Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Forecasting Corporate ESG Performance Based onExplainable Wide and Deep Forest

Yunduo Han
Hangzhou Dianzi University
Wanan Liu
wanan@hdu.edu.cn
Hangzhou Dianzi University
Ming Qian
Wuhan University
Xingyu Lan
Donghua University
Congyuan Pang
Donghua University
Naixi Chen
Liaoning Shihua University
Yao zou
Huainan Normal University
Feifei Wang
Hangzhou Dianzi University
Keyi Zhu
Donghua University

Research Article

Keywords: ESG rating, Wide and Deep Forest, Performance-interpretability dilemma

Posted Date: June 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9747080/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Forecasting Corporate ESG Performance Based on Explainable Wide and Deep Forest

Yunduo Han[1], Wanan Liu[2*], Ming Qian[3], Xingyu Lan[4], Congyuan Pang[4], Naixi Chen[5], Yao Zou[6], Feifei Wang[2], Keyi Zhu[4]

[1]School of Economics, Hangzhou Dianzi University, Hangzhou, 310018, China.
[2]School of Management, Hangzhou Dianzi University, Hangzhou, 310018, China.
[3]State Key Laboratory LIESMARS, Wuhan University, Wuhan, 430072, China.
[4]Glorious Sun School of Business and Management, Donghua University, Shanghai, 200051, China.
[5]School of Economics and Management, Liaoning Petrochemical University, Fushun, 113001, China.
[6]School of Economics and Management, Huainan Normal University, Huainan, 232038, China.

Abstract

Environmental, Social, and Governance (ESG) ratings serve as a fundamental component of sustainable finance; however, their credibility is substantially compromised by persistent methodological limitations, including the inherent rating performance-interpretability dilemma, lack of procedural transparency, and inconsistencies across rating agencies due to rating subjectivity. To this end, we introduce the Wide and Deep Forest (WDForest), a new objective ESG rating framework developed to enhance both interpretability and objectivity through structural decoupling of the rating process into an intra-layer "wide" enhancement module and a cross-layer "deep" cascade architecture. Depth is achieved via a sequential residual calibration mechanism, wherein each layer iteratively learns and rectifies unexplained residuals from its predecessor, while width is expanded through parallel boosted trees, which collectively capture intricate data patterns to strengthen the model's expressive power. Unlike conventional ESG rating models with fixed structures, WDForest employs a validation-based early stopping

*Corresponding author(s). E-mail(s): Wanan Liu, wanan@hdu.edu.cn. Contributing author(s) E-mail(s): Yunduo Han, yunduohan@hdu.edu.cn; Ming Qian, mingqian@whu.edu.cn; Xingyu Lan, lanxingyu@mail.dhu.edu.cn; Congyuan Pang, pcy@mail.dhu.edu.cn; Naixi Chen, chennaixi@lnpu.edu.cn; Yao Zou, zouyaodhu@mail.dhu.edu.cn; Feifei Wang, wangfei518@hdu.edu.cn; Keyi Zhu, 2231340@mail.dhu.edu.cn.

mechanism that directly responds to data complexity, WDForest autonomously determines its optimal architectural depth; this dynamic self-adaptation makes it a highly accurate and scalable solution for ESG rating. Comprehensive empirical evaluations across environmental, social, governance, and aggregated ESG domains demonstrate that WDForest consistently surpasses benchmark models such as XGBoost and random forests across key metrics, including mean squared error (MSE), coefficient of determination ($\boldsymbol{R^2}$), and Spearman rank correlation (SRCC), confirming superior rating performance and robustness. Furthermore, the hierarchical architecture of WDForest inherently supports multi-faceted interpretability, enabling global feature importance analysis, partial dependence curves, and interaction visualizations, which collectively mitigate the black-box constraints typical of conventional machine learning approaches. By unifying high predictive performance with transparent and traceable decision logic, WDForest offers a reliable analytical tool for investors, corporate managers, and policymakers involved in sustainable investment decision-making.

Keywords: ESG rating, Wide and Deep Forest, Performance-interpretability dilemma

1 Introduction

ESG ratings have become a central metric for evaluating corporate non-financial performance and play an increasingly critical role in sustainable finance. As a supplement to conventional financial analysis, these ratings are widely employed by investors to assess long-term risks and sustainability prospects, with surveys confirming that a large majority of institutional investors systematically integrate ESG scores into their decision-making processes (Amel-Zadeh and Serafeim 2018). By incorporating environmental, social, and governance dimensions typically absent from traditional financial reports, credible ESG ratings contribute to improved portfolio risk management and support the growth of socially responsible investing.

However, mainstream ESG ratings have long been questioned for their low reproducibility and significant divergence across agencies, owing to methodological dependence on expert judgment and subjective interpretation (Berg et al. 2022). This evaluation system generally involves agencies independently defining assessment frameworks and indicator weights, leading to an opaque, black-box-like decision process that lacks transparency and common standards. Compounding this issue, corporate sustainability data are characterized by disparate sources, inconsistent metrics, and voluntary disclosure practices, yielding fragmented and highly variable input quality (Christensen et al. 2022). These factors frequently result in conflicting ratings for the same entity across providers, substantially impairing the credibility and practical utility of ESG assessments. In response to these challenges, a growing body of research has advocated for the development of data-driven, objective ESG rating methodologies that leverage quantitative modeling and machine learning techniques to mitigate subjectivity and enhance reproducibility. These approaches aim to extract meaningful sustainability signals from corporate data,by employing transparent, algorithmically determined feature weights and interaction patterns (Gibson Brandon et al. 2021).

By reducing reliance on expert-defined frameworks and incorporating automated, reproducible evaluation processes, objective rating systems offer the potential to establish greater consistency, comparability, and methodological transparency across ESG assessments, thereby facilitating more reliable cross-firm and cross-sector evaluations (Khan et al. 2016).

Although data-driven ESG rating methodologies have substantially improved objectivity and reproducibility, their dependence on machine learning, particularly on complex models such as deep learning, often compromises interpretability and obscures the rationale behind rating decisions (Stocco 2023; Suarez Gifi 2024). This lack of transparency creates a "black box" problem that limits the ability of investors, firms, and regulators to identify the concrete factors influencing an ESG score, ultimately reducing the method's credibility and practical value. Even when such models attain high rating performance by capturing nonlinear or high-dimensional feature relationships, they typically fail to provide granular, actionable insight into a company's specific environmental, social, or governance strengths or weaknesses, thereby restricting their utility in sustainable investment and policy formulation.

Post-hoc interpretability techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are widely used to provide retrospective explanations for pretrained black-box models (Salih et al. 2023; Ribeiro et al. 2016; Lundberg and Lee 2017). They estimate feature importance and furnish local approximations of model behavior, enabling practical insight without altering the original model. However, these explanations remain external approximations rather than native reasoning: they can be unstable under correlated features or distribution shifts, incur nontrivial computational overhead at ESG-rating scale, and do not faithfully expose the mechanisms the model actually uses, risking misinterpretation. These limitations have encouraged inherently interpretable that embed transparency into the architecture,seeking to retain predictive performance while tying each output to traceable, semantically meaningful rationales. However, in the context of objective ESG rating, where accurate risk assessment depends on capturing complex, non-linear interactions across environmental, social, and governance indicators, these methods often yield oversimplified explanations. They commonly fail to represent high-order feature interdependencies, lack granular pillar-level attribution, and offer limited support for audit-critical requirements such as uncertainty quantification and temporal reproducibility. These gaps call for an approach that couples strong performance with structured, ESG-aware explanations.

To address the above limitations, we propose a Wide-and-Deep Forest (WDForest) for interpretable objective ESG rating. WDForest employs a residual cascade architecture with dynamic depth optimization, allowing it to adapt structurally to varying data complexity and maintain robustness across rating contexts. Each layer acts as a regularized boosting ensemble that iteratively refines feature representations and learns residual patterns not captured by prior layers, thereby progressively modeling high-order nonlinear interactions across ESG dimensions. Moreover, the model's inherently transparent, additive structure supports both global and local interpretability,

enabling clear attribution of the final rating to individual features and their interactions. By delivering not only high accuracy but also traceable reasoning, WDForest offers actionable insights into the drivers of corporate sustainability performance.

The contribution of this work is organized as follows:

(1) WDForest introduces a dynamic depth optimization mechanism that autonomously calibrates model complexity in response to ESG data characteristics. This adaptive growth algorithm progressively allocates forest depth based on feature dimensionality and data scale, enabling scalability across rating scenarios of varying scale without manual reconfiguration. The self-adaptation balances expressive capacity for capturing sustainability signals with robustness against overfitting, ensuring reliable performance across diverse sectors and reporting frameworks.

(2) Building upon this adaptive foundation, WDForest introduces a dual-pathway architecture that synergistically combines wide and deep components. The wide component, through intra-layer regularized parallel boosting, effectively captures broad feature interactions across ESG dimensions. Meanwhile, the deep component leverages inter-layer residual learning to hierarchically refine complex nonlinear representations. This cooperative design enables the model to jointly learn both memorization and generalization for robust predictive performance.

(3) WDForest's cascade structure motivates its additive interpretability by supporting both global identification of material ESG drivers and local decomposition of firm-level scores into transparent main and interaction effects. This resolves the "black box" limitations of conventional machine learning in ESG assessment, offering stakeholders traceable decision support alongside high rating performance, which is a crucial advancement for regulatory compliance and investor communications.

The paper is structured as follows. Section 2 establishes the theoretical foundation through a review of ESG rating methodologies and interpretable machine learning. Section 3 introduces the WDForest framework, detailing its architecture and learning algorithm. Section 4 describes the experimental setup, including datasets, benchmarks, metrics, and implementation. Section 5 presents empirical results, beginning with comparative performance analysis and followed by a multidimensional interpretability assessment at global, local, and interaction levels. Finally, Section 6 synthesizes key findings, discusses implications, acknowledges limitations, and suggests future research directions.

2 Related Work

This section presents a comprehensive review of the literature relevant to interpretable objective ESG rating methods, structured around three interconnected works. We begin by establishing the core problem through an examination of the historical and persistent issues in traditional ESG ratings, particularly rating divergence, methodological opacity, and their adverse market effects. We then review machine learning algorithms as a promising solution. Finally, we address the essential demand for interpretability in sustainable finance, critically assessing existing Explainable AI (XAI) methods and justifying the need for an inherently interpretable model design.

2.1 Subjective ESG rating methods

The integration of ESG criteria into investment analysis has shifted from a niche, values-based screening method to a mainstream financial strategy influencing trillions in assets under management. This shift has spurred the rapid expansion of ESG rating agencies, with over 70 providers worldwide supplying various forms of ESG scoring data. These agencies play a critical role by processing extensive corporate information to deliver standardized assessments that support investor decision-making in portfolio selection, risk management, and corporate engagement (Amel-Zadeh and Serafeim 2018).

However, rather than yielding a clear and consistent measure of corporate sustainability, this proliferation has resulted in a complex and often contradictory informational landscape, widely recognized as ESG rating divergence (Berg et al. 2022). Substantial academic research confirms that ratings for the same company often vary considerably across agencies, a situation that should be understood as a fundamental feature of the current ESG ecosystem rather than a minor statistical anomaly (Christensen et al. 2022).For instance, Berg et al. (2022) analyzed major rating providers, including KLD, Sustainalytics, and MSCI, and found a marked divergence in their pairwise correlations. This contrasts sharply with the remarkable consensus observed in credit ratings among agencies such as Moody's and S&P. Such divergence, commonly quantified using correlation coefficients, reveals a persistent lack of consensus and poses substantial challenges for investors reconciling conflicting signals and companies seeking to improve their sustainability profiles (Berg et al. 2022).

The academic consensus, largely shaped by the quantitative framework of Berg (Berg et al. 2022), attributes ESG rating divergence to three interconnected sources. Scope divergence stems from fundamental disagreements among agencies regarding which attributes are material under the E, S, and G pillars (Kotsantonis and Serafeim 2019). Measurement divergence, identified as the primary source of inconsistency, arises when agencies employ distinct indicators and sources to assess the same attribute (Berg et al. 2022). In contrast, weight divergence represents a far more limited source of variance, reflecting subjective judgments in aggregating indicators into composite scores (Berg et al. 2022). These drivers collectively reflect deeper systemic issues, including insufficient standardization, methodological opacity, and data reliability concerns (Amel-Zadeh and Serafeim 2018). Moreover, unlike traditional financial metrics, ESG factors tend to be qualitative and intangible, rendering them inherently resistant to objective measurement (Cort and Esty 2020).

This environment has fostered a landscape in which rating agencies develop proprietary methodologies, thereby preventing external stakeholders from comprehensively understanding what is being measured and how (Chatterji et al. 2009). Such methodological subjectivity is further compounded by severe data limitations, as a substantial portion of input data originates from unstandardized, self-reported corporate disclosures that are frequently incomplete, unaudited, or based on estimates rather than verified facts (Hughes et al. 2021).

The resulting absence of a clear and consistent ESG signal carries significant adverse economic implications. This divergence introduces substantial noise and uncertainty, which in turn influences corporate behavior and investor decision-making.

Empirical studies indicate that it elevates stock price volatility and information asymmetry (Di Giacomo et al. 2021), increases firms' cost of capital and suppresses stock liquidity (Li and Yang 2024), and distorts market valuations (Dorfleitner et al. 2015). Moreover, the lack of a consistent benchmark can generate perverse incentives, encouraging certain firms to resort to "greenwashing" or superficial compliance. This behavior raises agency costs and undermines the credibility of corporate sustainability claims (Hu et al. 2023).

2.2 Objective ESG rating approaches

In response to the widespread subjectivity of conventional ESG ratings and their adverse market effects, researchers have increasingly adopted machine learning (ML) to build more data-driven and reproducible objective ESG rating frameworks (D'Amato et al. 2022). These approaches employ algorithmic solutions to identify complex patterns in large datasets, complementing or replacing manual analysis. Below, we summarize ML-based ESG rating methods in two parts: individual machine learning-based ESG rating methods and ensemble methods.

2.2.1 Individual ML-based ESG rating approaches

This track centers on standalone models that learn ESG scoring functions directly from structured and unstructured inputs, highlighting semantic context and stable nonlinear relationships. To mitigate opacity and reliance on voluntary disclosures, NLP techniques are used to analyze corporate sustainability reports and news articles (Siering 2023). Transformer-based architectures such as BERT, when fine-tuned for ESG-specific tasks, move beyond keyword counting to provide contextual and sentiment-aware assessments, yielding more nuanced and potentially less biased evaluations (Lee et al. 2022). When the feature space is moderate and the objective is to capture semantic context or smooth nonlinear relationships while avoiding large ensembles, single-model solutions offer favorable efficiency and controllability.

2.2.2 Ensemble-based methods

For tabular and mixed-type ESG data, tree-based ensembles have become the dominant predictive paradigm, explicitly targeting complex interactions and strong nonlinear effects. Random Forest (Breiman 2001; D'Amato et al. 2022) and Gradient Boosting (Friedman 2001; Zheng et al. 2023), with advanced variants such as XGBoost (Chen and Guestrin 2016; Abdullah et al. 2023) and LightGBM (Ke et al. 2017; Kim and Kim 2023), typically outperform traditional linear baselines. Their advantages stem from robust handling of features and missing values, effective modeling of interaction structures, and regularization, column sampling, and early-stopping mechanisms that promote strong generalization (Abdullah et al. 2023). Consequently, ensemble methods are widely used as powerful baselines or production-grade models to enhance performance and stability in practical rating systems.

2.3 Interpretable machine learning

Although predictive performance serves as a primary driver for machine learning adoption, it is insufficient in high-stakes financial domains where model output must be trusted and verifiable to ensure practical utility (Babaei et al. 2025; Solanke 2023). ESG rating models in particular must provide transparent decision logic to meet the needs of diverse stakeholders, including investors, regulators, and corporations, who require interpretable outputs to make informed decisions, verify fairness, and comply with regulatory standards.

The field of XAI comprises two principal methodological paradigms for rendering machine learning models interpretable: post-hoc explanation techniques, which apply model-agnostic analyses to elucidate pre-trained black-box systems, and intrinsically interpretable approaches, which employ self-explaining glass-box models whose internal structure and decision logic constitute the explanation itself. (Arrieta et al. 2020; Rudin 2019). The growing empirical evidence reveals fundamental limitations in their reliability (Ribeiro et al. 2016; Lundberg and Lee 2017; Alqaraawi et al. 2023).The persistent limitations of post-hoc interpretability methods, including their inherent instability, potential for explanatory misinterpretation, and disconnection from underlying model mechanisms, reveal a fundamental need for new approaches to trustworthy machine learning. (Rudin 2019; Alqaraawi et al. 2023).

These challenges collectively underscore the theoretical and practical necessity of developing models with intrinsic interpretability, wherein predictive accuracy and decision transparency are jointly optimized within a unified architectural framework. This imperative has catalyzed significant advances in interpretable machine learning research, shifting the paradigm from explaining pre-existing black-box systems to constructing models that are inherently self-explanatory through their structure and parameters. Such architectures ensure that explanations are faithful, consistent, and semantically aligned with the actual reasoning process, thereby addressing critical requirements in high-stakes domains such as healthcare, finance, and autonomous systems where auditability and accountability are paramount. Generalized Additive Models (GAMs), pioneered by Hastie and Tibshirani (Hastie and Tibshirani 1986), constitute a powerful class of glass-box models that preserve interpretability by extending linear models with non-linear feature relationships while maintaining a strictly additive structure. This structure enables precise understanding of each feature's effect through visualization of its univariate shape function (Lou et al. 2013). Modern extensions have substantially narrowed the performance gap with black-box models, including tree-based ensembles like Explainable Boosting Machines (EBMs) (Nori et al. 2021) and deep learning approaches such as Neural Additive Models (NAMs) (Agarwal et al. 2021), which leverage separate neural networks to learn individual shape functions while preserving the additive framework's intelligibility (Agarwal et al. 2021).

3 Methodology

3.1 Overview of WDForest

In response to the modeling challenges posed by complex interactions among multidimensional corporate indicators, pronounced structural heterogeneity, and significant inter-sample variability in ESG ratings, this study proposes WDForest algorithm. The method is motivated by the limitations of conventional ensemble models in capturing highly nonlinear and hierarchically structured dependency patterns inherent in ESG assessments. To address this, we design a multilayer residual cascade architecture with an adaptive depth control mechanism, enabling multigranular and self-adjusting modeling of rating determinants. By extending the deep forest's hierarchical structuring to ESG-type prediction tasks and integrating dynamic early stopping with incremental ensemble learning, WDForest constitutes a mathematically rigorous, data-driven analytical framework. Its rating function can be formulated as:

$$f\left(x\right)=\sum_{l=1}^{L^{*}}g^{(l)}\left(x^{(l)}\right), \tag{1}$$

where L^{*} is the optimal cascade depth dynamically determined by validation set performance, $g^{(l)}\left(\cdot\right)$ is the ensemble prediction function for layer l, and $x^{(l)}$ is the feature-reconstructed for l-th layer. This architecture degenerates to a standard forest when $L^{*}=1$, but forms a deep augmented structure using the residual learning paradigm when $L^{*}>1$.

Figure 1 shows that WDForest through layer-by-layer feature reconstruction and integrated prediction, effectively identifying the complex nonlinear dependencies between a company's environmental, social, and governance dimensions. The residual connection mechanism deepens WDForest while ensuring efficient transmission of gradient information, enhancing its adaptability to company samples. The overall structure maintains broad perception at shallow levels (Wide) while achieving fine-grained feature refinement at deeper levels (Deep), thereby improving the robustness and interpretability of rating without over-reliance on pre-set assumptions.

3.2 Adaptive cascade-depth optimization

To address the uncertainty in modeling complexity caused by industry characteristics, enterprise size, and regional differences in ESG ratings, WDForest algorithm introduces dynamic depth control and early stopping mechanisms to achieve adaptive integrated architecture optimization. This mechanism effectively balances model expressiveness and generalization risk by monitoring validation set performance changes as the layers grow. Assume the validation-set MSE after the l-th cascade is:

$$V^{(l)}=\frac{1}{N_{\mathrm{val}}}\sum_{i=1}^{N_{\mathrm{val}}}\left(y_{i}-\sum_{k=1}^{l}g^{(k)}\left(x_{i}^{(k)}\right)\right)^{2}, \tag{2}$$

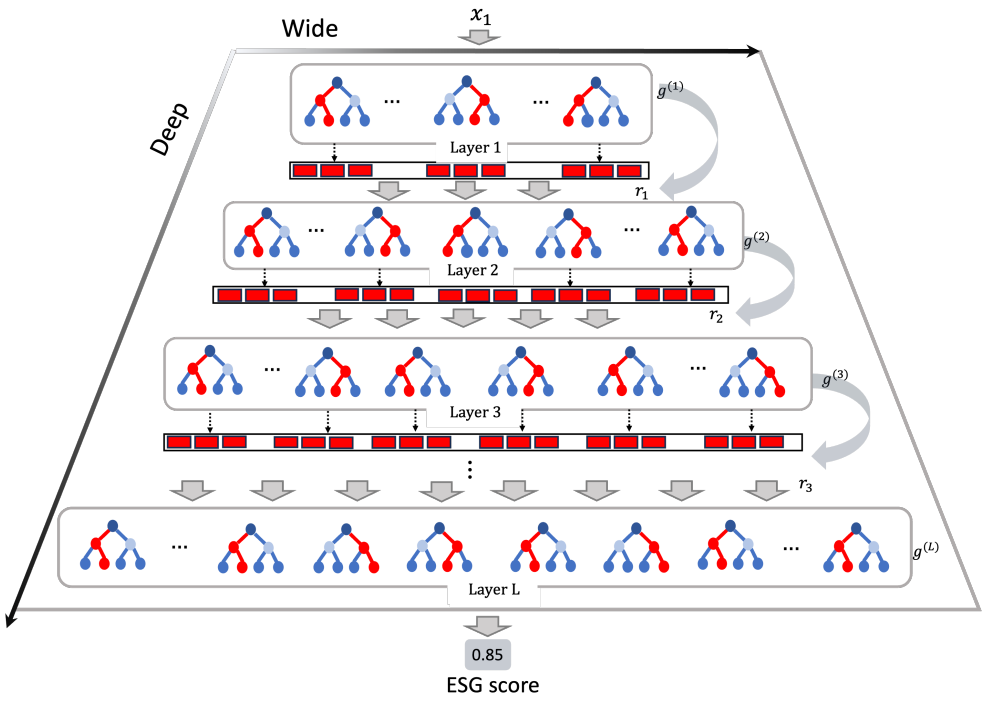


Fig. 1: Illustration of WDForest

the early stopping condition is defined as the existence of $V^{(l+\delta)} > V^{(l)}$ for τ consecutive occurrences with $\delta \in [1, \tau]$, at which point training is terminated and the optimal depth is taken as:

$$L^* = \underset{l \in [1,\, L_{\max}]}{\arg\min}\ V^{(l)}. \tag{3}$$

This mechanism is equivalent to solving the following optimization problem:

$$\min_{L^*} \mathbb{E}_{(x,y)\sim P_{\text{val}}} \left[L\left(y, \sum_{l=1}^{L^*} g^{(l)}\left(x^{(l)}\right) \right) \right], \tag{4}$$

where P_{val} is the validation set data distribution. This dynamic depth control allows WDForest to adapt to data complexity: for data dominated by simple linear relationships, $L^* = 1$; for data with complex interactions (such as the cross-cutting impact of governance and innovation in technology companies), multi-layer cascades are automatically triggered ($L^* \geq 2$).

3.3 Progressive feature learning based on residual propagation

Due to the complex nonlinear dependencies between different companies' environmental, social, and governance indicators, and the highly contributions of each

dimension to the final score, traditional single-level models often struggle to fully capture these multi-layered influencing mechanisms. To address this, WDForest incorporates a feature space reconstruction and residual propagation mechanism. By gradually enhancing feature representation and iteratively correcting prediction errors, it achieves a progressive approximation of the ESG scoring mechanism.

Define a feature mapping function $\phi^{(l)} : \mathbb{R}^P \to \mathbb{R}^{P+1}$, which recursively constructs the ESG indicators for each layer. Specifically, for layer l, the ESG indicators for sample i, denoted $x_i^{(l)}$, can be expressed as:

$$x_i^{(l)} = \begin{cases} x_i, & l = 1 \\ \phi^{(l-1)}(x_i), & l \geq 2 \end{cases}, \tag{5}$$

where $x_i \in \mathbb{R}^P$ represents the original ESG feature vector of sample i, and P is the feature dimension. When $l \geq 2$, $\phi^{(l-1)}$ represents the feature mapping function for layer $l-1$, which expands the low-dimensional feature space to a higher dimension to accommodate the new representations learned by the previous layers. The residual term $r_i^{(l-1)}$ is defined as the difference between the current cumulative forecast and the true ESG score y_i:

$$r_i^{(l-1)} = y_i - \sum_{k=1}^{l-1} g^{(k)}\left(x_i^{(k)}\right), \tag{6}$$

where $g^{(k)}$ is the ensemble prediction function at the k-th level. This residual reflects the portion of score variation unexplained by the first $l-1$ layers and serves as the learning target for the l-th level. By fitting the residuals of the previous stage layer by layer, WDForest decomposes the complex ESG scoring problem into a series of subproblems, sequentially learning the underlying patterns not captured at each level.

From the perspective of the function space, this process constructs a nested sequence of subspaces $\mathcal{H}^{(1)} \subset \mathcal{H}^{(2)} \subset \cdots \subset \mathcal{H}^{(L^*)}$, where the dimension of each subspace $\mathcal{H}^{(l)}$ satisfies:

$$dim\left(\mathcal{H}^{(l)}\right) = dim\left(\mathcal{H}^{(l-1)}\right) + 1. \tag{7}$$

This structure corresponds to the gradual expansion of a series of orthogonal basis functions in the Reproducing Kernel Hilbert Space (RKHS) (Aronszajn 1950), where each layer of basis functions focuses on fitting the orthogonal components spanned by the prediction residuals of the previous layer (Friedman 2001). In the context of ESG ratings, this means that WDForest can gradually identify and separate influencing factors at different levels: the shallow layer may capture the linear impact of basic indicators, while the deep layer focuses on learning high-order interaction effects, thus mathematically ensuring the multi-scale analysis capability of the scoring formation mechanism. The mathematical essence of residual propagation is to construct a series of orthogonal basis functions in RKHS, where each layer focuses on learning the residual components that the previous layer cannot explain.

3.4 Hierarchical boosting ensemble and regularization

To address common issues such as noise, multicollinearity, and skewed sample distributions in ESG rating data, WDForest algorithm introduces an ensemble learner based on the XGBoost framework in each cascade structure, supplemented by strict regularization control to ensure that WDForest maintains robust generalization performance as it deepens layer by layer. Specifically, the prediction function of the integrator at layer l is defined as:

$$g^{(l)}(x) = \sum_{t=1}^{T^{(l)}} \eta \cdot h_t^{(l)}(x). \tag{8}$$

Here, $T^{(l)} = T^{(l-1)} + \Delta$ with $\Delta \in \mathbb{Z}^+$ indicates that the number of CART trees used in this layer increases with each level to gradually enhance WDForest's expressive power; $h_t^{(l)}$ is the CART tree generated in round t, and η is the learning rate. This is used to control the contribution strength of each tree to avoid overly aggressive single-layer fitting. This design enables WDForest to adapt to various local patterns by gradually increasing complexity when addressing structural differences in ESG indicators across different companies. To further enhance the robustness of the tree model, the gain function used when splitting nodes is given the following regularization form:

$$\mathcal{G}(S) = \frac{1}{2}\left[\frac{\| \mathbf{G}_L \|_2^2}{\mathbf{H}_L + \lambda} + \frac{\| \mathbf{G}_R \|_2^2}{\mathbf{H}_R + \lambda} - \frac{\| \mathbf{G}_L + \mathbf{G}_R \|_2^2}{\mathbf{H}_L + \mathbf{H}_R + \lambda}\right] - \gamma, \tag{9}$$

where $\mathbf{G}_L = \sum_{i \in I_L} g_i$ and $\mathbf{H}_L = \sum_{i \in I_L} h_i$ represent the sum of the gradients and the sum of the Hessians of the samples in the left child node, respectively. The same applies to $\mathbf{G}_R$ and $\mathbf{H}_R$. λ is the penalty coefficient used to suppress excessive fluctuations in the node output value, and γ is the splitting threshold parameter. Mathematically, this gain function has the properties of γ-strong convexity and λ-smoothness, which can ensure that WDForest converges to the global optimal solution in each splitting step, effectively avoiding the local optimal trap caused by the inherent noise or sparsity of the ESG indicator.

From the perspective of statistical learning theory, the generalization error of this integrated structure is constrained by the following bounds (Vapnik 1998):

$$R\left(g^{(l)}\right) \leq \hat{R}_S\left(g^{(l)}\right) + \sqrt{\frac{T^{(l)} \cdot \log N + \log\left(\frac{1}{\delta}\right)}{2N}}, \tag{10}$$

where $R\left(g^{(l)}\right)$is the expected risk, $\hat{R}_S\left(g^{(l)}\right)$ is the empirical risk, N is the number of training samples, and N is the confidence parameter. By controlling the number of trees per layer, $T^{(l)}$, to grow at a rate of $O\left(\sqrt{L^* \cdot T_{max}}\right)$, the generalization error bound of the overall model can be controlled within a certain range.

By controlling the growth rate of $T^{(l)}$, the overall error bound maintains a convergence rate of $O\left(\sqrt{L^* \cdot T_{max} \cdot N^{-1/2}}\right)$. This theoretical guarantee ensures that WDForest model maintains high predictive stability and interpretability even with limited and highly corporate ESG data, providing a reliable computational foundation for rating analysis.

3.5 Complexity analysis

To assess the theoretical generalization capacity and structural complexity of WDForest algorithm in the context of ESG ratings, a rigorous analysis of its representational learning capabilities is essential. Given that ESG data inherently integrates continuous environmental indicators, discrete governance variables, and high-dimensional social features, the richness and controllability of WDForest's hypothesis space become critical for balancing expressive power against overfitting risks.

Let $F^{(L)}$ denote the cascade hypothesis space with a total number of layers L. Its VC dimension (Vapnik-Chervonenkis dimension) is used to characterize the theoretical expressive power of WDForest and satisfies the following relationship:

$$VCdim\left(F^{(L)}\right) = O\left(\sum_{l=1}^{L} T^{(l)} \cdot 2^{d^{(l)}}\right), \tag{11}$$

where $T^{(l)}$ is the number of CART trees integrated in layer l. $d^{(l)}$ is the maximum depth of a single tree in layer l. This equation shows that the overall model complexity accumulates with the number of layers L, and the contribution of each layer depends on the product of the number of trees and its depth. This reflects the mechanism by which WDForest algorithm gradually expands its representational capacity through a multi-layer cascade structure.

In ESG rating practice, if model complexity grows unconstrained, it will easily lead to overfitting on highly corporate data. To this end, WDForest uses an early stopping strategy to dynamically determine the optimal cascade depth L^*, thereby constraining the actual hypothesis space to within the range $F^{(L)}$. This control effectively limits WDForest's VC dimension to $O\left(\sum_{l=1}^{L} T^{(l)} \cdot 2^{d^{(l)}}\right)$, avoiding the situation where excessive representational capacity exceeds the information capacity supported by the data itself. In particular, when the internal rules of ESG data are relatively simple, the early stopping mechanism will prompt L^* to take a smaller value, thereby reducing the complexity term; while for complex rating scenarios involving multi-dimensional nonlinear interactions, WDForest allows L^* to increase moderately to capture deeper feature interaction effects, while still ensuring that complexity is controlled through the upper bound of L^*.

3.6 Interpretability of WDForest

3.6.1 Global interpretability

To provide a transparent and quantitative basis for ESG rating decisions, WDForest employs a feature importance metric grounded in functional variance decomposition. This method quantifies the contribution of each ESG indicator by decomposing the model's prediction function into additive components representing both individual feature main effects and their interaction effects. Specifically, the importance of a feature subset is measured by its portion of the total variance explained, captured through the following conditional expected variance formulation:

$$
\begin{aligned}
I(x_j) &= V_{x\setminus j}\left[\mathbb{E}_{x_j}\left(H_j(x_j) \mid x_{\setminus j}\right)\right] \\
&= \mathbb{E}_{x_j}\left[V_{x\setminus j}\left(H_j(x_j) \mid x_{\setminus j}\right)\right].
\end{aligned} \tag{12}
$$

To quantify the independent influence of each ESG indicator, the cumulative main effect of feature j is defined as $H_j(x_j) = \sum_{l=1}^{L^*}\sum_{t=1}^{T^{(l)}} \eta \cdot h_{t,j}^{(l)}(x_j)$, representing the aggregated predictions from all decision trees associated with feature j across the cascade levels. This formulation reflects that the importance of feature j is determined by the volatility of WDForest's conditional expectation as x_j varies, where x_j denotes the set of all features excluding j. In practical ESG rating contexts, a high conditional expectation variance for a feature indicates its strong independent impact on the final rating outcome, correspondingly yielding a large importance value $I(x_j)$.

3.6.2 ESG feature interaction

To further reveal the synergistic or counteracting effects between ESG indicators, the interaction strength is quantified using the H statistic proposed by Friedman and Popescu (Friedman and Popescu 2008):

$$
I_{jk} = \frac{\mathbb{E}[(H_{jk}(x_j, x_k) - H_j(x_j) - H_k(x_k))^2]}{\mathbb{V}[H_{jk}(x_j, x_k)]}. \tag{13}
$$

Here, $H_{jk}(x_j, x_k)$ denotes the joint effect function of features j and k, and $\mathbb{V}[H_{jk}(x_j, x_k)]$ represents the variance of this joint effect. This statistic quantifies the degree to which the combined influence of two features deviates from the linear addition of their individual main effects, thereby capturing non-additive interaction behavior. For instance, in analyzing the impact of "R&D investment" (feature j) and "board independence" (feature k) on ESG ratings of technology firms, a significantly positive I_{jk} value indicates a meaningful interaction effect, suggesting that the governance structure moderates how innovation activities influence the final rating outcome.

The inherent complexity and nonlinear interactions among corporate environmental (E), social (S), and governance (G) indicators often render global feature importance insufficient for fully explaining individual rating outcomes. To address this, WDForest leverages its explicitly additive structure to provide a hierarchical local interpretability framework, enabling sample-level predictions to be decomposed into traceable per-feature and interaction components. This decomposition permits the identification of key indicators and their interactive effects that drive the rating of a specific company, with local predictions expressed in the following additive form:

$$
f(x_i) = \alpha_0 + \sum_{j=1}^{P} H_j(x_{ij}) + \sum_{1\le p<q\le P} H_{pq}(x_{ip}, x_{iq}). \tag{14}
$$

Here, the global bias term α_0 represents the average rating level across the training set, serving as the baseline prediction. The cumulative main effect function of feature

j, denoted $H_j(x_{ij})$, aggregates the outputs of all tree models associated with this feature throughout the cascade levels, capturing its independent contribution to the enterprise rating. Similarly, the interaction effect function $H_{pq}(x_{ip}, x_{iq})$ quantifies the non-additive, synergistic effect between features p and q, reflecting how their joint presence influences the rating beyond their individual main effects.

The inherent interpretability of WDForest framework stems directly from its forward computation mechanism, which integrates two synergistic processes: a layer-by-layer residual learning architecture decomposes complex prediction tasks into a sequence of additive corrections, with each layer's output $g^{(l)}\left(x^{(l)}\right)$ explicitly modeling the residual from its predecessor; concurrently, the tree-based learners enable granular traceability of feature effects by analyzing the decision path such as the sequence of splitting rules for each sample, thereby allowing precise attribution of predictions to specific features and their interactions.

4 Experimental settings

4.1 ESG Datasets

The analysis employs ESG data sourced from the Refinitiv (2018) database, selected for its global coverage, comprehensive indicator system, and methodologically consistent rating framework. Covering the period from 2002 to 2023, the dataset captures the historical evolution of ESG principles, including key developments such as the adoption of the UN Sustainable Development Goals. Separate prediction tasks are formulated for the Environmental, Social, and Governance pillars, with each pillar's score serving as the prediction target. Following systematic data preprocessing, the final analytical samples comprise 89,066 observations for the E pillar, 108,756 for the S pillar, and 108,808 for the G pillar.

To ensure robust data quality, a rigorous preprocessing pipeline was implemented, involving duplicate removal, elimination of features with excessive missing values ($>70\%$) or near-zero variance, and mitigation of multicollinearity by discarding highly correlated features. Predictive feature subsets were further refined through Lasso regression, while Z-score standardization and normalization were deliberately avoided to preserve native data scales and semantic interpretability. The final datasets were partitioned into fixed training (70%), validation (15%), and testing (15%) sets to ensure reproducibility and rigorous evaluation.

4.2 ESG rating models

To rigorously evaluate the performance and interpretability of the proposed WDForest framework, we constructed a comprehensive benchmark ensemble encompassing a broad spectrum of methodological paradigms,ranging from classical statistical approaches and modern additive frameworks to high-performance black-box ensembles and state-of-the-art interpretable techniques and enabling a systematic and robust assessment of its comparative advantages. The evaluation begins with interpretable additive models, incorporating three variants of Generalized Additive Models (GAMs) that preserve intrinsic interpretability while capturing non-linear feature effects. To

establish a predictive performance upper bound, high-performance ensembles such as Gradient Boosting Decision Trees (GBDT) are included, positioning WDForest relative to leading opaque systems. Furthermore, WDForest is contextualized within the contemporary explainable AI landscape through comparison with prominent glass-box models like EBM (Nori et al. 2021) and NAM (Agarwal et al. 2021), as well as hybrid interpretable architectures including DeepGBM (Ke et al. 2019) and GBDT-PL (Shi et al. 2019). A critical benchmark is also established against CRABoost and its foundational component, the Additive Boosting Model (ABM) (Schulte and Rügamer 2025), ensuring direct comparison with another recent state-of-the-art interpretable framework.

4.3 ESG scoring metrics

To holistically and rigorously evaluate the performance of WDForest, we employ a comprehensive suite of six metrics for comprehensive comparison. Predictive performance is primarily quantified through three error-based metrics, with the Mean Squared Error (MSE) measuring the average squared difference between true values and predicted values. The MSE is formally defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2\,. \tag{15}$$

The Root Mean Squared Error (RMSE) is derived as the square root of the MSE, restoring the metric to the original scale of the target variable for more intuitive interpretation Plevris et al. (2022). In contrast, the Mean Absolute Error (MAE) provides a more robust assessment by averaging the absolute differences between true and predicted values, thereby assigning equal linear weight to all errors. The MAE is formally expressed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}_i|\,. \tag{16}$$

The goodness-of-fit of objective ESG rating models is quantified using the coefficient of determination (R^2), which measures the proportion of variance in corporate sustainability performance scores explained by the model relative to a simple mean-prediction baseline . The R^2 metric is formally defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{17}$$

Finally, as the accurate ordinal ranking of entities based on sustainability performance constitutes a primary objective of any objective ESG rating system, we evaluate the ranking capability of objective ESG rating models using two specialized metrics. The normalized Gini coefficient quantifies the model's discriminatory power by measuring the concordance between predicted and actual performance rankings, where higher values reflect a superior ability to stratify entities by ESG performance. Complementarily, the Spearman rank correlation assesses the monotonic alignment between

predicted and true values based on their ranked order, with the difference in ranks for each observation, denoted as di, serving to quantify the consistency of the generated ranking sequence:

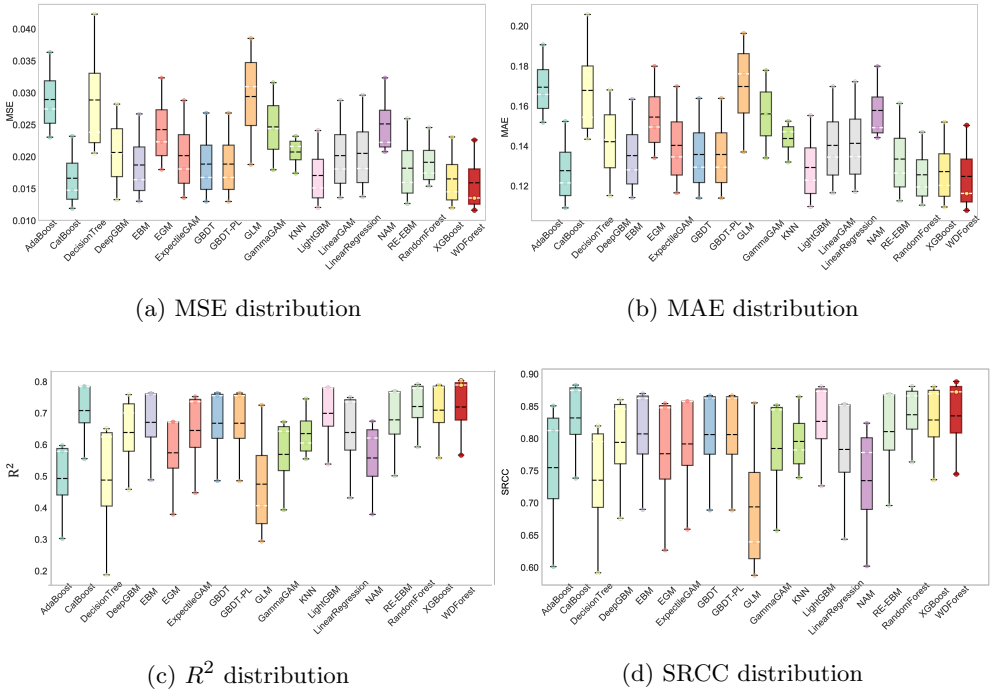
$$\mathrm{SRCC} = 1 - \frac{6 \sum d_i^2}{n\left(n^2 - 1\right)}.$$

4.4 Implementation details

A unified, reproducible experimental framework is devised to rigorously assess WDForest's predictive performance and interpretability for ESG rating. It encompasses data preprocessing, model and baseline configuration, evaluation metrics, implementation details, and hyperparameter tuning; targeted sensitivity analyses are further conducted to validate the model's adaptive layer depth and overall robustness. Experiments are carried out across four tasks,including Environmental, Social, Governance,and aggregate ESG. To minimize cross-model comparison bias and enhance reproducibility, data processing is held strictly consistent: numerical features are imputed using the median or mean according to distributional properties, categorical features are imputed by mode, and native missing-value handling is retained in tree-based models to improve robustness; continuous variables are standardized column-wise, categorical variables are one-hot encoded, and extreme values are treated via quantile clipping (Winsorization) to mitigate outlier influence.

The evaluation spans approaches from high interpretability to strong predictive performance by covering three model families, with WDForest serving as the focal method of assessment. Traditional machine learning comprises Logistic Regression (LR) (Nelder and Wedderburn 1972), Decision Trees (DT) (Quinlan 1986), k-Nearest Neighbors (KNN) (Cover and Hart 1967), and Random Forests (RF) (Breiman 2001); ensemble learning and gradient boosting encompass XGBoost (Chen and Guestrin 2016), LightGBM (Ke et al. 2017), and classical Gradient Boosting (GB) (Friedman 2001); in addition, complex ensembles and interpretable additive models are incorporated to better capture nonlinearities and interaction structures, including DeepGBM (Ke et al. 2019), EBM (Nori et al. 2021), and NAM (Agarwal et al. 2021). All models are trained and evaluated under identical data splits, preprocessing pipelines, and assessment protocols to ensure fairness of comparison.

Reproducibility and stability are ensured through standardized training and implementation practices. All models employ the same train/validation/test split, optimization targets validation-set MSE, and the test set is reserved exclusively for final reporting; each experiment is repeated 30 times to assess stochastic variability, with the primary performance tables reporting test-set metrics and the box plots summarizing repeated-trial distributions. Fixed random seeds enforce consistent partitioning and sampling, and tree models adopt default batching strategies to avoid artificial discrepancies. Hyperparameter search ranges are set to a learning rate $[0.001, 0.1]$ and a regularization coefficient $\lambda \in [0.5, 1.5]$. All models strictly adhere to an 70:15:15 split into training, validation, and test sets; the validation set (15%) is dedicated to hyperparameter optimization and determination of the early-stopping policy, while the test set remains untouched to guarantee the impartiality and comparability of the final evaluation. WDForest adopts an adaptive growth mechanism with



(a) MSE distribution

(b) MAE distribution

(c) R^2 distribution

(d) SRCC distribution

Fig. 2: Performance comparison of ESG rating algorithms across evaluation metrics

an early-stopping patience of 3 layers; training stops when the validation MSE fails to improve for 3 consecutive layers.

5 Experimental results

This section presents a systematic evaluation of WDForest framework for ESG rating across three critical dimensions: rating performance, sensitivity analysis, and interpretability. The comprehensive assessment begins by quantifying overall predictive performance on environmental, social, governance, and composite ESG tasks using the metrics MSE, MAE, R^2, and SRCC, while also examining pillar-specific effectiveness and the consistency between the distributions of predicted and actual scores. We further analyze WDForest's sensitivity to key hyperparameters such as cascade depth, number of base learners, and training set size, alongside an in-depth interpretability analysis derived from global feature importance, instance-level cascade tracing, and interaction visualization. A defining feature of the framework is its data-driven adaptive growth architecture, where cascade depth is dynamically determined through a validation-based early stopping mechanism rather than preset.

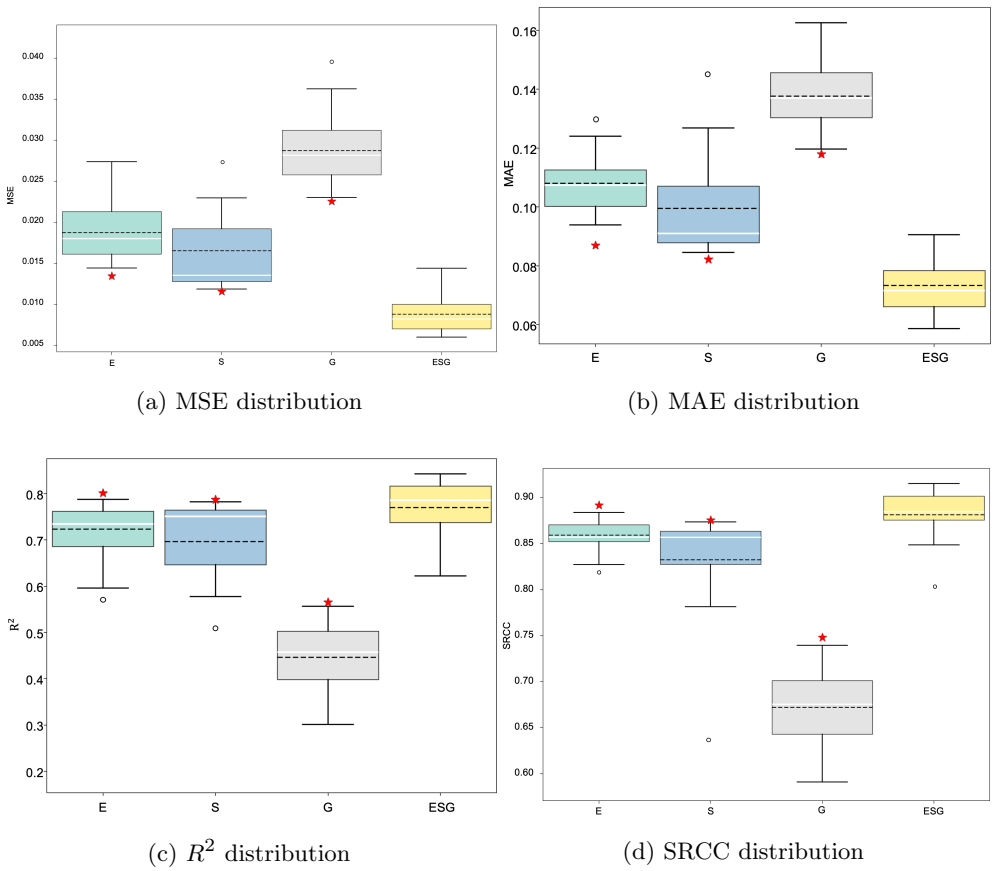
5.1 Performance comparison of ESG rating approaches

5.1.1 Performance comparison of ESG rating algorithms

To rigorously assess the advantage of WDForest over benchmark algorithms, we design a performance oriented box plot evaluation framework. As shown in Figure 2, WDForest is compared with nineteen widely used baselines whose model structures range from simple to intricate, including linear models, explainable boosting machine models, and gradient boosted decision tree models such as XGBoost. The comparison uses four key metrics, with mean squared error and mean absolute error serving as error based measures, and the coefficient of determination R^2 and SRCC serving as measures of explanatory power and rank consistency. This combination enables us to examine the control of numerical errors and the ability to capture structural relationships in the target variable, including the proportion of variance explained and the preservation of rank order, thereby visualizing differences across algorithms in rating performance, robustness, and consistency. These visualizations depict the distribution, central tendency, and variability of each model's performance over repeated trials, where the median line reflects central performance and the interquartile range (IQR) (Tukey 1977) conveys prediction stability with a narrower box indicating lower volatility. Analysis of the error metrics reveals a clear performance hierarchy, as WDForest consistently attains the lowest values in both MSE and MAE distributions, demonstrating its superior performance in minimizing rating residuals. In measures of goodness-of-fit and rank preservation, WDForest again achieves top performance, with R^2 nearing 0.80 and SRCC nearing 0.90, demonstrating its capacity to explain a substantial portion of variance while accurately maintaining prediction order, which is a critical requirement in rating tasks. Collectively, the results robustly validate the efficacy of WDForest architecture, which excels not only in rating performance and explanatory power but also in operational stability.

5.1.2 Performance variation across ESG components

Figure 3 comparatively assesses the distribution of predictive performance achieved by WDForest across E, S, G, and composite ESG rating tasks, revealing performance patterns that align closely with the inherent characteristics of each sustainability domain. WDForest demonstrates consistently low error rates and high predictive performance in both Environmental and Social dimensions, reflecting its capacity to effectively represent relevant features and capture measurable sustainability activities in these areas. In contrast, Governance ratings exhibit broader error dispersion and reduced explanatory power, which may be attributed to the qualitative nature, lower observed variability, and often delayed materialization of governance-related metrics. Notably, WDForest delivers optimal performance in the composite ESG rating task, yielding the narrowest error distribution and highest correlation stability, demonstrating its ability to synthesize multi-dimensional sustainability information and compensate for domain-specific weaknesses, particularly those associated with governance-related complexities, thereby providing a robust, reliable, and highly consistent foundation for objective ESG assessment.



(a) MSE distribution

(b) MAE distribution

(c) R^2 distribution

(d) SRCC distribution

Fig. 3: Comparative performance on overall ESG versus component-specific (E, S, G) prediction tasks

Tables 1–4 present a comparative performance evaluation of 18 ESG rating algorithms spanning simple linear models to complex glass-box and neural network architectures. Among these, WDForest demonstrates consistent and marked superiority in the aggregated ESG rating task, achieving a test-set R^2 of 0.8490 alongside low error metrics (MSE: 0.0058, RMSE: 0.0759, MAE: 0.0578) and superior ranking performance reflected in an N-Gini of 0.9209 and Spearman rank correlation of 0.9183. These results indicate that WDForest not only captures global ESG patterns with high predictive performance but also maintains strong ordinal consistency and correlation fidelity, highlighting its ability to learn multi-dimensional sustainability signals into a coherent and reliable rating output. The performance advantage stems from WDForest's unique wide-and-deep architecture, which incorporates residual cascading and tree-based learning. This enables successive layers to dynamically correct residuals from prior predictions while refining the feature space, thereby systematically reducing cumulative errors throughout the rating process. Such behavior aligns closely with the

Table 1: Algorithm performance on dataset E

Algorithm	MSE	RMSE	MAE	R^2	N-Gini	SRCC
LR	0.0181	0.1344	0.1075	0.7335	0.858	0.8567
GLM	0.0187	0.1368	0.1115	0.7238	0.8558	0.8545
DT	0.0237	0.1541	0.1077	0.6496	0.8232	0.8189
KNN	0.0173	0.1317	0.0998	0.7441	0.8652	0.8641
RF	0.0161	0.1265	0.1022	0.7639	0.8746	0.8733
AdaBoost	0.0274	0.1655	0.1402	0.5959	0.8575	0.8500
GBDT	0.0167	0.1292	0.1031	0.7538	0.8682	0.8660
XGBoost	0.0144	0.1200	0.0939	0.7874	0.8863	0.8840
LightGBM	0.0150	0.1226	0.0969	0.7783	0.8817	0.8797
CatBoost	0.0147	0.1212	0.0950	0.7833	0.8841	0.8821
ExpectileGAM	0.0180	0.1342	0.1074	0.7342	0.8584	0.8571
GammaGAM	0.0243	0.1560	0.1211	0.6411	0.8526	0.8510
EGM	0.0223	0.1492	0.1240	0.6715	0.8549	0.8536
NAM	0.0222	0.1489	0.1187	0.6728	0.8284	0.8272
EBM	0.0163	0.1277	0.1017	0.7592	0.8709	0.8689
GBDT-PL	0.0167	0.1292	0.1030	0.7538	0.8682	0.8660
DeepGBM	0.0204	0.1429	0.1137	0.6987	0.8559	0.8449
WDForest	0.0135	0.1160	0.0870	0.8015	0.8944	0.8918

model's embedded residual learning mechanism, layer-wise performance boosting, and adaptive early stopping strategy, all contributing to its robust and stable ESG scoring capability. Besides, WDForest significantly outperforms ensemble solutions such as Random Forest and gradient boosting models on the composite ESG task, establishing a clear advantage in optimizing rating performance for objective rating applications. Overall, these findings validate WDForest's strong generalization capacity and ranking stability across multi-faceted ESG contexts. Even within the challenging setting of composite ESG assessment, its residual-cascade and wide-deep fusion mechanisms enhance representational power, yielding state-of-the-art performance that offers a transparent and quantitative foundation for downstream applications.

5.2 ESG scoring distribution consistency analysis

The validation of alignment between model-predicted ESG scores and subjectively assessed ground-truth ratings constitutes a critical test for any data-driven rating framework, as it directly measures objective rating models' capacity to replicate human expert judgment across complex, multi-dimensional sustainability concepts. In objective ESG rating systems, merely achieving high numerical accuracy is insufficient; the model must also capture the nuanced distributional characteristics, such as central tendency, dispersion, and tail properties, inherent in each pillar and their composite representation. This analysis rigorously evaluates whether the predicted scores not only minimize errors but also reproduce the semantic and structural patterns of the underlying subjectively constructed benchmarks, thereby ensuring that the automated ratings remain interpretable, fair, and actionable for downstream applications in sustainable finance and regulatory oversight.

Table 2: Algorithm performance on dataset S

Algorithm	MSE	RMSE	MAE	R^2	N-Gini	SRCC
LR	0.0137	0.117	0.0915	0.7483	0.8611	0.8551
GLM	0.0385	0.1961	0.1618	0.2927	0.5913	0.5871
DT	0.0205	0.1432	0.1036	0.6230	0.8045	0.7944
KNN	0.0215	0.1468	0.1121	0.6039	0.7839	0.7815
RT	0.0130	0.1142	0.0908	0.7587	0.8681	0.8617
AdaBoost	0.0230	0.1516	0.1268	0.5774	0.8295	0.8115
GBDT	0.0129	0.1137	0.0887	0.7622	0.868	0.862
XGBoost	0.0119	0.1092	0.0845	0.7807	0.8783	0.8726
LightGBM	0.012	0.1094	0.085	0.7798	0.8779	0.8724
CatBoost	0.0118	0.1088	0.0846	0.7822	0.8792	0.8736
ExpectileGAM	0.0135	0.1163	0.091	0.7512	0.8628	0.8569
GammaGAM	0.0179	0.1338	0.1032	0.6709	0.85	0.8437
EGM	0.0179	0.1339	0.1107	0.6703	0.8532	0.847
NAM	0.0207	0.1439	0.1147	0.6192	0.7886	0.7816
EBM	0.013	0.1139	0.0884	0.7616	0.8674	0.8614
GBDT-PL	0.0129	0.1137	0.0887	0.7622	0.868	0.862
DeepGBM	0.0132	0.1149	0.0893	0.7571	0.8653	0.8596
WDForest	0.0116	0.1076	0.0822	0.7872	0.8821	0.8757

To holistically evaluate the alignment between predicted ESG scores and expert-assigned benchmarks across Environmental, Social, Governance, and Composite dimensions, we visualize the comparative distributions of objective model outputs against subjective rating distributions using a multi-panel density plot.

As can be seen from Figure 4, it presents a comparative assessment of four models against the true score distribution across E, S, G, and composite ESG dimensions. WDForest demonstrates the closest alignment with the true distribution in both shape and tail behavior, accurately capturing the central peak while smoothly adhering to the left and right tails. This distributional fidelity reflects the model's capacity to balance local precision with global consistency, avoiding overconfidence in high-density regions while maintaining sensitivity to extreme yet meaningful performance variants, which is a critical capability in ESG ratings where both typical and outlier behaviors carry material significance. By contrast, the other three advanced ensemble-based benchmark models exhibit systematic limitations that collectively undermine their practical efficacy: excessive mid-range conservatism collapses diverse performance levels toward the median and erases critical differentiators; localized overfitting produces spurious fluctuations that mask authentic underlying patterns; and inadequate tail coverage fails to represent low-frequency yet high-stake sustainability performances.

In the environmental pillar scores illustrated in Figure 4(a), the histogram and the reference line reveal a pronounced concentration of values in the lower performance range, together with a tail that gradually rises across the middle to upper scoring spectrum. WDForest, shown by the black curve, demonstrates exceptional distributional fidelity by closely tracking the true data profile, smoothly capturing transitional trends in the 0.2 to 0.4 interval and maintaining consistent alignment along the right tail. In contrast, Random Forest introduces artifactual volatility through multiple

Table 3: Algorithm performance on dataset G

Algorithm	**MSE**	**RMSE**	**MAE**	R^2	**N-Gini**	**SRCC**
LR	0.0296	0.1720	0.1414	0.4302	0.6535	0.6472
GLM	0.0309	0.1758	0.1475	0.4047	0.6454	0.6386
DT	0.0423	0.2056	0.1539	0.1859	0.5948	0.5908
KNN	0.0232	0.1522	0.1197	0.5540	0.7423	0.7385
RT	0.0275	0.1660	0.1381	0.4629	0.6893	0.6837
AdaBoost	0.0363	0.1904	0.1626	0.3013	0.6169	0.6005
GBDT	0.0268	0.1636	0.1345	0.4841	0.6936	0.6881
XGBoost	0.023	0.1517	0.1219	0.5566	0.7435	0.7395
LightGBM	0.024	0.155	0.126	0.5371	0.7301	0.7256
CatBoost	0.0231	0.1521	0.1222	0.5543	0.7417	0.7377
ExpectileGAM	0.0288	0.1696	0.139	0.4461	0.6646	0.6585
GammaGAM	0.0315	0.1776	0.1437	0.3924	0.6627	0.6567
EGM	0.0323	0.1797	0.1516	0.3779	0.633	0.6261
NAM	0.0323	0.1797	0.1484	0.3781	0.6121	0.6054
EBM	0.0266	0.1632	0.1324	0.4871	0.6942	0.6892
GBDT-PL	0.0268	0.1636	0.1345	0.4841	0.6936	0.6881
DeepGBM	0.0282	0.1679	0.137	0.4571	0.6801	0.6754
WDForest	0.0226	0.1502	0.118	0.5655	0.7518	0.7483

local peaks that amplify noise rather than underlying patterns; GBDT and XGBoost exhibit premature attenuation in the upper tail and a systematic overestimation of mid to high range scores, reflecting a bias toward central tendencies at the expense of distributional extremities.

5.3 Dynamic learning process of WDForest

WDForest is its data-driven, self-adapting cascade mechanism that dynamically determines optimal model complexity, it is essential to empirically validate how this automatic depth selection impacts predictive performance across different ESG rating tasks. Therefore, we further assess the efficacy of the adaptive growth architecture by tracing performance metrics throughout the cascading process. Figure 5 traces the layer-wise evolution of predictive performance, quantified by MSE and R^2, as the WDForest architecture progressively deepens its cascade across Environmental, Social, Governance, and composite ESG rating tasks.

Figure 5(a) illustrates the variation of MSE as the number of WDForest layers increases across four rating dimensions. The validation MSE exhibits a consistent decreasing trend in all tasks, demonstrating the effectiveness of the residual-enhanced architecture in error reduction. A clear hierarchy emerges among the dimensions: the composite ESG curve achieves the lowest error, indicating higher learnability; environmental and social dimensions follow similar trajectories; while governance maintains the highest MSE, reflecting its inherent complexity and noisiness. The descent rate gradually slows in intermediate layers as curves approach task-specific plateaus, with each dimension reaching a distinct minimum corresponding to its optimal depth. An early stopping patience of 3 was implemented to halt expansion when no improvement occurs over three consecutive layers, validating MSE minimization as an effective

Table 4: Algorithm performance on dataset ESG

Algorithm	**MSE**	**RMSE**	**MAE**	R^2	**N-Gini**	**SRCC**
LR	0.0064	0.0798	0.063	0.8329	0.9126	0.9104
GLM	0.0088	0.094	0.0749	0.7684	0.8785	0.8755
DT	0.0084	0.0919	0.0727	0.7786	0.8832	0.8804
KNN	0.0069	0.083	0.0657	0.8193	0.9055	0.9032
RT	0.0095	0.0974	0.0786	0.7511	0.8735	0.8702
AdaBoost	0.0123	0.1107	0.0906	0.6786	0.8573	0.8519
GBDT	0.0144	0.12	0.0878	0.6223	0.8108	0.806
XGBoost	0.0063	0.0792	0.0621	0.8356	0.9142	0.9118
LightGBM	0.0081	0.09	0.0716	0.7877	0.8883	0.8856
CatBoost	0.006	0.0774	0.0587	0.8427	0.918	0.9155
ExpectileGAM	0.0082	0.0904	0.0716	0.7856	0.8874	0.8847
GammaGAM	0.0101	0.1005	0.079	0.7351	0.8789	0.8762
EGM	0.0112	0.1056	0.0865	0.7073	0.8796	0.8766
NAM	0.0107	0.1034	0.0825	0.7195	0.8515	0.8486
EBM	0.0073	0.0853	0.0675	0.8092	0.9002	0.8978
GBDT-PL	0.0081	0.09	0.0716	0.7877	0.8883	0.8856
DeepGBM	0.0097	0.0985	0.0766	0.7452	0.8711	0.8685
WDForest	0.0058	0.0759	0.0578	0.8490	0.9209	0.9183

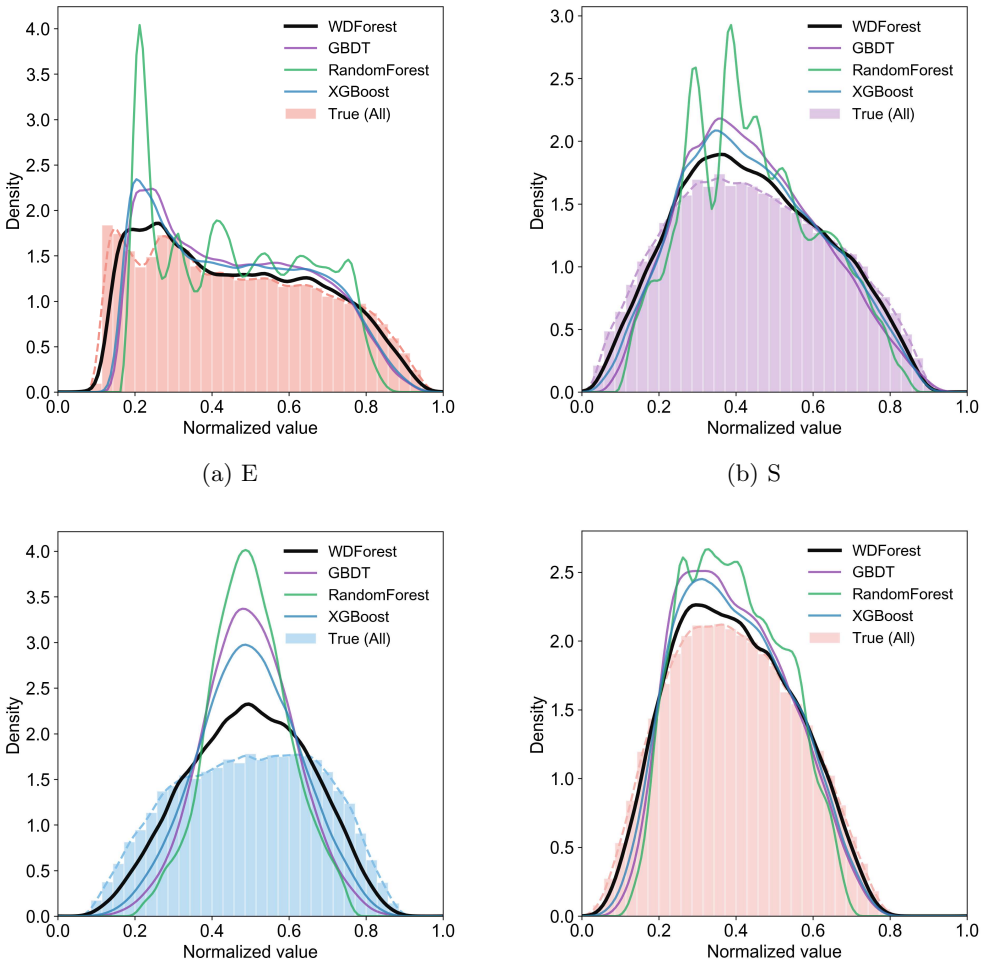
stopping criterion. This progression supports selecting depths near plateau regions for optimal performance-efficiency trade-offs, as further deepening yields diminishing returns.

Figure 5(b) analyzes the progression of R^2 as WDForest's layer depth increases, showing steady improvement in validation R^2 across all rating dimensions as the model's explanatory power and fit strengthen with each added layer. The ESG dimension consistently achieves the highest R^2 values, reflecting its stronger learnability, while environmental and social curves follow similar intermediate trajectories, and governance remains lowest due to its inherent complexity. As layers extend into intermediate ranges, the rate of R^2 growth gradually slows and curves approach saturation, indicating diminishing returns from further depth expansion.

5.4 Sensitivity analysis

5.4.1 Sensitivity analysis of the number of base learners

Figure 6 reveals critical patterns in the generalization performance of the WDForest during its adaptive growth, offering insights into the capacity-generalization trade-off that directly informs its application to ESG rating tasks. A clear architectural transition occurs around Layer 6, where the generalization gap remains stable within a narrow range across all decision tree configurations, indicating that the model effectively captures essential ESG data structures without compromising generalization. Beyond Layer 7, however, the gap increases sharply, thereby reflecting a shift toward fitting dataset-specific noise and introducing a significant risk in ESG contexts, where rating robustness depends on generalizable sustainability patterns. The effect of tree

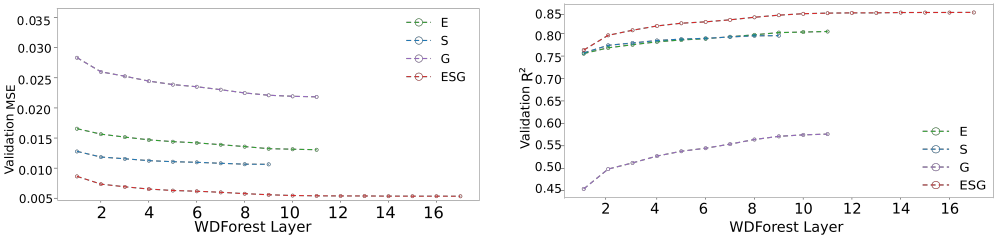


(a) E (b) S

(c) G (d) ESG

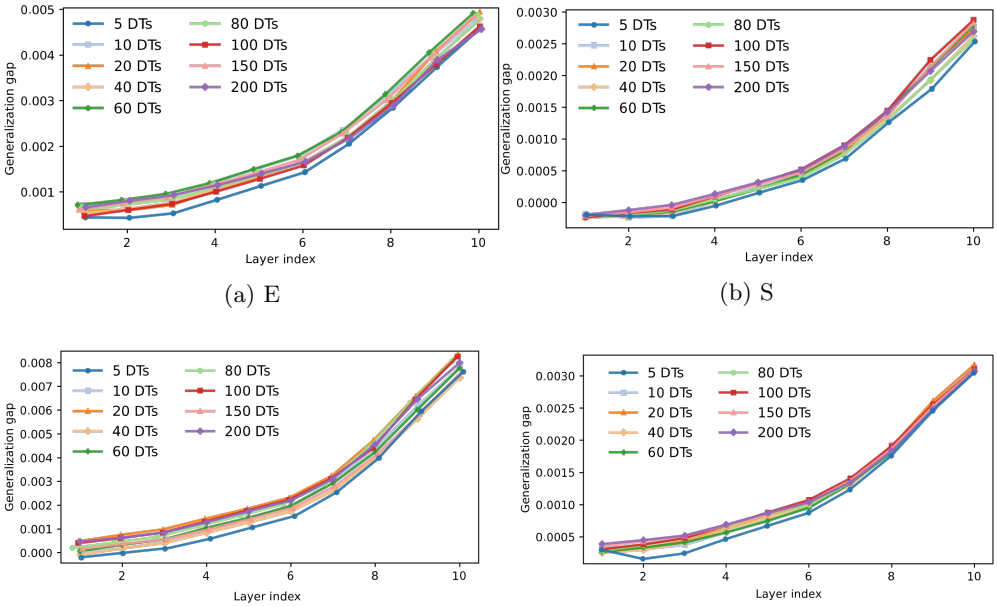
Fig. 4: Comparative analysis of objective versus subjective ESG pillar scores

count varies with depth: shallow layers show negligible differences between configurations, suggesting capacity redundancy, while deeper layers exhibit a strong positive correlation between tree number and gap magnitude, highlighting the overfitting danger of excessive model complexity in capturing nuanced governance or social metrics. This behavior stems from WDForest's residual learning mechanism, in which minor fitting errors in early layers amplify through successive stages, accelerating generalization decay. These observations suggest treating Layer 6 as a critical checkpoint for initiating capacity constraints or stronger regularization, integrating the generalization gap directly into the optimization strategy rather than focusing solely on validation error minimization. Ultimately, the results demonstrate diminishing returns



(a) The effect of layer depth on MSE (b) The effect of layer depth on R^2

Fig. 5: Adaptive layer growth process



(a) E (b) S

(c) G (d) ESG

Fig. 6: The impact of the number of base learners on generalization gap

to depth in ESG rating applications, where excessive architectural complexity not only fails to improve predictive accuracy but also undermines rating consistency and trustworthiness through error propagation and capacity oversaturation.

Figure 6(a) demonstrates the Environment dimension, where the curves display shallow stability in the early layers, gradual change through the middle layers, and clear separation in the deep layers. Sensitivity to the number of trees emerges mainly in the deep region. From Layers 7 to 10 the curve rises progressively as the tree count increases from 60 to 200, with a peak close to 0.005. This pattern indicates that deep

expansion of capacity lifts the generalization gap in a roughly linear manner while the magnitude remains moderate.

Figure 6(b) reveals the Social dimension, where the overall magnitude is lower and the peak is around 0.003. Layers 1 to 7 reveal almost no difference attributable to the number of trees, and only Layers 8 to 10 show mild separation. Higher tree counts yield slightly steeper slopes than lower tree counts, yet the divergence is far weaker than in Environment and Governance. In practice, 40 to 60 DTs typically suffice to keep the slope stable in the deep layers.

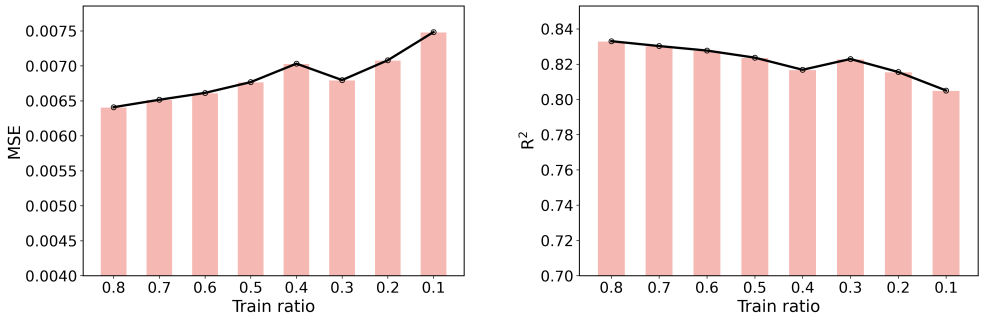
Figure 6(c) shows the Governance dimension, which is the most sensitive to the number of trees and the main driver of generalization decay. Signals that stratify with tree count are visible from approximately Layer 3, and separation becomes most pronounced in Layers 7 to 10. Curves with 100 to 200 DTs bend upward in a near exponential manner, with the gap approaching 0.008, while 5 to 60 DTs maintain relatively controlled growth. This pattern reflects the complexity of governance indicators and multiple sources of noise. In this dimension, the number of trees determines both the onset of the turning point and the steepness of the subsequent increase.

Figure 6(d) integrates the aggregated ESG dimension, where the stability of Environment and Social partially dilutes the instability of Governance. The overall shape is similar to Social and the peak is around 0.003. Even so, after Layer 7 there is an orderly separation with increasing tree count. Slopes for at least 100 DTs are significantly higher than those for at most 60 DTs, which indicates that deep high capacity can carry the overfitting risk of governance submodules into the overall rating.

5.4.2 Performance on different scale of ESG datasets

WDForest employs a data-driven approach, dynamically adjusting model complexity through deep expansion and residual learning to effectively scale to datasets of varying sizes and complexities. This adaptive mechanism not only ensures robust performance under large sample conditions but also provides a viable pathway for building reliable ESG rating models in data-scarce scenarios, where traditional methods often suffer from overfitting or instability. To systematically evaluate its sample efficiency and generalization stability under constrained data conditions, we conducted a training ratio ablation experiment.

Figure 7 systematically depicts the performance trajectory of WDForest at different training ratios and its relationship with sample size. In Figure 7(a), MSE monotonically increases with decreasing training ratio, but exhibits a clear piecewise nonlinearity: when the ratio decreases from 80% to 40%, the increase in MSE is relatively gradual; in the range of 40% to 30%, MSE shows a slight decrease, possibly because 30% falls within a more suitable complexity range, and at 30%, the MSE for a single evaluation shows a slight decrease due to the influence of sample size and sampling noise. Around 20%, MSE accelerates again. Consistent with the data, when the training ratio decreases from 80% to 10%, MSE only increases from 0.0063 to 0.0074, with an absolute increase of approximately 0.0011, still within a controllable range. This model demonstrates that although sample reduction often amplifies variance and adversely alters the bias-variance tradeoff, WDForest, with its ensemble structure and



(a) MSE (b) R^2

Fig. 7: Model performance on different ratio of training sets

inherent regularization properties, maintains low error under small sample conditions, exhibiting good small-data adaptability.

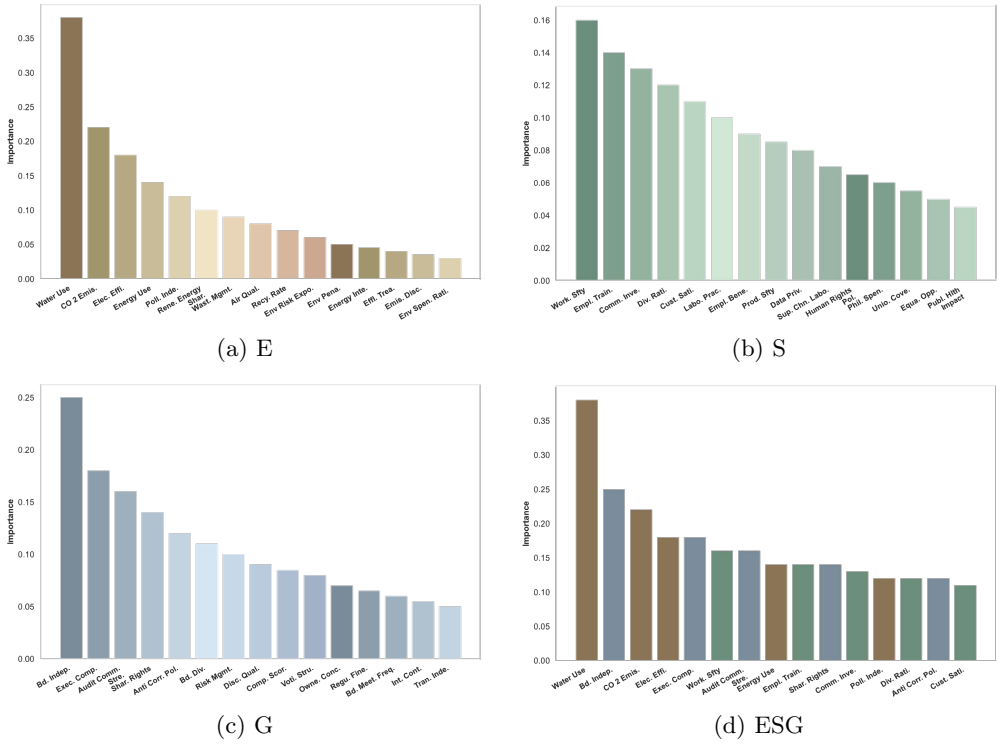
Figure 7(b) shows the variation of R^2 with the training sample size, revealing its typical non-linear decay characteristic. The metric gradually decreases between 80% and 40%, with a slight increase between 40% and 30%, which is typically a normal fluctuation resulting from the combined effect of data splitting and implicit regularization. The decline accelerates below the 30% threshold. Quantitatively, R^2 decreases from approximately 0.8369 to 0.8050, an absolute decrease of only 0.032, but even under extremely scarce data conditions (70% reduction in training samples), its value remains above the 0.80 threshold, demonstrating excellent generalization robustness.

Combining the observations from both subgraphs, performance degradation exhibits a significant non-linear sensitivity to sample size reductions, with the 30% training ratio region serving as a critical inflection point where system stability becomes particularly sensitive to perturbations. This finding suggests that when operating below this threshold, auxiliary techniques such as data augmentation, strategic resampling, or weighted ensemble mechanisms may be necessary to mitigate information loss. In summary, these results confirm the architectural robustness of WDForest in small- to medium-scale data environments, highlighting its practicality in resource-constrained ESG rating scenarios where labeled samples are inherently limited.

5.5 Interpretation of WDForest

5.5.1 Global interpretation

Figure 8 summarizes the gain-based feature importance aggregated in WDForest, with split gains from each layer normalized to ensure total importance sums to unity. As shown in Figure 8(a), WDForest on the environmental model pillar dataset exhibits a sharply skewed importance distribution: the leading variable, water use or intensity, accounts for approximately forty percent of total influence, substantially exceeding a



(a) E (b) S

(c) G (d) ESG

Fig. 8: Feature importance rankings across different dimension

secondary tier comprising energy consumption, emission policy, and energy efficiency indicators. This pronounced skew underscores the role of water-related metrics as the most discriminative global drivers for the E pillar, consistently selected across layers due to their efficacy in reducing residual variance. Secondary features, including waste management, air quality, and operational energy efficiency, form a coherent cluster that complements the dominant water signal by capturing both policy-level commitments and operational performance. Lower-ranked variables, such as environmental disclosure and spending ratios, contribute only marginal improvements in later layers, reflecting their weaker predictive utility or stronger collinearity with dominant water and energy metrics.

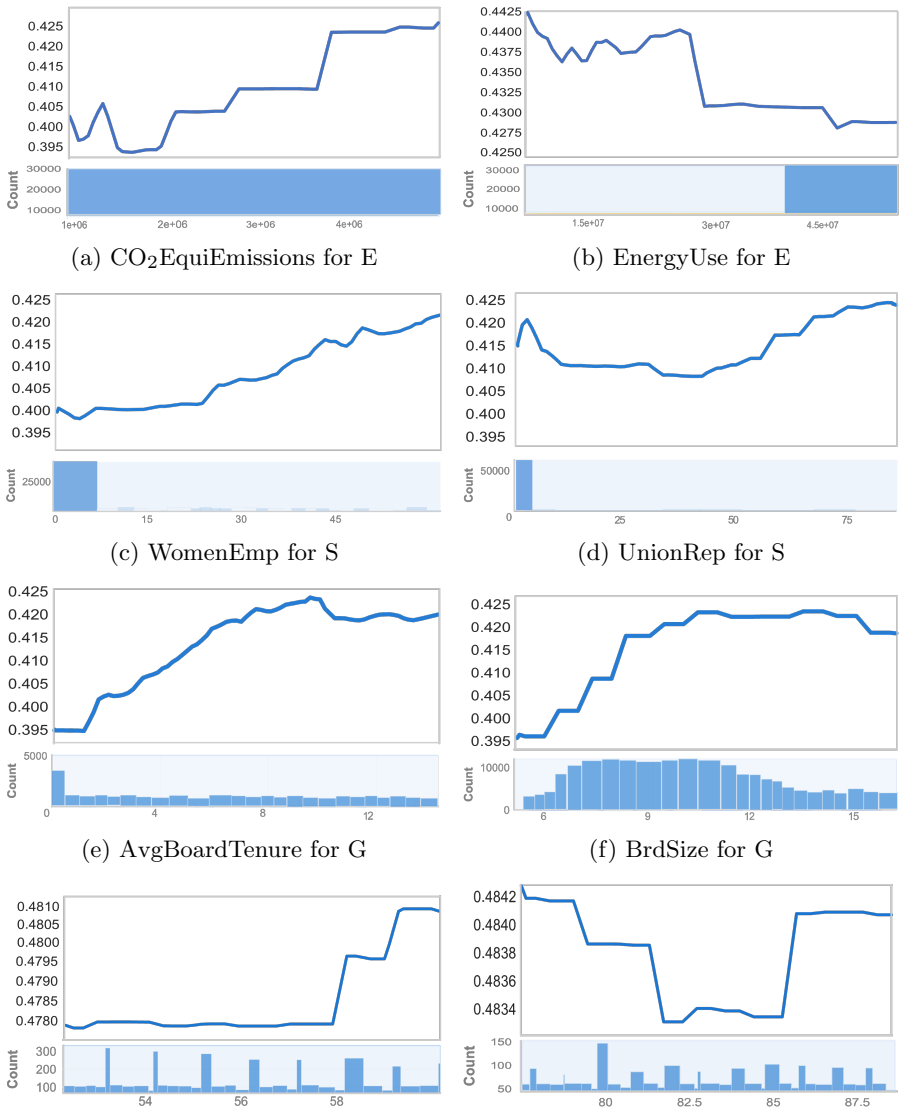
Figure 8(b), corresponding to the social model, reveals a more balanced distribution of importance across the top-ten features, including employees health and safety team, training and development policy, corporate responsibility awards, which appear in a gently descending order of influence. In contrast to the sharply skewed environmental profile, WDForest on the social pillar dataset exhibits a flatter long-tail distribution, indicating a more diffusely structured signal space. WDForest's cascade architecture learns these social signals progressively by first capturing broad governance-related aspects, while workforce-specific attributes gain explanatory power

only after accounting for baseline policy heterogeneity. The absence of a single dominant feature underscores the inherently composite and multi-stakeholder nature of social performance, wherein firms can achieve comparable outcomes through diverse configurations of practices, a complexity that WDForest effectively represents through its additive residual correction mechanism without relying on any single anchoring predictor.

Figure 8(c) shows WDForest demonstrates performance in governance pillar. WDForest on the governance model pillar dataset exhibits a significant importance gradient, with board independence dominating with an importance score of approximately 0.25, far surpassing other characteristics. Executive compensation and the audit committee follow closely behind, forming a distinct three-tiered ladder structure. Unlike the balanced distribution of social pillar dataset, WDForest on the governance model pillar dataset exhibits a more concentrated pattern of feature dependence, reflecting the critical role of core institutional arrangements in corporate governance assessment. WDForest's effectively models the hierarchical nature of governance structures by first capturing the foundational governance signal of board independence and then gradually refining the impact of compensation mechanisms and oversight systems. This steep importance distribution indicates that within the governance dimension, a few core institutional features carry the majority of predictive information, reflecting the structured nature of governance quality assessments.

Figure 8(d) integrates all pillars into the overall ESG target, revealing a synthesized importance structure where environmental water use remains the dominant global driver. Board independence dominates the second tier, followed by carbon dioxide emissions, electricity efficiency, and executive compensation in subsequent ranks. This hierarchy demonstrates that composite ESG scoring transcends simple pillar aggregation; WDForest's cascade prioritizes variables exhibiting consistent explanatory power across dimensions, particularly those systematically reducing residual variance in integrated score prediction. The framework implicitly calibrates cross-pillar interactions: robust governance mechanisms enhance policy adoption and disclosure quality, thereby strengthening environmental and social performance. Conversely, tangible environmental outcomes provide concrete evidence reinforcing governance and social commitments, elevating their importance in overall ratings.

From a managerial perspective, the figure offers clear guidance. For environmental performance, prioritize water stewardship and energy efficiency programs, then formalize emissions reduction targets and transparent reporting, as these deliver the largest predictive gains and likely broader impact. For social dimensions, combine strong employees health and safety team, training and development policy and corporate responsibility awards, since improvements are cumulative and a balanced program outperforms single-focus interventions. For governance, ensure empowered board independence, executive compensation, and r audit committee, as these form the backbone that supports credible execution in other areas. At the aggregate ESG level, policies and structures are necessary but insufficient: tangible, measurable performance, especially in environmental domains, decisively influences the overall score.



(a) CO_2EquiEmissions for E

(b) EnergyUse for E

(c) WomenEmp for S

(d) UnionRep for S

(e) AvgBoardTenure for G

(f) BrdSize for G

(g) WomenEmp for ESG

(h) UnionRep for ESG

Fig. 9: Key individual feature effect curves

5.5.2 Individual feature effects

To leverage WDForest's capacity to characterize nonlinear and context-specific dependencies between ESG features and ESG ratings, we move beyond global importance metrics toward granular functional insights that directly inform strategic intervention points across environmental, social, and governance domains. By mapping how specific ESG indicators, ranging from emissions to governance metrics, nonlinearly drive

rating results, this work uncovers the hidden mechanisms of performance sensitivity. It empowers precise resource allocation towards interventions that yield the greatest sustainability return, moving ESG management from generalized compliance to targeted, evidence-based optimization.

Figure 9(a) indicates that, in the E model, CO_2 Equivalents Emissions (Direct Scope) are positively correlated with the E score, presenting an overall stepwise upward pattern. The economically expected relation with the E score is monotonic negative: higher absolute emissions imply weaker environmental performance and greater transition risk. In our sample, the curve exhibits a stepwise upward profile, which is atypical under a pure-level interpretation. A plausible reading is that the scoring rubric rewards credible transition trajectories,firming with high baselines but material reductions, more complete disclosure, or verified offsets gain disproportionately in the low-to-medium score range, with marginal improvements flattening once low-cost abatement options are exhausted. Industry normalization and intensity scaling can also invert the apparent slope. This panel should therefore be interpreted as "transition intensity" rather than raw emissions level; the plateau at higher scores is consistent with diminishing returns along the abatement cost curve. Robust inference should focus on the main density band and verify whether the variable is recorded as absolute level or intensity/reduction.

Figure 9(b) documents a pronounced negative association between the energy use rate and the E score, with the steepest descent occurring from low to medium efficiency. The expected relation with the E score is monotonic negative. The steep decline from low to medium efficiency reflects high-Return on Investment (ROI) retrofits and process optimization. Beyond a medium-efficiency threshold, marginal improvement weakens as projects become capital-intensive with longer payback, consistent with the law of diminishing returns. Tail regions at extremely high intensity are sparsely populated; conclusions should be drawn from the core density area, controlling for industry mix and revenue deflators that affect intensity denominators.

Figure 9(c) reveals that, for the S score, the share of women employees maintains a monotonic positive relation with the S score, rising more sharply in the low-to-medium band. A monotonic positive relation with the S score is expected and observed. The steeper slope in the low-to-medium range signals "critical-mass effects": moving from token representation to substantive inclusion yields outsized gains in engagement, retention, and workplace safety. At higher proportions, slope moderates as institutions reach a steady state; further improvements depend less on headcount ratios and more on leadership pipelines, equal-pay enforcement, and culture programs. Given limited samples in middle-to-high bands, inference should anchor on the main density region.

Figure 9(d) highlights an overall positive relationship for union representation, with the largest marginal improvement emerging as organization moves from low to medium levels.The overall positive relation is financially intuitive: moving from low to medium organization improves process discipline and incident reporting, yielding the largest marginal gains. At high organization levels, benefits taper as coordination costs and bargaining frictions increase, producing diminishing returns. Sparse observations in extreme bands warrant caution; results should be conditioned on sectoral labor regimes and local regulation.

Figure 9(e) demonstrates an inverted-U pattern for average board tenure, pointing to an optimal tenure band where gains peak before reversing. The inverted-U relation indicates an optimal governance window: short-to-medium tenures strengthen oversight quality and responsiveness, while very long tenures risk path dependency, groupthink, and weakened challenge to management. The peak reflects efficient governance where institutional memory and independence are balanced. As tenure extends, the marginal effect turns negative, consistent with rising entrenchment costs. Interpretation should focus near the peak density and control for staggered boards and director refreshment policies.

Figure 9(f) depicts an inverted-U relation for board size, strengthening from small to medium boards and tapering beyond the optimal range. The inverted-U shape is expected: moving from small to medium boards adds domain expertise (risk, audit, sustainability), improving governance capability; beyond the optimum, decision latency and diluted accountability erode effectiveness. Benefit saturation around the optimal size is followed by negative contributions as committees proliferate without net control improvement. Conclusions should primarily reference the medium-size region; extreme small/large boards are under-sampled.

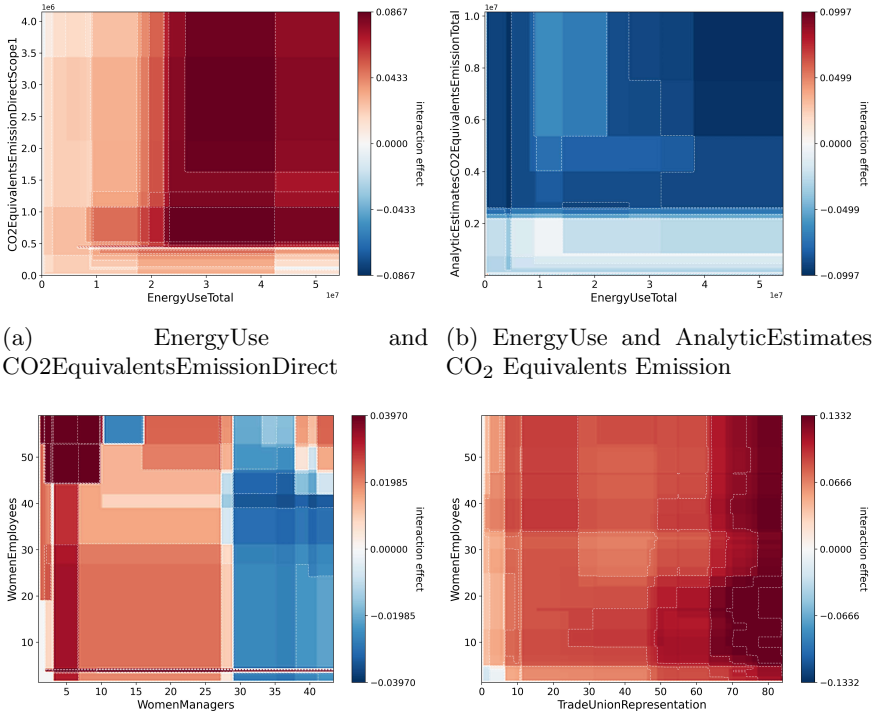
Figure 9(g) evidences a stronger positive correlation between women employees and the composite ESG score than in the S dimension alone, unfolding as a stepped ascent followed by a plateau.Inclusive cultures correlate with safer operations (E), better compliance and risk oversight (G), and talent advantages (S). The low-to-medium band is a key uplift interval where governance and environmental externalities of diversity become salient; at high shares, marginal gains slow as systems reach a steady state. The stepped ascent followed by a plateau fits a synergy-to-consolidation path. Low density in the highest band suggests focusing on the main steps and testing sector heterogeneity.

Figure 9(h) portrays a piecewise effect for union representation, turning positive in the medium range, and easing at high levels. In early stages, setup costs and frictions may offset benefits; once processes are established, integrated compliance and safety systems deliver net gains across E and G; at very high levels, diminishing returns set in as additional organization adds limited incremental control. Robust interpretation should center around the inflection band and adjust for differences in regulatory context and sector union norms.

5.5.3 ESG feature interaction

Beyond individual feature effects, pairwise interactions play a central role in WDForest's ESG assessments by capturing synergistic and antagonistic relationships. These patterns are visualized through heatmaps that are shown in Figure 10, where warm colors represent positive interactions and cool colors denote negative ones.

Figure 10(a) characterizes the interaction between total energy consumption and direct CO_2 equivalent emissions, integrating scale-dependent usage with source-level environmental burden to test whether operational intensity reinforces or offsets primary emissions risks. The heatmap exhibits a broad warm plateau across energy use $E \in [2.0 \times 10^7, 4.8 \times 10^7]$ and direct emissions $D \in [0.6 \times 10^6, 3.0 \times 10^6]$, where WDForest detects a positive interaction effect peaking at ≈ 0.0867 consistent with



(a) EnergyUse and CO2EquivalentsEmissionDirect

(b) EnergyUse and AnalyticEstimates CO_2 Equivalents Emission

(c) WomenManagers and WomenEmployees

(d) TradeUnionRepresentation and WomenEmployees

Fig. 10: Key feature interaction effect plots

efficiency alignment at higher operating scales. A near-neutral bright band appears at mid-range levels,indicating minimal interplay when neither dimension dominates. Cooler pockets toward the low-energy/low-emission corner reflect weak or negative interactions, typical of under-utilization and diffuse impact. The pattern foregrounds the emissions–intensity balancing problem: robust environmental performance depends on coordinated management of energy and direct-emission trajectories, where joint improvements yield stronger ESG signals than isolated changes.

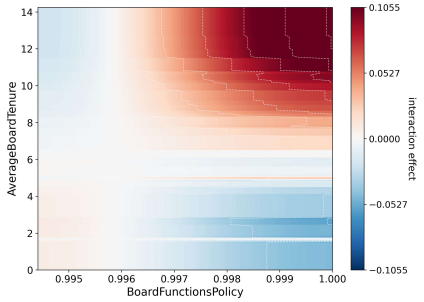
Figure 10(b) characterizes the interaction between total energy consumption and analytically estimated total CO_2 equivalent emissions, integrating scale dependent usage with system-wide modeled burden to evaluate whether operational intensity amplifies or attenuates portfolio emissions risk. The heatmap is dominated by cool tones across $E \in [1.0{\times}10^7,\ 5.0{\times}10^7]$ and total estimated emissions $T \in [0.4{\times}10^7,\ 1.0{\times}10^7]$, where WDForest detects a negative interaction effect down to -0.0997 consistent with compounding environmental pressure rather than efficiency gains. A near-neutral bright band emerges around $T \in [2.0 \times 10^6,\ 2.5 \times 10^6]$ and $E \in [0.8 \times 10^7,\ 2.2 \times 10^7]$, indicating minimal interplay when aggregate emissions remain constrained and energy scale is moderate. Sparse, warmer hints at very low energy do not coalesce into

a favorable corridor, underscoring that isolated intensity reductions are insufficient without concurrent emissions abatement. Overall, the pattern spotlights a system-level decarbonization constraint: ESG improvement hinges on coordinated management of energy demand and total CO_2 equivalent output, with joint reductions yielding stronger sustainability signals than unilateral changes.

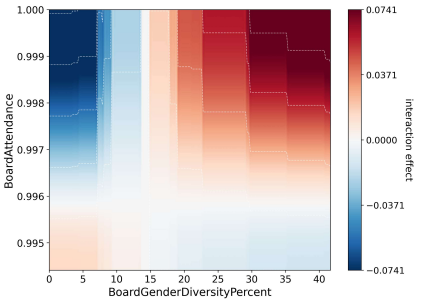
Figure 10(c) characterizes the interaction between the share of women in management and the share of women employees, linking leadership representation with organization wide inclusion to test whether diversity efforts strengthen social-pillar performance. The heatmap exhibits two dominant regimes: a warm corridor where WomenManagers are roughly 5 to 25 and WomenEmployees are roughly 28 to 55, where WDForest identifies a positive interaction with a maximum magnitude around 0.0397, indicating synergistic gains when expansion in the managerial ranks is accompanied by parallel progress among employees. In contrast, a cool basin appears where WomenManagers are roughly 30 to 43 and WomenEmployees are roughly 22 to 40, the interaction turns negative, suggesting that advances at the leadership level have not diffused to the broader workforce, the ESG signal weakens and inconsistency risks emerge. A near-neutral bright band is visible when the employee share is very low from about 5 to 12, overall representation is insufficient and the joint effect of the two dimensions is close to zero. Small warm pockets in the upper right corner, where both management and employee shares are high, indicate that strong representation on both tracks can recover synergy. Overall, robust social performance depends on coordinated progress across management and workforce, growth on one side while the other lags erodes the interaction advantage.

Figure 10(d) explores the interaction between trade union representation and the share of women employees, combining collective bargaining coverage as a proxy for worker voice with gender participation as an inclusion signal to assess whether stronger representation amplifies equity outcomes recognized by WDForest as favorable for social pillar performance. The heatmap reveals a dominant high-interaction region as union representation rises into the 65–85 range while women employees span roughly 10–30, with warmer tones indicating reinforcement where organized labor engagement and gender inclusion move together to deliver system-level gains. A broad warm field persists across mid levels of union representation around 15–50 with women employees 25–50, suggesting solid yet more diffuse synergy. A transitional bright band near mid-range values, centered around union representation 45–55 and women employees 20–35, signals minimal interaction when neither dimension clearly leads, reflecting balanced but underpowered participation dynamics. In the lower-left corner, small cool pockets with union representation 0–8 and women employees 3–8 mark negative interaction where weak bargaining coverage coexists with sparse female presence, pointing to compounded social risk. This pattern highlights the representation–inclusion complementarity in social pillar evaluation, where coordinated strengthening of bargaining coverage and gender participation is essential for achieving high sustainability ratings.

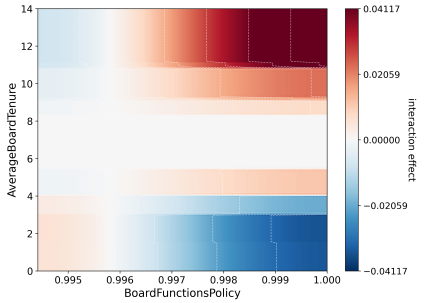
Figure 10(e) investigates the interaction between BoardFunctionsPolicy and AverageBoardTenure, combining the robustness of governance processes as a signal of oversight strength with tenure length as a measure of experience continuity to assess whether a stronger policy framework complements accumulated expertise recognized



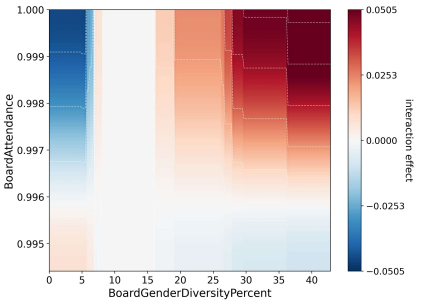
(e) BoardFunctionPolicy and AverageBoardTenure for G



(f) BoardGenderDiversity and BoardAttendance for G



(g) BoardFunctionPolicy and AverageBoardTenure for ESG



(h) BoardGenderDiversity and BoardAttendance for ESG

Fig. 10: Key feature interaction effect plots

by WDForest as favorable for governance performance. The heatmap reveals a dominant high-interaction region as BoardFunctionsPolicy approaches 0.999–1.000 and AverageBoardTenure lies between 10–14, with warmer tones indicating synergy where structured oversight and seasoned board composition reinforce each other to deliver system-level gains. A transitional bright band around tenure 5–7 and policy 0.996–0.998 suggests minimal interaction when both dimensions sit at mid levels, reflecting balanced yet underpowered governance dynamics. In the lower-left corner, cool pockets with policy 0.994–0.996 and tenure 0–3 mark negative interaction where weak formalization coexists with limited experience, pointing to compounded governance risk. Another cool strip at low tenure 1–3 even under very high policy near 0.999–1.000 shows that policy strength alone does not offset shallow experience. Overall, the pattern highlights the policy–tenure complementarity in governance pillar evaluation, where coordinated reinforcement of rule execution and board stability is essential for achieving high sustainability ratings.

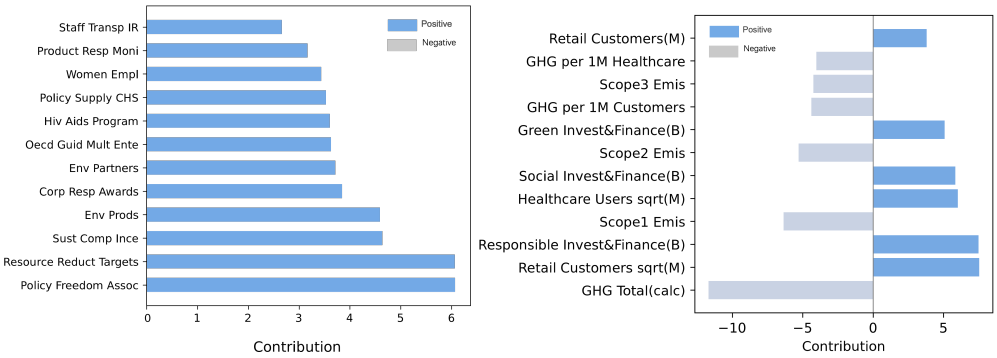
Figure 10(f) characterizes the interaction between board attendance and the percentage of women on the board, treating diligence as a signal of rule execution and gender diversity as a structural indicator of inclusion to assess whether higher

attendance amplifies the governance benefits of diversity recognized by WDForest as favorable for the governance pillar. The heatmap reveals a dominant high-interaction region when attendance rises to about 0.998–1.000 and diversity lies between 25–40, with warmer tones indicating that diligent oversight and diverse composition reinforce each other to deliver system-level governance gains. A bright band near mid-range values, centered around attendance 0.996–0.997 and diversity 12–18, signals weak interaction when both dimensions sit at moderate levels, reflecting balanced yet underpowered governance dynamics. In the upper-left corner, cool patches appear where attendance is very high at 0.999–1.000 but diversity is very low at 0–8, indicating negative interaction: high attendance alone cannot offset the imbalance caused by homogeneity, which tends to limit discussion quality and constrain oversight perspectives. The pronounced horizontal gradient suggests that diversity is the dominant driver while attendance acts as an amplifier; once diversity crosses roughly the 18–20 threshold, the joint effect strengthens rapidly. This configuration echoes Figure 10(e): the previous figure highlights that strong policy needs seasoned tenure to produce robust governance, whereas Figure 10(f) emphasizes that high attendance requires sufficient diversity to translate into substantive oversight quality. To achieve higher sustainability ratings within the governance pillar, attendance must improve and diversity must increase.

Figure 10(g) shows how codified board functions and tenure continuity combine to turn environmental and social progress into durable, governance anchored ESG performance recognized by WDForest. A broad warm plateau appears when board functions policy approaches 0.999 to 1.000 and average board tenure reaches 10 to 14, indicating strong synergy: structured oversight and seasoned directors reinforce one another, increasing the system's capacity to absorb and transmit E and S gains into composite ESG outcomes. A near-neutral bright band around tenure 5 to 7 with policy 0.996 to 0.998 signals ESG inertia: governance is orderly but underpowered, so cross-pillar benefits do not fully materialize. Cool pockets at tenure 0 to 3 under low policy 0.994 to 0.996 mark negative interaction, where weak formalization and limited experience compound governance risk and dampen ESG throughput. A further cool strip along tenure 0 to 3 even under very high policy close to 0.999 to 1.000 shows that policy strength alone cannot offset shallow experience; misalignment between rule intensity and institutional memory reduces ESG realization.

Figure 10(h) is read through an ESG lens as an enabling channel that turns governance quality into composite ESG gains. It characterizes how reliable meeting participation and meaningful gender representation work together to transmit environmental and social progress into durable ESG performance recognized by WDForest. A broad warm plateau emerges when diversity reaches roughly 28–42 percent and attendance rises to 0.998–1.000, signaling an ESG amplifier zone where inclusive deliberation and steady presence align board decisions with long-horizon sustainability goals. A near-neutral bright band around 10–17 percent diversity with attendance near 0.996–0.999 indicates ESG inertia, showing that mid-level governance delivers order but does not fully mobilize cross-pillar benefits. The deep blue pocket at very low diversity of 0–6 percent, even under near-perfect attendance, marks an ESG bottleneck where homogeneous perspectives limit the conversion of E and S improvements

into governance-anchored outcomes.A faint cool shelf at low attendance combined with single-digit diversity further depresses composite performance, highlighting a dual drag on ESG throughput.



(a) Hangzhou Bank 2020 (b) Ping An Insurance 2022

Fig. 11: Prediction contribution plots in ESG ratings

5.5.4 Local interpretation

To demonstrate WDForest's capacity for transparent, instance-level ESG assessment, we conduct a local interpretability analysis that deconstructs the rating for individual entities into traceable feature contributions. This approach moves beyond aggregate performance metrics to provide actionable insights into the specific drivers of sustainability scores for specific companies under specific conditions. Figure 11 presents a local interpretability analysis for Hangzhou Bank (2020) and Ping An Insurance (2022) using WDForest performance decomposition, quantifying the total contributions of the top 12 features in composite ESG ratings. Blue bars extending rightward indicate positive feature contributions, while leftward extensions denote negative impacts.

Figure 11(a) depicts Hangzhou Bank's ESG prediction contribution profile for 2020. In this study, bars represent marginal contributions to the overall ESG score, blue indicates positive values, gray indicates negative values, and the x axis measures contribution. Governance and social dimensions provide the primary uplift. On governance, disclosure and transparency, product responsibility monitoring, risk management and compliance processes, anti corruption and accountability, partnerships, and corporate responsibility recognition improve compliance quality, reduce regulatory penalties and operational risk, and sustain reputation and customer trust. On social, employee structure and training programs, women employees and inclusion practices, consumer protection and after sales service, employee relations and career development enhance customer experience and business resilience, lower mis selling and complaints, and stabilize deposits and lending relationships, thus generating larger positive contributions in the model. Environmental items are also positive but smaller in magnitude, driven by green products and services, resource efficiency and reduction targets, and energy

upgrades in branches and offices. For banks, direct emissions are relatively low, and environmental performance is reflected mainly through business policy and product design and through the management of clients' environmental risks rather than the bank's own absolute emissions. This produces a contribution profile where governance and social dimensions are the main drivers and the environmental dimension adds supplementary gains. To further strengthen the environmental dimension, the bank can raise the share of green credit, set binding reduction targets, tighten environmental risk screening in lending and investment, improve energy efficiency of data centers and offices, and align management incentives with environmental outcomes to expand environmental contribution without increasing compliance risk.

Figure 11(b) shows Ping An Insurance's ESG prediction contribution profile for 2022. The pattern is environmental drag with support from social and governance factors. Environmental bars, including total greenhouse gas emissions and Scope 1, Scope 2 and Scope 3, are strongly negative, which indicates that both direct and indirect emissions remain the main pressure on the rating. For an insurance company, the largest driver of negative environmental impact is financed emissions that outweigh operational emissions. Under Partnership for Carbon Accounting Financial(PCAF) and International Financial Reporting Standards(IFRS) accounting, exposures in the investment and underwriting portfolio to high carbon sectors such as power, coal chemicals, steel, transport and construction are attributed to financed emissions, and these typically exceed Scope 1 and Scope 2 by an order of magnitude. Extensive office networks, data centers, cloud platforms and call centers add to electricity-related emissions, and consolidation of entities in the healthcare ecosystem such as hospitals and clinics can further raise Scope 2. In contrast, social and governance indicators such as customer coverage across retail and healthcare users and responsible, green and social financing deliver notable positive contributions, reflecting the breadth of the financial and healthcare ecosystem and a responsible stance in capital allocation. From a finance perspective, this mix supports long-term value and risk diversification through service expansion and responsible investment. To improve the aggregate ESG score, the company needs systematic decarbonization: set science based targets aligned with PCAF for investment and underwriting, increase green and renewable assets, tighten policies for high carbon sectors, reduce Scope 2 through renewable procurement and efficiency upgrades in data centers and buildings, and manage supply chain and travel related long tail emissions to lower both absolute and intensity metrics, thereby offsetting the environmental negatives.

6 Conclusion

ESG ratings have become a critical tool for assessing corporate sustainability, guiding investment decisions, and promoting transparent corporate practices globally. However, reliance on current subjective ESG rating methods produces inconsistent results and limited transparency. To address these limitations, this study developed a data-driven objective ESG rating method. The inherent constraints of expert-dependent approaches were overcome by systematically resolving challenges such as unobservable

variables and the performance-interpretability trade-off. Specifically, WDForest framework was designed to reconcile performance with transparency through an adaptive residual cascade and a widening pathway that progressively enriched feature representation. The depth component adapted to data complexity via sequential residual learning, with each layer correcting errors from its predecessor. Simultaneously, the wide component employed a tree ensemble to generate structured, leaf-based feature representations. These representations enhanced the model's capacity to fit complex target functions, while their deterministic nature ensured the composite prediction remained fully decomposable and traceable across all ESG assessment contexts.

Empirical evaluations across the E, S, G pillars and overall ESG tasks demonstrated WDForest's dual strengths. The framework consistently delivered strong predictive validity, reflected by balanced error metrics, R^2, SRCC, and distributional fidelity. Equally important, its multi-layer interpretability scaffolding through layer-wise feature importance, individual feature-effect curves, and per-tree contribution tracing revealed domain-specific drivers and nonlinear response shapes. These analyses uncovered patterns such as diminishing returns and inverted U-shaped effects in governance-related indicators, as well as interaction structures in which complementary sustainability practices jointly yielded outsized improvements. Case-level examinations further showed that WDForest could generate granular, auditable narratives linking observed outcomes to specific features, layers, and trees, thereby turning opaque scores into actionable, verifiable explanations.

Building on this robust foundation, several avenues for future research emerged. First, extending WDForest to diverse markets, datasets, and rating taxonomies would further test its generalizability under alternative methodological conventions. Second, integrating unstructured disclosures through advanced text-processing pipelines could enrich the "wide" pathway with semantic signals while preserving auditability. Third, adapting the framework toward causal inference and policy impact evaluation would open the door to counterfactual analysis, stress testing, and fairness audits across ESG dimensions. These directions promised to deepen WDForest's contribution to the ESG analytics field, advancing both scientific rigor and practical utility.

Authors' Contributions Yunduo Han contributed to conceptualization, methodology, software development, formal analysis, visualization, and writing of the original draft. Wanan Liu contributed to conceptualization, supervision, funding acquisition, and review and editing of the manuscript. Ming Qian, Xingyu Lan, Congyuan Pang, Naixi Chen, Yao Zou, Feifei Wang, and Keyi Zhu contributed to data curation, investigation, validation, and review and editing of the manuscript. All authors read and approved the final manuscript.

Data availability and access The data that support the findings of this study are available from Refinitiv. Data may be available from the corresponding author upon reasonable request and with permission from the data provider and the authors' institution.

Declarations

Competing interest The authors have no relevant financial or non-financial interests to disclose.

Ethical and informed consent for data used Not applicable to this paper.

References

Alqaraawi, A., Al-Qurishi, M., Al-Rakhami, M.: On the Reliability of Post-hoc Explainable AI Methods. arXiv preprint arXiv:2305.18738 (2023)

Abdullah, A., Chowdhury, T., Roni, M.: Green innovation and esg performance: A machine learning approach. Sustainability **15**(15), 11578 (2023)

Arrieta, A.B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Herrera, F.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion **58**, 82–115 (2020)

Agarwal, R., Melamed, L., Frosst, N., Zhang, X., Caruana, R., Hinton, G.: Neural additive models: Interpretable machine learning with neural nets. In: NeurIPS, pp. 4200–4211 (2021). NeurIPS

Aronszajn, N.: Theory of reproducing kernels. Transactions of the American Mathematical Society **68**(3), 337–404 (1950) https://doi.org/10.2307/1990404

Amel-Zadeh, A., Serafeim, G.: Why and how investors use esg information: Evidence from a global survey. Financial Analysts Journal **74**(3), 87–103 (2018)

Babaei, S., Chen, Y., Liu, X., Liu, Y., Rajan, D.: SAFE-ML: A Framework for Robust and Explainable Machine Learning in Finance. arXiv preprint arXiv:2402.09458 (2025)

Berg, F., Kölbel, J.F., Rigobon, R.: Aggregate confusion: The divergence of esg ratings. Review of Finance **26**(6), 1315–1344 (2022)

Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001) https://doi.org/10.1023/A:1010933404324

Cort, T., Esty, D.C.: ESG Standards: A Path to Convergence. Yale Center for Business and the Environment (2020)

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, pp. 785–794. ACM, San Francisco, CA, USA (2016). https://doi.org/10.1145/2939672.2939785 . ACM

Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory **13**(1), 21–27 (1967) https://doi.org/10.1109/TIT.1967.1053964

Chatterji, A.K., Levine, D.I., Toffel, M.W.: How well do social ratings actually measure corporate social responsibility? Journal of Economics & Management Strategy **18**(1), 125–169 (2009)

Christensen, H.B., Serafeim, G., Sikochi, A.: Why is corporate virtue in the eye of the beholder? the case of esg ratings. The Accounting Review **97**(1), 147–176 (2022)

D'Amato, V., D'Ecclesia, R.L., Levantesi, S.: ESG score prediction through random forest algorithm. Baikal School of BRICS Working Paper Series (2022)

Di Giacomo, D., D'Adamo, P., Stanam, B.: The divergence of esg ratings: An analysis of italian listed companies. Economies **9**(4), 177 (2021)

Dorfleitner, G., Halbritter, G., Nguyen, M.: Measuring the level and risk of corporate responsibility–an empirical analysis from a corporate finance perspective. Journal of Banking & Finance **56**, 113–129 (2015)

Friedman, J.H., Popescu, B.E.: Predictive learning via rule ensembles. The Annals of Applied Statistics **2**(3), 916–954 (2008) https://doi.org/10.1214/07-AOAS124

Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics **29**(5), 1189–1232 (2001)

Gibson Brandon, R., Krueger, P., Schmidt, P.: Esg rating disagreement and stock returns. Financial Analysts Journal **77**(4), 104–127 (2021)

Hu, Y., Chen, S., Hu, M.: The impact of esg rating divergence on corporate greenwashing. Finance Research Letters **55**, 103947 (2023)

Hughes, A.T., Lonie, A.A., Paolucci, F.: Esg and financial performance: The role of disclosure. Accounting & Finance **61**(3), 3945–3972 (2021)

Hastie, T.J., Tibshirani, R.J.: Generalized additive models. Statistical Science **1**(3), 297–318 (1986)

Kim, J.H., Kim, M.K.: A study on the esg evaluation prediction model based on lightgbm. The Journal of the Korea Institute of Information and Communication Engineering **27**(7), 883–890 (2023)

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (2017). NeurIPS. Also available at arXiv:1711.04416

Kotsantonis, S., Serafeim, G.: Four things no one will tell you about esg data. Journal of Applied Corporate Finance **31**(2), 50–58 (2019)

Khan, M., Serafeim, G., Yoon, A.: Corporate sustainability: First evidence on materiality. The Accounting Review **91**(6), 1697–1724 (2016) https://doi.org/10.2308/accr-51383

Ke, G., Xu, Z., Zhang, J., Bian, B., Liu, T.-Y.: Deepgbm: A deep learning framework distilled by gbdt for online prediction tasks. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2574–2583. ACM, Anchorage, AK, USA (2019). https://doi.org/10.1145/3292500.3330668 . ACM

Lou, Y., Caruana, R., Gehrke, J., Hooker, G.: Accurate intelligible models with pairwise interactions. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 623–631 (2013). ACM

Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017) (2017). NeurIPS

Lee, J., Lee, S., Lee, J.: Predicting ESG Scores with Text Data using ESG-BERT. arXiv preprint arXiv:2209.05811 (2022)

Li, Y., Yang, M.: The impact of esg rating divergence on stock liquidity. In: Proceedings of ICEDBM 2023, pp. 235–242. Atlantis Press, Dordrecht, Netherlands (2024). ICEDBM

Nori, H., Caruana, R., Bu, Z., Shen, J.H., Kulkarni, J.: Accuracy, interpretability, and differential privacy via explainable boosting. In: ICML, pp. 8227–8237. PMLR, Virtual Event (2021). ICML

Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. Journal of the Royal Statistical Society: Series A (General) **135**(3), 370–384 (1972) https://doi.org/10.2307/2344614

Plevris, V., Solorzano, G., Bakas, N.P., Ben Seghier, M.E.A.: Investigation of performance metrics in regression analysis and machine learning-based prediction models. Applied Sciences **12**(19), 9799 (2022) https://doi.org/10.3390/app12199799

Quinlan, J.R.: Induction of decision trees. Machine Learning **1**(1), 81–106 (1986) https://doi.org/10.1007/BF00116251

Refinitiv: Refinitiv Eikon. Refinitiv. Financial information and data platform providing real-time market data, news, and analytics (2018)

Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016). ACM

Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)

Suarez Gifi, F.: Unveiling the blackbox within ESG ratings' blackbox: Toward a framework for analyzing AI adoption and its impacts. Preprint (2024)

Siering, M.: Natural language processing for esg analysis of corporate reports. In: Proceedings of the 56th Hawaii International Conference on System Sciences (2023). HICSS

Shi, Y., Li, J., Li, Z.: Gradient boosting with piece-wise linear regression trees. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI) (2019). IJCAI. IJCAI 2019

Solanke, B.: The Role of Explainable AI in Forensic Cybersecurity. In: Explainable AI in Healthcare. Springer, Cham (2023)

Schulte, M., Rügamer, D.: Additive model boosting: New insights and path(ologie)s. In: Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR, Mai Khao, Thailand (2025). PMLR

Salih, A.M., Raisi-Estabragh, Z., Boscolo Galazzo, I., Radeva, P., Petersen, S.E., Lekadir, K., Menegaz, G.: A perspective on explainable artificial intelligence methods: Shap and lime. arXiv (2023) https://doi.org/10.48550/arXiv.2305.02012 arXiv:2305.02012 [stat.ML]

Stocco, D.: Esg ratings explainability through machine learning techniques. Annals of Operations Research (2023)

Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley, Reading, MA (1977)

Vapnik, V.N.: Statistical Learning Theory. Wiley, New York, NY, USA (1998)

Zheng, G., Zhang, Y., Zeng, Y., Xu, Z.: Predicting corporate esg performance using machine learning. Sustainability **15**(13), 10118 (2023)