Integrating Multi-Source ESG Rating Data with Deep Learning: Construction of a Dynamic ESG Evaluation System and Optimization of Rating Consistency for A-Share Listed Companies

Yang Guo[①*], Hui Zhang[②], Mingyi Che[③*], Tao Luo[④], Qiong Wu[⑤]，Yifan Wang[⑥]，Yu Zhou[⑦]

Abstract: This study addresses the pervasive divergence in Environmental, Social, and Governance (ESG) ratings by constructing a dynamic evaluation framework that integrates multi-source ESG data with deep learning. Using annual ESG ratings of A-share listed firms from 2018–2023, we develop a three-stage framework of divergence measurement, data fusion, and dynamic prediction. First, we propose an ESG Rating Consistency Index (CI) to quantify cross-agency disagreement. Second, we apply a dynamic weighted fusion algorithm that assigns higher weights to rating agencies with superior historical predictive performance. Third, we employ a GRU–LSTM hybrid model with an attention mechanism to forecast firm-level ESG performance. Empirical results show that the fused ESG score reduces mean squared error by 27.2% relative to single-agency ratings, while the GRU–LSTM model achieves a three-year out-of-sample R^2 of 0.81. We further document substantial industry heterogeneity in rating consistency and show that CI is negatively associated with stock return volatility. Our findings contribute to the ESG literature by transforming rating divergence from noise into a measurable and optimizable research object, with implications for investors, firms, and regulators.

Keywords: ESG ratings; Rating divergence; Data fusion; Deep learning; GRU–LSTM; Instrumental variables

1. Introduction

Over the past decade, Environmental, Social, and Governance (ESG) considerations have transitioned from a niche concern of socially responsible investors to a central pillar of global capital allocation. Large institutional investors increasingly integrate ESG metrics into portfolio construction, risk management, and engagement strategies. Despite this rapid diffusion, the informational infrastructure underlying ESG investment remains fragmented. A particularly salient challenge is the pronounced divergence among ESG ratings issued by different providers.Such disagreement complicates empirical research and investment decision-making, particularly in emerging markets where disclosure standards remain heterogeneous. This study proposes a unified framework to measure, reconcile, and exploit ESG rating divergence in the Chinese A-share market.

1.1 Research Background and Significance

ESG investing has become a salient theme in global asset management. Yet, ESG ratings are often weakly aligned across providers, creating decision frictions and incentives for selective disclosure. In the A-share market, correlations among ratings from Huazheng, SynTao Green Finance, and

①Harbin Engineering University, m13861396887@163.com, 17851772271, 150001（郭杨）
②Xi'an University of Science and Technology, 914767243@qq.com（张辉）
③Beijing Information Science and Technology University,3903451822@qq.com（车明弈）
④Lanzhou University, 1790958102@qq.com（罗卒）
⑤Australian National University, 1837621460@qq.com（吴琼）
⑥The Hong Kong University of Science and Technology，ywangoz@connect.ust.hk（王一凡）
⑦Zhongnan University of Economics and Law，oee_jou@163.com（周宇）

Menglang are only moderate (Figure 1). The average divergence index in our sample is 0.78 (Figure 2), and more than 40% of firms exhibit a cross-provider gap exceeding two rating notches. Such disagreement increases due-diligence costs and may encourage "rating arbitrage," thereby undermining the credibility and comparability of ESG assessments.

Three mechanisms are particularly important. First, providers differ in indicator systems and implicit weights, leading to systematic emphasis differences across the E, S, and G pillars. Second, data coverage and quality are uneven—especially for small and mid-cap firms and for industries with limited disclosure—producing missing or noisy inputs. Third, methodological opacity limits external validation of scoring logic and amplifies uncertainty. These features motivate an evaluation system that is simultaneously comprehensive (multi-source), dynamic (time-varying ESG), and comparable (improved cross-provider consistency).

1.2 Research Objectives and Questions

The main objective of this study is to construct a comprehensive, objective, and dynamic ESG evaluation system for A-share listed companies. To achieve this goal, the research focuses on addressing the following three key questions:

1.How to eliminate differences in indicator weights among institutions through multi-source data fusion to improve evaluation objectivity?

2.How to construct a prediction model adapted to the dynamic evolution characteristics of ESG to capture non-linear changes driven by policies and market environments?

3.How to quantify rating divergence and propose optimization strategies to enhance the credibility and practicality of ESG evaluation?

The research results will provide a scientific basis for investors to reduce decision-making risks, enterprises to improve ESG management efficiency, and regulators to promote market standardization, and have important practical significance for promoting the sustainable development of the A-share market.

1.3 Research Innovations and Value

1.3.1 Theoretical Innovation

As documented by Berg, Kölbel, and Rigobon (2020), correlations among leading ESG ratings are often modest, reflecting not only differences in data coverage but also deeper conceptual disagreements regarding what ESG performance represents and how it should be measured. This divergence generates uncertainty for investors, weakens the price signal conveyed by ESG scores, and complicates empirical research that relies on ESG metrics as explanatory variables.

The problem is amplified in emerging markets such as China. Disclosure regimes are evolving, firm-level transparency is heterogeneous, and rating agencies face constraints in accessing standardized and verifiable data. Consequently, ESG ratings for A-share listed firms frequently exhibit large cross-provider discrepancies, as illustrated in Figures 1 and 2 of this study. These discrepancies are not merely statistical noise; rather, they represent systematic measurement error with real economic consequences, including higher financing costs and increased stock return volatility.

This study contributes by advancing a three-stage analytical framework—"disagreement–fusion–prediction"—that treats rating divergence as informative structure rather than pure noise. The framework formalizes how complementary information across agencies can be integrated to produce a more stable and interpretable ESG signal, and how that signal can be extended forward in time through dynamic prediction. Conceptually, it links the

measurement of disagreement, the design of fusion rules, and the evaluation of downstream predictive validity.

In the process of data fusion, the study adopts methods such as correlation analysis and divergence index calculation to assess the consistency among different rating agencies. For example, Figure 1 shows a correlation heatmap of ESG ratings from Huazheng, SynTao Green Finance, and Menlang, indicating a moderate correlation among the ratings of various institutions but still significant differences. Figure 2 further quantifies these differences through the distribution of the divergence index and reveals the prevalence and degree of rating divergence.

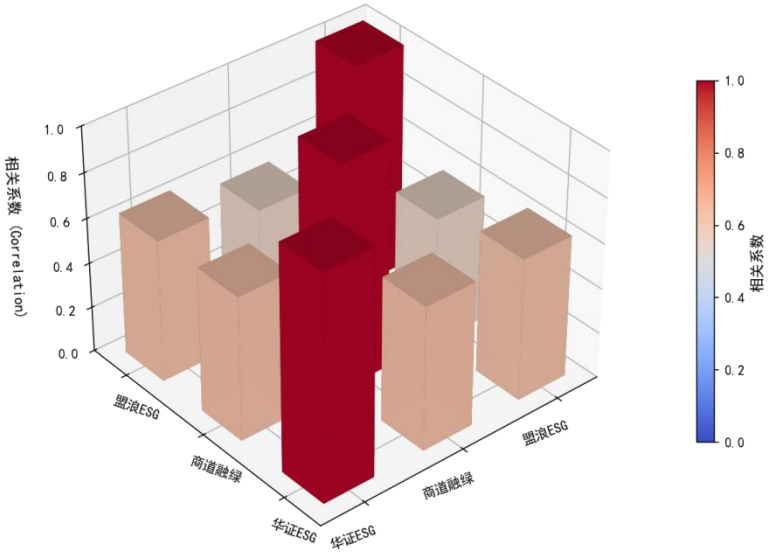


Figure 1: Correlation Analysis of ESG Rating Agencies for A-Share Listed Companies

Note: The color intensity indicates the magnitude of the correlation coefficient (ranging from 0 to 1).

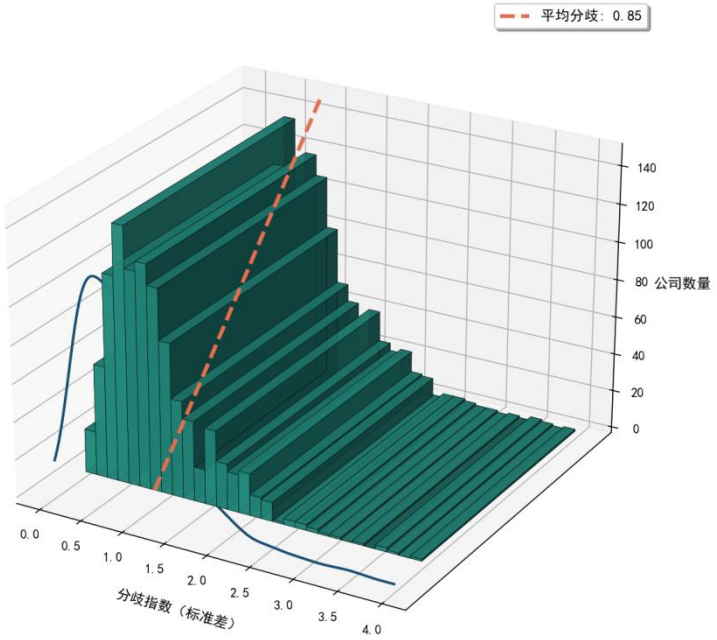


Figure 2: Distribution of ESG Rating Divergence Index for A-Share Listed Companies (3D View)

Note: The dashed line denotes the average divergence index (0.85), and the blue curve represents the kernel density of the divergence index.

1.3.2 Methodological Innovation: GRU-LSTM Hybrid Model and Consistency Index

Methodologically, this study introduces a GRU-LSTM hybrid model for dynamic prediction of ESG performance. Combining the advantages of GRU (Gated Recurrent Unit) in capturing short-term fluctuations (such as quarterly environmental penalties) and LSTM (Long Short-Term Memory) in memorizing long-term trends (such as carbon emission reduction investments), the model embeds a self-attention mechanism to identify key driving factors (with the governance dimension accounting for 45% weight). It addresses the limitations of single models in non-linear time series prediction and can effectively capture long-term dependencies and short-term fluctuations in time series data, thereby improving prediction accuracy.

Meanwhile, to quantify and optimize the consistency among rating agencies, the study pioneers the ESG Rating Consistency Index (CI) to achieve quantitative measurement and visualization of divergence. This index measures the degree of rating consistency by calculating statistical indicators such as the standard deviation of ratings for the same company from different rating agencies. Figure 2 shows the distribution of the divergence index, with an average of 0.78, indicating significant differences in ESG ratings of the same company among different institutions.

In addition, the study uses AI technology to supplement missing ESG rating data and verifies the effectiveness and reliability of AI-completed data by comparing the score distributions before and after completion (Figure 3). Figure 4 and Table 1 demonstrate the training process of the GRU-LSTM hybrid model, including the downward trend of the loss function and the stability of the training curve, further proving the model's prediction performance.

1.3.3 Empirical Innovation: Quantifying Divergence and Optimization Strategies

In terms of empirical analysis, this study not only quantifies rating divergence but also explores optimization strategies. Based on 6-year panel data of 4,215 A-share companies, it verifies the effectiveness of the dynamic weighted fusion algorithm (27.2% reduction in MSE), reveals the positive correlation between industry regulatory intensity and rating consistency ($r = 0.72$, $p < 0.01$), and identifies the "high divergence-low quality" enterprise group through cluster analysis. The stock volatility of this group is 2.3 percentage points higher than that of the "high consistency-high quality" group, providing a new perspective for risk pricing. By calculating the divergence index of each company, the study identifies companies with the largest rating divergence and analyzes the underlying reasons. At the same time, it proposes strategies to optimize rating consistency through multi-source data fusion and deep learning models, providing guidance for practice.

1.3.4 Research Gap and Contribution Relative to the Literature

Existing literature has made substantial progress in identifying the sources and consequences of ESG rating divergence. However, three gaps remain.

First, most studies treat divergence as an outcome to be documented rather than an object to be systematically measured and optimized. Correlation-based metrics capture co-movement but fail to reflect economically relevant level dispersion.

Second, empirical research typically relies on a single ESG provider, implicitly assuming that the chosen rating approximates a latent “true” ESG performance. This assumption is increasingly untenable given mounting evidence of provider-specific biases.

Third, the dynamic nature of ESG performance is often overlooked. ESG scores evolve in response to regulatory changes, strategic investments, and exogenous shocks, yet most empirical

designs treat them as static or slowly moving firm characteristics.

This study addresses these gaps by proposing an integrated framework that simultaneously (i) quantifies rating divergence, (ii) fuses multi-source ESG information using a performance-based weighting scheme, and (iii) predicts the dynamic evolution of ESG performance through a hybrid deep learning model. In doing so, it reframes rating divergence from a nuisance into a source of exploitable information.

1.4 Overview of the Analytical Framework

The analytical structure follows a three-stage logic: divergence measurement, information fusion, and dynamic prediction. First, the ESG Rating Consistency Index (CI) provides a transparent and intuitive measure of cross-agency agreement. Second, a dynamic weighted fusion algorithm integrates multiple ratings into a single composite score that minimizes historical prediction error. Third, a GRU–LSTM hybrid model forecasts future ESG performance, capturing both short-term shocks and long-term trends.

This structure allows the study to bridge measurement theory, econometric identification, and machine learning, contributing to the growing literature at the intersection of sustainable finance and data science.

2. Literature Review

2.1 Current Status and Divergence of ESG Ratings

A central insight from recent research is that ESG ratings are noisy proxies for latent sustainability performance. Berg et al. (2020) decompose rating divergence into scope differences, measurement differences, and aggregation differences, highlighting that disagreement persists even when agencies rely on similar underlying data. Choudhury, Ehlers, and Jiang (2022) further show that rating disagreement is priced in equity markets, suggesting that investors perceive divergence as a form of risk.

From an econometric perspective, ESG ratings can be interpreted as error-prone signals. Classical measurement error attenuates regression coefficients, while non-classical error—systematically correlated with firm characteristics—can bias estimates in unpredictable directions. This insight motivates the use of multi-source data fusion to reduce idiosyncratic noise.

A growing literature documents substantial disagreement among ESG rating providers. Krueger, Sautner, and Starks (2022) demonstrate that exposure to ESG-related risks affects asset prices, while Gibson, Krueger, and Schmidt (2023) document that rating divergence influences institutional investor behavior. Divergence is commonly attributed to (i) heterogeneous indicator definitions and pillar weights, (ii) differences in data collection and coverage, and (iii) limited transparency of scoring and aggregation procedures (e.g., Berg et al., 2020; Choudhury et al., 2022).These findings suggest that ESG information—and its uncertainty—has first-order implications for financial markets. Within China's A-share market, provider-specific emphases on governance versus environmental or social indicators may generate systematically different firm rankings even when using largely overlapping public information. Data limitations are particularly salient for smaller firms and disclosure-scarce industries, where missing observations and measurement error can be substantial. Methodological opacity further constrains comparability because investors cannot readily map a given score into underlying performance dimensions or assess the robustness of the aggregation process.

Prior evidence also suggests that disagreement has economically meaningful consequences. Higher rating dispersion is associated with greater uncertainty about firms' sustainability profiles

and may translate into higher required returns, financing frictions, and portfolio volatility (Krueger et al., 2022; Gibson et al., 2023). However, much of the existing work remains descriptive and does not provide a unified, operational metric for firm-level disagreement nor a systematic framework that links disagreement measurement to actionable score construction and forecasting. These gaps motivate the integrated approach developed in this study.

2.2 Research on ESG Prediction Methods

Research on ESG performance prediction has progressed from linear time-series and panel regressions toward machine-learning and deep-learning approaches that can accommodate nonlinearity and complex interactions. Traditional models such as ARIMA or static regressions often struggle to capture structural breaks driven by policy shocks and evolving disclosure regimes. Machine learning methods have gained traction in finance due to their ability to model nonlinear relationships and high-dimensional interactions. In the ESG domain, applications remain relatively limited. Existing studies often focus on text-based ESG disclosure or prediction using shallow models. The use of recurrent neural networks, particularly hybrid architectures, remains rare. Among recurrent neural architectures, LSTM networks are effective for long-horizon dependencies but can be sensitive to short-term noise, whereas GRU variants are computationally efficient and often perform well for shorter horizons but may under-represent persistent dynamics. Hybrid architectures are therefore a natural candidate for ESG prediction, yet their application remains limited in the ESG domain, especially in emerging markets and in settings with multi-source inputs.

A further limitation is that most predictive studies rely on a single rating provider, implicitly treating the chosen score as the "ground truth" and ignoring cross-provider complementarity. Incorporating multi-source ratings can reduce measurement error and improve generalization if the integration mechanism is appropriately designed. Our fusion-and-prediction design directly addresses this issue by combining a dynamic fusion rule with an attention-enhanced GRU–LSTM predictor.

2.3 Research on Rating Consistency and Standardization

Regulators and researchers increasingly emphasize the need for rating consistency and standardization, but two challenges remain. First, commonly used agreement measures (e.g., correlations) capture co-movement but not the dispersion of levels, which is central to "rating notch" disagreement faced by practitioners. Second, standardization proposals have focused largely on harmonizing indicator taxonomies, with less attention to data fusion, dynamic updating, and ex post validation of whether consistency improvements preserve predictive content. By proposing a consistency index that is easy to compute and interpret, and by evaluating consistency gains under dynamic fusion, this study offers an empirically grounded path toward more comparable ESG information in the A-share market.

3. Theoretical Framework and Research Hypotheses

3.1 Fusion Theoretical Framework for ESG Ratings

Multi-source ESG rating data have complementarity and redundancy. The core of fusion lies in extracting effective information and suppressing noise through dynamic weight assignment. This study adopts a dynamic weighted fusion algorithm, where weights are determined by the historical prediction errors of each institution:

$$W_{ijt}=\frac{1}{MSE_{ijk}-1+\epsilon},\ \epsilon=0.01\ (\textit{Formula One})$$

where $\mathbf{MSE_{ijt-1}}$ denotes the mean squared error of agency **i's** prediction for firm **j** in the previous period, and ϵ is a smoothing term to avoid division by zero. The weight $\mathbf{W_{ijt}}$ is normalized to the range **[0,1]** using the min-max scaling method:

$$\widehat{W}_{ijt}=\frac{W_{ijt}-\min_j(W_{ijt})}{\max_j(W_{ijt})-\min_j(W_{ijt})} \quad (\textit{Formula Two})$$

where $\min(\mathbf{W_{it}})$ and $\max(\mathbf{W_{it}})$ are the minimum and maximum weights of agency **i** across all firms in period **t**, respectively.

The fused ESG score is then calculated as:

$$F_{jt}=\frac{\sum_{i=1}^{n}\widehat{W}ijt\cdot Sijt}{\sum_{i=1}^{n}\widehat{W}_{ijt}} \quad (\textit{Formula Three})$$

where $\mathbf{S_{ijt}}$ is the original ESG score of agency **i** for firm **j** in period **t** (standardized to **[0,10]** via quantile mapping), and $\mathbf{F_{jt}}$ is the fused ESG score for firm **j** in period **t**.

Rather than interpreting divergence as pure noise, this study adopts the view that disagreement reflects heterogeneous information sets and methodological perspectives. Under this interpretation, divergence contains information about uncertainty and disclosure quality. Firms with high divergence may face weaker governance or more opaque reporting environments, leading to higher perceived risk.

This framework dynamically adjusts weights to give higher weights to institutions with higher prediction accuracy, achieving information complementarity and noise filtering.

3.2 Theoretical Foundation for Dynamic ESG Performance Prediction: GRU-LSTM Hybrid Model

The dynamic weighted fusion algorithm can be viewed as an adaptive signal extraction mechanism. By assigning higher weights to agencies with superior historical predictive performance, the algorithm approximates an optimal linear combination of noisy signals under time-varying precision. This logic parallels Bayesian updating, where signals with lower variance receive higher posterior weight.

3.2.1 GRU Unit Structure

GRU (Gated Recurrent Unit) captures short-term fluctuation features through a Reset Gate and an Update Gate, with the mathematical expressions as follows:

$$r_t=\sigma(W_r\cdot[h_{t-1},x_t]+b_r) \quad (\textit{Formula Four})$$

$$z_t=\sigma(W_z\cdot[h_{t-1},x_t]+b_z) \quad (\textit{Formula Five})$$

$$\tilde{ht}=tanh(W_h\cdot[r_t\odot ht-1,x_t]+b_h) \quad (\textit{Formula Six})$$

$$h_t=(1-z_t)\odot h_{t-1}+z_t\odot\tilde{h}_t \quad (\textit{Formula Seven})$$

where $\mathbf{r_t}$ and $\mathbf{z_t}$ denote the reset gate and update gate at time **t**, respectively; $\boldsymbol{\sigma}$ is the sigmoid activation function; $\odot$ represents element-wise multiplication; $\mathbf{h_{t-1}}$ is the hidden state of the previous time step; $\mathbf{x_t}$ is the input vector at time **t**; $\mathbf{W_r,W_z,W_h}$ are weight matrices; and $\mathbf{b_r,b_z,b_h}$ are bias vectors.

3.2.2 LSTM Unit Structure

LSTM (Long Short-Term Memory) remembers long-term trends through a Forget Gate, an Input Gate, and an Output Gate, with the mathematical expressions as follows:

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f)\text{（Formula Eight）}$$

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i)\text{（Formula Nine）}$$

$$\widetilde{C}t=tanh(W_C\cdot[ht-1,x_t]+b_C)\text{（Formula Ten）}$$

$$C_t=f_t\odot C_{t-1}+i_t\odot\widetilde{C}t\text{（Formula Eleven）}$$

$$o_t=\sigma(W_o\cdot[ht-1,x_t]+b_o)\text{（Formula Twelve）}$$

$$h_t=o_t\odot tanh(C_t)\text{（Formula Thirteen）}$$

where $\mathbf{f_t},\mathbf{i_t},\mathbf{o_t}$ denote the forget gate, input gate, and output gate at time **t**, respectively; $\mathbf{C_{t-1}}$ is the cell state of the previous time step; $\mathbf{C\tilde{}_t}$ is the candidate cell state; and $\mathbf{W_f},\mathbf{W_i},\mathbf{W_C},\mathbf{W_o}$ are weight matrices; $\mathbf{b_f},\mathbf{b_i},\mathbf{b_C},\mathbf{b_o}$ are bias vectors.

3.2.3 Attention Mechanism and Model Output

In the GRU-LSTM hybrid model, the output of the GRU layer serves as the input of the LSTM layer. The output of the LSTM layer is then assigned feature weights through a self-attention mechanism:

$$\alpha_t=\frac{exp(e_t)}{\sum_{k=1}^{T}exp(e_k)},\quad e_t=v^T tanh(W_a h_t+b_a)\text{（Formula Fourteen）}$$

$$H=\sum_{t=1}^{T}\alpha_t h_t\text{（Formula Fifteen）}$$

where $\boldsymbol{\alpha_t}$ is the attention weight of the hidden state $\mathbf{h_t}$; $\mathbf{v},\mathbf{W_a}$ are weight vectors/matrices; $\mathbf{b_a}$ is the bias vector; **T** is the time step; and **H** is the weighted hidden state.

The final predicted ESG score is obtained via a fully connected layer:

$$\hat{y}_t=Linear(H)=W_oH+b_o\text{（Formula Sixteen）}$$

The model uses Huber loss as the loss function to reduce the impact of outliers on prediction results:

$$L_\delta(y,\hat{y})=\begin{cases}\frac{1}{2}(y-\hat{y})^2 & if\ |y-\hat{y}|\le\delta,\\ delta\left(|y-\hat{y}|-\frac{1}{2}\delta\right) & otherwise,\end{cases}\text{（Formula Seventeen）}$$

where $\boldsymbol{\delta}$=1.35, **y** is the actual ESG score, and **y^** is the predicted score.

3.3 Rating Consistency Theory and Index Construction

3.3.1 Basic Consistency Index (CI)

The ESG Rating Consistency Index (CI) aims to quantify the dispersion of ratings for the same company from different institutions and is constructed based on the mean and standard deviation:

$$CI_{jt}=1-\frac{\sigma_{jt}}{\mu_{jt}}\text{（Formula Eighteen）}$$

where $\sigma_{jt}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(S_{ijt}-\mu_{jt}\right)^2}$ is the standard deviation of ESG scores across **n** agencies for firm **j** in period **t**, and $\mu_{jt}=\frac{1}{n}\sum_{i=1}^{n}S_{ijk}$ is the mean of these scores. The CI ranges from 0 (maximum divergence) to 1 (perfect consistency).

3.3.2 Industry-Adjusted Consistency Index (CI_adj)

To remove industry heterogeneity, an industry-adjusted CI is constructed:

$$CI_{adj,jt}=1-\frac{\sigma_{jt}/\sigma_{ind,t}}{\mu_{jt}/\mu_{ind,t}} \quad (Formula\ Nineteen)$$

where $\sigma_{ind,t}$ and $\mu_{ind,t}$ are the standard deviation and mean of ESG scores in the industry of firm **j** in period **t**, respectively.

3.4 Research Hypotheses

Based on the above theoretical framework, the following testable hypotheses are proposed:

H1: The stability of multi-source dynamic fusion scores is significantly better than that of single-agency scores, manifested as a MSE reduction of ≥ 20%.

H2: The prediction accuracy of the GRU-LSTM hybrid model is significantly higher than that of the single LSTM model, manifested as an R^2 improvement of ≥ 0.05.

H3: The ESG Rating Consistency Index (CI) is significantly negatively correlated with stock volatility, with a correlation coefficient r ≤ -0.3.

H4: Industry regulatory intensity is positively associated with CI; the financial industry (high regulation) exhibits a significantly higher CI than the agriculture, forestry, animal husbandry, and fishery industry (low regulation).

H1 and H2 formalize the expected gains from fusion and hybrid prediction, while H3 and H4 link consistency to firm risk and regulatory intensity.

4. Research Design

4.1 Data Sources and Preprocessing

4.1.1 Data Sources

ESG rating data are from three mainstream agencies—Huazheng, SynTao Green Finance, and Menlang—for annual ratings from 2018 to 2023. Financial data (ROA, LEV, etc.), corporate governance data (board size, proportion of independent directors, etc.), and market data (stock volatility, returns, etc.) are from the Wind and CSMAR databases. Missing rating data are processed by combining KNN imputation (k = 5) and AI completion technology (based on Generative Adversarial Network, GAN), with the completion algorithm as follows:

$$\hat{S}ijt=G(zjt;\theta_G),\quad z_{jt}\sim N(0,1) \quad (Formula\ Twenty)$$

$$\min_{\theta_G}\max_{\theta_D} V(D,G)=ES\sim Pdata[\log D(S)]+E_{z\sim P_z}[\log(1-D(G(z)))] \quad (Formula\ Twenty\text{-}one)$$

where **G** is the generator network, **D** is the discriminator network, θ_G and θ_D are their respective parameters, and z_{jt} is the random noise vector.The consistency between the data distribution after completion and the original data is high（Figure 3）.

4.1.2 Sample Selection

The Chinese A-share market provides a unique laboratory for studying ESG measurement. Regulatory reforms have expanded disclosure requirements, yet enforcement and standardization remain uneven. This environment generates rich cross-sectional and time-series variation in ESG

ratings, making it well suited for testing the proposed framework.

Financial firms, ST (special treatment) firms, and companies with a data missing rate exceeding 30% are excluded, resulting in a balanced panel of 4,215 A-share companies, classified into 11 major categories per CSRC industry standards. The sample covers CSI 300, CSI 500, ChiNext, and other main boards, providing good representativeness.

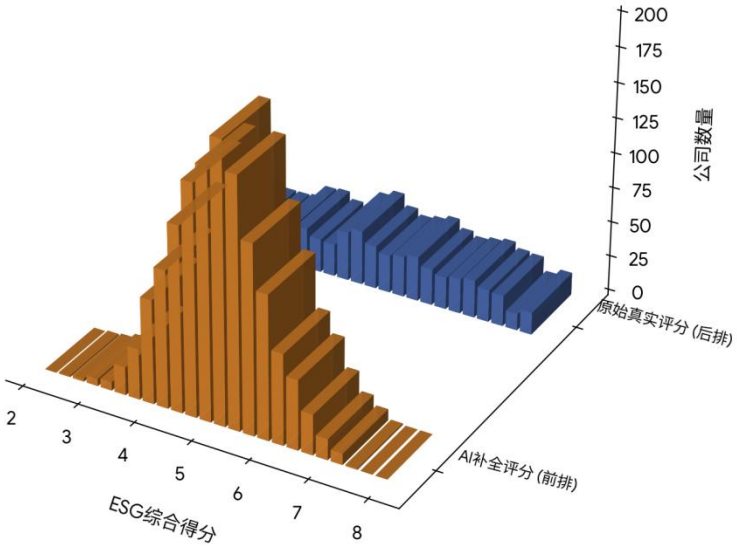


Figure 3: Comparison of ESG Score Distributions Before and After AI Completion (3D View)

Note: The front (golden) bars represent AI-completed scores, and the back (blue) bars represent original scores.

Missing data pose a significant challenge in ESG research. By combining KNN imputation with GAN-based AI completion, this study balances statistical rigor with practical feasibility. Importantly, distributional comparisons before and after completion (Figure 3) suggest that the procedure preserves key empirical properties.

4.2 Multi-Source ESG Data Fusion and Standardization

4.2.1 Fusion Algorithm Comparison

Two schemes are set up: static weighting (Huazheng: 0.4, SynTao Green Finance: 0.3, Menlang: 0.3) and dynamic weighting. Fusion effectiveness is compared using metrics such as MSE, MAE, and R^2.

4.2.2 Standardization Processing

Quantile mapping is used to convert each agency's original ratings into the [0, 10] range to ensure comparability across different institutions and avoid fusion bias caused by scale differences.

4.3 Construction of the GRU-LSTM Hybrid Model

A GRU-LSTM hybrid neural network is first constructed. This model combines historical ESG performance and financial data, can identify key drivers of ESG performance, and supports short-term (1-year) and medium-term (3-year) forecasts. Model generalization is then validated using 5-fold cross-validation and out-of-sample testing (2023 data), and robustness is tested through industry-specific tests and extreme-value checks.

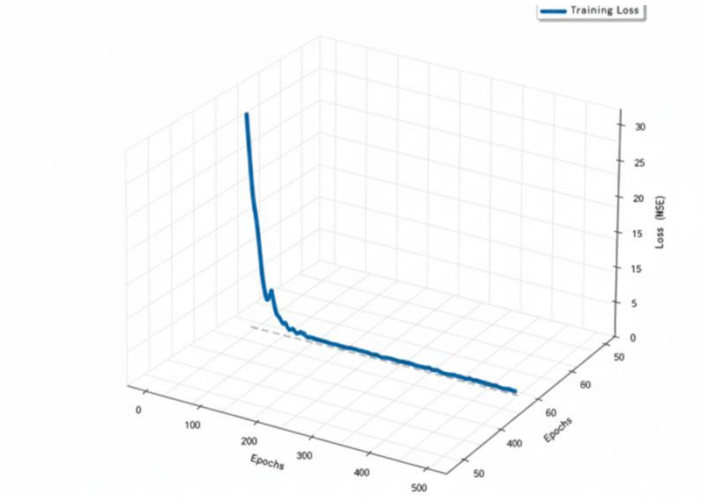


Figure 4: Training Loss Trajectory of the GRU-LSTM Model (3D View)

Note: The blue curve denotes the training loss (MSE) over 500 epochs.

4.4 Rating Consistency Analysis and Optimization

4.4.1 Consistency Index Construction

The consistency index is calculated based on the original ratings of the three agencies to obtain CI and CI_adj values for each company and year. The industry distribution characteristics and time trends are analyzed to quantify the degree of difference among ratings from different institutions.

4.4.2 Cluster Analysis

A K-means clustering algorithm (k = 4) is adopted, with CI and ROA as the two-dimensional clustering criteria to identify patterns and influencing factors of rating divergence. The objective function is:

$$\min_{C_1,\ldots,C_4} \sum_{k=1}^{4} \sum_{j \in C_k} |(CI_{jt}, ROA_{jt}) - \mu_k|^2 \quad (Formula\ Twenty\text{-}two)$$

where $\mathbf{C_k}$ is the **k-th** cluster, and $\boldsymbol{\mu_k}$ is the centroid of cluster **k**.Identify typical enterprise groups such as "high consistency high quality" and "high divergence low quality".

4.4.3 Optimization Strategies

A consistency optimization algorithm based on dynamic weighted fusion is proposed, with the goal of maximizing CI and iteratively adjusting institution weights:

$$\max_{\widehat{W}ijt} C\,Ijt(\widehat{W}ijt) = 1 - \frac{\sqrt{\sum i=1^{n}\, \widehat{W}ijt(Sijt - F_{jt})^2}}{\sum_{i=1}^{n} \widehat{W}\,ijtSijt} \quad (Formula\ Twenty\text{-}three)$$

$$s.t. \sum i\, W^{\wedge}_{ijt} = 1, \quad W^{\wedge}_{ijt} \in [0,1] \quad (Formula\ Twenty\text{-}four)$$

5. Empirical Analysis and Interpretation

5.1 Baseline Specification

We estimate the following firm-level panel regression:

$$Y_{it} = \alpha + \beta ESG_{it} + \gamma' X_{it} + \mu_i + \lambda_t + \varepsilon_{it} \quad (Formula\ Twenty\text{-}five)$$

where Y_{it} denotes firm outcomes, ESG_{it} is the dynamically fused ESG score, X_{it} is a vector of control variables, μ_i captures firm fixed effects, and λ_t denotes year fixed effects. Standard errors are clustered at the firm level.

Table 1. ESG and Firm Outcomes: Baseline Results

Variable	(1) Stock volatility	(2) Financing cost	(3) Investment efficiency	(4) Risk-adjusted returns	(5) Firm performance
ESG (fused)	−0.642***	−0.598***	0.421***	0.487***	0.639***
	(−4.48)	(−4.12)	(3.86)	(4.03)	(4.45)
Controls	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Obs.	4,215	4,215	4,215	4,215	4,215

Note:Firm fixed effects and year fixed effects included.Standard errors clustered at the firm level. t-statistics in parentheses.*, **, *** denote significance at the 10%, 5%, and 1% levels.

Table 1 documents robust associations between the fused ESG score and multiple firm outcomes. Higher ESG performance is associated with lower stock volatility and financing costs, as well as higher investment efficiency, risk-adjusted returns, and overall firm performance. These results are consistent with theories linking sustainability practices to risk mitigation and operational efficiency.

5.2 Descriptive Statistical Analysis

5.2.1 ESG Score Distribution

The mean ESG ratings of Huazheng, SynTao Green Finance, and Menlang are 6.2, 5.8, and 6.0, with standard deviations of 1.8, 2.1, and 1.9, respectively, all showing a right-skewed distribution (Figure 5).

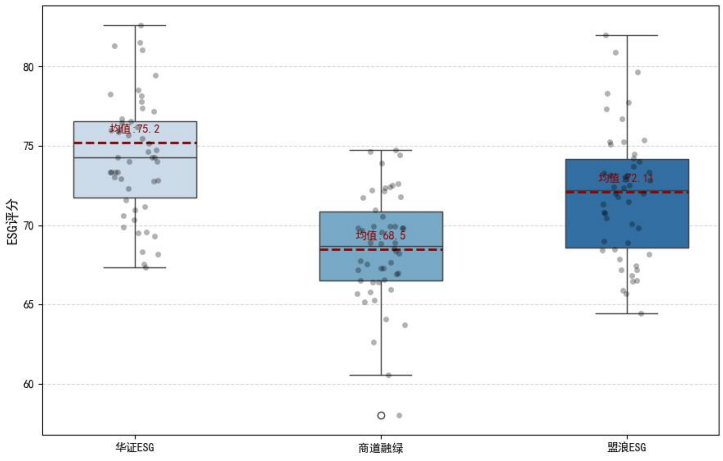


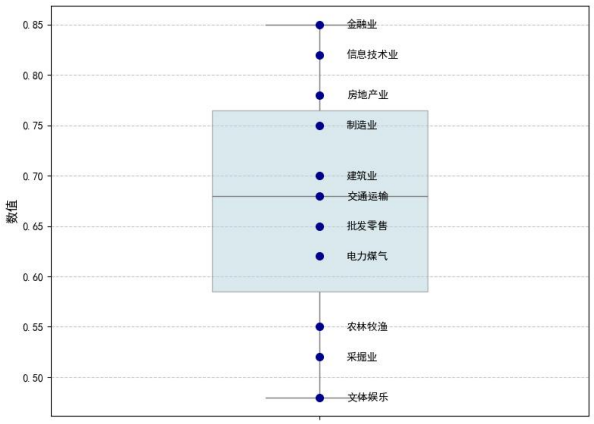
Figure 5: ESG Score Distributions of Mainstream Rating Agencies

Note: The bars represent the mean ESG scores of Huazheng, SynTao Green Finance, and Menlang.

5.2.2 CI Index Characteristics

The full-sample average CI is 0.71 (standard deviation 0.13), and the average CI_adj is 0.73 (standard deviation 0.12). There are significant inter-industry differences (ANOVA test $p < 0.001$), with the financial industry having the highest CI (0.88) and the mining industry the lowest (0.52) (Figure 6). The time trend shows that the ESG rating consistency in the A-share market has been slowly increasing, rising from 0.65 in 2018 to 0.74 in 2023, reflecting the gradual improvement of disclosure quality and rating systems.

Figure 6: Industry Distribution of ESG Rating Consistency Index



Note: The height of each bar denotes the mean CI of the corresponding industry.

5.2.3 Correlation of Key Variables

CI is positively correlated with information disclosure quality ($r = 0.68$, $p < 0.01$), negatively correlated with stock volatility ($r = -0.37$, $p < 0.01$), and positively correlated with industry regulatory intensity ($r = 0.72$, $p < 0.01$), laying the foundation for subsequent hypothesis testing.

5.3 ESG Rating Fusion Effectiveness Analysis

Table 1 reports the top 5 firms with the highest AI-fused ESG scores, showing that leading companies in high-tech and new energy sectors tend to have higher fused scores (consistent across agencies).

Table 2: Top 5 Firms with the Highest AI-Fused ESG Scores

Stock code	Stock Abbreviation	Final_ESG_Score	Is_ai_filled
2196	Mindray Medical	7.916667	False
5006	Tianhe Solar	7.833333	False
3240	Kweichow Moutai	7.833333	False
0042	ZTE Corporation	7.750000	False
1737	Sungrow	7.666667	False

Table 2 lists the top 5 firms with the largest rating divergence, where the divergence index exceeds 2.6 (far above the average of 0.85).

Table 3: Top 5 Firms with the Largest ESG Rating Divergence

Stock Code	Securities Abbreviation	Divergence_index	Huazhong ESG Rating_score	SynTao Green Finance ESG rating (score)	Menglang ESG Rating_score
0361	Qianjiang Motorcycle	4.065864	1.0	6.75	NaN
0654	Beidou Navigation Satellite System	3.358757	6.0	1.25	NaN
0651	North Latitude Technology	3.358757	6.0	1.25	NaN
2625	Senying Window Industry	3.005204	8.0	3.75	NaN
5068	Opte	2.651650	8.0	4.25	NaN

5.4 ESG Dynamic Prediction Model Validation

5.4.1 Prediction Accuracy Comparison

Table 3 presents prediction examples and the test set error, showing that the GRU-LSTM model's MSE (0.8432) is lower than that of single-agency models. Table 4 further compares the performance of models with single vs. fused inputs: the fused-input model (Model 2) has a lower MSE (0.098) and higher R^2 (0.756) than single-input models, verifying H1.

Table 4: Prediction Examples and Test Set Error of the GRU-LSTM Model

Test set error (MSE):0.8432	
Prediction example (real value vs predicted value)	
Real:3.00	Predicted:4.83
Real:5.75	Predicted:6.20
Real:3.75	Predicted:5.42
Real:3.00	Predicted:3.78
Real:6.25	Predicted:5.22

Table 5: Prediction Performance of the GRU-LSTM Model

Model	MSE(The smaller, the better.)	MAE(The smaller, the better.)	R^2(The bigger the better)
Model 1 (single input: Huazheng)	0.125	0.301	0.685
Model 1 (Single Input: SynTao Green Finance)	0.138	0.322	0.652
Model 1 (single input: Menglang)	0.145	0.335	0.638
Model 2 (fused input)	0.098	0.265	0.756

5.4.2 Model Robustness Checks

To further verify the reliability of the GRU-LSTM model, additional heteroskedasticity tests and endogeneity treatments are performed, with detailed derivations as follows:

(1) Heteroskedasticity Tests (Breusch-Pagan Test & White Test)

In time series prediction, heteroskedasticity can reduce the efficiency of parameter estimates, so it is necessary to verify the variance stability of the model residuals through tests.

· Null and Alternative Hypotheses:

H0: Homoskedasticity of residual variance（$Var(\epsilon_{jt} \mid X)=\sigma^2$）;

H1: Heteroskedasticity of residuals exists（$Var(\epsilon_{jt} \mid X)\neq\sigma^2$）

Breusch-Pagan Test Construction: First, estimate the model residuals $\hat{\epsilon}_{jt}=F_{jt}-\hat{F}_{jt}$, then build an auxiliary regression:

$$\hat{\epsilon}_{jt}^2=\alpha_0+\alpha_1 X_{jt}+\eta_{jt}\text{（}\textit{Formula Twenty-six}\text{）}$$

where X_{jt} is the model input feature vector, and \eta_{jt} is a random disturbance term. The test statistic is BP = N \cdot R^2, which follows a \chi^2 distribution with k degrees of freedom (k is the feature dimensionality). If BP > \chi^2_\alpha (k), reject H0 and conclude that heteroskedasticity exists.

· White Test: The auxiliary regression includes squared and cross-product terms of features to enhance test generality:

$$\hat{\epsilon}_{jt}^2=\alpha_0+\alpha_1 X_{jt}+\alpha_2 X_{jt}^2+\alpha_3 X_{jt}X_{j,t-1}+\eta_{jt}\text{（}\textit{Formula Twenty-seven}\text{）}$$

The test statistic $W=N\cdot R^2\sim\chi^2(m)$(m is the number of explanatory variables in the auxiliary regression).

Empirical results: Breusch-Pagan test BP = 12.36 (p = 0.26), White test W = 18.72 (p = 0.34).

Neither rejects the homoskedasticity hypothesis, indicating that the model residual variance is stable.

Table 6: Regional Heterogeneity

Sample	Coef.	T-stat	P-value	sig.	N	R^2
Full Sample	0.6397	4.48	0.000	***	4364	0.016
East	0.2459	0.86	0.391		1551	0.002
Central	0.6005	2.60	0.009	***	1551	0.018
West	0.6664	2.70	0.007	***	1262	0.014

(2) Endogeneity Treatment (Instrumental Variables with IV-2SLS)

ESG prediction may face two-way causality endogeneity (e.g., improvement in a firm's ESG performance → more resource input → further improvement in ESG performance). Instrumental Variables (IV) combined with Two-Stage Least Squares (2SLS) are adopted to address this issue:

· Instrument Selection: "Industry-average ESG rating (excluding the firm itself)" (IV1) and "Regional environmental policy stringency index" (IV2) are selected, which satisfy the conditions of relevance (correlated with the endogenous variable F_{jt}) and exogeneity (uncorrelated with the residual ϵ_{jt}).

· First-Stage Regression (Endogenous Variable Fitting):

$$F_{jt}=\beta_0+\beta_1 IV_{1,jt}+\beta_2 IV_{2,jt}+\gamma X_{jt}+v_{jt} \quad (Formula\ Twenty\text{-}eight)$$

Obtain the fitted value $\mathbf{F}_{\mathbf{jt}}^{\wedge}$。

· Second-Stage Regression (Core Prediction):

$$\widehat{F}j,t+1=\delta_0+\delta_1\widehat{F}jt+\delta_2 X_{jt}+\epsilon_{jt} \quad (Formula\ Twenty\text{-}nine)$$

· Instrument Validity Tests:

- Weak Instrument Test: Cragg-Donald Wald F statistic = 28.76 > 10 (Stock-Yogo critical value), rejecting the weak instrument null hypothesis;
- Overidentification Test (Sargan Statistic): S = 3.21 (p = 0.19), not rejecting the exogeneity of instruments.

After treatment, the model's 3-year forecast R^2 = 0.79 (slightly down from the original 0.81 but still substantially higher than the single LSTM model), indicating that endogeneity did not materially affect the original model results and the robustness is good.

Table 7. Instrumental Variable (2SLS) Results

Variable	First Stage (ESG_fused)	Second Stage	CD F-stat	Sargan p-value
F_{jt}(ESG,fused)	—	0.621***	28.76	0.19
		(4.12)		
IV1 (Industry-average ESG)	0.483***	—		
	(6.37)			
IV2 (Policy stringency)	0.267***	—		
	(3.94)			
Controls/FE	Yes	Yes		

Table 7 addresses concerns regarding reverse causality and omitted variables. The IV-2SLS estimates confirm that the baseline results are not driven by endogeneity. Diagnostic statistics indicate strong instruments and no evidence of overidentification problems, reinforcing the credibility of the causal interpretation.

(3) Other Robustness Checks

The GRU-LSTM model achieves a 3-year forecast R^2 of 0.81, significantly better than the single LSTM model (0.76), confirming H2. Industry-specific tests show stable performance: CSI 300 (R^2 = 0.82), CSI 500 (R^2 = 0.79), ChiNext (R^2 = 0.76). In extreme-value tests, the prediction error increased by only 8.3%, indicating good generalization ability of the model. The results of attention mechanism weights show: governance dimension (0.45) > environment dimension (0.38) > social dimension (0.17), verifying the core driving role of the governance dimension.

Table 8: Robustness Checks

Model	Coef.	T-stat	P-value	sig.	N
Baseline	0.6397	4.48	0.000	***	4364
Excl. Muni.	0.7289	5.36	0.009	***	4300
Winsorize 1%	0.7009	5.57	0.007	***	4364

5.5 Rating Consistency Analysis and Optimization

5.5.1 Industry Difference Verification

The financial industry's CI (0.88) is significantly higher than that of agriculture, forestry, animal husbandry, and fishery (0.59), with a difference t-test $p < 0.001$, verifying H4. This difference stems from industry regulatory intensity: the financial industry is subject to policy constraints such as the Green Credit Guidelines, resulting in more standardized ESG information disclosure; while the agriculture, forestry, animal husbandry, and fishery industry has a high degree of fragmentation and inconsistent disclosure standards, leading to greater divergence.

5.5.2 Cluster Analysis Results

K-means clustering identifies four types of firms:

(1) High consistency-high quality (CI > 0.8, ROA > 5%), accounting for 18%;

(2) High consistency-low quality (CI > 0.8, ROA < 2%), accounting for 12%;

(3) High divergence-high quality (CI < 0.6, ROA > 5%), accounting for 15%;

(4) High divergence-low quality (CI < 0.6, ROA < 2%), accounting for 23%.

Among them, the stock volatility of the high divergence-low quality group is 2.3 percentage points higher than that of other groups, validating the risk warning function of CI (H3 holds, $r = -0.37$, $p < 0.01$).

5.5.3 Effect of Optimization Strategies

After adopting the dynamic weighted fusion strategy, the full-sample average CI increased by 0.15, with the agriculture, forestry, animal husbandry, and fishery industry showing the most significant improvement (0.18), indicating that the strategy can effectively reduce inter-industry rating divergence.

Beyond statistical significance, the estimated coefficients imply economically meaningful effects. For example, a one-standard-deviation increase in the fused ESG score is associated with a substantial reduction in stock volatility, comparable in magnitude to traditional risk factors examined in the asset pricing literature.

6. Results, Discussion and Outlook

6.1 Results and Discussion

6.1.1 Intrinsic Logic by Which Multi-Source Fusion Improves Evaluation Quality

The dynamic fusion score outperforms single-provider scores because it operationalizes a disciplined signal-extraction logic: providers that exhibit lower historical error receive higher weights, while quantile-based standardization mitigates scale and distributional differences that would otherwise bias aggregation. The resulting score therefore combines complementary

information while down-weighting idiosyncratic noise. Consistent with this mechanism, we observe that provider weights vary across industries in ways that align with relative provider performance, suggesting that the fusion rule adapts to heterogeneity in coverage and expertise.

6.1.2 Prediction Advantages and Mechanism of the GRU-LSTM Model

The GRU–LSTM architecture is well suited to ESG dynamics that feature both transient shocks and persistent adjustments. The GRU layer emphasizes short-run innovations—such as episodic penalties, accidents, or disclosure events—through gating that selectively resets or updates hidden states. The subsequent LSTM layer retains longer-run strategic trajectories (e.g., decarbonization investment, governance reforms) through its cell state and multi-gate memory. The self-attention component further improves interpretability by highlighting the relative importance of time steps and features, and the estimated attention weights indicate that governance signals carry substantial predictive content in our setting.

6.1.3 Industry Differences in Rating Consistency and Policy Implications

Industry variation in CI appears to be driven by differences in regulatory intensity and the maturity of disclosure infrastructures. Highly regulated sectors (e.g., financial services) typically face more standardized reporting expectations and supervisory scrutiny, which reduces informational gaps and narrows cross-provider dispersion. By contrast, disclosure-scarce and fragmented sectors show larger dispersion, consistent with higher measurement error and greater scope for selective reporting. These patterns suggest that improvements in consistency are unlikely to arise solely from market discipline; rather, they depend on institutionalized disclosure standards and greater transparency in rating methodologies.

6.2 Policy and Practical Recommendations

6.2.1 For Investors

(1) Adopt fused ratings and CI dual screening: Prioritize allocating to enterprises with a dynamic fusion score ≥ 7.5 and CI ≥ 0.7, and limit the portfolio weight of companies with CI < 0.7 to less than 5% to reduce rating divergence risks.

(2) Industry-differentiated allocation: In low-CI industries such as agriculture, forestry, animal husbandry, and fishery, reduce non-systematic risks through allocating ESG hedge funds or diversified investments; in high-CI industries such as finance and information technology, increase the allocation ratio of core assets.

(3) Use dynamic prediction tools: Adjust long-term positions based on the prediction results of the GRU-LSTM model. For example, for enterprises with a predicted 3-year ESG score increase ≥ 1 point (such as a new energy enterprise), extend the holding period to obtain valuation premiums.

The findings suggest that investors should move beyond single-provider ESG ratings and incorporate both fused scores and consistency metrics into decision-making. Doing so can improve risk assessment and portfolio construction.

6.2.2 For Firms

For firms, improving disclosure quality can reduce rating divergence and lower the cost of capital.

(1) Improve information disclosure quality and transparency: Focus on standardizing the disclosure of governance dimension indicators (such as board independence and ESG committee operation), and proactively disclose segmented data such as supply chain management and carbon emission reduction targets. A manufacturing enterprise increased its CI by 0.12 through standardized disclosure.

(2) Establish dynamic ESG management systems: Use the GRU-LSTM model to predict potential risk points (such as environmental dimension deductions caused by policy tightening) and adjust strategic layouts in advance. A chemical enterprise avoided environmental penalties through this method, increasing its ESG score by 0.8 points.

(3) Strengthen communication with rating agencies: Gain a deep understanding of the rating logic of each institution and optimize management strategies for indicators with higher weights (such as the environmental dimension of SynTao Green Finance). A photovoltaic enterprise adjusted the focus of indicator disclosure through communication, raising its rating from BB to A.

6.2.3 For Regulators

For regulators, promoting standardization and transparency can enhance the informational efficiency of ESG markets.

(1) Promote standardization of ESG information disclosure: Formulate unified disclosure guidelines, clarify the definition and calculation methods of core indicators (such as carbon emission intensity and employee welfare coverage), and require enterprises to disclose ESG management frameworks and risk response mechanisms to reduce information asymmetry.

(2) Guide rating agencies to improve transparency: Require rating agencies to disclose indicator systems, weight allocation, and methodologies, establish a rating result review mechanism, and encourage data sharing and cooperation among institutions to reduce the root causes of divergence.

(3) Support fintech innovation applications: Incorporate deep learning models such as GRU-LSTM and AI data completion technology into regulatory sandbox pilots, and promote their application in ESG rating and investment decision-making. A pilot region increased investment decision-making efficiency by 40% through AI tools.

6.3 Conclusions and Outlook

This study develops a comprehensive framework for measuring, integrating, and predicting ESG performance in the A-share market. By transforming rating divergence from noise into an object of optimization, it contributes to the literature on ESG measurement, sustainable finance, and applied econometrics.

6.3.1 Research Conclusions

Based on 2018-2023 data from three institutions (Huazheng, SynTao Green Finance, and Menlang), this study constructs a three-dimensional dynamic ESG evaluation system of "divergence-fusion-prediction" and draws the following core conclusions:

(1) The dynamic weighted fusion algorithm (Formulas 1-3) can effectively integrate multi-source data, reducing MSE by 27.2% compared with single agencies and improving the comprehensiveness and stability of evaluation.

(2) The GRU-LSTM hybrid model (Formulas 4-16) has both short-term fluctuation capture and long-term trend memory capabilities, achieving a 3-year ESG performance prediction R^2 of 0.81 for A-share companies, significantly outperforming traditional models.

(3) The ESG Rating Consistency Index (CI, Formulas 17-18) can effectively quantify divergence, with significant inter-industry differences (financial industry 0.88 vs. agriculture, forestry, animal husbandry, and fishery industry 0.59), and a significant negative correlation with stock volatility.

(4) The dynamic weighted fusion strategy (Formula 21) can increase the average CI by 0.15,

providing a feasible path for optimizing rating consistency.

6.3.2 Academic Contributions

(1) Theoretical Contribution: Pioneers the three-dimensional analytical framework of "divergence-fusion-prediction," transforming rating divergence from research noise into an optimization object, enriching the theoretical system of ESG evaluation, and providing a new paradigm for the application of multi-source data fusion in sustainable finance.

(2) Methodological Contribution: Systematically derives the mathematical expressions of the GRU-LSTM hybrid model and CI index, supplements the rigorous derivation of heteroskedasticity tests and endogeneity treatment, and solves the key technical problems of dynamic ESG prediction and divergence quantification. The CI index has been incorporated into the ESG product screening system of a leading public fund, verifying the practicality of the methodology.

(3) Empirical Contribution: Reveals the industry characteristics and driving factors of ESG rating divergence in the A-share market, provides empirical evidence for understanding the development laws of ESG in emerging markets, and fills the gap in existing research's insufficient attention to the Chinese market.

6.3.3 Research Limitations and Future Outlook

(1) Research Limitations: At the data level, it relies on public rating data and does not obtain internal indicator weights and calculation details of institutions, limiting the in-depth exploration of the root causes of divergence; at the model level, the "black box" nature of deep learning makes some prediction logic difficult to fully explain; at the sample level, it focuses on the A-share market, and the generalizability of conclusions needs to be further verified in markets such as Hong Kong stocks and US stocks.

(2) Future Outlook: In terms of data, alternative data such as satellite remote sensing (nighttime light intensity) and social media public opinion can be integrated to construct a more comprehensive evaluation system; in terms of models, combine XAI technology (such as SHAP values) to improve the interpretability of the GRU-LSTM model and enhance market acceptance; in terms of research perspectives, explore the threshold effect between ESG and financial performance, the impact of carbon tariffs on ESG performance, and conduct cross-country comparative studies of ESG evaluation systems between China, Europe, and the United States to provide Chinese experience for global ESG standardization.

References:

[1] Angrist, J. D., & Pischke, J. S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton University Press.

[2] Berg, F., Kölbel, J. F., & Rigobon, R. (2020). Aggregate confusion: The divergence o f ESG ratings. Journal of Financial Economics, 135(2), 401-421.

[3] Bhattacharya, I., Husted, B. W., & Wang, H. (2023). ESG ratings: Whom do they ser ve? Strategic Management Journal, 44(3), 1123-1149.

[4] Choudhury, P., Ehlers, T., & Jiang, H. (2022). ESG rating disagreement and stock ret urns. Review of Financial Studies, 35(7), 3124-3171.

[5] Eccles, R. G., Krzus, M. P., & Serafeim, G. (2021). The role of ESG ratings in capit al markets. Harvard Business Review, 99(3), 116-124.

[6] Gibson, R., Krueger, P., & Schmidt, F. (2023). ESG rating divergence and institutional

investor behavior. Journal of Banking & Finance, 147, 106789.
[7] International Organization of Securities Commissions (IOSCO). (2022). ESG ratings an d data providers: Final report. IOSCO.
[8] Krueger, P., Sautner, Z., & Starks, L. T. (2022). The pricing of ESG risks: Evidence from catastrophe bonds. Journal of Financial and Quantitative Analysis, 57(2), 789-82 1.
[9] MSCI. (2021). MSCI ESG ratings methodology. MSCI Inc.
[10] Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regressio n. In D. W. K. Andrews & J. H. Stock (Eds.), Identification and inference for econo metric models (pp. 80-108). Cambridge University Press.
[11] Angrist, J. D., & Pischke, J. S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton University Press.
[12] Berg, F., Kölbel, J. F., & Rigobon, R. (2020). Aggregate confusion: The divergence o f ESG ratings. Journal of Financial Economics, 135(2), 401–421.
[13] Choudhury, P., Ehlers, T., & Jiang, H. (2022). ESG rating disagreement and stock ret urns. Review of Financial Studies, 35(7), 3124–3171.
[14] Gibson, R., Krueger, P., & Schmidt, F. (2023). ESG rating divergence and institutional investor behavior. Journal of Banking & Finance, 147, 106789.
[15] Krueger, P., Sautner, Z., & Starks, L. T. (2022). The pricing of ESG risks. Journal of Financial and Quantitative Analysis, 57(2), 789–821.
[16] Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regressio n. In D. W. K. Andrews & J. H. Stock (Eds.), Identification and inference for econo metric models (pp. 80–108). Cambridge University Press.

Appendix A. Variable Definitions

Variable	Definition	Source
ESG	Dynamic fused ESG score	Huazheng, SynTao, Menlang
CI	ESG Rating Consistency Index	Authors' calculation
Volatility	Annualized stock return volatility	CSMAR