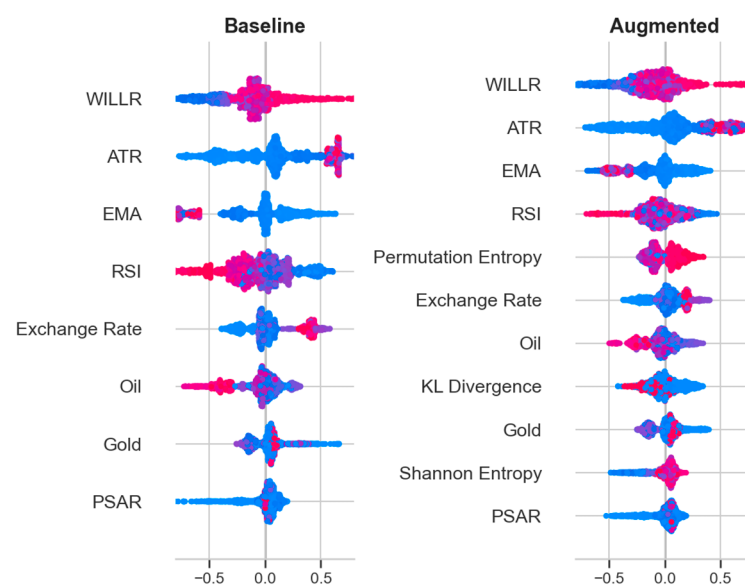


Figure 6. SHAP summary plots for the baseline and augmented models.

The SHAP-based interpretability analysis confirms Williams %R as the most influential predictor, with other technical indicators like ATR, EMA, and RSI also ranking as top contributors. These oscillators jointly capture momentum dynamics and exhibit a clear mean-reversion pattern: high values (overbought) are associated with negative SHAP contributions (predicting down), while low values (oversold) drive positive contributions (predicting up). Volatility measures (ATR) show symmetric, dispersed SHAP distributions, consistent with their role as risk amplifiers rather than direct directional signals. Macroeconomic variables (exchange rate, oil, gold) exhibit moderate, plausible effects; notably, exchange rate depreciation corresponds to downward ESG pressure, aligning with risk-off dynamics in the Turkish market.

Information-theoretic predictors show modest and theoretically consistent SHAP effects. Higher entropy values, indicating greater uncertainty, reduce the model's confidence in an up forecast, while larger KL divergences capture structural regime shifts. Despite their lower marginal contributions, these information-theoretic variables enhance calibration and temporal stability, aligning with their theoretical role as information-state descriptors rather than direct return predictors.

4.5. Model Sensitivity (Entropy Window Parameter)

To test the framework's temporal sensitivity, an entropy-window experiment was conducted under identical nested cross-validation and calibration protocols, varying the window length to capture different information regimes. The 7-day and 21-day configurations represent lower-window sensitivity and upper-window stability regimes, respectively. The former captures short-term, high-frequency dynamics with greater noise exposure, whereas the latter reflects more smoothed and stable structural patterns.

The NCV results for both entropy-window configurations (Supplementary Table S3) show that the entropy-augmented models consistently outperformed or matched their baseline counterparts under calibrated conditions, generally exhibiting reduced fold-to-fold variance and confirming the temporal robustness of the proposed framework.

Complementary robustness diagnostics (Tables S4 and S5) summarize fold-level calibration contrasts (Brier and ECE), stability measures (ΔCV %), and prediction-level calibration results. Across both alternative configurations ($T = 7$ and $T = 21$), the entropy-augmented models tended to achieve lower calibration errors and reduced fold-to-fold variability, particularly pronounced for the 21-day window, suggesting that the observed gains are systematic rather than incidental to the chosen window length.

The held-out test results further confirmed this pattern: the augmented model achieved higher BAcc under both settings (0.564 vs. 0.546 for $T = 7$ and 0.606 vs. 0.538 for $T = 21$), and McNemar's exact tests ($p = 0.0225$ and $p = 0.0032$, respectively) indicated significant differences in directional classification favoring the augmented specification.

5. Discussion

Navigating the non-linear and regime-dependent dynamics of ESG markets poses a persistent challenge, particularly in emerging economies where volatility, liquidity constraints, and abrupt policy shifts are commonplace. This study introduces a complexity-aware forecasting framework that integrates information-theoretic diagnostics into a hybrid machine-learning design, computed in a strictly time-respecting (leakage-aware) manner. Applied to Türkiye's BIST Sustainability Index as a natural stress test, the framework demonstrates that explicitly modeling the market's informational state and transitions leads to more stable, better-calibrated, and more reliable forecasts under shifting market regimes. These results highlight the practical importance of capturing informational structure and regime dynamics, elements that are often overlooked by conventional macro-technical predictors, and they provide a solid foundation for the interpretive, theoretical, and applied insights presented in the following sections, as well as for the study's concluding discussion of limitations and future research.

5.1. Interpretation of Findings

The augmented model's advantage stems from its ability to represent the ESG market's informational state and its transitions, aspects that conventional macro-technical predictors do not explicitly encode. Entropy- and divergence-based descriptors (SE, PE, and KL) condense key regime characteristics: uncertainty compression, ordinal simplification, and distributional shift. In practical terms, this means that the model perceives not only recent price movement but also how information itself is organized and transmitted through the market. By receiving signals aligned with regime change rather than raw volatility, the learner stabilizes its internal decision scale across heterogeneous conditions, which naturally enhances calibration and generalization.

From an economic perspective, this mechanism indicates that markets characterized by higher informational diversity (i.e., higher entropy) tend to exhibit more adaptive and resilient dynamics, as diverse information sources and heterogeneous expectations foster stability

through decentralized adjustment. Conversely, periods of reduced or compressed entropy reflect phases in which market participants act upon similar information, leading to diminished informational heterogeneity and potential inefficiency. In this context, incorporating information-theoretic diagnostics into the modeling framework enables the learner to capture transitions from information-rich and stable regimes to stressed and synchronized ones, where predictive uncertainty becomes nonlinear. This ability allows the model to adapt to changing market conditions more realistically, improving both interpretability and calibration.

This interpretation is empirically supported by both the NCV and the held-out test results. Across validation folds, the augmented model achieved not only higher predictive accuracy but also markedly greater stability; its performance-to-stability ratio (R) more than doubled for key imbalance-robust criteria, indicating reliability gains far beyond simple level effects. Furthermore, the framework achieved statistically significant gains in probabilistic quality, producing better-calibrated forecasts at both the fold and prediction levels. This superiority was decisively confirmed on the strictly held-out test set, where the informational diagnostics led to a statistically significant reduction in misclassifications and concentrated its largest gains in the same imbalance-robust criteria ($B_{Acc} \approx +12.8\%$, $MCC \approx +38.5\%$). Taken together, the evidence substantiates the central claim that incorporating information-theoretic diagnostics enhances not only point discrimination but also the stability, calibration, and interpretability of ESG index forecasts under heterogeneous and regime-dependent conditions.

5.2. Theoretical Implications

Theoretically, this study contributes to bridging the conceptual gap between information theory and sustainable finance by showing that entropy- and divergence-based measures represent a distinct and meaningful source of predictive information in ESG markets. Traditional macro-financial and technical indicators primarily reflect observable price dynamics but provide limited insight into the informational mechanisms that generate those dynamics. By conceptualizing ESG indices as complex adaptive systems in which agents interact under evolving beliefs, policy shocks, and heterogeneous information flows, the framework offers a reinterpretation of market predictability. It advances the perspective that predictability arises not solely from historical price patterns but from the market's underlying informational diversity and structural adaptability.

Within this theoretical framework, the non-linear dynamics of ESG markets are not statistical noise to be filtered out but quantifiable reflections of the system's adaptive state, challenging the assumptions of perfect informational efficiency. These information-theoretic diagnostics transform the abstract notion of market adaptability into measurable constructs that capture how information is organized, transmitted, and restructured over time. SE serves as a direct proxy for the market's informational diversity and resilience: a market with high SE reflecting a broad distribution of beliefs, strategies, and expectations is typically more adaptive and resistant to shocks. Conversely, periods of declining SE and PE reveal a contraction in informational variety, often associated with consensus-driven or herd-like behavior. Such states of compressed entropy and high synchronization are informationally fragile and prone to the structural realignments that KL divergence is designed to detect. Collectively, SE, PE, and KL provide a coherent theoretical bridge linking market complexity, behavioral coordination, and structural change within sustainable finance systems.

Extending beyond the empirical context, entropy as a measure of informational richness represents a universal principle of complex adaptive systems rather than a market-specific artefact. Its theoretical foundation in information theory underpins its applicability across diverse financial environments. Nonetheless, the expression of these informational

dynamics is shaped by contextual factors such as volatility, liquidity, and institutional structure that influence how information is assimilated and reorganized. In this sense, the emerging market context represented by the Turkish ESG index serves as a natural stress test. Its regime-shifting, high-volatility environment amplifies the informational signals that entropy metrics are designed to detect. The augmented model's superior performance in this setting provides strong empirical evidence that these signals become most critical precisely when market complexity is high and conventional models falter. This is demonstrated not only by its asymmetric advantage in correctly classifying discordant predictions but also by its superior stability and calibration. This robustness under stress supports their generalizability and underscores their relevance for understanding market complexity in both emerging and developed economies.

5.3. Practical Implications

From a practical perspective, the proposed framework provides an implementable methodology for investors and policymakers, particularly in emerging ESG markets where volatility, policy shifts, and informational asymmetries are pervasive. Its practical utility lies in delivering forecasts that are not only demonstrably more stable across regimes but also better calibrated. This emphasis on calibration is particularly crucial; the quality of probabilistic predictions is a fundamental determinant of their reliability [119], and achieving accurate calibration remains a known challenge for modern, high-capacity classifiers like the one used in this study [120]. Beyond these forecasting improvements, the underlying information-theoretic diagnostics also serve as a complementary analytical layer. These measures help to illuminate the market's underlying informational dynamics (e.g., shifts in consensus, herd-like coordination, or structural breaks), thereby offering a perspective that conventional indicators alone do not fully capture.

For investors and risk managers, the practical utility of this framework is empirically supported by its superior performance across both the NCV and the strictly held-out test set. The observed improvement, reflected in greater stability, better calibration, and fewer misclassifications on imbalance-robust metrics, results from the model's ability to interpret underlying informational signals, providing a deeper contextual understanding than price-based indicators alone. Accordingly, investors and risk managers can use these diagnostics as interpretable, non-price-based risk indicators. A decline in sequential complexity (PE) may serve as an early warning of herd-like market dynamics, while an increase in distributional divergence (KL) indicates a structural break that may require recalibration of risk models dependent on recent historical data. Integrating these insights, either directly through diagnostic dashboards or indirectly through the model's calibrated probabilistic outputs, supports more adaptive asset allocation, regime-aware hedging, and proactive risk governance.

For policymakers and regulators, the framework provides a quantitative tool for monitoring market integrity and evaluating the systemic effects of new ESG-related interventions. Because the diagnostics capture how informational diversity and structural dynamics evolve over time, they can be used to assess how policy actions, disclosure mandates, or regulatory adjustments influence market stability and efficiency. A sustained decline in SE may signal a contraction in informational diversity, reflecting lower transparency or growing consensus among market participants conditions that can precede reduced resilience or potential coordination risk. Likewise, persistently low PE can indicate synchronized trading behavior and diminished informational efficiency, warranting closer supervisory attention. In contrast, a sharp rise in KL divergence may reveal that a policy announcement or external event has materially altered the return-generating structure, signaling the onset of a regime transition. By integrating these information-theoretic di-

agnostics into ongoing monitoring and evaluation frameworks, policymakers can better identify early signs of stress, design more adaptive regulatory responses, and promote the long-term transparency and resilience of ESG financial systems.

5.4. Limitations and Future Research

Despite its robust design and compelling findings, this study has several limitations that motivate future work. First, the analysis was confined to a single emerging-market ESG index within the observed study period. While this provided an effective stress test for the types of volatility and regime shifts present in the sample (including COVID-19 and post-2022 turbulence), future research should validate the framework's generalizability across a broader portfolio of ESG indices. This would also test the model's robustness against systemic shocks or liquidity disruptions not observed within the current data window. Second, the design focused on three information-theoretic signals (SE, PE, KL) used as features; expanding to alternative complexity measures (e.g., Tsallis entropy, multifractal descriptors) may yield additional diagnostic or predictive value. Third, the macro-financial and technical baseline was intentionally parsimonious. This design limits a broader assessment of the model's sensitivity to other categories of macro shocks (e.g., sudden interest rate hikes or major geopolitical events) that are only partially captured within the study period. Evaluating richer factor sets and optimized technical indicators would provide a stronger comparator and may uncover interactions with the information-theoretic layer. Fourth, the empirical strategy relied on a calibrated XGBoost classifier, which is widely regarded as the state-of-the-art approach for tabular learning [121,122]. Future research could extend this framework by embedding the proposed entropy-based signals into sequence-oriented architectures such as LSTMs, Temporal Convolutional Networks, or Transformer-based models to examine whether long-range temporal dependencies provide additional predictive value beyond the lag-structured representation adopted here. Finally, extending the targets beyond direction to return magnitude and volatility would further test decision relevance.

6. Conclusions

This study demonstrates that a complexity-aware, leakage-aware forecasting framework that embeds information-theoretic diagnostics can materially improve the reliability of next-day ESG index direction predictions in an emerging-market setting. By augmenting a macro-technical baseline with SE, PE, and KL divergence, the framework delivers forecasts that are not only more accurate but also demonstrably more stable and better calibrated, highlighting the value of making informational states and transitions explicit in regime-dependent markets. Empirical evidence from NCV and held-out testing supports this conceptual argument.

The augmented model exhibited superior discrimination on the strictly held-out test set, achieving a +38.5% gain in the MCC and a +12.8% improvement in BAcc. Beyond point discrimination, model reliability and calibration were both enhanced: nested cross-validation analyses showed that the performance-to-stability ratio more than doubled (>100%) across key metrics, while Brier losses decreased significantly ($\Delta = -0.014$). The improvement was further supported by a fold-stratified McNemar's exact test ($p < 0.001$), indicating consistently fewer misclassifications. Collectively, these findings confirm that incorporating explicit informational diagnostics strengthens not only predictive accuracy but also the temporal stability and probabilistic reliability of ESG index forecasts under varying market regimes. Taken together, these results demonstrate a generalizable and practically relevant modeling principle.

Beyond its immediate application to Türkiye's BIST Sustainability Index, the approach suggests a portable and potentially generalizable blueprint for regime-sensitive prediction

in sustainable finance, advancing methodological depth while maintaining applied relevance. Future applications may extend this framework to other national or global ESG indices and integrate it into real-world trading or risk-monitoring systems that require regime-aware and probabilistically calibrated decision support. In doing so, the proposed approach provides a scalable foundation for linking theoretical advances in information theory with the practical demands of sustainable investment and financial stability analysis. In an era when sustainable finance is rapidly gaining importance yet remains fraught with volatility and uncertainty, this framework provides a principled pathway for navigating one of modern finance's most complex and consequential frontiers.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/e27111164/s1>, Table S1. Hyperparameter search space for the Time-Respecting Calibrated XGBoost model; Table S2. Final Hyperparameters; Table S3. NCV results for baseline and augmented specifications with calibrated probabilities (7-day and 21-day entropy windows); Table S4. Summary of fold- and prediction-level robustness analyses (7-day entropy window); Table S5. Summary of fold- and prediction-level robustness analyses (21-day entropy window).

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

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

Acknowledgments: 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.

Appendix A. Information-Theoretic Feature Definitions

The formal definitions of Shannon entropy (SE), permutation entropy (PE), and Kullback–Leibler (KL) divergence used in this study are given below:

Shannon Entropy (SE):

$$H_{SE} = -\sum_{i=1}^K p_i \ln(p_i)$$

Permutation Entropy (PE):

$$H_{PE} = -\frac{1}{\log_2 d!} \sum_{j=1}^{d!} p(\pi_j) \log_2(p(\pi_j))$$

Kullback–Leibler Divergence (KL):

$$D_{KL}(P \parallel Q) = \sum_{i=1}^K p_i \ln\left(\frac{p_i}{q_i}\right)$$

Parameters used in this study:

- Number of bins $K = 10$;

- Numerical offset $\varepsilon = 10^{-12}$ (for stability);
 - Embedding dimension $d = 3$;
- Time lag $\tau = 1$.

Appendix B. Time-Respecting Calibration Protocol

Let the development span (training+validation) contain indices $t = 1, \dots, T_{dev}$, and let $\alpha \in (0, 1)$ denotes the calibration fraction of this span (α). The cut-off point is defined as $t_c = \lfloor (1 - \alpha)T_{dev} \rfloor$. The calibration protocol proceeded as:

1. Fit the base XGBoost $f \{ (X_t, y_t) \}_{t=1}^{t_c}$.
2. Score the later, disjoint calibration holdout $\{ (X_t, y_t) \}_{t=t_c+1}^{T_{dev}} : s_t = f(X_t)$ (uncalibrated margins/probabilities as implemented).
3. Fit a mapping g on $\{ (s_t, y_t) \}_{t=t_c+1}^{T_{dev}}$ and apply it forward, yielding calibrated probabilities $\hat{p}_t = g(f(X_t))$.

Appendix C. Evaluation Metrics

Directional-accuracy metrics:

$$BAcc = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

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

Discrimination metric:

$$AUC = \frac{1}{n_1 n_0} \sum_{i: y_i=1} \sum_{j: y_j=0} (1\{\hat{s}_i > \hat{s}_j\} + \frac{1}{2} 1\{\hat{s}_i = \hat{s}_j\})$$

References

1. Engle, R.F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* **1982**, *50*, 987–1007. [CrossRef]
2. Bollerslev, T.P. Generalized Autoregressive Conditional Heteroskedasticity. *J. Econom.* **1986**, *31*, 307–327. [CrossRef]
3. Hamilton, J.D. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica* **1989**, *57*, 357–384. [CrossRef]
4. Cont, R. Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues. *Quant. Financ.* **2001**, *1*, 223–236. [CrossRef]
5. Global Sustainable Investment Alliance (GSIA). *Global Sustainable Investment Review 2022*; GSIA: London, UK, 2023; Available online: <https://www.gsi-alliance.org/members-resources/gsir2022/> (accessed on 12 September 2025).
6. Tucker, J.J., III; Jones, S. Environmental, Social, and Governance Investing: Investor Demand, the Great Wealth Transfer, and Strategies for ESG Investing. *J. Financ. Serv. Prof.* **2020**, *74*, 56.
7. Morningstar. *Global Sustainable Fund Flows Quarterly Data*; Morningstar: Chicago, IL, USA, 2025; Available online: <https://www.morningstar.com/business/insights/blog/funds/global-sustainable-fund-flows-quarterly-data> (accessed on 12 September 2025).
8. MSCI. *Sustainability and Climate Trends to Watch for 2025*; MSCI ESG Research LLC: New York, NY, USA, 2025; Available online: <https://www.msci.com/documents/1296102/51277550/2025%2BSustainability%2Band%2BClimate%2BTrends%2BPaper.pdf> (accessed on 12 September 2025).
9. Vo, N.N.; He, X.; Liu, S.; Xu, G. Deep Learning for Decision Making and the Optimization of Socially Responsible Investments and Portfolio. *Decis. Support Syst.* **2019**, *124*, 113097. [CrossRef]

10. Xidonas, P.; Essner, E. On ESG Portfolio Construction: A Multi-Objective Optimization Approach. *Comput. Econ.* **2024**, *63*, 21–45. [CrossRef] [PubMed]
11. De Lucia, C.; Pazienza, P.; Bartlett, M. Does Good ESG Lead to Better Financial Performances by Firms? Machine Learning and Logistic Regression Models of Public Enterprises in Europe. *Sustainability* **2020**, *12*, 5317. [CrossRef]
12. Sorathiya, A.; Saval, P.; Sorathiya, M. Data-Driven Sustainable Investment Strategies: Integrating ESG, Financial Data Science, and Time Series Analysis for Alpha Generation. *Int. J. Financ. Stud.* **2024**, *12*, 36. [CrossRef]
13. Sabbaghi, O. The Impact of News on the Volatility of ESG Firms. *Glob. Financ. J.* **2022**, *51*, 100570. [CrossRef]
14. Suprihadi, E.; Danila, N. Forecasting ESG Stock Indices Using a Machine Learning Approach. *Glob. Bus. Rev.* **2024**, 09721509241234033. [CrossRef]
15. Bhandari, H.N.; Pokhrel, N.R.; Rimal, R.; Dahal, K.R.; Rimal, B. Implementation of Deep Learning Models in Predicting ESG Index Volatility. *Financ. Innov.* **2024**, *10*, 75. [CrossRef]
16. Gunduz, H.; Cataltepe, Z. Borsa Istanbul (BIST) Daily Prediction Using Financial News and Balanced Feature Selection. *Expert Syst. Appl.* **2015**, *42*, 9001–9011. [CrossRef]
17. Sadorsky, P. Modeling Volatility and Conditional Correlations between Socially Responsible Investments, Gold and Oil. *Econ. Model.* **2014**, *38*, 609–618. [CrossRef]
18. De Oliveira, E.M.; Cunha, F.A.F.S.; Cyrino Oliveira, F.L.; Samanez, C.P. Dynamic Relationships between Crude Oil Prices and Socially Responsible Investing in Brazil: Evidence for Linear and Non-Linear Causality. *Appl. Econ.* **2017**, *49*, 2125–2140. [CrossRef]
19. Maraqa, B.; Bein, M. Dynamic Interrelationship and Volatility Spillover among Sustainability Stock Markets, Major European Conventional Indices, and International Crude Oil. *Sustainability* **2020**, *12*, 3908. [CrossRef]
20. Guo, T.; Jamet, N.; Betrix, V.; Piquet, L.A.; Hauptmann, E. ESG2Risk: A Deep Learning Framework from ESG News to Stock Volatility Prediction. *arXiv* **2020**, arXiv:2005.02527. [CrossRef]
21. Zunino, L.; Zanin, M.; Tabak, B.M.; Pérez, D.G.; Rosso, O.A. Forbidden Patterns, Permutation Entropy and Stock Market Inefficiency. *Physica A* **2009**, *388*, 2854–2864. [CrossRef]
22. Bariviera, A.F.; Guercio, M.B.; Martinez, L.B.; Rosso, O.A. A Permutation Information Theory Tour through Different Interest Rate Maturities: The Libor Case. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2015**, *373*, 20150119. [CrossRef]
23. World Commission on Environment and Development. *Our Common Future: Report of the World Commission on Environment and Development*; Oxford University Press: Oxford, UK, 1987.
24. Cassen, R.H. Our Common Future: Report of the World Commission on Environment and Development. *Int. Aff.* **1987**, *64*, 126. [CrossRef]
25. Harris, J.M. Sustainability and Sustainable Development. *Int. Soc. Ecol. Econ.* **2003**, *1*, 1–12. Available online: <https://isecoeco.org/pdf/susdev.pdf> (accessed on 15 July 2025).
26. S&P Dow Jones Indices. *S&P ESG Indices Methodology*; S&P Global: New York, NY, USA, 2024; Available online: <https://www.spglobal.com/spdji/en/documents/methodologies/methodology-sp-esg%2B-indices.pdf> (accessed on 14 September 2025).
27. Robiyanto, R.; Nugroho, B.A.; Huruta, A.D.; Frensidy, B.; Suyanto, S. Identifying the Role of Gold on Sustainable Investment in Indonesia: The DCC-GARCH Approach. *Economies* **2021**, *9*, 119. [CrossRef]
28. Sahoo, S. Harmony in Diversity: Exploring Connectedness and Portfolio Strategies among Crude Oil, Gold, Traditional and Sustainable Index. *Resour. Policy* **2024**, *97*, 103222. [CrossRef]
29. Shaikh, I. On the Relationship between Policy Uncertainty and Sustainable Investing. *J. Model. Manag.* **2022**, *17*, 1504–1523. [CrossRef]
30. Özçim, H. BİST Sürdürülebilirlik Endeksi ve Makroekonomik Veriler Arasındaki İlişkinin GARCH Modelleri Çerçevesinde İncelenmesi. *Pamukkale Univ. J. Soc. Sci. Inst.* **2022**, *49*, 115–126.
31. Kaya, M. BİST Sürdürülebilirlik Endeksi ile Fosil Yakıt Fiyatları Arasındaki İlişkinin Analizi. *Abant Soc. Sci. J.* **2023**, *23*, 1475–1495.
32. Umar, Z.; Abrar, A.; Zaremba, A.; Teplova, T.; Vo, X.V. Network connectedness of environmental attention—Green and dirty assets. *Financ. Res. Lett.* **2022**, *50*, 103209. [CrossRef]
33. Umar, Z.; Gubareva, M. The relationship between the COVID-19 media coverage and the environmental, social and governance leaders equity volatility: A time–frequency wavelet analysis. *Appl. Econ.* **2021**, *53*, 3193–3206. [CrossRef]
34. Akhtaruzzaman, M.; Boubaker, S.; Umar, Z. COVID-19 media coverage and ESG leader indices. *Financ. Res. Lett.* **2022**, *45*, 102170. [CrossRef]
35. Pina, V.; Bachiller, P.; Ripoll, L. Testing the reliability of financial sustainability: The case of Spanish local governments. *Sustainability* **2020**, *12*, 6880. [CrossRef]
36. Rodríguez Bolívar, M.P.; López Subires, M.D.; Alcaide Muñoz, L.; Navarro Galera, A. The financial sustainability of local authorities in England and Spain: A comparative empirical study. *Int. Rev. Adm. Sci.* **2021**, *87*, 97–114. [CrossRef]

37. Santis, S. The demographic and economic determinants of financial sustainability: An analysis of Italian local governments. *Sustainability* **2020**, *12*, 7599. [\[CrossRef\]](#)
38. Benito, B.; Guillamón, M.D.; Ríos, A.M. The sustainable development goals: How does their implementation affect the financial sustainability of the largest Spanish municipalities? *Sustain. Dev.* **2023**, *31*, 2836–2850. [\[CrossRef\]](#)
39. Alshubiri, F.N. Analysis of financial sustainability indicators of higher education institutions on foreign direct investment: Empirical evidence in OECD countries. *Int. J. Sustain. High. Educ.* **2021**, *22*, 77–99. [\[CrossRef\]](#)
40. Bui, T.D.; Nguyen, T.P.T.; Sethanan, K.; Chiu, A.S.; Tseng, M.L. Natural resource management in Vietnam: Merging circular economy practices and financial sustainability approach. *J. Clean. Prod.* **2024**, *480*, 144094. [\[CrossRef\]](#)
41. Yi, H.; Kim, K.N. Transforming aid-funded renewable energy systems: A case study of policy-driven financial sustainability in rural Bangladesh. *Renew. Energy* **2025**, *246*, 122752. [\[CrossRef\]](#)
42. Maenuddin, M.K.; Hamid, S.A.; Nassir, A.M.; Fahlevi, M.; Aljuaid, M.; Jermisittiparsert, K. Measuring the financial sustainability and its influential factors in microfinance sector of Pakistan. *SAGE Open* **2024**, *14*, 21582440241259288. [\[CrossRef\]](#)
43. Awaworyi Churchill, S. Microfinance financial sustainability and outreach: Is there a trade-off? *Empir. Econ.* **2020**, *59*, 1329–1350. [\[CrossRef\]](#)
44. Atichasari, A.S.; Ratnasari, A.; Kulsum, U.; Kahpi, H.S.; Wulandari, S.S.; Marfu, A. Examining non-performing loans on corporate financial sustainability: Evidence from Indonesia. *Sustain. Futures* **2023**, *6*, 100137. [\[CrossRef\]](#)
45. Zabolotnyy, S.; Wasilewski, M. The concept of financial sustainability measurement: A case of food companies from Northern Europe. *Sustainability* **2019**, *11*, 5139. [\[CrossRef\]](#)
46. Githaiga, P.N. Revenue diversification and financial sustainability of microfinance institutions. *Asian J. Account. Res.* **2022**, *7*, 31–43. [\[CrossRef\]](#)
47. Najam, H.; Abbas, J.; Alvarez-Otero, S.; Dogan, E.; Sial, M.S. Towards green recovery: Can banks achieve financial sustainability through income diversification in ASEAN countries? *Econ. Anal. Policy* **2022**, *76*, 522–533. [\[CrossRef\]](#)
48. Kong, Y.; Donkor, M.; Musah, M.; Nkyi, J.A.; Ampomg, G.O.A. Capital structure and corporates financial sustainability: Evidence from listed non-financial entities in Ghana. *Sustainability* **2023**, *15*, 4211. [\[CrossRef\]](#)
49. Githaiga, P.N.; Soi, N.; Buigut, K.K. Does Intellectual Capital Matter to MFI's Financial Sustainability? *Asian J. Account. Res.* **2023**, *8*, 41–52. [\[CrossRef\]](#)
50. Gleißner, W.; Günther, T.; Walkshäusl, C. Financial Sustainability: Measurement and Empirical Evidence. *J. Bus. Econ.* **2022**, *92*, 467–516. [\[CrossRef\]](#)
51. Berg, F.; Lo, A.W.; Rigobon, R.; Singh, M.; Zhang, R. Quantifying the returns of ESG investing: An empirical analysis with six ESG metrics. MIT Sloan Research Paper No. 6930-23. *SSRN* **2023**. [\[CrossRef\]](#)
52. Tao, J.; Shan, P.; Liang, J.; Zhang, L. Influence Mechanism between Corporate Social Responsibility and Financial Sustainability: Empirical Evidence from China. *Sustainability* **2024**, *16*, 2406. [\[CrossRef\]](#)
53. Wong, S.M.H.; Chan, R.Y.K.; Wong, P.; Wong, T. Promoting Corporate Financial Sustainability through ESG Practices: An Employee-Centric Perspective and the Moderating Role of Asian Values. *Res. Int. Bus. Financ.* **2025**, *75*, 102733. [\[CrossRef\]](#)
54. Ur Rahman, R.; Ali Shah, S.M.; El-Gohary, H.; Abbas, M.; Haider Khalil, S.; Al Altheeb, S.; Sultan, F. Social Media Adoption and Financial Sustainability: Learned Lessons from Developing Countries. *Sustainability* **2020**, *12*, 10616. [\[CrossRef\]](#)
55. Muneer, S.; Singh, A.; Choudhary, M.H.; Alshammari, A.S.; Butt, N.A. Does Environmental Disclosure and Corporate Governance Ensure the Financial Sustainability of Islamic Banks? *Adm. Sci.* **2025**, *15*, 54. [\[CrossRef\]](#)
56. Nabipour, M.; Nayyeri, P.; Jabani, H.; Shahab, S.; Mosavi, A. Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data: A comparative analysis. *IEEE Access* **2020**, *8*, 150199–150212. [\[CrossRef\]](#)
57. Bonello, J.; Brédart, X.; Vella, V. Machine learning models for predicting financial distress. *J. Res. Econ.* **2018**, *2*, 174–185. [\[CrossRef\]](#)
58. Molina-Gómez, N.I.; Rodríguez-Rojas, K.; Calderón-Rivera, D.; Díaz-Arévalo, J.L.; López-Jiménez, P.A. Using machine learning tools to classify sustainability levels in the development of urban ecosystems. *Sustainability* **2020**, *12*, 3326. [\[CrossRef\]](#)
59. Ting, T.; Mia, M.A.; Hossain, M.I.; Wah, K.K. Predicting the financial performance of microfinance institutions with machine learning techniques. *J. Model. Manag.* **2025**, *20*, 322–347. [\[CrossRef\]](#)
60. Shi, Y.; Charles, V.; Zhu, J. Bank financial sustainability evaluation: Data envelopment analysis with random forest and Shapley additive explanations. *Eur. J. Oper. Res.* **2025**, *321*, 614–630. [\[CrossRef\]](#)
61. Lee, O.; Joo, H.; Choi, H.; Cheon, M. Proposing an integrated approach to analyzing ESG data via machine learning and deep learning algorithms. *Sustainability* **2022**, *14*, 8745. [\[CrossRef\]](#)
62. Raman, N.; Bang, G.; Nourbakhsh, A. Mapping ESG trends by distant supervision of neural language models. *Mach. Learn. Knowl. Extract.* **2020**, *2*, 453–468. [\[CrossRef\]](#)
63. Lin, S.-L.; Jin, X. Does ESG predict systemic banking crises? A computational economics model of early warning systems with interpretable multi-variable LSTM based on mixture attention. *Mathematics* **2023**, *11*, 410. [\[CrossRef\]](#)
64. Maasoumi, E. A compendium to information theory in economics and econometrics. *Econom. Rev.* **1993**, *12*, 137–181. [\[CrossRef\]](#)
65. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [\[CrossRef\]](#)

66. Zhou, R.; Cai, R.; Tong, G. Applications of entropy in finance: A review. *Entropy* **2013**, *15*, 4909–4931. [\[CrossRef\]](#)
67. Shternshis, A.; Mazzarisi, P.; Marmi, S. Measuring Market Efficiency: The Shannon Entropy of High-Frequency Financial Time Series. *Chaos Solitons Fractals* **2022**, *162*, 112403. [\[CrossRef\]](#)
68. Scrucca, L. Entropy-Based Volatility Analysis of Financial Log-Returns Using Gaussian Mixture Models. *Entropy* **2024**, *26*, 907. [\[CrossRef\]](#)
69. Bandt, C.; Pompe, B. Permutation entropy: A natural complexity measure for time series. *Phys. Rev. Lett.* **2002**, *88*, 174102. [\[CrossRef\]](#)
70. Huang, X.; Shang, H.L.; Pitt, D. Permutation Entropy and Its Variants for Measuring Temporal Dependence. *Aust. N. Z. J. Stat.* **2022**, *64*, 442–477. [\[CrossRef\]](#)
71. Olbryś, J. Entropy of Volatility Changes: Novel Method for Assessment of Regularity in Volatility Time Series. *Entropy* **2025**, *27*, 318. [\[CrossRef\]](#) [\[PubMed\]](#)
72. Hou, Y.; Liu, F.; Gao, J.; Cheng, C.; Song, C. Characterizing Complexity Changes in Chinese Stock Markets by Permutation Entropy. *Entropy* **2017**, *19*, 514. [\[CrossRef\]](#)
73. Henry, M.; Judge, G. Permutation Entropy and Information Recovery in Nonlinear Dynamic Economic Time Series. *Econometrics* **2019**, *7*, 10. [\[CrossRef\]](#)
74. Siokis, F. High Short Interest Stocks Performance during the COVID-19 Crisis: An Informational Efficacy Measure Based on Permutation-Entropy Approach. *J. Econ. Stud.* **2023**, *50*, 1570–1584. [\[CrossRef\]](#)
75. Fan, Y.; Yang, Y.; Wang, Z.; Gao, M. Instability of Financial Time Series Revealed by Irreversibility Analysis. *Entropy* **2025**, *27*, 402. [\[CrossRef\]](#)
76. Maasoumi, E.; Racine, J. Entropy and predictability of stock market returns. *J. Econom.* **2002**, *107*, 291–312. [\[CrossRef\]](#)
77. Martín, M.T.; Plastino, A.; Rosso, O.A. Generalized statistical complexity measures: Geometrical and analytical properties. *Physica A* **2006**, *369*, 439–462. [\[CrossRef\]](#)
78. Li, J.; Shang, P. Time Irreversibility of Financial Time Series Based on Higher Moments and Multiscale Kullback–Leibler Divergence. *Physica A* **2018**, *502*, 248–255. [\[CrossRef\]](#)
79. Ishizaki, R.; Inoue, M. Short-Term Kullback–Leibler Divergence Analysis to Extract Unstable Periods in Financial Time Series. *Evol. Inst. Econ. Rev.* **2024**, *21*, 227–236. [\[CrossRef\]](#)
80. Ponta, L.; Carbone, A. Kullback–Leibler Cluster Entropy to Quantify Volatility Correlation and Risk Diversity. *Phys. Rev. E* **2025**, *111*, 014311. [\[CrossRef\]](#)
81. Zunino, L.; Tabak, B.M.; Pérez, D.G.; Garavaglia, M.; Rosso, O.A. Inefficiency in Latin-American market indices. *Europhys. Lett.* **2007**, *84*, 60008. [\[CrossRef\]](#)
82. Fernández Bariviera, A.; Zunino, L.; Guercio, M.B.; Martinez, L.B.; Rosso, O.A. Revisiting the European sovereign bonds with a permutation-information-theory approach. *Eur. Phys. J. B* **2013**, *86*, 509. [\[CrossRef\]](#)
83. Eichengreen, B.; Rose, A.K.; Wyplosz, C. Exchange market mayhem: The antecedents and aftermath of speculative attacks. *Econ. Policy* **1995**, *10*, 249–312. [\[CrossRef\]](#)
84. Baur, D.G.; McDermott, T.K. Is gold a safe haven? International evidence. *J. Bank. Financ.* **2010**, *34*, 1886–1898. [\[CrossRef\]](#)
85. Hamilton, J.D. Causes and consequences of the oil shock of 2007–08. *Natl. Bur. Econ. Res.* **2009**, *40*, 215–283.
86. Yin, L.; Yang, Q. Predicting the oil prices: Do technical indicators help? *Energy Econ.* **2016**, *56*, 338–350. [\[CrossRef\]](#)
87. Dai, Z.; Zhu, H.; Kang, J. New technical indicators and stock returns predictability. *Int. Rev. Econ. Financ.* **2021**, *71*, 127–142. [\[CrossRef\]](#)
88. Huang, J.Z.; Huang, W.; Ni, J. Predicting bitcoin returns using high-dimensional technical indicators. *J. Financ. Data Sci.* **2019**, *5*, 140–155. [\[CrossRef\]](#)
89. Padhi, D.K.; Padhy, N.; Bhoi, A.K.; Shafi, J.; Ijaz, M.F. A fusion framework for forecasting financial market direction using enhanced ensemble models and technical indicators. *Mathematics* **2021**, *9*, 2646. [\[CrossRef\]](#)
90. Ayala, J.; García-Torres, M.; Noguera, J.L.V.; Gómez-Vela, F.; Divina, F. Technical analysis strategy optimization using a machine learning approach in stock market indices. *Knowl. Based Syst.* **2021**, *225*, 107–116. [\[CrossRef\]](#)
91. McHugh, C.; Coleman, S.; Kerr, D. Technical indicators for energy market trading. *Mach. Learn. Appl.* **2021**, *6*, 100–110. [\[CrossRef\]](#)
92. Das, A.K.; Mishra, D.; Das, K.; Mishra, K.C. A feature ensemble framework for stock market forecasting using technical analysis and Aquila optimizer. *IEEE Access* **2024**, *12*, 187899–187918. [\[CrossRef\]](#)
93. Anggono, A.H. Investment strategy based on exponential moving average and count back line. *Rev. Integr. Bus. Econ. Res.* **2019**, *8*, 153–161.
94. Murphy, J.J. *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*; Penguin: New York, NY, USA, 1999.
95. Panchal, M.; Gor, R.; Hemrajani, J. A hybrid strategy using mean reverting indicator PSAR and EMA. *IOSR J. Math.* **2020**, *16*, 11–22.

96. Yazdi, S.H.M.; Lashkari, Z.H. Technical analysis of forex by Parabolic SAR indicator. In Proceedings of the International Islamic Accounting and Finance Conference, Kuala Lumpur, Malaysia, 19–21 November 2012.
97. Jiang, Z.; Ji, R.; Chang, K.-C. A Machine Learning Integrated Portfolio Rebalance Framework with Risk-Aversion Adjustment. *J. Risk Financ. Manag.* **2020**, *13*, 155. [\[CrossRef\]](#)
98. Panigrahi, A.K.; Vachhani, K.; Chaudhury, S.K. Trend identification with the relative strength index (RSI) technical indicator—A conceptual study. *J. Manag. Res. Anal.* **2021**, *8*, 159–169. [\[CrossRef\]](#)
99. Naved, M.; Srivastava, P. Profitability of oscillators used in technical analysis for financial market. *Adv. Econ. Bus. Manag.* **2015**, *2*, 925–931. [\[CrossRef\]](#)
100. Yamanaka, S. Average true range. *Stock. Commod.* **2002**, *20*, 76–79.
101. Wilder, J.W., Jr. *New Concepts in Technical Trading Systems*; Trend Research: Greensboro, NC, USA, 1978.
102. Cohen, G. Trading cryptocurrencies using algorithmic average true range systems. *J. Forecast.* **2023**, *42*, 212–222. [\[CrossRef\]](#)
103. Riedl, M.; Müller, A.; Wessel, N. Practical considerations of permutation entropy: A tutorial review. *Eur. Phys. J. Spec. Top.* **2013**, *222*, 249–262. [\[CrossRef\]](#)
104. Öz, E.; Aşkın, Ö.E. Classification of hepatitis viruses from sequencing chromatograms using multiscale permutation entropy and support vector machines. *Entropy* **2019**, *21*, 1149. [\[CrossRef\]](#)
105. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [\[CrossRef\]](#)
106. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [\[CrossRef\]](#)
107. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794.
108. Niculescu-Mizil, A.; Caruana, R. Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on MACHINE Learning, Bonn, Germany, 7–11 August 2005; ACM: New York, NY, USA, 2005; pp. 625–632.
109. Platt, J.C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*; Smola, A.J., Bartlett, P., Schölkopf, B., Schuurmans, D., Eds.; MIT Press: Cambridge, MA, USA, 1999; pp. 61–74.
110. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002; ACM: New York, NY, USA, 2002; pp. 694–699.
111. Varma, S.; Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **2006**, *7*, 91. [\[CrossRef\]](#)
112. Bergmeir, C.; Benítez, J.M. On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* **2012**, *191*, 192–213. [\[CrossRef\]](#)
113. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **1950**, *78*, 1–3. [\[CrossRef\]](#)
114. Naeini, M.P.; Cooper, G.; Hauskrecht, M. Obtaining well calibrated probabilities using Bayesian binning. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29, pp. 2901–2907.
115. Hodges, J.L.; Lehmann, E.L. Estimates of location based on rank tests. *Ann. Math. Stat.* **1963**, *34*, 598–611. [\[CrossRef\]](#)
116. Efron, B. Better bootstrap confidence intervals. *J. Am. Stat. Assoc.* **1987**, *82*, 171–185. [\[CrossRef\]](#)
117. Stouffer, S.A.; Suchman, E.A.; DeVinney, L.C.; Star, S.A.; Williams, R.M., Jr. *The American Soldier, Vol. 1: Adjustment During Army Life*; Princeton University Press: Princeton, NJ, USA, 1949.
118. Hedges, L.V.; Olkin, I. *Statistical Methods for Meta-Analysis*; Academic Press: Cambridge, MA, USA, 2014.
119. Gneiting, T.; Balabdaoui, F.; Raftery, A.E. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B* **2007**, *69*, 243–268. [\[CrossRef\]](#)
120. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, Australia, 6–11 August 2017; Volume 3.
121. Shwartz-Ziv, R.; Armon, A. Tabular Data: Deep Learning Is Not All You Need. *Inf. Fusion* **2022**, *81*, 84–90. [\[CrossRef\]](#)
122. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data? *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 507–520.

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.