Journal of
Risk and Financial Management

MDPI

Systematic Review

Predicting, Using, and Assessing ESG Signals: A Tripartite Systematic Review of Machine Learning in Sustainable Finance

Imane El Imami [1,*], Abdelkader El Alaoui [2], Bassma Guermah [3], Said Ouatik El Alaoui [1] and Miklos Vasarhelyi [4]

[1] Engineering Sciences Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra 14000, Morocco; ouatikelalaoui.said@uit.ac.ma
[2] Rabat Business School, International University of Rabat, Rabat 11100, Morocco; abdelkader.elalaoui.2024@gmail.com
[3] TICLab, International University of Rabat, Rabat 11100, Morocco; bassma.guermah@uir.ac.ma
[4] Rutgers Continuous Auditing and Reporting Lab (CARLab), Rutgers Business School, Rutgers University, Newark, NJ 07102, USA; miklosv@business.rutgers.edu
* Correspondence: imane.elimami@uit.ac.ma

Abstract

Environmental, Social, and Governance (ESG) ratings increasingly shape capital allocation, corporate strategy, and regulatory oversight, yet their credibility is constrained by methodological opacity, rating divergence, and greenwashing risk. Prior reviews treat machine learning (ML) in ESG as a prediction problem. We identify an emerging research trajectory in which ML is increasingly used not only to consume ESG signals but also to verify their construction and credibility. Drawing on signaling theory, we conduct a PRISMA-guided systematic review of 127 peer-reviewed studies from Scopus and Web of Science to examine how machine learning (ML), deep learning (DL), Natural Language Processing (NLP), and Explainable AI (XAI) are transforming ESG rating analysis. We develop a tripartite framework classifying studies by the functional role of the ESG score: predicted (n = 29), used (n = 57), or assessed (n = 41). Our central contribution is the first synthesis of the methodological-assessment stream, organized into four clusters: XAI reverse-engineering of proprietary scoring functions, divergence reconciliation, greenwashing detection, and unsupervised industry-materiality clustering. The evidence assembled in this stream indicates that ESG ratings weight low-cost aspirational disclosure heavily relative to costly performance evidence, suggesting that greater reliance on aspirational disclosure relative to performance evidence may increase greenwashing risk, consistent with signaling-theory concerns. A study-level validation appraisal further shows that the most extreme fit statistics often arise in target-proximal reconstruction or non-temporal validation settings, cautioning against interpreting high R^2 as evidence of transferable out-of-time forecasting.

Keywords: ESG; machine learning; deep learning; explainable AI; sustainable finance; NLP; greenwashing

Check for updates

Academic Editor: Thanasis Stengos

Received: 17 July 2026
Revised: 30 August 2026
Accepted: 31 August 2026
Published: 8 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Over the past few years, Environmental, Social, and Governance (ESG) factors have rapidly gained significant interest as a key component and a core standard of economic stability, corporate strategy, investment processes (van Duuren et al., 2016; Vannoni & Ciotti, 2020), and strategic planning across firms (H. Lee et al., 2024b). In the US, assets

managed under ESG mandates increased from $3 trillion in 2010 to more than $17 trillion in 2020, demonstrating the rising demand from regulators, investors, stakeholders, and corporate leaders to prioritize corporate sustainability (Read, 2023; Tsalis et al., 2023), social responsibility, and transparent governance practices (Annesi et al., 2025; Zhan, 2023). The purpose of these ratings is to give investors a consistent and comprehensive snapshot of how companies rank across a wide spectrum of ESG dimensions, from carbon emissions to labor practices and board diversity. Companies can also analyze their ESG performance to identify potential avenues for improvement, thereby enhancing their market competitiveness and public investment decisions. ESG ratings play a key role in promoting green innovation and enhancing financial performance (e.g., mitigating risk to reduce bond credit spreads) (J. Li et al., 2023; J. Wang et al., 2023). This is a result of international efforts. The notion of ESG was first introduced in the United Nations, The Global Compact (2004)[1], an influential report that emphasized the integration of ESG issues into long-term financial analysis and short-term investment processes; they subsequently established the Principles for Responsible Investment (PRI) in 2006, then adopted a 2030 Agenda for Sustainable Development Goals (SDGs) in 2015 (Friede, 2019; Friede et al., 2015), suggesting standard metrics for sustainability disclosures.

Traditional ESG evaluation methods have been widely criticized due to the obscurity of the calculation methods (H. Zhang et al., 2024). Traditional scoring methods have been described as static, opaque, non-transparent, inconsistent, and inefficient in dealing with large and timely data (Berg et al., 2022; Dorfleitner et al., 2015), usually relying on company-reported and annualized data collected manually and evaluated by expert opinion. These constraints lead to a high level of heterogeneity and discrepancies in ESG assessments across different rating agencies (Billio et al., 2021). For example, a company might have a lower ESG score in one industry but a higher score in another, or receive different ratings from various agencies (Dwivedi et al., 2023), making it difficult to compare scores issued by various rating providers (e.g., Refinitiv, MSCI, S&P, LSEG, Wind), especially with market fluctuations (Chatterji et al., 2016). This variation demonstrates the subjective nature of many evaluations (Gyönyörová et al., 2023), undermining their credibility and eroding the confidence of stakeholders who depend on these ratings for important decisions (Sariyer et al., 2024). Hence, the need for more trustworthy, comprehensible, and unbiased assessment methodologies is crucial (Amir & Serafeim, 2018; H.-Y. Lin & Hsu, 2023).

In order to mitigate these challenges, emerging technologies of artificial intelligence (AI), such as ML and DL, enable the assessment and prediction of ESG ratings (e.g., analyzing the informational content of balance sheet variables) more reliably (D'Amato et al., 2024). As pressure on investment managers to incorporate ESG into their portfolios increases, ML is identified as a major enabler of large-scale sustainable investment that uses advanced data analysis to assess and emphasize essential data (Macpherson et al., 2021). On the one hand, ML algorithms can be applied to extract relevant information from unstructured text and identify patterns and anomalies in large corpora, thereby enhancing and increasing the efficiency of ESG assessments (Xu, 2024). On the other hand, DL models—especially NLP and transformer-based architectures—are useful for extracting large bodies of text in company reports, social media posts, news articles and earnings calls (Kantos et al., 2023) to interpret ESG-related content and extract sensitive and hidden narratives of ESG signals that could impact the sustainability profile of the firm (Serafeim, 2020). This integration of AI and ML is reshaping the ESG data management landscape, offering the potential to streamline data collection, improve accuracy, and provide more timely insights, thus laying the groundwork for more sophisticated methodologies for ESG scoring.

Although there is increasing interest, current available reviews and studies are either narrow or lack a systematic analysis. Very few studies systematically synthesize the practi-

cal effectiveness of various ML and DL models in the ESG context, encompassing the types of models employed, the data sources and performance metrics utilized, the challenges and future research directions. Most reviews focus on the conveniences of ESG criteria incorporation without explicitly investigating how certain ML and DL models improve the predictive quality and accuracy of ESG scores in distinct sectors and regions (Gangi et al., 2019). Although interest in ML- and DL-based ESG analysis is growing, existing reviews approach the literature through different but complementary lenses. Khaw et al. (2024) synthesize the factors influencing ESG performance through bibliometric and thematic analysis; Binzaiman et al. (2024) focus specifically on ML-based ESG-score prediction; and Seow (2025) provides a broader synthesis through bibliometric, thematic, and TCCM analyses. These reviews provide valuable insights into research themes, methods, contexts, and predictive applications, but they do not explicitly distinguish studies according to the functional role of the ESG signal within the empirical design. Berg et al. (2022) document substantial divergence across ESG rating providers, highlighting the methodological inconsistency and limited comparability of ESG signals. Yet the growing body of ML research that responds to this credibility problem—through XAI-based interrogation of rating methodologies, divergence reconciliation, alternative ESG-signal construction, and greenwashing detection—remains fragmented across existing review categories rather than synthesized as a distinct analytical stream. Section 2.2 positions the proposed tripartite framework against these prior approaches and develops the theoretical basis for this distinction. The purpose of this study is to fill this structural gap by conducting a comprehensive PRISMA-guided systematic review of 127 peer-reviewed studies (Page et al., 2021). The final corpus was coded through a structured, author-validated AI-assisted classification protocol and organized into a Tripartite Framework based on the functional role of ESG scores in each study (as illustrated in Figure 1). This procedure produced three mutually exclusive streams: ESG scores as prediction targets, ESG scores as financial or operational inputs, and ESG scores as objects of methodological assessment.

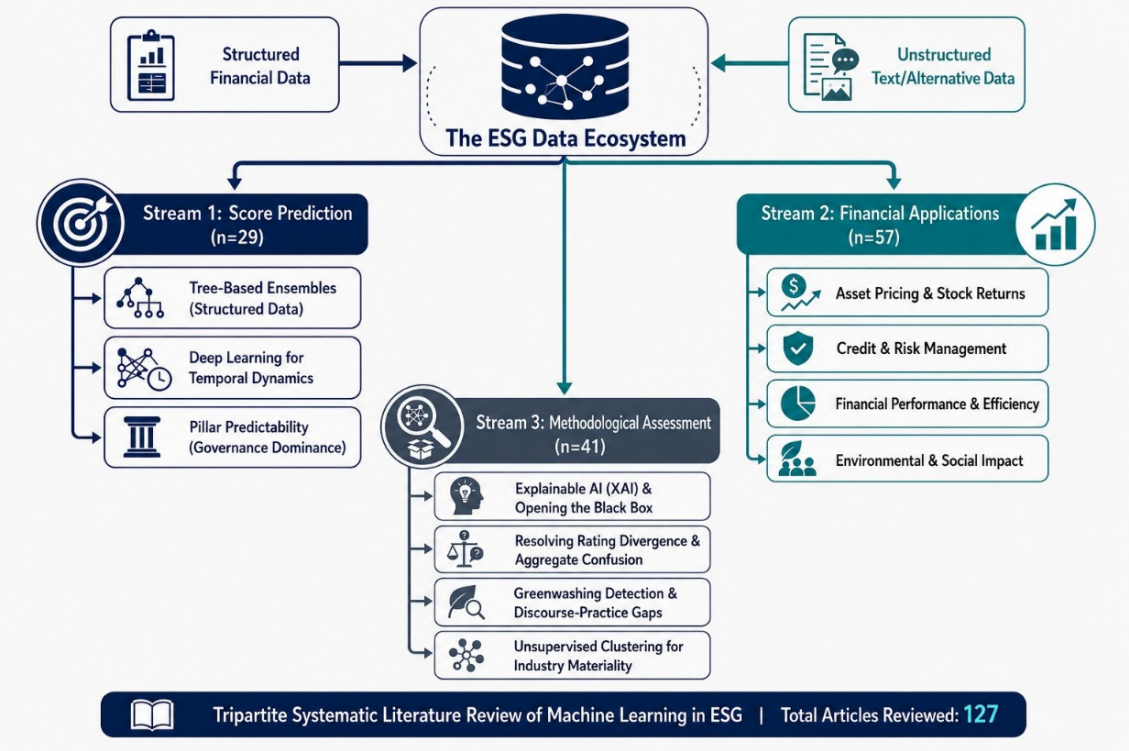


Figure 1. The tripartite framework of ML integration in ESG.

After classification, each stream was analyzed using a common extraction template covering model architecture, data source, ESG provider, prediction or assessment task, performance metric, geographic context, and methodological limitation. Our study aims to answer the following research questions (RQs):

- RQ1: To what extent do varying ML and DL architectures differ in their predictive performance and suitability for ESG scoring tasks?
- RQ2: How does the functional role of ESG scores—whether as a predictive target or an explanatory input—dictate the methodological choice?
- RQ3: How does the integration of unstructured alternative data vis-à-vis structured financial data enhance the resolution of ESG signals?
- RQ4: Which ESG pillars (E, S, or G) exhibit the highest machine-readability, and what structural factors contribute to these variances?
- RQ5: What are the critical barriers to model adoption, specifically regarding 'black box' interpretability and the detection of greenwashing?
- RQ6: Does the emergence of ML-based ESG methodological assessment indicate an increasing shift toward signal verification, and what does the composition of that literature reveal about the credibility of ESG ratings?

Accordingly, this review makes four contributions. First, it provides the first dedicated synthesis of the ESG methodological-assessment stream (n = 41), organizing it into four analytically distinct clusters—XAI-based black-box reverse-engineering, divergence reconciliation, greenwashing detection, and unsupervised industry-materiality clustering. Second, it introduces a methodological taxonomy of NLP architectures for ESG diagnostics that maps unstructured data sources to algorithmic applications and their characteristic failure modes. Third, it establishes an empirically grounded pillar machine-readability hierarchy and explains it structurally rather than descriptively. Fourth, it embeds these findings within a signaling-theory-based tripartite framework that distinguishes the ESG score as a prediction target, an explanatory input, and an object of methodological scrutiny. The framework, therefore, does more than organize applications: it differentiates whether ESG research reconstructs, relies on, or verifies the underlying signal, making assumptions about signal validity explicit across the literature.

The paper is structured as follows: Section 2 reviews the ESG and ML/DL background. Section 3 describes the review protocol and screening procedure. Section 4 details the data analysis process through statistical results, figures and tables. Section 5 discusses the limitations and challenges, then proposes future directions, and lastly, Section 6 concludes the paper.

2. ESG and ML/DL Background

The theoretical material in this section is drawn from outside the PRISMA corpus. The review protocol described in Section 3 governs the 127 empirical studies synthesized in Section 4; the signaling, legitimacy, and decoupling literatures discussed below were identified separately and serve as an interpretive scaffold for that synthesis rather than as review objects in their own right.

2.1. Theoretical Framing: ESG Ratings as Signals

ESG ratings exist to resolve an information asymmetry. Firms possess private information about their sustainability conduct; investors do not. Signaling theory (Spence, 1978) holds that this asymmetry is resolved only when signals are costly—that is, when they are cheaper to send for high-quality types than for low-quality ones (Connelly et al., 2011). Signals that are equally cheap for all types cannot separate them and collapse into pooling equilibria.

ESG rating agencies function as signal intermediaries: they convert firm disclosure into a composite public signal. This intermediation introduces a second-order problem. If the intermediary weights cheap signals (e.g., policy commitments, target declarations, stated intentions) over costly ones (e.g., audited emissions reductions, verified labor outcomes), then the resulting rating cannot separate substantive performers from symbolic ones. Legitimacy and decoupling theory (Meyer & Rowan, 1977; Suchman, 1995) predicts exactly this: organizations adopt ceremonial structures that secure legitimacy without altering operations. This framing yields the organizing logic of our review. The three streams are not merely topical categories; they are three distinct epistemic postures toward the ESG signal:

- Stream 1—Predicting the signal. ML reconstructs agency scores from firm attributes. The agency rating is treated as ground truth; the research problem is approximation error.
- Stream 2—Using the signal. ML consumes ratings as explanatory inputs to financial outcomes. The rating is treated as a valid measure; the research problem is economic association.
- Stream 3—Assessing the signal. ML is turned on the rating itself. The rating is treated as an object of suspicion; the research problem is signal credibility.

Streams 1 and 2 share an assumption that Stream 3 rejects. Both are signal-taking: they presuppose that the rating measures what it claims to measure. Stream 3 is signal verification. The emergence and growth of Stream 3 since 2023 is consistent with a predicted response to a credibility problem: when signal validity is in doubt, verification technology acquires economic value. We treat this correspondence as an interpretive frame rather than as a demonstrated causal account. The next subsection explains why this signal-oriented distinction is not captured by the principal organizational frameworks used in prior ESG reviews; Section 4.5 then assesses what the composition of the corpus can and cannot establish about the emergence of verification-oriented research.

2.2. Positioning the Tripartite Framework Against Prior Review Frameworks

Existing reviews organize ESG and ML research through complementary but analytically different lenses. Seow (2025), for example, applies the TCCM framework to examine the literature according to theory, context, characteristics, and methodology, alongside bibliometric and thematic analyses. This provides a multidimensional account of how ESG–ML research is theoretically grounded, where it is conducted, and which methods it employs. The tripartite framework addresses a different question: what epistemic role does the ESG signal play within the empirical design? Studies employing similar methods and operating in similar contexts may nevertheless treat an ESG rating as a prediction target, an explanatory input, or an object whose validity itself requires scrutiny.

Prediction-centered reviews impose a different limitation. Binzaiman et al. (2024) explicitly organize their review around the use of ML to predict ESG scores, with model accuracy, datasets, methods, and prediction performance forming the principal analytical focus. This perspective provides valuable evidence on predictive techniques but, by design, gives limited visibility to studies in which ESG scores are consumed as inputs or in which ML is used to investigate rating construction, divergence, or credibility rather than to reproduce the score itself. Bibliometric and thematic classifications capture yet another dimension of the literature. Khaw et al. (2024), for instance, identify bibliometric clusters and seven substantive themes explaining ESG performance, including digital transformation, financial strategies, regulatory mechanisms, governance, board diversity, and organizational excellence. Such classifications organize studies according to shared subject matter and determinants of ESG performance. They do not, however, distinguish

between studies that accept an ESG signal for prediction or application and studies that question how that signal was produced.

The tripartite framework therefore complements rather than replaces these prior approaches. Its contribution is to shift the unit of classification from the attributes or topic of a study to the functional status of the ESG signal. Grounded in the signaling perspective developed in Section 2.1, the framework distinguishes signal reconstruction (Stream 1), signal use (Stream 2), and signal verification (Stream 3). This distinction has theoretical consequences: it separates research that provisionally accepts the rating as an informative signal from research that treats the validity of the signal itself as an empirical problem. It thereby makes visible the methodological-assessment literature and permits the review to examine whether the growth of verification-oriented research is consistent with increasing concern over ESG signal credibility rather than merely cataloging another application of ML. Table 1 summarizes these differences and shows that the proposed framework is distinguished not primarily by database coverage or review technique but by its organizing unit: the functional and epistemic role assigned to the ESG signal within each study.

Table 1. Positioning against prior systematic reviews of ML in ESG context.

Feature	Seow (2025)	Binzaiman et al. (2024)	Khaw et al. (2024)	This Study
Framework	TCCM	PRISMA	Bibliometric + SLR	Tripartite (PRISMA)
Primary organizing	Theory	ESG-score	ESG-score	Functional
Logic	Context, characteristics, and methodology	Prediction	Performance determinants and thematic clusters	role of the ESG signal
Corpus size	57	21	216	127
Databases	WOS, Scopus	Google Scholar	WOS, Scopus	WOS, Scopus
ESG as dependent variable	Partially	Yes	No	Yes (n = 29)
ESG as independent variable	Partially	No	No	Yes (n = 57)
Assessment/XAI/ Greenwashing	No	No	No	Yes (n = 41)
Pillar-level predictability	No	No	No	Yes
NLP taxonomy for diagnostics	No	No	No	Yes

2.3. Discrepancies in ESG Scores Across Providers

ESG ratings are designed to summarize firm-level sustainability performance across environmental, social, and governance dimensions. However, unlike financial accounting measures, ESG ratings are not produced through a universally standardized measurement architecture. Rating providers differ in indicator selection, data sources, treatment of missing information, industry materiality assumptions, and aggregation rules. Consequently, the same firm may receive substantially different ESG evaluations across providers, not necessarily because its sustainability performance changes but because the underlying rating methodologies encode different definitions of materiality and risk. Generally, the ESG score is determined based on a mix of quantitative data, qualitative assessment, and industry factors; it is a weighted sum of scores assigned to individual components as defined in Equation (1). While the aim is to evaluate a firm's performance, including environmental footprint, social responsibility and corporate governance[2], there is no single

standard formula for calculating these scores, since each rating agency has its own proprietary methodology. This produces substantial heterogeneity across providers. These discrepancies occur because of variations in methodologies, data sources, scoring systems, and subjective judgments used by different agencies.

$$S_{ESG} = \sum_{p \in \{E,S,G\}} (w_p \times S_p) \quad (1)$$

where S_p is the pillar score, and w_p is the weight of the pillar.

It is important to note that while Equation (1) represents the traditional linear compensatory approach, some advanced methodologies are shifting toward non-compensatory or geometric aggregations to prevent severe controversies in one pillar from being masked by outperformance in another. The score calculation begins at the data-collection stage. Agencies collect quantitative and numerical information on components such as greenhouse gas emissions, energy use, employee retention rates and diversity numbers. Additionally, they examine some qualitative data including company policies, management practices and governance structures that pertain to ESG issues. This information is collected from multiple sources like financial statements and sustainability reports, news outlets, NGO publications and public databases[3]. The collected data is then classified into E/S/G categories. Each category is assigned a weight according to its significance for the company's industry sector. Companies are then scored based on specific criteria, often employing a numerical or letter grade system depending on the rating agency. Lastly, the weights of each individual category are aggregated to form the overall ESG Score of a given company. The score is frequently benchmarked against other firms within the same industry to provide a relative ranking, allowing for comparative analysis.

Despite these common steps, the methodological variations among agencies are significant. Researchers like Berg et al. (2022) have rigorously demonstrated that discrepancies in ESG ratings are primarily attributable to differences in how indicators are measured (56%) and the scope of data considered (38%) rather than just differences in weighting. This means that agencies may emphasize different aspects of ESG performance, collect varying sets of data, or apply subjective interpretations to qualitative information, all contributing to the observed "aggregate confusion". For instance, one agency might heavily penalize a company for its carbon footprint, while another might prioritize its diversity initiatives, leading to divergent overall scores for the same entity. Such inconsistencies pose a challenge for investors and stakeholders seeking clear, comparable insights into corporate sustainability (Agosto & Tanda, 2025). Appendix A depicts the differences in ESG score scales across six major rating agencies.

3. Research Methodology

3.1. Search Strategy and Literature Identification

Our study performs a Systematic Literature Review (SLR), which is commonly used to cover a research topic comprehensively and to provide evidence while minimizing subjectivity and bias (Boell & Cecez-Kecmanovic, 2015; Tranfield et al., 2003). The research is primarily quantitative in design and adopts the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol. This protocol improves the extensiveness, rigor, and reproducibility of literature reviews and presents a clear framework for the review process as a whole, as described in Figure 2.

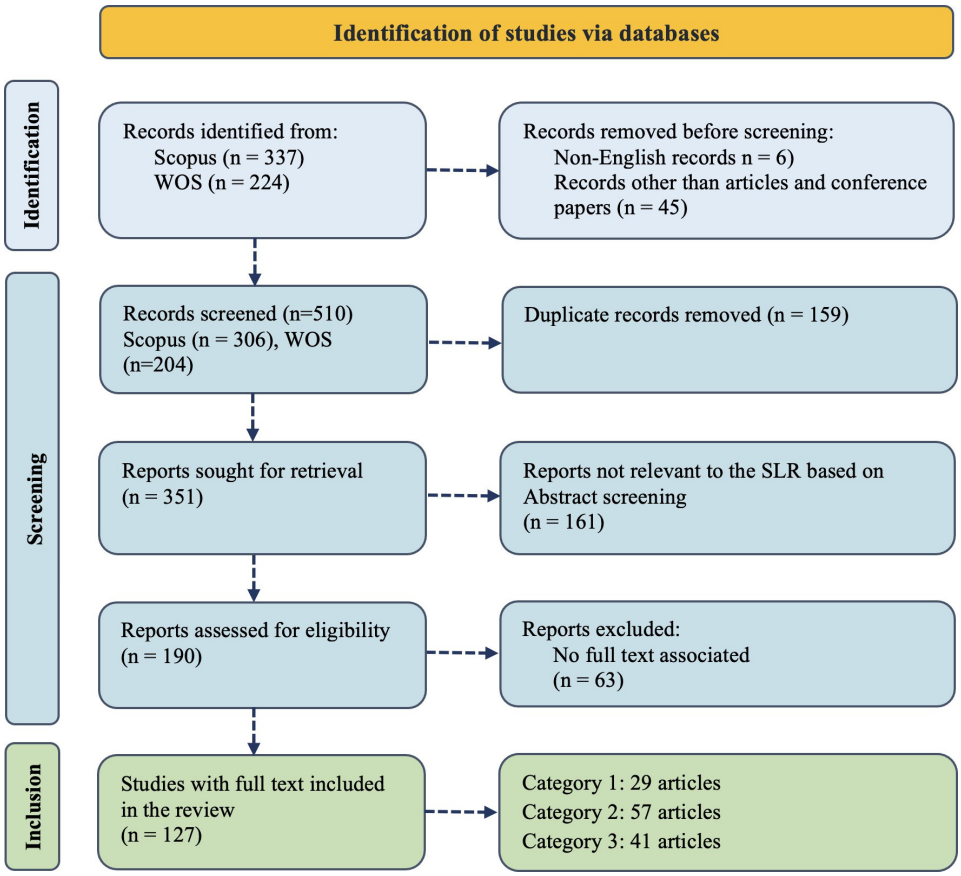


Figure 2. Preferred reporting items for systematic reviews and meta-analyses protocol.

The selection of keywords for this systematic literature review was strategically designed to capture the broad and emerging intersections between AI techniques, especially ML and DL, and their applications in predicting and assessing ESG ratings and scores or using them as inputs in financial applications such as credit scoring, portfolio optimization, and corporate social responsibility (CSR) performance. The keywords and terms that have been included in the search query "title, abstract and keywords" are: ("Artificial Intelligence" OR "AI" OR "Machine Learning" OR "Deep Learning" OR "Neural Network*" OR "Random Forest" OR NLP OR "XGBoost" OR "BERT") AND ("ESG rating*" OR " ESG scoring" OR "ESG score*" OR "ESG evaluation" OR "environmental social governance rating"). This strategy proved useful because some important articles were missed when field tags such as "title" or "title and abstract" were used. Additionally, the combination of keywords using the Boolean operators "AND" and "OR" was informed by prior research emphasizing the pivotal role of advanced computational methods in addressing sustainability challenges, particularly in improving the transparency, accuracy, and predictive power of ESG assessments. The database search was conducted in Scopus and Web of Science, with the final extraction completed in April 2026. No lower publication-year restriction was imposed. This led to the identification of 337 documents from Scopus and 224 from WOS. The complete database-specific search strings, searched fields, filters, and retrieval counts for Scopus and Web of Science are reported in Appendix E.

Scopus and Web of Science were selected because of their broad coverage of peer-reviewed business, finance, and management research and their structured metadata, which support reproducible screening and deduplication. Google Scholar was excluded because its broader coverage is accompanied by less consistent filtering, result stability, and export functionality, which limits reproducibility in a PRISMA-guided review. Preprints were excluded because they have not undergone formal peer review and may change

substantially across versions. We acknowledge that these choices may omit some recent AI-related research, particularly studies disseminated rapidly through preprint repositories or specialized databases such as IEEE Xplore and the ACM Digital Library. This limitation is acknowledged in Section 5. Before screening, 6 non-English records and 45 records other than articles and conference papers were removed (see Table 2 for inclusion/exclusion criteria). This yielded 510 records screened (Scopus n = 306, WOS n = 204). After eliminating 159 duplicate records via Zotero, 351 reports were sought for retrieval. Of these, 161 were excluded based on abstract screening for relevance to the SLR's scope. The remaining 190 reports were assessed for eligibility; 63 were excluded because no full text was associated with the record.

Table 2. Inclusion and exclusion criteria for article selection.

Inclusion Criteria	Exclusion Criteria
– English language publications – Peer-reviewed journal articles and conference papers – Primary focus on ESG scores or ratings – Empirical ML or DL experimentation – Utilizes quantitative ESG data or relevant textual corpora – Indexed in Scopus or Web of Science – No lower publication-year restriction; records eligible if available at the final search date in April 2026	– Non-English publications – Book chapters, editorials, commentaries, and preprints – ESG mentioned only peripherally, or pure CSR without quantitative scoring – Purely theoretical or conceptual studies (no computational modeling) – Lacks empirical data – Grey literature and non-indexed sources

Note: Purely theoretical and conceptual studies were excluded because the tripartite framework classifies studies according to the functional role of ESG scores within an empirical design. Without an empirical application, this role cannot be consistently coded using the same classification criteria. This exclusion applies only to the systematic corpus synthesized in Section 4; relevant theoretical literature is incorporated in Section 2 to provide the conceptual foundation for the framework.

The search was intentionally centered on explicit ESG scores, ratings, and assessments because the unit of analysis in the tripartite framework is the functional role assigned to an ESG score or rating within an empirical design. Broader literatures on sustainability performance, green finance, climate risk, or responsible investment were therefore not included unless they explicitly operationalized an ESG score, rating, pillar score, or comparable ESG assessment variable. Including studies without such a variable would expand the review into adjacent sustainability domains that cannot be classified consistently under the same functional-role criterion.

3.2. Data Extraction and AI-Assisted Coding Protocol

After the final corpus of 127 studies was established, each article was subjected to a structured data-extraction and coding procedure. For every study, we extracted the following bibliographic and methodological variables: author(s), year of publication, journal or conference outlet, database source, geographic context, ESG data provider, type of ESG variable, role of the ESG score in the empirical design, ML or DL model used, data type, prediction or assessment task, performance metrics, key findings, and stated limitations. The central coding decision concerned the functional role assigned to ESG scores or ratings in each study. Based on this criterion, the articles were classified into three mutually exclusive streams. Stream 1, "ESG score prediction," includes studies in which ESG scores, ESG ratings, or E/S/G pillar values are treated as the dependent variable or prediction target. Stream 2, "financial application," includes studies that use ESG scores or ratings as independent variables, explanatory variables, or features to predict financial, market, credit, operational, or environmental outcomes. Stream 3, "methodological assessment," includes studies that evaluate, explain, audit, reconcile, or challenge existing ESG rating

methodologies, as well as studies that develop alternative ESG assessment frameworks or signals in response to limitations of conventional scoring approaches.

To strengthen coding consistency, we used a two-stage AI-assisted classification procedure. First, the title, abstract, keywords, and, when necessary, the methodology and results sections of each article were processed using two large language models (LLMs), Gemini 3.1 and GPT-5.5, with the same standardized classification prompt reported in Appendix B. The LLM outputs were treated as preliminary coding suggestions rather than final classifications. Second, the authors validated the classifications against the full text, with particular attention to cases in which the two models disagreed. For these cases, final assignment was based on the study's primary analytical objective and the functional role of the ESG variable. A study remained in Stream 1 when an externally defined ESG score, rating, pillar, grade, or category constituted the principal supervised prediction or reconstruction target, even when XAI was subsequently used to interpret its drivers. By contrast, a study was assigned to Stream 3 when prediction or classification served primarily as one component of a broader effort to evaluate, redesign, compare, or construct an ESG assessment methodology or alternative ESG signal. This rule preserves mutually exclusive final classifications while explicitly recognizing genuine Stream 1/Stream 3 boundary cases. The final author-validated distribution was 29 studies in Stream 1, 57 in Stream 2, and 41 in Stream 3.

Across the 127 studies, Gemini 3.1 and GPT-5.5 assigned the same stream to 115 studies (90.6%). The final author classifications coincided with Gemini 3.1 for 118 studies (92.9%) and with GPT-5.5 for 124 studies (97.6%); all three classification sources were aligned for 115 studies (90.6%). The remaining twelve cases were therefore examined individually and are reported in Appendix C Table A4 with their preliminary labels, final author decision, and adjudication rationale. Because final author validation was conducted with the preliminary LLM outputs visible, these percentages are treated solely as descriptive coding concordance rather than independent inter-rater reliability. Cohen's κ is therefore not reported in the revised analysis.

The use of LLMs as preliminary coders also introduces limitations related to prompt sensitivity, model-specific bias, interpretation of compressed methodological descriptions, and reproducibility as model versions evolve. We mitigated these risks by applying the same prompt to both models and retaining all final classification decisions under author control following full-text review. The LLMs therefore functioned as an auxiliary consistency-checking layer rather than autonomous reviewers. Three studies, namely Del Vitto et al. (2023), Ferro et al. (2025) and Storchevoy et al. (2025) (see Appendix C Table A4), were retained as explicit S1/S3 boundary cases. As a sensitivity check, we recomputed the distribution by assigning all three cases to Stream 1 and, alternatively, all three to Stream 3. Under these bounding scenarios, Stream 3 ranges from 40 studies (31.5%) to 43 studies (33.9%), compared with 41 studies (32.3%) in the final classification. The overall distribution is therefore not materially dependent on the treatment of these three boundary cases.

To assess the predictive validity of Stream 1 more systematically, we conducted a targeted methodological appraisal of all 29 studies. For each study, we recorded whether prior-period ESG scores or ratings were used as predictors, whether E/S/G pillar scores or closely related provider scoring components entered the prediction design, the principal reported performance metric, the presence of a held-out or cross-validated evaluation, and whether longitudinal validation respected temporal ordering. Importantly, we distinguish autoregressive forecasting from target leakage: using information available at t to predict an ESG rating at t + 1 is a legitimate forecasting design when evaluation is performed on a chronologically later holdout. By contrast, reconstructing an aggregate ESG score from contemporaneous pillar scores or provider-defined scoring components is treated as

target-proximal reconstruction, and its goodness-of-fit is not interpreted as independent forecasting performance. The study-level appraisal is reported in Appendix D, Table A5.

4. Results and Discussion

Our analysis of the 127 reviewed studies reveals patterns in the application of ML and DL models, the data sources, the performance metrics, and the most influential or influencing ESG pillar(s). In this section, we begin with an overview of the principal data sources used and then turn to the three streams of research we identify: predicting the ESG scores as output, using the ESG scores as inputs to predict financial performance and assessing the ESG ratings' methodologies. One overarching finding of this review is that we observe a strong inter-relationship between the robustness and originality of ML/DL applications in the ESG domain and the diversity and quality of the data sources they rely on. As Figure 3 shows, major ESG rating providers—LSEG-Refinitiv, Bloomberg, SynTao Green Finance, MSCI, Wind, SP Global, Sustainalytics, CSI, Huazheng, and FTSE Russell—are the most frequent and serve as primary sources for ESG scores. LSEG-Refinitiv leads, appearing in 35 of the 127 studies (27.6%), followed by Bloomberg (23 studies, 18.1%), MSCI (16 studies, 12.6%), and Wind (14 studies, 11.0%). Taken as a category, ESG rating agencies and data providers account for 46.2% of all recorded data-source uses across the corpus (Table 3). Because a single study may draw on several sources, this category share should not be read as a share of studies.

Table 3. Data sources, typology, and extraction protocols in the reviewed literature.

Data Source Category	Type of Data	Extraction Protocol/Mechanics	Examples from Literature	Freq
ESG Rating Agencies & Data Providers	Structured ESG Scores, Raw ESG Metrics	• Bulk historical panel extraction via proprietary terminal APIs. • Point-in-time annual rating downloads.	Bloomberg, LSEG/Refinitiv, MSCI, S&P Global, Wind, SynTao Green Finance, Sustainalytics, Huazheng, Asset4	46.2%
Financial & Corporate Databases	Indices, Financial Ratios, Historical Pricing Data	• ISIN/Ticker matching protocols across multi-year windows. • Automated web scraping of exchange registries.	S&P 500 Index, EuroStoxx-600, CSI, FTSE Russell, Moody's Orbis, Morningstar, CSMAR, Compustat, CRSP, Yahoo Finance	31.0%
Unstructured & Alternative Data	News, Transcripts, Social Media, MD&A Reports	• LexisNexis/Factiva API queries. • Direct PDF parsing of annual reports. • Python/BeautifulSoup scraping of media and Reddit/Twitter APIs.	GDELT GKG, Reuters, Twitter, Reddit, NYTimes, Novethic, ESGToday, LexisNexis, Seeking Alpha, Corporate MD&A	13.9%
Public/Governmental Sources	Regulatory Frameworks, Official Reporting Standards	• Manual extraction from open-source government repositories. • NLP dictionary baseline construction.	UNGC, GRI, SASB, KCGS, China Statistical Yearbook, SIPO	8.8%

Note. Percentages denote each category's share of all recorded data-source uses across the 127 studies, not the share of studies. Because a study may draw on more than one source, frequencies sum to more than 127.

Yet, these scores are often combined in many studies with traditional, structured corporate and financial data retrieved from databases such as Yahoo Finance, Moody's Orbis, CSMAR, etc., generally used to track firms' performance, thus building confidence with

stakeholders and facilitating risk management. Remarkably, the presence of unstructured and governmental data sources shows the importance of NLP and Transformer-based models in capturing ESG signals that are timely and context-rich, from news articles (the Independent, the Guardian, NYTimes), annual reports (MD&A), earnings call transcripts (LexisNexis, Seeking Alpha), and social media (Reddit, Twitter). For reference, Table 3 shows all the data sources used, along with their frequency within the papers studied. It is important to note that a given article may use one or more data sources; the total of the frequencies therefore exceeds the number of reviewed studies.

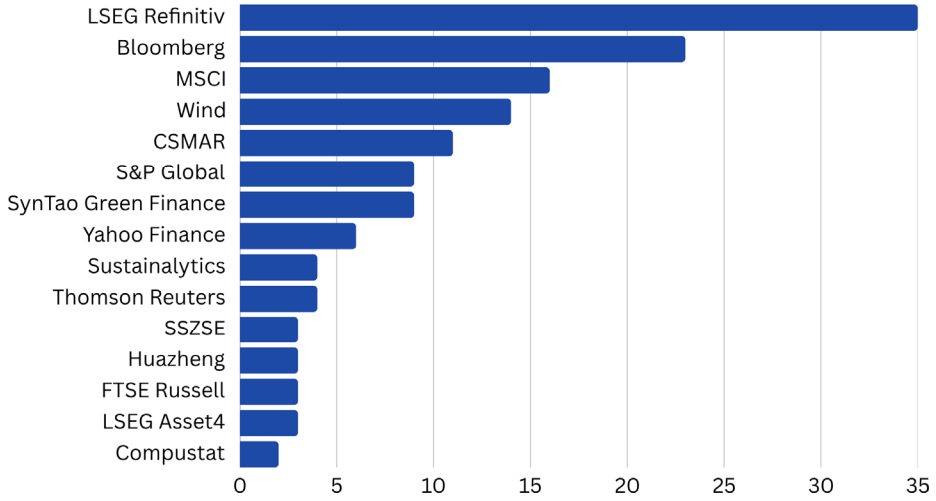


Figure 3. Top 15 data sources used in the reviewed literature.

4.1. ESG Scores as Dependent Variables

Twenty-nine peer-reviewed studies predicted ESG scores directly as the target variable for ML and DL architectures. These studies have adopted a range of approaches yet collectively highlight significant trends in the field of ESG scoring automation, investing decisions and rating transparency. In total, 22 models were applied in this stream, ranging from simple models such as Linear Regression (LR), Logistic Regression, Decision Tree (DT), and Naïve Bayes (NB), through intermediate-complexity models like Random Forest (RF), XGBoost, and LightGBM, to complex DL architectures such as CNN, RNN, ANN, LSTM, Stacked Generalization (S. L. Lin et al., 2025), Extreme Learning Machine (H.-Y. Lin & Hsu, 2023), and the SGM-BRR model. The evidence does not converge on a single algorithmic champion; it converges on a principle: the optimal architecture is determined by the structure, scale, and temporal properties of the input data, not by any universal algorithmic ranking. Three methodological clusters organize this evidence.

4.1.1. Structured Financial Data and Tree-Based Ensembles

Across Stream 1, reported predictive performance is strongly conditioned by both target proximity and validation design (Appendix D, Table A5). Some of the highest fit statistics arise when models reconstruct ESG ratings from information closely related to the target: Belkhiria et al. (2025), for example, obtain $R^2 = 0.988$ when the aggregate ESG score is modeled directly from its E, S, and G pillars, while David et al. (2025) report cross-validated R^2 approaching 0.996 using lagged ESG and ESG-derived components. However, extreme performance is not confined to such designs; H.-Y. Lin and Hsu (2023) report near-perfect R^2 under a random train-test split without lagged ESG or pillar-score predictors. We therefore distinguish score reconstruction, autoregressive forecasting, and out-of-time prediction rather than treating high goodness-of-fit as equivalent evidence of predictive generalization.

For AI-enabled FinTech applications combining financial data with social media sentiment, Gradient Boosting Machines implemented via H_2O AutoML similarly achieve Accuracy = 89% alongside real-time anomaly detection capabilities unavailable to static ensemble methods (Choudhari et al., 2025). J. Zhang and Zhao (2026) deploy XGBoost-SHAP with SMOTE-ENN on 15 structured indicators for Chinese A-share companies (2013–2022; Huazheng/CSMAR), achieving an Accuracy of 0.91. Two specialized sub-models serve distinct investment decision objectives: a level definition model (Accuracy = 0.92) for portfolio positioning and a risk identification model (Accuracy = 0.97) for downside ESG screening. Gradient Boosting also demonstrates cross-domain robustness when SDG macro-indicators replace firm-level financial as inputs, achieving Accuracy = 0.92 (A. Jain et al., 2025), confirming that the architecture's non-linear interaction detection is effective regardless of whether the feature set is micro-level balance-sheet data or macro-level sustainability indicators.

For researchers seeking to improve predictive accuracy when analyzing industries over time and dealing with complex, structured datasets, advanced Stacked Generalization Models (SGM-BRR) would be a powerful tool since they would intelligently combine prediction results from multiple base learners (RF, GBDT, XGBoost and LightGBM), correcting their weaknesses, and employing a random search algorithm to optimize models' parameters using a meta-learner such as Bayesian Ridge Regression. See (Z. Wang et al., 2024). This hybrid model demonstrates superior predictive performance on Chinese A-share data from 2012 to 2020, especially in the industrial, information technology, materials, energy, and health care industries, explaining 80% of the ESG information. Interestingly, in (H.-Y. Lin & Hsu, 2023), ELM and SVM outperformed XGBoost and RF when predicting ESG ratings of non-financial firms in Taiwan during the entire study period (2018–2021), including pandemic and non-pandemic periods, demonstrating a strong performance R^2 of 0.98 and RMSE of 0.8364. SVM is so effective because it can always find a hyperplane to separate data regardless of whether the space is very high-dimensional or not through "The kernel trick".

4.1.2. Deep Learning for Temporal ESG Dynamics

The static, annually refreshed nature of most agency ESG scores is a structural artifact of data collection cycles, not a reflection of how corporate sustainability evolves. (S. Chen et al., 2026) provide the most comprehensive architecture comparison for temporal ESG classification, benchmarking logistic regression, MLP, LSTM, RNN, GRU, Transformer, and Mamba on China A-share data (2014–2021). GRU achieves Accuracy = 83.64% and F1 = 0.7295, outperforming all competitors. The GRU's gating mechanism is well-matched to the structural pattern of ESG rating data, while the Transformer's under-performance reflects quadratic attention cost over long sequences and training data sparsity that annual ESG records cannot resolve. Gearing (debt-to-asset ratio) and total assets are the two most important predictors, confirming that leverage-based financial risk, not earnings, encodes the primary ESG signal accessible from financial statements. (S.-L. Yang, 2025) applies the GRU-LSTM hybrid to CDP carbon disclosure ratings (2020–2022), achieving Accuracy = 0.74 and F1 = 0.71, outperforming both logistic regression and Random Forest on all metrics. The model's advantage is located in its capacity to track sequential GHG emissions reporting quality trajectories, where prior-year disclosure behavior predicts next-year rating migration. (Y. Chen & Cheng, 2025) extend temporal modeling across a weighted ensemble (GM(1,1), ARIMA, LSTM) for Yangtze River Delta enterprises, where the ensemble achieves MAPE = 4.5%, outperforming LSTM alone (5.1%). The ensemble's advantage reflects complementary bias-variance profiles across different sub-periods.

Interestingly, traditional ML models remain insufficient for large-scale unstructured textual data and for predicting ESG ratings from news articles. In this regard, deep CNN

alongside transformer-based models can be particularly suitable, especially in capturing the dynamic correlations between events, sentiments, and topics in a large corpus of 3.7 million news articles covering 3000 US firms across approximately 200 countries (Aue et al., 2025). The model achieves a Mean Absolute Difference of 11.86 against Refinitiv Asset4 ratings while outperforming Transformer and DistilBERT variants. Additionally, KoBERT is increasingly used to automate fine-grained pillar classification (E, S, G) of earnings-call transcripts and news articles, as well as sentiment polarity scorings (H. Lee et al., 2025a), to be used later as features in a regression model in order to effectively compare the predicted ESG ratings to agency-assigned ones. Remarkably, complex models are not always better. K-NN demonstrated this theory by predicting the ESG performance of Turkish firms, solely based on past ESG grades (Taskin et al., 2025). The model achieved superior results, reflecting strong persistence and clustering in ESG ratings without overfitting. Figure 4 presents the distribution of the main models used in these works.

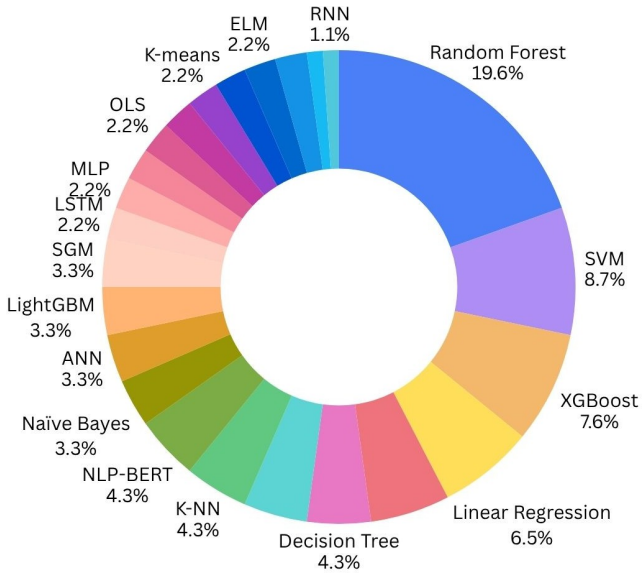


Figure 4. Distribution of ML and DL models used for predicting ESG scores. Colors are used solely to visually distinguish model categories and do not represent an additional classification.

4.1.3. Pillar Predictability: The Structural Hierarchy of E, S, and G

The prediction target was multifaceted; the aim of some studies is to predict the overall combined ESG score, while some focus on separate E/S/G pillars to detect the most predictable pillar and to reveal what features have the most influence. Reported pillar predictability appears contradictory across studies, but the disagreement is an artifact of differing input regimes rather than a genuine empirical conflict: Governance is most accessible where standardized disclosure data dominate; Environment is most predictable given rich operational data yet least reconstructible from financial statements alone; and Social remains intractable under all regimes. We develop each case below.

The Governance Pillar

Several recent studies make a strong case that the Governance (G) pillar is computationally transparent and predictable within the ESG framework, which makes it more amenable to ML and DL analysis. For instance, Ang et al. (2023) reported that their fixed-effects panel models yielded the best results when estimating Governance scores, simply because governance data such as audit committee composition, executive compensation policy, and board independence are generally more stable over time and more quantitative, structured, and consistently reported than social or environmental metrics. This is

substantially supported by Ferro et al. (2025), who applied RF to separate general, forward-looking "aspirational" features from concrete, backward-looking "performance" features. The key finding was that the Governance category weight had the highest magnitude on these actual "performance" metrics, meaning it was inherently more rooted in observable, standardized, and auditable indicators, hence more predictable to ML models. (Z. Wang et al., 2024) SGM-BRR achieves the largest single-model improvement for the G score (0.23 RMSE reduction).

Furthermore, the research by Del Vitto et al. (2023) indicates the Environmental (E) pillar to be the most remarkably predictable ESG pillar overall, revealing the highest accuracy in replication across all of the industries, namely Manufacturing, information and financial sectors; the Governance (G) pillar also demonstrates a significant contribution and exhibits strong predictability in US companies, notably achieving the lowest RMSE of 0.075. The predictability of the Governance score is influenced by specific features that vary from one country to another. For example, in the US, the independence of the chairman from the CEO role is a major factor, while in Europe, board member attendance rates and the percentage of non-executive board members are rewarded, and finally, improvement tools for employee compensation and audit committee definition play a major role in Chinese firms.

The Environmental Pillar

Environmental pillar predictability is context-dependent in a theoretically informative way. When operational emissions data (e.g., carbon intensity, energy consumption, GHG trajectories, water use) are included as inputs alongside financial ratios, the E pillar achieves the highest R^2 values within ensemble frameworks: the SGM-BRR model used in (Z. Wang et al., 2024) excelled in capturing energy consumption-related features in E dimension, supporting that its drivers are intrinsically connected to the financial and operational data they involved. Similarly, H. Lee et al. (2025a) demonstrated strong performance (R^2 of 0.92) when forecasting the E pillar using random-effect regression. This implies two assumptions: Firstly, environmental performance, particularly around metrics such as resource usage and carbon emissions, is being steadily standardized and has an effect on a firm's balance sheet; secondly, it seems that alternative data sources such as news are specifically well-suited to quickly capturing high-profile and dynamic environmental events and controversies. Yet K. Lee (2026) identifies the Environmental pillar as the least reconstructible from the International Financial Reporting Standards (IFRS) financial accounts alone, which is a finding that is entirely reconcilable with the above. The E pillar is simultaneously the most predictable given rich environmental inputs and the least predictable given only financial statements. This dichotomy directly operationalizes the empirical case for satellite imagery, Carbon Disclosure Project (CDP)-linked operational datasets, and geospatial environmental monitoring as primary E-pillar prediction infrastructure. Alternative data sources such as news are specifically well-suited to capturing high-profile environmental events and controversies that surface before annual disclosure cycles (Aue et al., 2025), while time-series architectures are essential for tracking GHG trajectories: (S.-L. Yang, 2025)'s GRU-LSTM demonstrates that sequential carbon disclosure patterns carry temporal signal (prior-year CDP grade predicting next-year grade) that no cross-sectional model architecture can access by design.

The Social Pillar

The Social pillar remains the most persistently intractable prediction target across all architectural families and data environments. The structural cause is categorical: social metrics span human rights practices, supply-chain labor standards, gender pay equity, employee satisfaction, and community relations, which are domains with no standard-

ized mandatory disclosure obligation, where content is predominantly qualitative and self-reported, and where independent verification is largely absent, unlike the G/E pillars, making them possibly the least transparent and challenging components of ESG. Consistent with this characterization, no study in the current corpus achieves high predictive performance for the Social pillar using financial statements alone. NLP approaches partially bridge this gap. (S. L. Lin et al., 2025) demonstrate that negative media attention directed at Human Rights sub-pillars produces statistically significant ESG score deterioration, positioning controversy detection in news streams as the most viable current S-pillar prediction pathway. (Del Vitto et al., 2023) confirm that Social replication requires more features than Environmental from an equivalent starting pool, yielding a systematically higher noise-to-signal ratio (RMSE = 0.089–0.096 for S vs. 0.071–0.084 for E) that persists across all sector-region combinations tested. This places the S pillar as an important frontier for future research, especially in generating new data sets and modeling methodologies to be used to drive our understanding of its complex dynamics.

The methodological appraisal also qualifies cross-study comparisons of predictive performance. Eight of the 29 Stream 1 studies explicitly use prior ESG scores or ratings as inputs, while three ESG-target studies directly incorporate E/S/G pillar scores, and several others reconstruct ratings from provider-defined scoring components. Although 25 studies report some form of holdout evaluation or cross-validation, only three clearly use a chronologically separated future-period test. Accordingly, high R^2 or accuracy should not be interpreted uniformly across Stream 1: a model reproducing a provider score from its components, a model forecasting next year's score from its historical value, and a model predicting an unseen future rating from independent information answer materially different predictive questions.

4.2. ESG Scores as Independent Variables

Unlike in the first stream, ESG scores in this section act as independent variables alongside other inputs such as unstructured and financial data in order to predict financial outcomes. From a global perspective, we have noticed that tree-based models, namely RF, XGBoost, and DT, are widely used across the 57 studies in this stream, as illustrated in Figure 5.

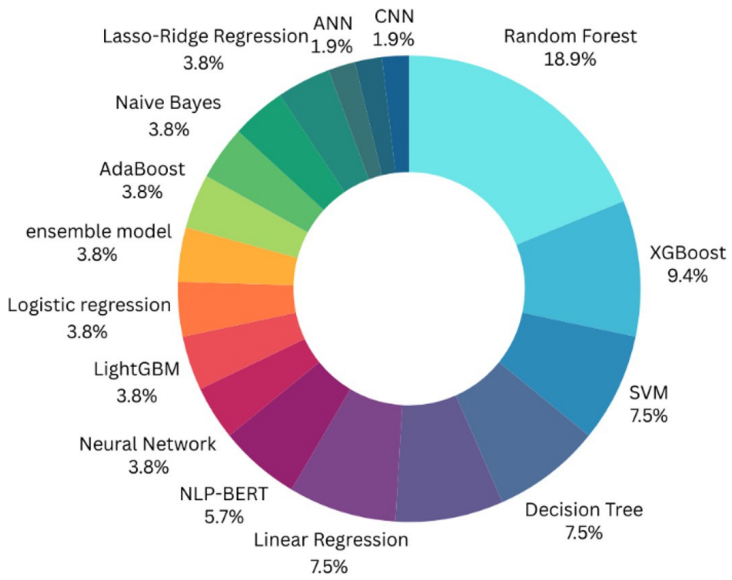


Figure 5. Distribution of ML and DL models used to predict financial outcomes with ESG scores as independent variables. Colors are used solely to visually distinguish model categories and do not represent an additional classification.

Given that Stream 2 encompasses the largest sub-corpus (n = 57) with the most heterogeneous outcome variables, we conducted a Multiple Correspondence Analysis (MCA) through the Biblioshiny module of the Bibliometrix R package (version 5.2.1) to ensure a rigorous and replicable thematic synthesis; Streams 1 and 3, with smaller and more methodologically coherent corpora (n = 29 and n = 41, respectively), were organized through iterative close-reading and conceptual grouping rather than statistical clustering. MCA is a matrix that combines co-word analysis with a two-dimensional map of the most frequently mentioned phrases in the related scientific documents. Based on the frequency of each keyword and the collective use of its references in the recovered document, the MCA approach clusters the keywords in the recoverable scientific documents (Ragazou et al., 2022; Ying et al., 2023). The conceptual structure map visually illustrates the semantic proximity of the key terms often extracted from article titles, abstracts, and keywords (see Figure 6), revealing four distinct conceptual clusters. These terms are distributed across the factorial plane, with Dimension 1 (21.79%) and Dimension 2 (19.68%) being the largest proportions of variance in the data, reflecting the primary axes along which the concepts differentiate.

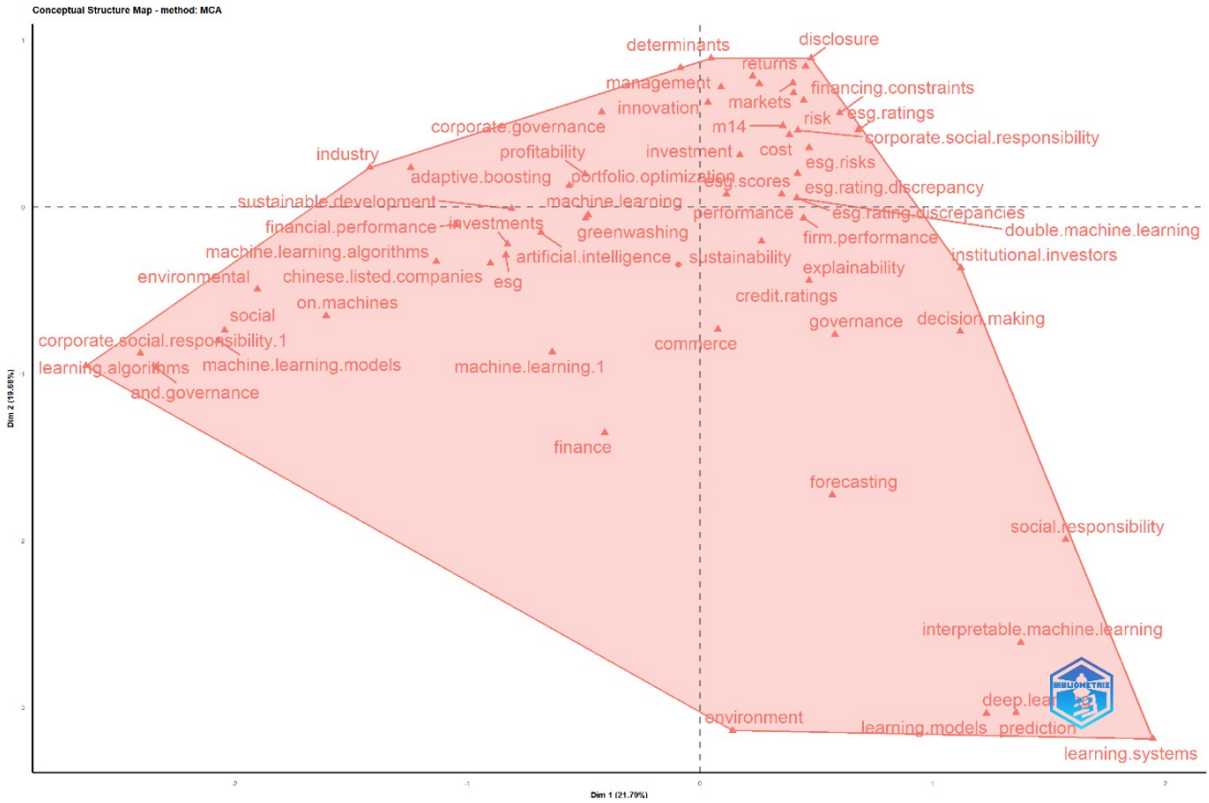


Figure 6. Multiple Correspondence Analysis (MCA). Source: Biblioshiny.

For example, a prominent upper-right quadrant clusters terms like "markets," "returns," "portfolio optimization," "performance" "cost," and "financing constraints," which are associated with the theme of Asset Pricing, Stock Returns and Market Performance, while adjacent terms such as "credit ratings," "ESG risks," "risk," and "governance" are aggregated to form the theme of Credit and Risk management. Additionally, the lower-right quadrant, dominated by terms like "environment," "social responsibility," "sustainability," "learning systems," "prediction," and "greenwashing", clearly aligns with Environmental and Social Impact Prediction. Finally, recurring terms in the central and upper-left regions, such as "industry," "corporate governance," "profitability," "decision making," and "firm performance", support the inclusion of the standalone theme "Financial perfor-

mance and operational Efficiency". The spatial segregation and semantic cohesion of these keyword groupings offer empirical support for the four thematic categories, which are further discussed and summarized below.

4.2.1. Asset Pricing, Stock Returns, and Market Performance

This first theme affirms that the integration of ESG scores into ML and DL architecture significantly improves the prediction of stock movement, portfolio optimization, and systematic risk identification. To do so, ESG scores were operationalized through two primary channels: structured static data from rating agencies and dynamic unstructured sentiment extracted via NLP. Methodologically, tree-based models have emerged as the preferred tools for analyzing equity returns. LightGBM, for example, has successfully delineated the relationship between the sign of excess price returns and ESG scores within the European market, achieving an accuracy of 67.9% (Assael et al., 2023). However, frontier research warns against blindly trusting ML return predictions. B. Chen (2026) provides a critical case study on the Chinese A-share market, using XGBoost to uncover a highly significant ESG alpha. Yet, through XAI, the study reveals this as an "alpha illusion"—the ML model's predictive power stems from non-linearly capturing traditional risk factors (size, illiquidity) rather than pure ESG fundamentals. Furthermore, the volatility of these returns is highly sensitive to external shocks; the SVM classifiers highlight that negative ESG-related news severely impacts the Chinese stock market, generating a cumulative abnormal return (CAR) of −0.547% (Y. Chen et al., 2024). Beyond absolute returns, ML proves highly effective in parsing risk and institutional performance. RF confirms that ESG ratings are the main fundamental factors in explaining stock Betas, allowing investors to mitigate systematic risk (Di Ventura et al., 2026; Martín-Cervantes & Valls Martínez, 2023). This risk mitigation directly translates to institutional success; Momparler et al. (2025) and Climent et al. (2025) deploy XGBoost to confirm the predominant role of ESG scores in predicting annualized returns of US-registered equity funds and mutual funds outperformance (R^2 = 0.694), proving that sustainability acts as a definitive hedge in asset management.

The most profound methodological evolution in this stream is the transition towards hybrid DL frameworks for multi-objective portfolio construction. RF in early studies showed an impressive performance (R^2 = 0.913) when predicting ESG-valued returns (Z. Wu et al., 2023). However, there is a strong contemporary convergence towards dynamically extracting unstructured public sentiment using NLP (Sokolov et al., 2021) and integrating it into advanced neural networks. For example, (Truyols-Pont et al., 2024) integrate LSTM predictions into a Black–Litterman model (BL-LSTM), achieving a remarkable Sharpe ratio of 0.81. This is further pushed by Gaurav et al. (2025) and Martínez-Barbero et al. (2026), who deployed a three-stage AI-driven framework combining Multivariate Bidirectional Long Short-Term Memory (MBi-LSTM), Non-dominated Sorting Genetic Algorithm (NSGA-II) and Actor-Critic Reinforcement Learning for stock returns prediction, optimization and dynamic rebalancing (see Figure 7). The aim is to build intelligent and competitive portfolios that outperform the market and can adapt to dynamic, fluctuating markets without affecting the overall performance.

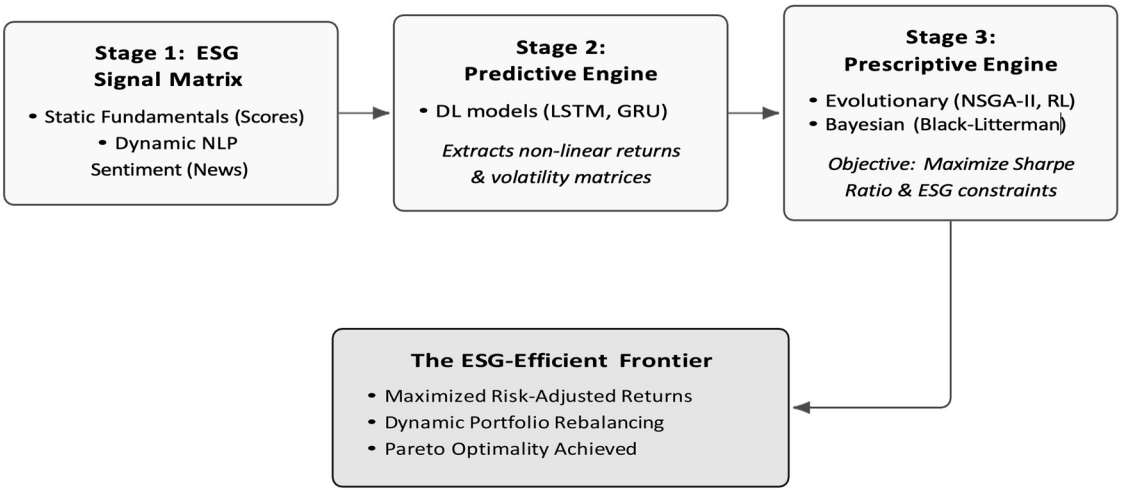


Figure 7. The predictive–prescriptive architecture for ESG portfolio optimization. Synthesized from recent frontier literature, modern sustainable asset management abandons single-stage models in favor of dual-engine frameworks. DL networks (e.g., LSTM) act as the predictive engine for asset returns, feeding directly into advanced prescriptive optimizers (e.g., Genetic Algorithms, Black–Litterman) to mathematically achieve the ESG-Efficient Frontier. (Conceptual framework synthesized by the authors).

4.2.2. Credit and Risk Management

The inclusion of ESG metrics in credit and risk management is essential for measuring the long-term, non-financial risks that significantly affect a firm's solvency and cost of capital. In this regard, ML models are increasingly used to study the complex and often non-linear relationship between ESG performance and the firm's credit and default risk. Throughout the literature surveyed, we observed an obvious shift toward the development of ensemble and hybrid models for various risk management scenarios. To decrease the corporate debt financing cost, Hu (2025) presents three combinations of models, namely RF + NN, SVM + GB and LSTM + XGBoost, with the LSTM + XGBoost hybrid model reaching the best recall, accuracy and F1 score of 95%, 90% and 90%, respectively. Interestingly, the negative correlation between ESG ratings and financing costs is obviously more pronounced in developed countries than in developing ones. In the scenario of bankruptcy forecasting for Chinese enterprises, Kaleem et al. (2024) identify RF as the most reliable model, consistently achieving 100% in both AUC and F1-score with or without ESG factors, followed by Logistic Regression and SVM, underscoring that ESG integration helps in reducing harsh Type II errors. This reduction mitigates the tendency to misclassify distressed companies as healthy, thus minimizing erroneous assessments of distressed firms. A perfect AUC and F1-score, while reported as achieved, is exceedingly rare in genuine out-of-sample financial prediction and may reflect class imbalance handling or limited test-set size rather than true generalization capability. Notably, the prediction of credit ratings seems more nuanced; on one hand, Agosto et al. (2025) propose a novel explainable-by-design Bayesian learning model to solve the issues of data inconsistency and rating divergence and create a "merged score". The model does not exhibit strong performance (AUROC = 0.59), reflecting the limited impact of ESG factors on credit ratings; it nonetheless outperforms XGBoost, Random Forest, and Bagging, which lack explainability by design. On the other hand, Giudici and Wu (2025) address this issue and introduce a multiple ensemble model. The Stacked Ensemble Model was the best in terms of accuracy since it has the ability to reveal the non-linear correlation between credit and ESG ratings, with GB showing superior fairness and explainability, through S.A.F.E methodology. Employing an interpretable FLAME algorithm, D. Li and Adriaens (2024) reveal that G and S pillars are essential components of

pricing benefits and play a pivotal role in reducing US corporate bond issuance costs by 12.7 basis points for Architecture, Engineering, and Construction (AEC)-related industries. This is consistent with Giudici and Wu (2025), who similarly identified the Governance pillar as the most significant determinant of credit rating, highlighting its continued importance in being a crucial signal for a firm's credit quality.

4.2.3. Financial Performance and Operational Efficiency

Whereas the prior subthemes focused on market perceptions and risk considerations, this category of studies goes a step further to examine firm-level questions: Is a strong ESG stance a definitive strategy for remaining profitable, efficient, and competitive? To answer this, researchers are shifting away from traditional linear models toward advanced ML architectures to capture complex corporate dynamics. Once again, a significant majority of studies deploy tree-based ensemble and DL models to forecast a firm's profitability as part of performance indicators (Terzioğlu et al., 2026). For example, D'Amato et al. (2024) and Ferrari et al. (2025) find RF to be the best model to predict Earnings Before Interests (EBIT) and Return on Investment (ROI), with R^2 = 0.884 and R^2 = 0.876, respectively. The authors indicate that good social and governance practices maximize profitability and decrease systematic risk in the tech industry, making the S pillar the most impactful. Similarly, Dossa et al. (2025) achieved superior predictive power (R^2 = 97.2%) using the Extra Trees model for Return on Assets (ROA), driven heavily by the Environmental pillar. Furthermore, DL architectures exhibit robust capacity within the domain of accounting profitability (ROA, ROE). For example, Hsu et al. (2025) demonstrated that Optuna-tuned LSTM networks outperform traditional baselines (R^2 = 0.72). The literature also exhibits a strong consensus around the deployment of ANNs and Feedforward models. Studies utilizing these architectures (Cheng et al., 2025; Herman et al., 2025) consistently reject the assumption of linearity. Instead, they empirically demonstrate a 'threshold effect' or J-Curve; initial ESG investments may stagnate or temporarily depress returns due to capital expenditure, but financial performance accelerates exponentially once a specific threshold of ESG structural maturity is reached (see Figure 8). Finally, one important observation of D'Amato et al. (2024) is that the relationship between ESG and profitability is decidedly U-shaped; their results suggest that a high-scorer in ESG manages to earn significant financial profits, while medium-scorers experience depressed profitability owing to compliance and implementation costs. Therefore, for ESG to create financial value, it must be a core structural element of corporate strategy rather than a marginal, performative activity.

Beyond basic profitability, a distinct cluster of literature focuses on the fact that the quality of the ESG signal is crucial for broader operational efficiency, measured via Total Factor Productivity (TFP). Recent studies (Z. Li and Yang (2025), Ma et al. (2026), and Xue et al. (2024)) uniformly applied XGBoost paired with SHAP for Chinese listed companies, adopting the same data sources, namely Wind, Syntao and CSMAR. The consensus here is twofold: first, as detailed in Table 4, XGBoost drastically outperforms traditional baseline models in predicting TFP by significantly improving the goodness-of-fit and reducing predictive error. Second, XAI reveals that while high aggregate ESG scores boost productivity, ESG rating disagreement between providers acts as a severe penalty, leading to a decreased TFP. The most rigorous methodological advancement in this stream is the transition from correlation to causal inference. Traditional studies suffer heavily from endogeneity (profitable firms naturally have more capital to invest in ESG disclosures). To isolate the true economic impact, recent frontier research deploys Double/Debiased ML (DDML) (Gaoli et al., 2025). By algorithmically filtering out nuisance parameters and omitted variable bias, these studies prove a definitive, causal relationship: optimized ESG structures directly stimulate digital technological innovation and takeover activities, providing evidence that

ESG is a causal driver of operational efficiency, not merely a byproduct of existing wealth. This is also complemented by B. Yang et al. (2025), who implemented a Dual ML framework in order to connect ESG ratings to external support and internal governance (i.e., improving digital innovation efficiency and enhancing human resources level) within firms. This indicates that the market values clear and consistent signals of sustainability and punishes the opacity incurred by rating divergence, which might be considered "greenwashing".

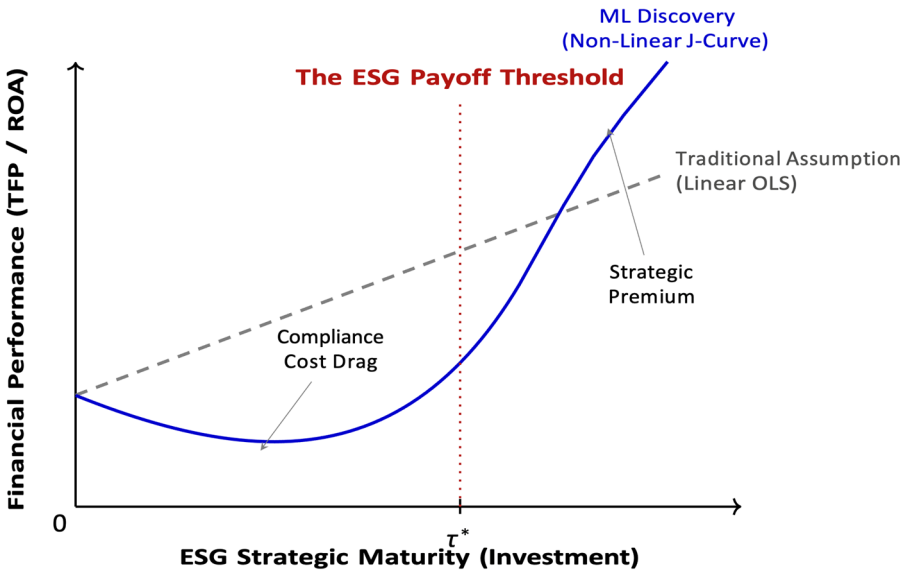


Figure 8. The non-linear profitability frontier. This illustrates how traditional linear models fail to capture the ESG-profitability nexus. Advanced ML architectures reveal a distinct J-curve: initial ESG integration incurs compliance costs that temporarily drag on performance, but crossing the structural maturity threshold (τ^*) unlocks exponential operational efficiency and premium financial returns. (Conceptual framework synthesized by the authors).

Table 4. Predictive performance of XGBoost vs. baseline models in forecasting Total Factor Productivity (TFP) for Chinese listed companies.

Study	Model	R^2	RMSE
Z. Li and Yang (2025)	Random Effect (Baseline)	0.5955	0.6876
	XGBoost	0.7045	0.5909
Ma et al. (2026)	Ridge Regression (Baseline)	0.5586	0.4357
	XGBoost	0.8007	0.3782
Xue et al. (2024)	XGBoost without ESG (Baseline)	0.6300	0.6420
	XGBoost with ESG	0.6590	0.6030

4.2.4. Environmental and Social Impact

While the preceding subthemes show that ESG can reduce financial risk and improve firm-level performance, the environmental and social impact cluster addresses a deeper question: whether ESG-related information translates into measurable real-economy sustainability outcomes. Across the seven studies in this subtheme, the evidence does not support a simple linear claim that "better ESG automatically produces better environmental performance." Rather, it reveals a more conditional mechanism in which ESG signals become impactful only when they are converted into innovation capability, digital transformation, governance discipline, or market pressure (see Figure 9). Methodologically, the literature is separated into two complementary families. First, XGBoost-based studies are mainly used for high-dimensional driver discovery. F. Liu et al. (2024) show that ESG

ratings are the strongest predictor of corporate green innovation, suggesting that ESG operates as a strategic resource that organizes firms' environmental actions, disclosure quality, and innovation capacity. Similarly, Clement et al. (2025) demonstrate that corporate CO_2 intensity is not explained only by conventional environmental inputs such as energy, water, and renewable energy consumption; social variables, particularly the proportion of women in the workforce, also emerge as important predictors. This finding is analytically important because it challenges the separation of the E and S pillars: environmental performance is partly shaped by organizational and social structures, not only by resource-use efficiency.

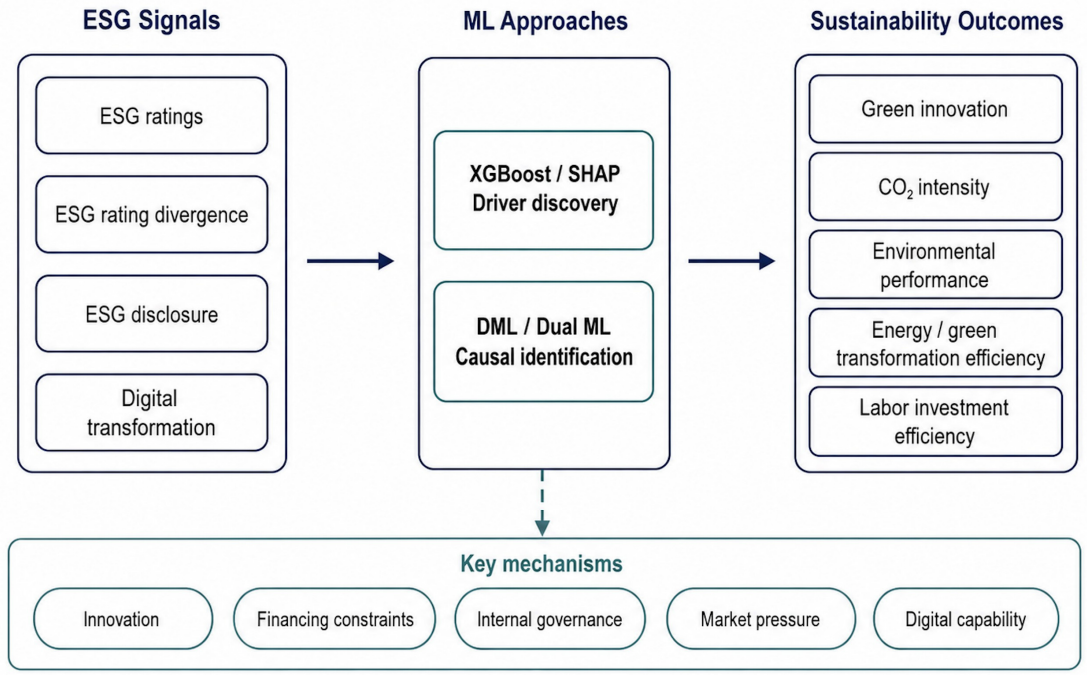


Figure 9. From ESG signals to sustainability outcomes: predictive and causal ML pathways.

Second, Double ML (DML) studies move the discussion from prediction to causal inference. This is a crucial methodological advance because environmental and social outcomes are vulnerable to endogeneity: firms with stronger governance, better access to financing, or higher innovation capacity may both receive better ESG evaluations and perform better environmentally. By using DML, Wan et al. (2025), Y. Wang et al. (2025), and L. Wu et al. (2025) attempt to isolate the causal effect of ESG rating divergence on energy efficiency, labor investment efficiency, and green transformation efficiency. Their findings reveal that disagreement over ESG ratings is neither uniformly harmful nor beneficial. In some contexts, divergence weakens resource allocation and suppresses energy-efficiency improvement by redirecting managerial attention toward rating management rather than substantive operational change. In other contexts, especially among energy firms facing high market visibility and stronger governance scrutiny, rating divergence can operate as an external pressure mechanism that accelerates green transformation through green innovation. B. Yang et al. (2025) further extend this logic by showing that ESG ratings support digital technological innovation through two channels: external market support and internal governance. Ren et al. (2023) complement this mechanism by demonstrating that digital transformation and its peer effects enhance environmental performance through innovation and the alleviation of financing constraints. Taken together, this subtheme suggests that ESG contributes to environmental and social impact not as a static score but as a dynamic signal that interacts with digital capability, innovation systems, governance structures, and stakeholder pressure. Therefore, the central contribution of this literature

is not merely identifying which model performs best but showing that ML/DL methods make visible the hidden pathways through which ESG is transformed into measurable sustainability outcomes. In summary, Table 5 provides a structured synthesis of the current literature landscape, capturing the best-performing models to predict key financial outcomes for each theme.

Table 5. Map of predicted financial outputs and best-performing ML/DL models across Stream 2 themes.

Theme	Predicted Output	Best-Performing ML/DL Model	Example Study
Credit and Risk Management	Credit ratings and creditworthiness	GB, Bayesian S.A.F.E., Stacked Ensemble Model	Agosto et al. (2025) and Giudici and Wu (2025)
	Bankruptcy, default, and financial distress risk	RF, XGBoost	Kaleem et al. (2024) and Ma et al. (2026)
	Corporate bond insurance spreads, bond pricing, and debt financing cost	ECS-XGBoost, LSTM + XGBoost, Extra trees, FLAME	Ding and Xu (2025), Feng et al. (2025), Hu (2025), D. Li and Adriaens (2024), and Zhu et al. (2025)
	Portfolio risk and resilience under ESG shocks	RF, DML + GAM, TSO-DSVM	S. Liu and Yu (2025), Wei and Zeng (2025), and J. Zhang and Wang (2025)
Financial Performance and Operational Efficiency	Corporate profitability and financial performance: EBIT, ROA, ROE, ROI	RF + SHAP, XGboost, LSTMOPT, CNNOPT, FNN	D'Amato et al. (2024), Dossa et al. (2025), Ferrari et al. (2025), Herman et al. (2025), Hsu et al. (2025), S.-W. Lin and Lin (2025), Nie and Wang (2025), and Terzioğlu et al. (2026)
	Total Factor Productivity (TFP)	XGboost + SHAP, EBM	Z. Li and Yang (2025), Ma et al. (2026), and Xue et al. (2024)
	Operational efficiency and digital innovation	RF, DML	Gaoli et al. (2025), B. Yang et al. (2025), and Y. Zhang et al. (2025)
	Financial predictability	SOM, ANN	Cheng et al. (2025) and Dincă et al. (2025)
Environmental and Social Impact Prediction	Carbon emissions and environmental performance	XGboost	Clement et al. (2025) and Ren et al. (2023)
	Energy efficiency and green transformation efficiency	XGboost, DML	Wan et al. (2025) and Y. Wang et al. (2025)
	Corporate green and digital innovation	XGboost, DML	F. Liu et al. (2024) and B. Yang et al. (2025)
	Social/labor efficiency	DML	(L. Wu et al., 2025)
Asset Pricing, Stock Returns, and Market Performance	Stock return prediction, ESG alpha, and abnormal returns	LightGBM, RF, Multi-layer Perceptron regressor, Multivariate LSTM + SHAP	Assael et al. (2023), B. Chen (2026), S. Chen et al. (2023), Y. Chen et al. (2024), Liang (2025), Nagesh et al. (2025), and Rosinus and Lansky (2025)
	Systematic risk and asset-pricing factor models	Shallow ANN/K4-RBFN with XAI, RF	Di Ventura et al. (2026)
	Mutual fund performance	XGBoost	Climent et al. (2025) and Momparler et al. (2025)
	Firm valuation and price multiples (P/E and P/B ratios)	XGBoost, RF, Ridge Regression	Joshi and Chauhan (2020)
	Portfolio construction and sustainable portfolio optimization	XGBoost, NLP, BL-LSTM, GRU-BiLSTM, MBi-LSTM, CNN, GA	Carlei et al. (2025), Gaurav et al. (2025), V. Jain and Sahay (2025), Martínez-Barbero et al. (2026), Sajadi et al. (2025), Sokolov et al. (2021), Truyols-Pont et al. (2024), and Z. Wu et al. (2023)

Note: S.A.F.E: Sustainability, Accuracy, Fairness, and Explainability. ECS-XGBoost: Efficient Chicken Swarm tuned XGboost. GAM: Generalized Additive Models. LSTMOPT: Long Short-Term Memory with Optuna. FNN: Feedforward Neural Network. EBM: Explainable boosting Machine. SOM: Self-organizing map. K4-RBFN: K4-Radial Basis Function Network. BL-LSTM: Black–Litterman model with LSTM. GRU-BiLSTM: Gated Recurrent Unit with Bidirectional LSTM. GA: Genetic Algorithm.

4.3. Assessing and Explaining ESG Scores Through ML and DL Models

Distinct from the predictive and forecasting applications discussed in the previous sections, this part of the review concentrates on a new trend of research that employs ML and DL models to evaluate, interpret, and ultimately enhance the transparency of existing ESG scoring methodologies. The 41 studies included in this last category deploy advanced analytical ML and DL techniques, not as forecasting instruments but as diagnostic tools, designed to unpack the "black box" of ESG itself. The main purpose is either to create new scores from scratch and compare them with the ones provided by agencies or to deconstruct,

validate, and interpret the existing ratings, in view of revealing the prevalent issues of methodological opacity, inconsistency, score discrepancies, and potential biases that currently undermine the consistency of ESG assessments. This analytical approach is essential for building trust among stakeholders and for developing more robust and meaningful ESG evaluation frameworks. At the heart of this analytical endeavor is the application of advanced NLP techniques, particularly Transformer-based models. This approach moves beyond the confines of structured datasets to analyze primary source documents and unstructured data that form a substantial portion of ESG-related information, such as legal proceedings, news articles, earnings call transcripts, and corporate sustainability reports. Figure 10 illustrates the conceptual framework of this diagnostic paradigm, demonstrating how diverse unstructured data streams are ingested and processed by hierarchical NLP architectures to generate independent ESG assessments.

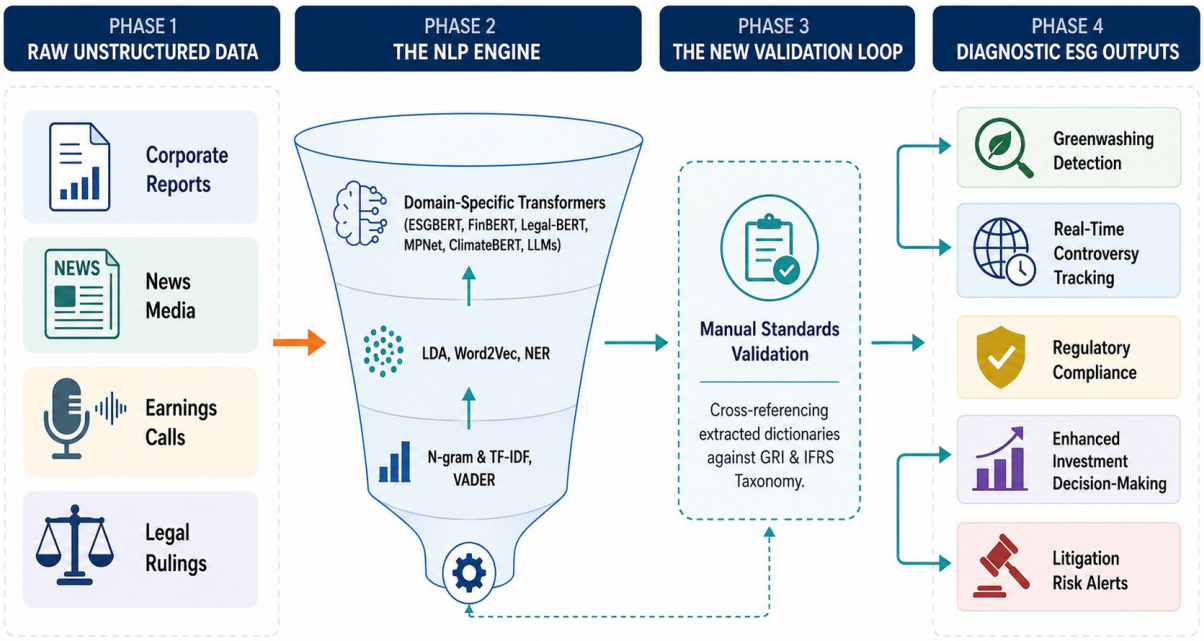


Figure 10. The unstructured data processing framework. Hierarchical NLP architectures operate as computational assessment engines, moving from foundational text mining to domain-specific Transformers to transform raw corporate and media text into diagnostic ESG outputs. (Conceptual framework synthesized by the authors).

A cornerstone of modern NLP is the Bidirectional Encoder Representations from Transformers (BERT) model and its derivatives (Devlin et al., 2019), a pre-trained model released by Google in 2018 that performs strongly across nearly all NLP tasks. In particular, approximately 60% of the included studies used BERT-derivative architectures in classifying raw news text into ESG issues (C.-C. Chen et al., 2023), providing daily analyses of news articles using the DynamicESG dataset (Tseng et al., 2023), analyzing large datasets, and processing media coverage to produce public-opinion-derived ESG scores (Fischbach et al., 2023). This has evolved into the development of bespoke, pre-trained language models such as E-BERT, designed to assign ESG rating classes (AAA, BBB, CCC, DDD) in order to rate ESG reports with high objectivity (M. Zhang et al., 2025), ESG-KIBERT and FinBERT for industry-specific classification (H. Lee et al., 2025b), and the ESG2PreEM framework for E, S, G categorization using textual data (H. Lee et al., 2024a). Interestingly, topic extraction has likewise evolved: BERTopic-based pipelines extracted seven thematic categories (e.g., management team, statements, financial, energy, emission, carbon, environmental) from earnings call transcripts and tracked their evolution between 2017 and 2023 (H. Lee et al., 2025c). Other studies leverage

sentence transformer models (Distilled BERT) to automate scoring by measuring the textual similarity between annual reports and established frameworks like the Global Reporting Initiative (GRI) standards (Ngee et al., 2024).

However, the deployment of fully automated NLP clustering, such as unsupervised LDA or BERTopic, introduces a critical vulnerability: the risk of capturing linguistic noise or corporate buzzwords rather than material sustainability practices. To prevent these models from generating mere statistical artifacts, a rigorous methodological imperative is emerging within the literature: the human-in-the-loop validation of NLP-generated dictionaries (see phase 3 in Figure 10). Extracted terminologies and thematic clusters must be manually cross-referenced and validated against established, globally recognized taxonomies, specifically GRI and IFRS. This vital verification layer ensures that the algorithmic outputs are not just mathematically significant but are substantively grounded in verifiable corporate sustainability commitments. Four methodologically distinct clusters organize the analytical objectives of this part of the review: (i) XAI applications that open the black box of proprietary scoring functions; (ii) reconciliation methodologies that address ESG rating divergence and "aggregate confusion"; (iii) greenwashing detection frameworks that triangulate discourse against practice; and (iv) unsupervised clustering approaches that establish industry-specific materiality. These clusters are not mutually exclusive: several studies operate at their intersections, but each cluster is defined by its primary analytical objective.

4.3.1. Cluster A: XAI and Opening the Black Box of Proprietary Ratings

The opacity of proprietary scoring functions used by major agencies (Refinitiv, MSCI, Bloomberg, LSEG, Sustainalytics) constitutes the central methodological problem this cluster addresses. Studies in this cluster do not seek to predict ESG scores accurately as much as to reverse-engineer the implicit weighting schemes, feature dependencies, and structural assumptions embedded within rating methodologies. Although models like XGBoost, GB, and RF are preferred for their high predictive accuracy, their performance comes at the cost of interpretability. The "black box" nature of models presents a critical challenge to researchers, policymakers, and investors when making informed sustainable investments and identifying risks. It is almost impossible to understand how the model arrives at its conclusions and which features have the most significant impact on the predictions by simply inspecting its structure. Therefore, XAI offers powerful techniques to avoid the "Black Box" issues and facilitate debugging and interpretation of any ML model despite its structure including Local Interpretable Model-agnostic Explanations (LIME); SHapley Additive exPlanations (SHAP), which was first introduced in game theory in 1953 Shapley (2016); Permutation Feature Importance (PFI), which was first applied on the RF model by Breiman (2001); Surrogate models; Partial Dependent Plot (PDP); Accumulated Local Effects (ALE) plots; and Individual Conditional Expectation (ICE). Del Vitto et al. (2023) and Ferro et al. (2025) are formally assigned to Stream 1 but are also discussed here as boundary evidence because their score-reconstruction models are subsequently used to examine the construction and interpretability of proprietary provider ratings. Their inclusion in this discussion does not alter the mutually exclusive stream counts.

The most consequential finding of this group directly challenges the substantive validity of agency ratings. By replicating LSEG's ESG scoring methodology using RF with iterative feature selection on the entire 186-metric granular database for the 2010–2023 period, approximately 60% of ESG score weights are demonstrated to depend on aspirational, forward-looking features (e.g., policy commitments, target declarations, disclosure intentions), while only roughly 40% reflect backward-looking performance metrics (Ferro et al., 2025). On reading through signaling theory, the Ferro et al. (2025) decomposition

is more than a descriptive imbalance. If roughly 60% of ESG score weight derives from aspirational, forward-looking features, then the majority of the signal is constructed from precisely the low-cost claims that signaling theory predicts cannot separate types. A firm with no intention of decarbonizing can declare a decarbonization target at negligible cost. Greenwashing, on this reading, is not an anomaly or an enforcement failure to be detected after the fact; it is the equilibrium outcome that a rating architecture weighted toward cheap signals should be expected to produce. This reframes the greenwashing-detection literature from a policing exercise into a diagnostic of a structural design flaw, and it predicts, correctly, the U-shaped relationship that Eckberg et al. (2026) observe: firms with the resources to score well also have the resources to signal selectively. This imbalance transforms ESG scores into instruments that may reward disclosure intent over substantive action and provides quantitative grounding for the long-standing critique that ESG scores can inflate without underlying sustainability improvements.

Additionally, Term Frequency—Inverse Document Frequency (TF-IDF), Legal-BERT, and SHAP were applied in the context of litigation exposure, with SHAP identifying regulatory sanctions as the dominant predictor of latent sustainability risk (Park, 2025). This finding addresses a critical blind spot in traditional ESG ratings: symbolic compliance (which is defined as the practice of meeting disclosure requirements while concealing material legal exposures) escapes detection by agency methodologies that rely on company-reported information. The integration of court-ruling data with conventional ESG features establishes that legal proceedings carry strong incremental signal about substantive ESG risk above and beyond what disclosed metrics capture. An XAI-enhanced framework combining XGBoost, SHAP, and human-in-the-loop validation in 18 firms in the banking, aviation, and chemical sectors over 2021–2023 achieves a 12.4% improvement in ESG-score consistency that meets standard fairness thresholds across protected sector groups (Ahmad & Ahmed, 2026). The human-in-the-loop layer addresses a fundamental limitation of fully automated XAI: SHAP attributions reflect what the model learned, not what should be true; domain validation closes that gap. For site-level critical-mineral ESG evaluation, a hybrid ESG assessor framework integrating expert-elicited knowledge with XGBoost and SHAP provides a nuanced cobalt-mining risk profile in Iran, addressing supply-chain ESG dimensions that aggregate firm-level ratings systematically obscure (Sarkheil et al., 2026).

Building on earlier replication work, white-box and black-box ML models combined with SHAP reverse-engineer the proprietary Refinitiv ESG issuance process, recovering ESG scores with RMSE = 0.078 for the Environmental score and RMSE = 0.075 for the Governance score in the US region using ANN (Del Vitto et al., 2023). Studies have leveraged SHAP to show that dynamic company news sentiment is a key predictor of a firm's ESG rating (Ang et al., 2023) and that board-related features are the most influential drivers of corporate governance ratings (Svanberg et al., 2022). Moreover, SHAP identifies environment-dedicated product lines and total CO_2-equivalent emissions as dominant E-pillar features, freedom of association policy as the most influential S-pillar variable, and board diversity as the principal G-pillar driver. Cheng et al. (2025) even created a novel ESG assessment tool called the "iScore", using SOM-ANN frameworks combined with bootstrapping, designed specifically to adjust the divergence of ratings provided by agencies. This highlights a powerful trend of using advanced ML not just for prediction but as an inferential tool to assess the firm-level characteristics that drive ESG performance measurements.

4.3.2. Cluster B: Resolving ESG Divergence and "Aggregate Confusion"

ESG rating divergence is no longer a peripheral methodological concern; it is the defining structural problem of contemporary ESG assessment. The empirical magnitude of this divergence is severe: the average correlation between major rating agencies is approximately

0.53, with pillar-specific correlations of 0.59 for Environment, 0.31 for Social, and near zero (0.02) for Governance (Berg et al., 2022). This shows that agencies are not merely weighting the same data differently; they are measuring fundamentally different constructs under the same label. Bi-level AI frameworks specifically designed to handle component divergence reconcile conflicting governance controversy scores across agencies, demonstrating that the divergence is structurally resolvable when the reconciliation is performed at the indicator level rather than the composite-score level (Jlifi et al., 2026). The real-world consequences of unresolved divergence are quantified using DML on ESG rating disagreement (ESGRD) and corporate labor investment efficiency: rating divergence statistically and economically impairs firm labor investment efficiency and amplifies information asymmetry between firms and stakeholders (L. Wu et al., 2025). The DML framework's advantage in this context is its capacity to disentangle the causal effect of rating divergence from the underlying ESG performance it ostensibly measures, which is a separation that conventional regression methods cannot achieve in the presence of strong covariate endogeneity.

Aggregation methodology itself is interrogated as a divergence-management strategy. Constructing a bottom-up national ESG rating by aggregating the distribution of listed firms' ESG scores across twelve developed economies between 2013 and 2022, four aggregation schemes (mean, median, Sen's inequality-adjusted index, dispersion-adjusted mean) yield highly consistent rankings, with CatBoost regression validating the median as the most robust headline measure for national sustainability (Hoffmann & D'Ecclesia, 2026). The methodological lesson generalizes beyond national aggregation: when firm-level scores diverge across agencies, central-tendency measures that are more robust than the arithmetic mean (which is sensitive to outlier ratings produced by methodologically extreme providers) provide more reliable consensus estimates. This finding has direct application to portfolio construction: an equally weighted median across agency scores will, on average, outperform a mean-aggregated composite at withstanding agency-specific methodological idiosyncrasy. Parallel work on the digitization of the ESG matrix builds standardized consolidated assessment frameworks that can be applied to companies across different industries using news articles, corporate reports, and social media, alongside ESG data being collected from various agencies (Telukdarie et al., 2024). The architectural argument is that the discrepancies between providers are partially attributable to unit-of-measure heterogeneity (Moody's measures total waste in tons, while MSCI measures it in pounds, for instance) and only partially to substantive methodological disagreement. The conjunction of adapted methodologies constitutes a coherent response to aggregate confusion: it does not eliminate divergence, but it operationalizes its quantification, isolates its causal consequences, and proposes robust aggregation rules that mitigate its impact on downstream decisions, hence providing stakeholders with a comprehensive view of a company's sustainability performance.

4.3.3. Cluster C: Greenwashing Detection and Discourse–Practice Gaps

Greenwashing is defined as the systematic misalignment between sustainability discourse and substantive practice. This cluster closes the empirical loop by directly measuring the gap between what firms say and what they do. The methodological logic is triangulation: when corporate ESG discourse, external market perception, and objective practice metrics are jointly analyzed, discrepancies become identifiable that no single data source can reveal. The strongest empirical demonstration of this triangulation logic uses three distinct data streams for Brazilian corporate sector firms over 2022–2023: ESG reports (quantified via ML textual analysis), Merco ESG reputation index, and objective greenhouse gas (GHG) emissions data. The econometric analysis reveals a statistically significant gap between sustainability claims and actual GHG practices, validating Decoupling Theory's

core proposition that corporate ESG discourse functions partially as symbolic compliance disconnected from operational reality (dos Santos et al., 2025). The empirical determinants of greenwashing behavior are systematically mapped using STOXX Europe 600 constituents: ESG and Environmental scores exhibit a U-shaped relationship with greenwashing, indicating that firms with both very low and very high (E)SG scores are most likely to engage in greenwashing (Eckberg et al., 2026). The low-score result is intuitive (firms under reputational pressure greenwash to appear acceptable); the high-score result is more theoretically consequential. Interestingly, firms with the resources to perform well in ESG also have the resources to selectively present that performance, and the marginal value of additional disclosure incentivizes increasingly favorable interpretation. ESG disclosure score, firm size, cash-to-assets ratio, and capital intensity are positively associated with greenwashing propensity. The implication for ML-based ESG screening is that high ESG scores cannot be treated as monotonically indicative of substantive sustainability; the marginal interpretive value of additional score points decreases at the upper tail, where greenwashing risk re-emerges.

Direct sentiment-based greenwashing measurement is operationalized through the Greenwashing Tendency Score (GTS), which uses VADER for sentiment polarity scoring and MPNet for semantic similarity assessment to quantify the discrepancy between a firm's reported sustainability efforts and its actual performance (Motz et al., 2025). The integration of greenwashing detection into investment-decision tooling is demonstrated by a real-time web application combining stock market data, ESG scores, and NLP-based greenwashing detection across 20 companies, generating a Greenwashing Risk Score that filters ethical investments at investment screening (Joglekar et al., 2025). For supply-chain ESG risk, an AI/big-data innovation system in tourism integrates network analysis, anomaly detection, NLP-based greenwashing detection, and predictive modeling on heterogeneous data sources (e.g., supply chain records, news articles, social media, public databases) to proactively identify latent ethical risks in decentralized supplier networks (Hong & Kim, 2025). Disclosure quality analysis itself uncovers structural patterns indicative of greenwashing. Textual structure analysis of climate-related financial disclosures in the European energy sector combines lexicometric diagnostics, BERT-based text classification, Latent Dirichlet Allocation (LDA) topic modeling, and panel regressions to evaluate alignment with TCFD recommendations; the analysis reveals that firms heavily prioritize metrics and strategy over governance and risk management in their disclosures, an asymmetric reporting pattern consistent with selective disclosure incentives (Chabot, 2025). To systematically categorize this shift toward text-driven diagnostics, Table 6 outlines the taxonomy of NLP architectures deployed across the literature. By mapping specific unstructured data sources to their respective algorithmic applications, the table highlights both the targeted diagnostic insights and the critical methodological limitations that future research must address.

Table 6. Methodological taxonomy and limitations of NLP architectures in ESG diagnostics.

Unstructured Data Source	Specific NLP Architecture(s)	Targeted ESG Diagnostic Insight	Key Methodological Limitation/Future Gap	Exemplar Studies
Corporate Reports & Disclosures	E-BERT, DistilBERT, Word2vec, LDA, Vanilla BERT, MPNet	Quantify greenwashing tendencies, assess managerial integrity and automate GRI/TCFD alignment from self-reported data.	Highly susceptible to corporate self-selection bias; models struggle to verify claims against off-balance-sheet externalities.	M. Zhang et al. (2025); Motz et al. (2025); Lv et al. (2025); Ngee et al. (2024); Schimanski et al. (2024); Chabot (2025); dos Santos et al. (2025)

Table 6. *Cont.*

Unstructured Data Source	Specific NLP Architecture(s)	Targeted ESG Diagnostic Insight	Key Methodological Limitation/Future Gap	Exemplar Studies
Global News Media	Longformer, VADER, ESG-Miner, Longformer	Extract real-time ESG controversies to bypass the temporal lag of annual rating agency assessments.	High noise-to-signal ratio; significant risk of algorithmic hallucination when clustering dynamic news events.	Tseng et al. (2023); Fischbach et al. (2023); C.-C. Chen et al. (2023); H. Lee et al. (2024c); Joglekar et al. (2025)
Earnings Call Transcripts	RoBERTa, FinBERT, ESG-KIBERT, ALBERT	Fine-grained classification of latent ESG signals directly from executive communications and Q&A sessions.	Executive audio is heavily rehearsed; algorithms capture linguistic optimism rather than true operational shifts.	H. Lee et al. (2024a, 2025b, 2025c)
Legal Proceedings & Supply Chain	Legal-BERT, TF-IDF, N-Gram, LLMs	Uncover hidden litigation risks and detect ethical supply chain violations obscured by symbolic compliance.	Significant temporal lag; court rulings and sanctions only become textually available after the ESG violation and subsequent market damage has occurred.	Park (2025); Hong and Kim (2025)
Social Media & Web Platforms	Nested FinBERT, MTL (ABSA), BERTopic, VADER	Predict market dynamics, screening ethical investments. Evaluate public sentiment polarization.	Extreme sentiment volatility. Metrics are severely skewed by bot-driven amplification and unverified retail claims.	Telukdarie et al. (2024); Saxena et al. (2024); van der Heever et al. (2024)

4.3.4. Cluster D: Unsupervised Clustering for Industry Materiality

The structural premise of major ESG rating methodologies, that a single composite score can meaningfully compare firms across radically different industries, is empirically untenable. Studies in this cluster confront this directly by deploying unsupervised learning to identify which ESG features are actually material for specific sectors, replacing the implicit "one-size-fits-all" assumption of mainstream rating methodologies with empirically derived, industry-conditioned materiality weights. Three clustering algorithms, namely Agglomerative Clustering, K-Means, and Gaussian Mixture Model (GMM), are applied to S&P 500 ESG scores from leading data providers to uncover sectoral patterns. Technology and healthcare firms achieve the highest ESG scores, particularly in the G and S dimensions, while industrial and energy sectors face the greatest environmental challenges (Ergenç & Aktaş, 2025). The implication is methodologically consequential: composite ESG scores systematically misprice sustainability risk by applying identical feature weights to firms whose business models concentrate ESG materiality in entirely different pillars. The clustering result operationalizes the case for sector-conditioned ratings as the default architecture for production ESG systems.

Cluster-based segmentation also operates at the firm-sentiment level. On one hand, distinct sentiment patterns are identified by integrating K-means and Hierarchical Clustering with Principal Component Analysis (PCA) on firm-specific ESG sentiment data. On the other hand, RF and XGBoost demonstrate that firms clustered into high ESG sentiment

groups experience significantly lower financial risk, establishing sentiment-derived ESG signal as economically material (Sultana & Zeya, 2025). For banks specifically, the materiality question is investigated using Forward Sequential Feature Selection, RF feature importance, and Mutual Information, applied to European banks in the Refinitiv/LSEG database for 2018–2022. The analysis identifies the exact, minimal subset of ESG indicators driving bank ratings, demonstrating that the majority of disclosed bank ESG metrics carry no incremental rating signal (Sikacz et al., 2025). The temporal evolution of industry-specific ESG materiality is mapped using 188,006 ESG score records from Chinese listed companies spanning 2009 through Q1 2024, with descriptive statistics and Analysis of Variance (ANOVA) confirming highly significant differences across industries and Prophet time-series forecasting validating projected industry-specific trajectories (Y. Zhang, 2026). Financial services and insurance sectors exhibit consistently higher ESG scores than traditional manufacturing. The scale of the dataset and the temporal extension establish that industry-conditioned materiality is not a static property but an evolving structural feature of how ESG assessment is operationalized in different sectors. For Italian SMEs, the inequality in ESG adoption is quantified using Gini-index analysis of residuals from ML models, revealing that structural barriers to ESG adoption are unequally distributed across firm size and sector, with quantifiable disparities that reveal the diffusion gaps in mainstream ESG ratings systematically obscure (Amendola et al., 2025).

4.4. Deconstructing Pillar Influence in Assessment Frameworks

While traditional literature often treats the Environmental, Social, and Governance pillars as co-equal components of corporate sustainability, diagnostic ML models reveal a stark structural imbalance. However, this SLR indicates that the dominant predictive pillar is not universally agreed upon; rather, it is highly sensitive to the underlying data architecture. To untangle this heterogeneity, Figure 11 maps the methodological pipeline of 17 key diagnostic studies, tracing the flow from raw data extraction to algorithmic processing and, ultimately, to the empirically dominant ESG pillar. A more nuanced viewpoint is that the "dominant" pillar is actually not the same for each firm but rather depends on a company's particular field of operations and business model. ESG-KIBERT operationalizes this view by assigning industry-specific weights to E, S, and G factors using the SASB standardized framework. As a result, environmental aspects are dominant for energy companies, social factors such as data security are vital for technology firms, while governance and systemic risk management are crucial for financial companies H. Lee et al. (2025b). Furthermore, in some studies, the pillars are not assessed individually. Prescriptive analytics found that the combination of factors, namely management (G1), is a crucial factor for companies at all levels of performance. However, highly sustainable companies that excel in E, S, G dimensions are likely to see the most significant improvement in their overall ESG score by improving their performance in the factors "emission" (E1), "human rights" (S1), and G1 (Sariyer et al., 2024).

Across the assessment-focused literature, the Governance factor emerges as a classical fundamental element, conditioning and fostering structured logic underpinning ESG rating frameworks. When analyzing corporate communications such as (earnings call transcripts), studies found that firms concentrate heavily on G components in their reports and disclosures, with the G component accounting for the highest thematic weight (WG = 3.1631 vs. WE = 1.3925 and WS = 1), and Governance terms being mentioned most frequently during automated analysis compared to E and S (H. Lee et al., 2024a, 2024c). Similarly, an examination of Chinese firms listed in the China Securities Index (CSI) reveals that corporate governance indicators were the dominant factors, followed by social responsibility, while environmental performance had a relatively low and less significant role (Shi et al., 2024).

This structural prioritization can be rationalized by the immediacy of governance-related risks: MSCI considers Governance the predominant pillar in short-term risk assessment because Governance failure can trigger immediate event-driven crises (Tseng et al., 2023). From a critical angle, this emphasis might also represent more procedural convenience than effective impact: Governance reforms (e.g., policy updates, changes to board structures) are often easier to implement and report than resource-intensive environmental or social initiatives (X. Wang, 2025), raising the question of whether Governance emphasis represents true progress or just the line of least resistance for companies seeking ESG certification.

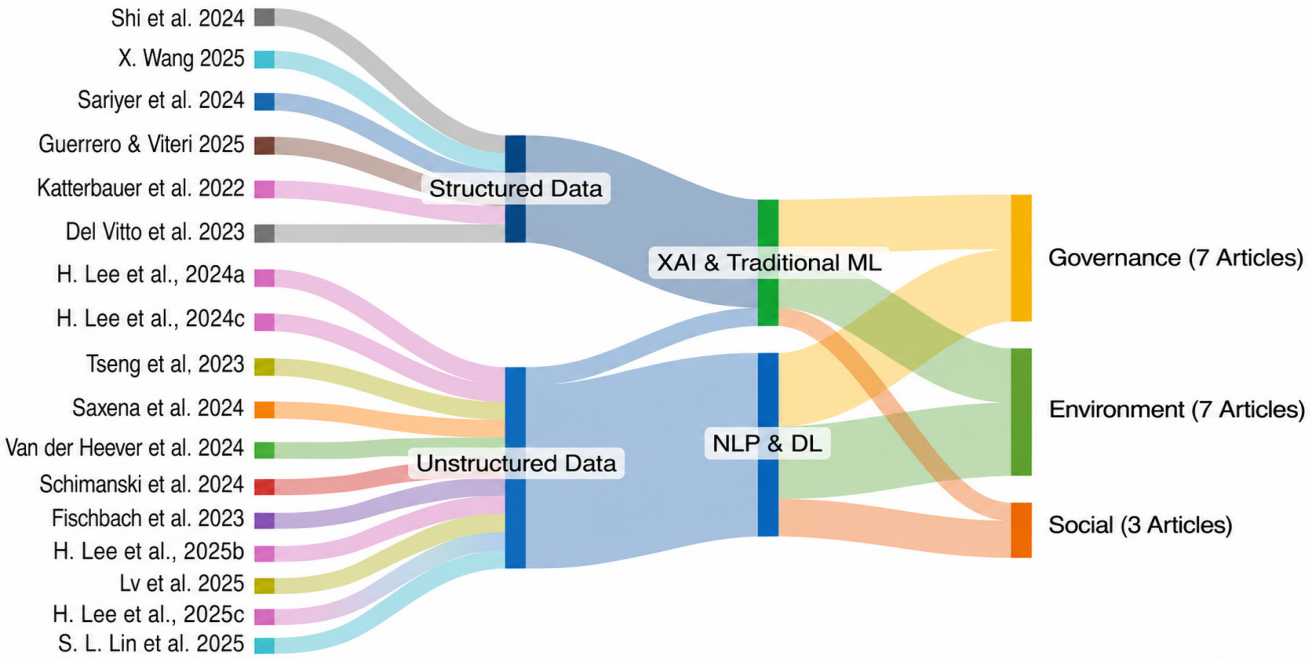


Figure 11. Methodological Sankey diagram mapping the pipeline from author dataset selection (structured vs. unstructured) through algorithmic processing to the resulting dominant ESG pillar (Shi et al., 2024; X. Wang, 2025; Sariyer et al., 2024; Guerrero & Viteri, 2025; Katterbauer et al., 2022; Del Vitto et al., 2023; H. Lee et al., 2024a, 2024c; Tseng et al., 2023; Saxena et al., 2024; van der Heever et al., 2024; Schimanski et al., 2024; Fischbach et al., 2023; H. Lee et al., 2025b; Lv et al., 2025; H. Lee et al., 2025c; S. L. Lin et al., 2025).

A parallel stream of research positions the Environmental pillar in a commanding position, arguing that climate change, greenhouse effect, hurricane frequency, and population expansion make E the foremost factor among ESG components (Saxena et al., 2024). This external pressure is also reflected in public perception: XAI analysis of Twitter sentiment finds public engagement with environmental topics highest, with a particularly strong positive sentiment of 50.4% compared with social or governance topics, and terms like greenwashing, green, transition, and climate strongly impact model sentiment predictions, according to Permutation Importance (van der Heever et al., 2024). This is evident from the superior performance of pillar-specific models like EnvRoBERTa, boasting a huge 95.7% accuracy rate when detecting environmental content, exceeding GovRoBERTa and SocRoBERTa (Schimanski et al., 2024), while ESG-Miner has shown strength in picking up environmental-themed headlines (Fischbach et al., 2023), indicating that environmental disclosure language is more distinguishable and machine-readable because it discusses structured factors such as carbon emissions and environmental fines. This is a property that, when combined with the disclosure-asymmetry finding, supports the inference that firms preferentially emphasize the pillar where standardized vocabulary best signals compliance. (Guerrero & Viteri, 2025). In addition, higher Environmental scores and financial

performance (e.g., return on equity) are aligned and positively correlated with Islamic environmental compliance in the context of Shariah-compliant corporations (Katterbauer et al., 2022).

The alluvial flow depicted in Figure 11 demonstrates a distinct methodological bifurcation. Studies relying on unstructured textual data (e.g., earnings calls, global news) predominantly deploy NLP and DL architectures, which frequently result in Governance and Environmental metrics emerging as the most heavily weighted predictive signals. Conversely, research utilizing structured, proprietary agency frameworks (e.g., Refinitiv, MSCI) relies heavily on XAI and traditional feature selection to crack open the 'black box,' yielding a tight split in dominance between Governance (driven by management and integrity indicators) and Environmental compliance. More importantly, across all methodological pipelines, the Social pillar remains systematically underrepresented as a primary predictive anchor. Despite less attention and focus being given to the Social pillar, it was highlighted in a few studies both as an essential mechanism influencing market performance and a genuine reflection of corporate integrity. The frequency of social topics in corporate earnings calls is the only factor causing a significant lagged causal effect on stock market indices such as the S&P 500 ESG index price (H. Lee et al., 2025c). Even when environmental topics are the most tackled in news articles, headlines, and earnings calls, issues with immediate human impact (e.g., workforce engagement, employee welfare, and local community relations) can immediately impact market sentiment in the short to medium term. The Social pillar is intrinsically connected with the environmental dimension: NLP-based management integrity analysis demonstrates that management trustworthiness positively influences modeled ES scores, making business operations a true manifestation of ethical commitment via the E and S dimensions Lv et al. (2025).

4.5. Synthesis of Research Questions

RQ1: Model Architectures. Tree-based ensembles (RF, XGBoost, LightGBM) dominate across all three streams when structured data is the primary input, achieving R^2 values of 0.71–0.83 for ESG score prediction. DL architectures (GRU, LSTM) demonstrate superiority only when temporal dynamics or large-scale unstructured text corpora are involved, with GRU achieving the highest accuracy (83.64%) for temporal ESG classification (S. Chen et al., 2026). The evidence converges on a principle rather than a single champion: the optimal architecture is determined by the structure, scale, and temporal properties of the input data.

RQ2: Functional Role. The functional role of the ESG score fundamentally dictates methodological choice. When ESG is the prediction target (Stream 1), ensemble methods with SHAP interpretability dominate. When ESG is an explanatory input (Stream 2), hybrid DL frameworks and causal inference methods (DML) are increasingly deployed. When ESG is an object of diagnostic scrutiny (Stream 3), NLP and transformer-based models (BERT derivatives) become the primary analytical tools, reflecting the shift toward unstructured data processing.

RQ3: Unstructured vs. Structured Data dilemma. The integration of unstructured alternative data (news, earnings calls, social media) materially enhances ESG signal resolution, particularly for the Environmental pillar, where news-derived sentiment captures high-profile controversies before annual disclosure cycles. For the Social pillar, NLP-derived controversy detection in media streams represents the most viable current prediction pathway. However, unstructured data introduces noise-to-signal challenges, and the review identifies a critical methodological imperative: human-in-the-loop validation of NLP-generated ESG dictionaries against established taxonomies (GRI, IFRS) to prevent algorithmic capture of corporate buzzwords rather than material sustainability practices.

RQ4: Pillar Machine-Readability. Pillar predictability is not universal but structurally conditioned by input data. Governance exhibits the highest machine-readability from structured disclosure data due to the binary, stable, and mandatory nature of its metrics (board independence, audit committees). Environment achieves the highest R^2 values when enriched with operational emissions data but is the least reconstructible from financial statements alone. Social remains the most intractable pillar across all architectures, owing to the absence of standardized mandatory disclosure, the qualitative and self-reported nature of its metrics, and the lack of independent verification infrastructure.

RQ5: Barriers—Black box and Greenwashing. The review reveals an extreme over-reliance on SHAP as the dominant (and often sole) XAI tool, creating a methodological monoculture that may bias interpretations. Greenwashing detection is emerging as a distinct analytical capability, operationalized through discourse–practice triangulation (comparing ESG reports, reputation indices, and objective performance data). A U-shaped relationship between ESG scores and greenwashing propensity means that high ESG scores cannot be treated as monotonically indicative of substantive sustainability.

RQ6: Signal-Taking to Signal Verification. Stream 3 comprises 41 studies, or 32.3% of the corpus, with most of this literature emerging in recent years. This pattern is consistent with an increasing research orientation toward verification and methodological scrutiny of ESG signals (see Figure 12), although the corpus alone cannot establish that it constitutes a structural transformation rather than part of the broader growth of AI-enabled ESG research. The sensitivity analysis, which places the Stream 3 share between 31.5% and 33.9%, further indicates that this pattern is not driven by the three identified S1/S3 boundary cases. We therefore interpret the evidence as an emerging verification-oriented research trajectory rather than a confirmed paradigm shift.

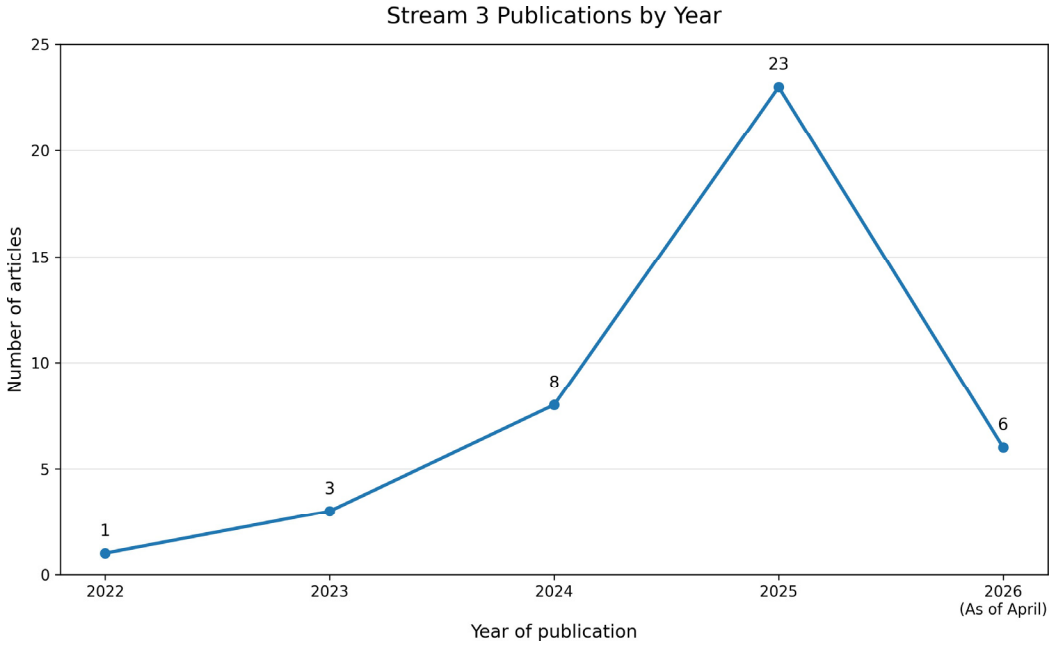


Figure 12. Number of publications per year—Stream 3. Note: 2026 is a partial year; the systematic search was completed in April 2026.

5. Limitations and Future Directions

Our SLR reveals that while the application of ML and DL to ESG analysis is a rapidly growing field, it is characterized by several critical gaps that present significant opportunities for future research. The current body of work, though foundational, primarily concentrates on a narrow set of models and data types, often overlooking the complex,

dynamic, and multifaceted nature of ESG factors. This section discusses these limitations and outlines promising future directions to build more robust, transparent, and impactful ESG models.

Before turning to the literature, we acknowledge the limitations of this review itself. First, the targeted methodological appraisal of the 29 Stream 1 studies reveals substantial heterogeneity in predictive validation (Appendix D, Table A5). Although most studies report some form of holdout testing or cross-validation, only three clearly employ a chronologically separated future-period test, while several high-fit results arise from target-proximal score reconstruction or non-temporal splitting of longitudinal data. The appraisal is intentionally focused on the leakage and generalization risks most relevant to ESG prediction and should not be interpreted as a comprehensive risk-of-bias assessment of all 127 studies. Second, the heterogeneity of tasks, metrics, and reporting conventions across streams precluded a formal meta-analysis, so our synthesis is thematic rather than quantitative, and the cross-study comparisons we present are descriptive. Third, the search was confined to Scopus and Web of Science, restricted to English-language records, and conducted at a single extraction point in April 2026; given the high share of IEEE- and ACM-indexed conference papers in the corpus, the exclusion of dedicated computer-science repositories may have left some relevant work unretrieved. Fourth, the evidence base is geographically concentrated, with a large share of studies—including many of the highest-performing structured-prediction results—drawn from Chinese A-share data, which conditions the external validity of our conclusions, including the pillar machine-readability hierarchy. As regards the coding design, the review did not employ independent blinded human double-coding; Gemini 3.1 and GPT-5.5 served as preliminary coding aids, while final author validation was conducted with their outputs visible. Accordingly, the agreement measures reported in Appendix C are descriptive coding-concordance measures rather than inter-rater reliability statistics; future reviews could strengthen this stage through independent blinded human coding.

A further limitation concerns data-provider concentration. Because a substantial share of the reviewed studies relies on LSEG/Refinitiv data, some findings—particularly the observed pillar machine-readability hierarchy—may partly reflect provider-specific scoring choices rather than universal ESG properties. Findings obtained from Refinitiv-based studies should therefore not be assumed to generalize unchanged to MSCI, Sustainalytics, S&P Global, or other providers. Future research should test whether model performance and pillar-level patterns remain stable across alternative rating systems. The published literature may also be affected by positive-result and publication bias, since successful prediction, explanation, or auditing applications are more likely to be reported than null or unsuccessful results. Consequently, the scarcity of reported failures should not be interpreted as evidence that ML-based ESG analysis is uniformly effective. This concern is reinforced by the limited use of standardized benchmark datasets, external validation, and cross-provider replication. Our conclusions therefore characterize the available published evidence rather than the full distribution of successful and unsuccessful applications.

A primary observation from our review is the prevalence of conventional ML algorithms such as RF, SVM, and standard ANN. Despite their strengths, the reviewed ML/DL models exhibit notable limitations. Most of the existing studies are constrained by their dependency on conventional algorithms that might overlook critical dynamic interdependencies between ESG indicators over time, potentially reducing their predictive efficacy under evolving market conditions. This identified gap indicates a pressing need to go beyond classical models and consider more sophisticated and evolved architectures that can exploit the interconnectedness and complexity of ESG pillars over time. For example, Graph Neural Networks (GNNs) are well-suited for modeling both the "Social"

and "Governance" factors. By translating supply chain labor dependencies, community relations, corporate ownership structures, and board interlocks into complex relational graphs, GNNs can capture the multi-tiered human and structural dynamics that standard models systematically miss. Given the temporal nature of ESG signals, TCNs offer a strong alternative to LSTMs, thanks to their ability to model long-term dependencies and computational efficiency. Furthermore, to address the interrelated nature of the three pillars, Multi-task Learning (MTL) architectures could be designed to forecast E, S, and G ratings simultaneously, allowing the models to learn shared representations and improve overall predictive performance.

Alongside the models mentioned above, few studies in the current literature applied ensemble learning, particularly stacked generalization models (Dwivedi et al., 2023; Giudici & Wu, 2025; Z. Wang et al., 2024). Therefore, we encourage future researchers to give them greater emphasis to obtain more accurate and robust predictions. This approach leverages the merits of various base learners by using multiple ESG ratings from each agency to train them. Methods such as bagging, stacking or boosting may be employed to derive a unified strong consensus score, which reduces the biases of individual models. Likewise, the Environmental dimension can be substantially strengthened by moving beyond textual, structured numerical and self-reported corporate data (S. L. Lin et al., 2025; Svanberg et al., 2022). For instance, CNN-trained geospatial data, such as satellite imagery, can be useful to verify and monitor corporate environmental impact, including deforestation or changes in land close to firm facilities. LSTM networks alongside the aforementioned TCNs are ideal for modeling time-series data related to energy consumption, carbon intensity, and GHG emissions. Finally, hybrid CNN-LSTM models can be particularly effective in combining spatial and temporal data to create a more comprehensive and verifiable assessment of a firm's environmental footprint. We further note, as a caution for interpretation, that the non-linear relationships reported across this literature are distinct and should not be conflated: a U-shaped relationship between firm ESG level and profitability (D'Amato et al., 2024), a separate U-shaped relationship between ESG level and greenwashing propensity (Eckberg et al., 2026), and a J-curve describing returns as ESG maturity increases over time (Cheng et al., 2025; Herman et al., 2025) each describe different variables.

Moreover, while our research indicates a potential bias towards the environmental and governance dimensions, the Social pillar remains empirically underexplored. The apparent disagreement in the literature over which pillar is "most predictable" is largely an artifact of differing input regimes—Governance is most accessible from standardized disclosure, the Environmental pillar is most predictable given rich operational data yet least reconstructible from financial statements alone, and the Social pillar remains the most intractable—rather than a genuine contradiction. Unlike Governance (binary metrics like board independence) or Environment (quantitative metrics like CO_2 emissions), Social metrics (e.g., human rights, human capital management, labor satisfaction) are complex, context-dependent, and lack standardized quantification, making them difficult for current ML models to process without advanced NLP. To deal with this, future research might combine diverse data sources and apply more sophisticated NLP models, namely FinBERT, RoBERTa, and VADER, to extract key insights from unstructured data. Analyzing sentiment and key themes from social media, detailed HR disclosures, and employee-review sites can provide dynamic, real-time metrics on employee satisfaction, corporate culture, and public perception, which are poorly captured by traditional S-scores. Moreover, the rapid emergence of LLMs, including GPT-5.5, LLaMA, and domain-specific variants, presents both an opportunity and a challenge for ESG assessment. LLMs can potentially automate ESG report analysis, detect greenwashing through nuanced reasoning, and generate consensus scores across agencies. However, their deployment raises concerns about hallucination risk,

reproducibility, and the opacity of proprietary model weights, which may paradoxically reintroduce the "black box" problem that XAI seeks to resolve. Future research should systematically benchmark LLM-based ESG analysis against established transformer architectures (BERT, RoBERTa) and evaluate their susceptibility to prompt-based manipulation in ESG scoring contexts.

The second significant gap is in interpretable models and rating divergence challenges. Our survey reveals an extreme preference towards SHAP for XAI. Although SHAP is one such promising method, this lack of variety and an over-reliance on only one XAI method are issues that require attention. The area of explainability provides a wider set of unexplored tools in the context of ESG. For example, other model-agnostic methods such as LIME, which explains individual predictions by generating a local, interpretable model, could provide an additional perspective. Likewise, as researchers begin to adopt neural networks, model-specific techniques like Integrated Gradients or Layer-Wise Relevance Propagation (LRP) will become essential for understanding DL models. This lack of methodological diversity suggests that our understanding of these complex models may be biased towards a single perspective. Depending only on SHAP might hide some interpretations that can be exposed by other XAI methods. There is therefore considerable scope for future ESG work to complement SHAP with other XAI techniques. This would yield not only multifaceted explanations but would also help to establish trust with stakeholders as well as support the deployment of ML-based ESG insights in practical financial decision-making.

Transparency is also necessary, given that divergence in ESG ratings across providers is a significant issue that can lead to considerable market confusion, as evidenced by the divergence in Tesla's varied and sometimes contradictory ESG ratings. While MSCI continues to laud Tesla as a top performer in the automobile sector for its environmental contributions in accelerating the transition to electric vehicles (Singla, 2025) by giving it an Average score "A", Sustainalytics penalizes it for concerns related to social and governance issues such as product safety controversies, board independence, labor practices, and executive compensation, giving it a Medium Risk score of 24.8 and positioning it in the middle of the pack for automakers (Mann, 2025). Refinitiv scores it around 63/100 for decent disclosure (Nampoothiri, 2022). In contrast, S&P Global brought attention to documented discrimination cases in Tesla factories and expressed high concerns regarding the company's corporate governance and public reporting practices; thus, Tesla was dropped from the S&P 500 ESG index in 2022[4]. Moreover, the unit of measure differs between agencies (Telukdarie et al., 2024). For instance, Moody's measured total waste in metric tons, while MSCI measured it in pounds. This divergence can lead to accusations of "greenwashing" and has tangible economic consequences. According to stakeholder theory, a firm's remedial green innovations may be seen as "greenwashing" when there is disagreement over ESG ratings, which will lead to higher costs of debt financing for corporations, ultimately bringing about a decline in TFP. This points to an interesting and potentially important opportunity: applying ML and DL not just to forecast ratings but to harmonize them. Advanced models could be developed to collect signals from various agencies, news sources, and corporate disclosures alongside other financial and macroeconomic data to produce a more stable, reliable, and consensus-based ESG score grounded in common norms and standards.

Lastly, there is a noticeable lack of studies about emerging markets and SMEs. The institutional, cultural, and regulatory contexts of these markets differ considerably from those of North America and Europe and from the large listed firms on which most current ESG research is focused. Moreover, data are generally more difficult to obtain, and their quality is challenging in these regions, while the nature of ESG risks and opportunities may be unique. We also observed that several of the included studies are constrained

by their dependence on a limited set of national regions or market indices, especially for NLP-based methods sensitive to language and cultural context, which might not be easily extendable to other markets such as Brazil, Turkey, Morocco, India, etc.; smaller firms; or alternative regulatory frameworks. The predominance of retail investors in China or the specific regulatory environment in Europe strongly affects how ESG information is priced. The need for broader and cross-regional analyses is crucial. Accordingly, one of the key areas for future research is the validation of existing models in the new context, along with developing tailor-made ones that would be responsive to the reality of emerging economies. This key area is not only important to global investors but it also serves a vital role in strengthening sustainable development, where it is most needed.

A related scope gap concerns heterogeneity within the financial sector. Models developed for traditional banks and other established financial institutions should not be assumed to transfer directly to FinTech firms, which differ in technological infrastructure, data environments, organizational structures, and governance arrangements. Recent comparative evidence published after our April 2026 search cutoff similarly emphasizes that ESG integration differs between FinTech and traditional financial institutions because sustainability practices are shaped by distinct technological capabilities, institutional structures, and governance systems (Sedghiani & Doğan, 2026). Future research should therefore test the cross-institutional transferability of ML-based ESG models and examine whether sector-specific ESG indicators, training data, and validation procedures are required for FinTech firms rather than applying unchanged models calibrated on traditional financial institutions.

6. Conclusions

The integration of cutting-edge computational technologies has become a necessity rather than a novelty in the ESG investing sector. This SLR of 127 peer-reviewed studies has explored the expanding field of ML and DL applications in ESG, a landscape simultaneously full of innovation opportunities and noticeable gaps. These articles were carefully classified and synthesized into three separate yet highly intertwined streams of the literature: (1) studies that predict ESG scores as a dependent variable to identify the key drivers of a company's ESG rating, thus enabling investors to make decisions about their portfolios and manage risks; (2) studies wherein ESG scores are used as independent variables to forecast financial outcomes; and (3) studies that employ ML models to benchmark, evaluate and explain the methodological opaqueness of the ESG scores. Our results verify the feasibility of the first two streams. When predicting ESG scores (Stream 1), RF, XGBoost and NN models have achieved high overall performance on both classification and regressing tasks, with pillar predictability conditioned by input regime rather than uniformly ordered: Governance is most accessible where standardized disclosure data dominate, the Environmental pillar attains the highest goodness-of-fit where rich operational data are available yet is the least reconstructible from financial statements alone, and the Social pillar remains intractable across input regimes. Conversely, when used as predictive input (Stream 2), ESG scores exhibit material and non-linear correlations with a wide range of financial outcomes, such as credit risk, portfolio investment, carbon intensity, mutual fund performance, stock returns and corporate Total Factor Productivity. Although tree-based models outperformed other techniques, we observed a strong emerging convergence towards ensemble learning methods (e.g., stacked generalization model) and hybrid frameworks that combine two or more ML techniques to improve the overall predictive capability and robustness. These applications confirm that, despite their flaws, ESG metrics do indeed capture financially relevant signals.

However, while the first two categories used ML to build predictive models on ESG scores, the third stream sheds light on the persistent challenge of ESG scores' discrepancies

between providers such as Refinitiv, MSCI, and Bloomberg, and deconstructs the opacity of traditional ESG ratings. DL techniques, especially advanced NLP and transformer-based models (e.g., VADER, BERT, c-TF-IDF, BERTopic, Word2vec), play a major role in giving an alternative solution by creating new ESG scores from company news articles, archives and sustainability reports and comparing them with the ones provided by agencies or validating the existing ratings, in view of revealing the prevalent issues of methodological opacity, inconsistency, and potential biases that currently undermine the consistency of ESG assessments. This central tension dictates a clear and compelling agenda for future research, moving the field from mere application to resolution. First, the "black box" problem must be met head-on by diversifying the use of XAI. The current over-reliance on SHAP, while useful, must be augmented with a broader suite of other XAI techniques to ensure that the interpretations are robust and not an artifact of a single method. Transparency must be the default, not an afterthought, if these models are to be trusted by regulators and investors. Second, the 'Social' pillar must be addressed. Its status as "empirically underexplored" is a direct result of a mismatch between data gathering and modeling. This gap calls for novel approaches. For example, combining NLP analysis of employee reviews and social media with GNNs to model corporate culture and supply chain labor practices—complex, relational data that traditional models miss. Third, the field must evolve beyond diagnosing divergence by leveraging advanced architectures, such as multi-task learning (MTL), to simultaneously model ratings from different agencies or develop consensus-based meta-models that integrate structured data, unstructured text, and satellite imagery to create a more robust, auditable, and reliable score.

Collectively, these findings provide explicit answers to the six research questions posed at the outset of this review. The optimal ML architecture is data-conditioned rather than universal (RQ1); the functional role of the ESG score dictates distinct methodological signatures across streams (RQ2); unstructured data materially enhances ESG signal resolution but requires human-in-the-loop validation (RQ3); pillar predictability is structurally hierarchical, with Governance and Environment leading and Social lagging (RQ4); the dual barriers of XAI monoculture and greenwashing opacity demand methodological diversification (RQ5); and the growing methodological-assessment stream indicates an emerging shift toward signal verification, while the available corpus does not yet establish a structural transformation of the field (RQ6). To sum up, this SLR finds that the integration of ML and DL has improved the ESG landscape. It has exposed the limitations of traditional ratings and armed researchers with the tools to challenge them. The distinct contribution of this review is to have isolated and structured that challenge—the assessment-and-resolution stream—as a field in its own right rather than leaving it dissolved within the broader literature on ESG prediction. The ultimate promise of AI in this domain is therefore not simply to build a better predictor of flawed scores but to forge a new generation of assessment tools that are, by design, transparent, reliable, and fundamentally more accountable.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jrfm19090708/s1. Supplementary File S1: PRISMA-stage screening records and final included-study records; Supplementary File S2: Study-level AI-assisted classification dataset reporting GPT-5.5, Gemini 3.1, and final author classifications.

Author Contributions: Conceptualization, I.E.I., A.E.A. and B.G.; Methodology, I.E.I., A.E.A. and B.G.; Formal Analysis, I.E.I.; Investigation, I.E.I.; Writing—Original Draft Preparation, I.E.I.; Writing—Review and Editing, A.E.A., B.G., S.O.E.A. and M.V.; Visualization, I.E.I. and B.G.; Supervision, A.E.A., B.G., S.O.E.A. and M.V.; Funding Acquisition, I.E.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Center for Scientific and Technical Research (CNRST), Rabat, Morocco, grant number 39UIT2023.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The screening records, study-level coding decisions, and final classification dataset are provided as Supplementary Materials accompanying this article. The standardized AI-assisted classification prompt and database-specific search strategies are reported in Appendixes B and E, respectively.

Acknowledgments: During the preparation of this work, the authors used GPT-5.5 and Gemini 3.1 as part of the AI-assisted classification consistency-checking procedure described in the methodology section. These tools were used to support the preliminary classification of reviewed studies into the three streams of the tripartite framework, while all final inclusion, exclusion, coding, interpretation, and validation decisions remained under the authors' control. The authors also used AI-assisted tools to support language refinement and improve readability. After using these tools, the authors reviewed, verified, and edited the output as needed and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Comparison of Final ESG Score Scales Across Rating Agencies

Rating Agency	ESG Score/Rating Scale	Description
MSCI	AAA to CCC	AAA (Leader). AA: Strong management. A—BB: Average. B—CCC (Laggard): Worst in class relative to peers, high exposure, and poor management of ESG risks.
Refinitiv	0 to 100	100: High transparency. 0: Very poor ESG performance.
Sustainalytics	0 to 100	0 (Lowest) to 100 (Highest). Low score = low risk.
S&P Global	0. to 100	Higher scores indicate better performance in managing material ESG risks, opportunities, and impacts, informed by their Corporate Sustainability Assessment (CSA).
ISS	A+ to D−	A+ (Excellent) to D− (Poor). Granted to industry leaders who meet demanding absolute performance expectations and are well-positioned to manage critical ESG risks. Also provides a decile rank (1–10).
Bloomberg	0 to 100	0 (Limited) to 100 (Comprehensive) ESG Disclosure. Primarily measures the extent of a company's public ESG disclosures, reflecting transparency.

Appendix B. AI-Assisted Coding Prompt

You are assisting with a systematic literature review on Machine Learning and Deep Learning applications in ESG ratings and finance. Classify the following article into one of three mutually exclusive streams based on the functional role of ESG scores or ratings.

Stream 1: ESG score prediction. Select this stream if ESG scores, ESG ratings, ESG rating classes, or E/S/G pillar scores are the dependent variable, target variable, or prediction output.

Stream 2: Financial application. Select this stream if ESG scores or ratings are used as independent variables, explanatory variables, input features, or predictors of financial, market, risk, credit, operational, or environmental outcomes.

Stream 3: Methodological assessment. Select this stream if the study evaluates, explains, audits, reconciles, or challenges ESG rating methodologies, including rating divergence, XAI—based explanation of rating drivers, greenwashing detection, ESG disclosure opacity, or provider disagreement.

Return the following fields:

1. Article title
2. Assigned stream
3. One-sentence justification
4. Evidence from the abstract or methodology
5. Confidence level: high, medium, or low
6. Any ambiguity requiring human review

Article information:
[Insert title, abstract, keywords, and methodology excerpt]

Appendix C. AI-Assisted Coding Agreement Statistics

This appendix reports the quantitative and case-level information supporting the AI-assisted coding procedure described in Section 3.2. Table A1 presents the stream distribution produced by each classification source, Table A2 reports descriptive coding concordance, Table A3a through Table A3c present the corresponding confusion matrices, and Table A4 documents the twelve non-unanimous cases and their final author adjudication. Because final validation was conducted with the preliminary LLM outputs visible, these statistics are descriptive and are not interpreted as independent inter-rater reliability measures.

Table A1. Stream distribution by classification source (N = 127).

Classifier	Stream 1	Stream 2	Stream 3
Gemini 3.1	35	57	35
GPT-5.5	27	56	44
Human validation	29	57	41

Note. Stream 1 = ESG score/rating prediction; Stream 2 = Financial, operational, or investment application; Stream 3 = Methodological assessment of ESG ratings. Each row sums to 127. Three-way unanimous agreement was obtained for 115/127 studies (90.6%).

Table A2. Observed coding concordance across classification sources.

Comparison	Agreement	Observed Agreement
Gemini 3.1 vs. GPT-5.5	115/127	90.6%
Gemini 3.1 vs. final author classification	118/127	92.9%
GPT-5.5 vs. final author classification	124/127	97.6%
Three-way unanimous agreement	115/127	90.6%

Table A3. Confusion matrices for pairwise comparisons (N = 127 per matrix). Rows = first rater; columns = second rater. S1 = Stream 1; S2 = Stream 2; S3 = Stream 3.

(a) Gemini 3.1 vs. GPT-5.5					(b) Gemini 3.1 vs. Human					(c) GPT-5.5 vs. Human				
G/GPT	**S1**	**S2**	**S3**	**Tot.**	**G/H**	**S1**	**S2**	**S3**	**Tot.**	**GPT/H**	**S1**	**S2**	**S3**	**Tot.**
S1	26	1	8	35	S1	28	1	6	35	S1	27	0	0	27
S2	0	55	2	57	S2	0	56	1	57	S2	0	56	0	56
S3	1	0	34	35	S3	1	0	34	35	S3	2	1	41	44
Tot.	27	56	44	127	Tot.	29	57	41	127	Tot.	29	57	41	127

Note. G = Gemini 3.1; GPT = GPT-5.5; H = final author classification. Diagonal cells represent agreement. Most off-diagonal cases involve the Stream 1/Stream 3 distinction, reflecting genuine overlap between supervised ESG-score prediction and methodological assessment.

Table A4. Author adjudication of the twelve non-unanimous classification cases.

Study	GPT-5.5	Gemini 3.1	Final Decision	Reason for Final Assignment
Del Vitto et al. (2023)	S3	S1	S1 Borderline S1/S3	Refinitiv E, S, and G ratings are the supervised prediction targets, and reconstruction accuracy is a central empirical endpoint. Explainability is subsequently used to interpret the fitted model; hence, the functional role remains prediction, while the transparency objective warrants the boundary flag.
Ferro et al. (2025)	S1	S3	S1 Borderline S1/S3	LSEG ESG scores are reconstructed from granular provider indicators before their drivers and materiality weights are interpreted. Because the existing provider rating remains the modeled output, the study is assigned to S1, while its explicit investigation of rating construction warrants the S1/S3 boundary designation.
S.-W. Lin and Lin (2025)	S2	S1	S2	ESG measures enter the broader efficiency framework, whereas the substantive outcome is firm operational, R&D, market, and overall efficiency in relation to CEO hubris.
Y. Chen et al. (2024)	S3	S2	S2	ESG-related news and sentiment function as explanatory information, while abnormal stock-market returns constitute the empirical outcome.
Amendola et al. (2025)	S3	S1	S3	ESG scores are explicit model outcomes, while ML performance and prediction residuals are used to understand differences in ESG adoption.
Censi et al. (2025)	S3	S1	S3	No existing provider rating is reproduced. The study develops and evaluates an integrated AI–Process Mining–RPA framework for ESG risk assessment, with ML classification serving as one component of the proposed assessment methodology.
C.-C. Chen et al. (2023)	S3	S1	S1	MSCI-derived ESG issue categories constitute the explicit classification targets. The MSCI framework supplies the labels but is not itself evaluated or challenged.

Table A4. *Cont.*

Study	GPT-5.5	Gemini 3.1	Final Decision	Reason for Final Assignment
Fischbach et al. (2023)	S3	S1	S3	The study constructs an alternative media-based ESG assessment architecture intended to generate ESG signals from external textual evidence.
Hassan Nassar et al. (2025)	S3	S1	S3	The study constructs a real-time NLP and knowledge-graph architecture for generating and interpreting alternative ESG signals in response to limitations of conventional assessments; no existing ESG-provider rating is the supervised target.
Katterbauer et al. (2022)	S3	S2	S3	The study constructs environmental-compliance classes through unsupervised learning and benchmarks them against conventional ESG/environmental scores and financial indicators. The primary contribution is an alternative environmental-compliance assessment framework rather than ESG-score prediction or financial-outcome prediction.
Storchevoy et al. (2025)	S3	S1	S3 Borderline S1/S3	This is a genuine boundary case because AI is trained to predict the RAEX G-component rating. However, prediction is instrumental to a broader methodological objective: the study compares alternative rating approaches, examines discrepancies between them, and evaluates how corporate-governance quality should be assessed for regional companies. The rating methodology itself is therefore the principal object of investigation.
van der Heever et al. (2024)	S3	S1	S3	The study uses XAI to construct and interpret an alternative ESG-provider score and neither predicts an existing ESG-provider score nor consumes ESG scores to predict a downstream financial outcome.

Note. These twelve studies are the cases in which the two preliminary LLM classifications and the final author classification were not unanimously aligned. Final adjudication was based on the study's primary analytical objective and the functional role of the ESG variable. Studies were assigned to Stream 1 when an externally defined ESG score, rating, pillar, grade, or category constituted the principal supervised target; to Stream 2 when ESG information primarily served as an explanatory input for a downstream outcome; and to Stream 3 when the principal contribution concerned the evaluation, redesign, comparison, or construction of an ESG assessment methodology or alternative signal. Del Vitto et al. (2023), Ferro et al. (2025), Storchevoy et al. (2025) were retained as explicit S1/S3 boundary cases.

Appendix D. Methodological Appraisal of Stream 1 Studies

Table A5 reports the study-level methodological appraisal conducted for the 29 studies classified in Stream 1. The appraisal focuses on predictor proximity to the ESG target, out-of-sample evaluation, and temporal validation, allowing score reconstruction, autoregressive forecasting, and genuine out-of-time prediction to be distinguished.

Table A5. Methodological appraisal of Stream 1 ESG prediction studies.

Study	Lagged ESG/ Rating Input	E/S/G Pillar Scores as Predictors	Principal Reported Predictive Result	Out-of-Sample Evaluation	Temporal Validation	Interpretation for Predictive Validity
Ang et al. (2023)	No	No	Best total-ESG RMSE ≈ 7.29	10-fold cross-validation	NR	Predicts ESG at $t + 1$, but the chronology of the cross-validation folds is not reported; therefore, genuine out-of-time generalization cannot be established.
Aue et al. (2025)	No	No	Mean absolute rating difference ≈ 11.85; exact-score accuracy relatively low	Train/test evaluation	No	Prediction relies on news-based information rather than target-proximal ESG inputs, but the evaluation does not use a chronological holdout.
Belkhiria et al. (2025)	No	Yes	R^2 = 0.988 for E/S/G → aggregate ESG; optimized financial-only RF test R^2 = 0.9175	80/20 train-test split	No	The R^2 = 0.988 result reflects direct reconstruction of the composite score from its constituent pillars and should not be interpreted as independent ESG forecasting.
Cai et al. (2025)	No	No	R^2 = 0.5312; rating-class accuracy = 93.33%	80/20 split plus cross-validation	No	Held-out evaluation is reported, but observations are not chronologically separated.
Y. Chen and Cheng (2025)	Yes	No	Ensemble MAPE ≈ 4.5%	70/30 train-test split	NR	Historical ESG information contributes to prediction, but the chronology of the train-test partition is not sufficiently reported.
Choudhari et al. (2025)	NR	No	Gradient Boosting accuracy = 92%; AUC = 0.94	Stratified k-fold CV and separate sector/location tests	NR	Additional sector/location testing strengthens robustness, but the temporal role of historical ESG information is insufficiently specified.
Cini and Ferrari (2025)	Yes	Yes	Accuracy ≈ 0.80 with ESG-related inputs versus ≈ 0.62 with financial-only inputs	Evaluation reported; split protocol incompletely specified	NR	Current ESG and pillar information is used to identify the subsequent rating class, creating substantial target proximity.
David et al. (2025)	Yes	Yes	Cross-validated R^2 ≈ 0.996	5-fold cross-validation	No/NR	Lagged ESG, pillar scores, and ESG-derived variables directly enter the model. The exceptionally high fit therefore reflects substantial autoregressive and target-proximal information.

Table A5. *Cont.*

Study	Lagged ESG/ Rating Input	E/S/G Pillar Scores as Predictors	Principal Reported Predictive Result	Out-of-Sample Evaluation	Temporal Validation	Interpretation for Predictive Validity
Dwivedi et al. (2023)	No	No	Gradient Boosting RMSE ≈ 3.47	5-fold cross-validation	N/A	Cross-sectional prediction uses financial and non-financial firm attributes; a temporal forecasting interpretation is not required.
Fabijańska et al. (2025)	No	No [a]	ANN R^2 ≈ 0.92–0.96	Random 80/20 train-test split	No	Although literal pillar scores are not predictors, provider-defined ESG scoring components are used to reproduce the provider score, making this a target-proximal reconstruction task.
Ferro et al. (2025)	No	No [a]	Provider aggregation recovered at approximately R^2 = 1.000	Reconstruction and holdout analyses	N/A	The near-perfect fit reflects reverse-engineering of the LSEG scoring architecture rather than forecasting of an unknown future ESG outcome.
A. Jain et al. (2025)	No	No	R^2 = 0.8678; MAPE = 7.91%	70/30 split plus 5-fold cross-validation	No	Prediction uses independent SDG-related variables, but no chronological future-period test is reported.
Kim and Lee (2025)	No	No	Random Forest R^2 = 0.826	80/20 split plus 5-fold cross-validation	No	The longitudinal panel is evaluated through non-temporal splitting, leaving out-of-time generalization untested.
K. Lee (2026)	No	No	Composite-rating F1 ≈ 0.652; change-prediction F1 ≈ 0.702	5-fold cross-validation with leakage checks	No	Predictor timing is aligned to reduce future-information leakage, but the evaluation does not employ a chronological holdout period.
H. Lee et al. (2025a)	Yes	No	2019 test MAE ≈ 0.732 (E), 0.637 (S), and 0.602 (G)	2016–2018 training; 2019 test	Yes	A stronger forecasting design: previous-year ratings are legitimate predictors because performance is evaluated on a chronologically later unseen year.
S. Chen et al. (2026)	Yes	No	GRU accuracy = 83.64%; F1 = 0.7295	Earlier observations used for training; later observations for testing	Yes	The model performs genuine autoregressive forecasting with chronologically separated training and testing observations.

Table A5. *Cont.*

Study	Lagged ESG/ Rating Input	E/S/G Pillar Scores as Predictors	Principal Reported Predictive Result	Out-of-Sample Evaluation	Temporal Validation	Interpretation for Predictive Validity
S. L. Lin et al. (2025)	No	Yes [b]	Stacking accuracy = 98.78%	Split details NR	NR	The reported 98.78% result concerns media-attention classification using ESG information and is not directly comparable with ESG-score prediction accuracy.
Del Vitto et al. (2023)	No	No [a]	R^2 up to ≈0.873 across pillar/settings	Repeated random train-test splits	No	Provider variables are used to reconstruct Refinitiv E/S/G ratings. The task is therefore score reconstruction rather than future-period forecasting.
C.-C. Chen et al. (2023)	No	No	F1 ≈ 0.69 for English and ≈0.7 for French	Fixed held-out test sets	N/A	The task is static NLP classification of ESG issues; longitudinal target leakage is not applicable.
H. Lee et al. (2024a)	No	No	BERT accuracy = 78.62%; pillar-label performance up to ≈88.4%	NR	N/A	ESG labels constitute prediction targets, but the train-test protocol is insufficiently reported for a strong generalization claim.
Taskin et al. (2025)	Yes	No	Test MAPE = 8.05%; implied prediction accuracy up to ≈91.95%	External future-year test	Yes	This is a legitimate persistence-forecasting design: historical ESG scores are used to predict a later ESG outcome and are evaluated on a future-period test set.
X. Wang (2025)	No	No	Accuracy ≈ 97.5%; F1 ≈ 0.96	80/20 split plus cross-validation	N/A	High classification performance is reported, although limitations in data construction and reporting constrain assessment of external generalizability.
Z. Wang et al. (2024)	Yes	No	Overall ESG R^2 = 0.8193; industry maximum ≈ 0.8971	Train-test evaluation plus 10-fold cross-validation	NR	Lagged ESG is an explicit predictor, but chronological separation between training and testing observations is not demonstrated.
C. Wang et al. (2025)	No	No	Overall ESG R^2 = 0.7373; E-pillar R^2 ≈ 0.94	Train/validation/test sets	NR	The study uses separate evaluation sets and training-only feature selection, although its simulated dataset and unspecified temporal structure limit temporal inference.

Table A5. *Cont.*

Study	Lagged ESG/ Rating Input	E/S/G Pillar Scores as Predictors	Principal Reported Predictive Result	Out-of-Sample Evaluation	Temporal Validation	Interpretation for Predictive Validity
S.-L. Yang (2025)	Yes	No	Accuracy = 0.74; F1 = 0.71; AUC = 0.78	Held-out protocol incompletely reported	NR	Predictors are lagged by one year, but a clearly separated future-period holdout is not documented; performance cannot therefore be treated as verified out-of-time accuracy.
J. Zhang and Zhao (2026)	No	No	Accuracy = 91.0%; F1 = 0.901; AUC = 0.977	80/20 split plus 3-fold CV and additional robustness tests	No	The study provides extensive robustness checks, including alternative markets and MSCI ratings, but does not employ a chronological test.
M. Zhang et al. (2025)	No	No	E-BERT accuracy ≈ 93%	*k*-fold/test evaluation	N/A	The task is static text-to-rating classification rather than longitudinal forecasting; temporal leakage is therefore not the principal concern.
H.-Y. Lin and Hsu (2023)	No	No	Test R^2 up to ≈ 0.9995	Random 70/30 train-test split	No	Near-perfect fit occurs without lagged ESG or pillar-score predictors, illustrating that non-temporal splitting of panel observations can itself inflate apparent generalization.
Martin-Melero et al. (2025)	No	No	Regression R^2 ≈ 0.60–0.80; classification accuracy ≈ 70–80%	70/30 split plus 5-fold tuning	N/A	Conventional held-out prediction uses financial or sector-exposure variables and does not make a longitudinal forecasting claim.

Note. NR = not reported with sufficient detail to verify; N/A = temporal validation is not applicable because the principal prediction task is static or cross-sectional rather than longitudinal forecasting. Lagged ESG/rating input refers to the explicit use of prior-period ESG scores or rating classes as predictors. "Temporal validation" is coded Yes only when the evaluation clearly separates earlier training observations from a later unseen test period. A No designation does not necessarily imply data leakage; it indicates that a chronological future-period test was not implemented. [a] No literal E/S/G pillar score is used as a predictor, but provider-defined underlying scoring components are used to reconstruct the provider rating. [b] ESG/pillar information is used as predictive input, but the reported 98.78% accuracy concerns the media-attention classification task rather than ESG-score prediction itself.

Appendix E. Database-Specific Search Strategies

To strengthen the transparency and reproducibility of the review, this appendix reports the database-specific search strategies applied in Scopus and WOS Core Collection. The same two conceptual keyword blocks were used in both databases: the first identifies AI and ML techniques, while the second identifies ESG ratings, scores, and evaluations. Searches were restricted to English-language journal articles and conference/proceedings papers, with no lower publication-year restriction. The final database search was completed in April 2026.

Appendix E.1. Scopus

The Scopus Advanced Search was conducted in the title, abstract, and keyword fields using the following query:

```
(TITLE-ABS-KEY (
  "artificial intelligence" OR "AI" OR "machine learning"
  OR "deep learning" OR "neural network*" OR
  "random forest" OR NLP OR "XGBoost" OR
  "BERT"
  )
  A
  N
  D
  TITLE-ABS-KEY (
  "ESG rating*" OR "ESG scoring"
  OR "ESG score*" OR "ESG
  evaluation"
  OR "environmental social governance rating"
  )
)
AND (
  LIMIT-TO (DOCTYPE, "ar")
  OR LIMIT-TO (DOCTYPE, "cp")
)
AND LIMIT-TO (LANGUAGE, "English")
```

Database: Scopus.
Fields searched: Title, Abstract, and Keywords (TITLE-ABS-KEY).
Document types: Article (ar) and Conference Paper (cp).
Language: English.
Publication-year restriction: None.
Final search date: April 2026.
Records retrieved: 337.

Appendix E.2. Web of Science Core Collection

The WOS Core Collection search applied the same keyword blocks to the title, abstract, and author-keyword fields:

```
(
  TI=("artificial intelligence" OR "AI" OR "machine learning"
      OR "deep learning" OR "neural network*" OR
      "random forest" OR NLP OR "XGBoost" OR
      "BERT")
  OR
  AB=("artificial intelligence" OR "AI" OR "machine learning"
      OR "deep learning" OR "neural network*" OR
      "random forest" OR NLP OR "XGBoost" OR
      "BERT")
  OR
  AK=("artificial intelligence" OR "AI" OR "machine learning"
      OR "deep learning" OR "neural network*" OR
      "random forest" OR NLP OR "XGBoost" OR
      "BERT")
)
A
N
D
(
  TI=("ESG rating*" OR "ESG scoring"
      OR "ESG score*" OR "ESG
      evaluation"
      OR "environmental social governance rating")
  OR
  AB=("ESG rating*" OR "ESG scoring"
      OR "ESG score*" OR "ESG
      evaluation"
      OR "environmental social governance rating")
  OR
  AK=("ESG rating*" OR "ESG scoring"
      OR "ESG score*" OR "ESG
      evaluation"
      OR "environmental social governance rating")
)
```

Database: Web of Science Core Collection.
Fields searched: Title (TI), Abstract (AB), and Author Keywords (AK).
Document types: Article and Proceedings Paper.
Language: English.
Publication-year restriction: None.
Final search date: April 2026.
Records retrieved: 224.

Notes

1 https://www.scirp.org/reference/referencespapers?referenceid=3033722 (accessed on 11 December 2025).
2 https://corporatefinanceinstitute.com/resources/esg/esg-environmental-social-governance/ (accessed on 30 November 2025).
3 https://www.msci.com/indexes/documents/methodology/ESG_Factors_In_Methodology_20241119.html?utm_source=chatgpt.com (accessed on 5 March 2026).
4 https://www.barrons.com/articles/tesla-sp-esg-index-sustainability-ratings-51653011864 (accessed on 25 September 2025).

References

Agosto, A., Cerchiello, P., & Giudici, P. (2025). Bayesian learning models to measure the relative impact of ESG factors on credit ratings. *International Journal of Data Science and Analytics*, *20*(2), 357–368. [CrossRef]

Agosto, A., & Tanda, A. (2025). Divergence and aggregation of ESG ratings: A survey. *Open Research Europe*, *5*, 28. [CrossRef]

Ahmad, I., & Ahmed, T. (2026). AI-enhanced ESG framework for sustainability: A multi-sectoral analysis through an explainable AI approach. *Sustainability*, *18*(2), 794. [CrossRef]

Amendola, A., Bernardelli, A. E., & Giudici, P. (2025). Measuring inequality in the adoption of ESG scores by small and medium enterprises. *International Review of Economics & Finance*, *103*, 104522. [CrossRef]

Amir, A.-Z., & Serafeim, G. (2018). Why and how investors use ESG information: Evidence from a global survey. *Financial Analysts Journal*, *74*(3), 87–103. [CrossRef]

Ang, G., Guo, Z., & Lim, E.-P. (2023). On predicting ESG ratings using dynamic company networks. *ACM Transactions on Management Information Systems*, *14*(3), 27. [CrossRef]

Annesi, N., Battaglia, M., Ceglia, I., & Mercuri, F. (2025). Navigating paradoxes: Building a sustainable strategy for an integrated ESG corporate governance. *Management Decision*, *63*(2), 531–559. [CrossRef]

Assael, J., Carlier, L., & Challet, D. (2023). Dissecting the explanatory power of ESG features on equity returns by sector, capitalization, and year with interpretable machine learning. *Journal of Risk and Financial Management*, *16*(3), 159. [CrossRef]

Aue, T., Jatowt, A., & Färber, M. (2025). Predicting company ESG ratings from news articles using multivariate timeseries analysis. In *Companion proceedings of the ACM on web conference 2025* (pp. 1774–1780). Association for Computing Machinery. [CrossRef]

Belkhiria, S., Lajmi, A., & Sayed, S. (2025). Predicting environmental social and governance scores: Applying machine learning models to French companies. *Journal of Risk and Financial Management*, *18*(8), 413. [CrossRef]

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

Billio, M., Costola, M., Hristova, I., Latino, C., & Pelizzon, L. (2021). Inside the ESG ratings: (Dis)agreement and performance. *Corporate Social Responsibility and Environmental Management*, *28*(5), 1426–1445. [CrossRef]

Binzaiman, F., Edhrabooh, K. M., Alromaihi, M., & AlShammari, M. (2024). Predicting environmental, social, and governance scores with machine learning: A systematic literature review. In *2024 5th international conference on data analytics for business and industry (ICDABI)* (pp. 117–122). IEEE. [CrossRef]

Boell, S. K., & Cecez-Kecmanovic, D. (2015). On being 'systematic' in literature reviews in IS. *Journal of Information Technology*, *30*(2), 161–173. [CrossRef]

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. [CrossRef]

Cai, B., Ye, Z., & Chen, S. (2025). Intelligent ESG evaluation for construction enterprises in China: An LLM-based model. *Buildings*, *15*(15), 2710. [CrossRef]

Carlei, V., Furia, D., Ceccarelli, A., & Cascioli, P. (2025). Outperforming ESG stocks portfolio: A machine learning ranking model with catboots regressor. *The North American Journal of Economics and Finance*, *80*, 102517. [CrossRef]

Censi, R., Campana, P., Bellini, F., Schettino, F., & De Pucchio, C. (2025). AI-driven process mining for ESG risk assessment in sustainable management. *Buildings*, *15*(23), 4260. [CrossRef]

Chabot, M. (2025). Textual and AI-based analysis of climate disclosures: Evidence from the European energy sector. *The British Accounting Review*, 101736. [CrossRef]

Chatterji, A. K., Durand, R., Levine, D. I., & Touboul, S. (2016). Do ratings of firms converge? Implications for managers, investors and strategy researchers. *Strategic Management Journal*, *37*(8), 1597–1614. [CrossRef]

Chen, B. (2026). Can machine learning uncover ESG alpha in the Chinese A-share market? An 'alpha illusion' case study. *Finance Research Letters*, *91*, 109462. [CrossRef]

Chen, C.-C., Tseng, Y.-M., Kang, J., Lhuissier, A., Day, M.-Y., Tu, T.-T., & Chen, H.-H. (2023). Multi-lingual ESG issue identification. In C.-C. Chen, H. Takamura, P. Mathur, R. Sawhney, H.-H. Huang, & H.-H. Chen (Eds.), *Proceedings of the fifth workshop on financial technology and natural language processing and the second multimodal AI for financial forecasting* (pp. 111–115). Association for Computational Linguistics (ACL). Available online: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85178086754&partnerID=40&md5=47df2b2d5b6eabb76895290a0c38cfb6 (accessed on 11 December 2025).

Chen, S., Han, X., Zhang, Z., & Zhao, X. (2023). ESG investment in China: Doing well by doing good. *Pacific Basin Finance Journal*, *77*, 101907. [CrossRef]

Chen, S., Zhang, J., & Ge, L. (2026). *Predicting ESG ratings with recurrent neural network models: Evidence from China's A-share listed companies*. Computational Economics. [CrossRef]

Chen, Y., & Cheng, S. (2025). Machine learning–based ESG score prediction: Evidence from the Yangtze river delta. In *Proceedings of the 2025 2nd international conference on digital economy and computer science* (pp. 832–835). Association for Computing Machinery. [CrossRef]

Chen, Y., Zheng, Y., Lv, G., Gu, Y., & Zhang, W. (2024). Quantifying the effect of ESG-related news on Chinese stock movements. *Journal of Environmental Management*, *354*, 120301. [CrossRef]

Cheng, L. T., Cheong, T. S., Wojewodzki, M., & Chui, D. (2025). The effect of ESG divergence on the financial performance of Hong Kong-listed firms: An artificial neural network approach. *Research in International Business and Finance*, *73*, 102616. [CrossRef]

Choudhari, A., Bhandari, R., Kakade, R., Ramteke, N., Tatake, V. V., & Raidas, A. (2025). Assessing the effectiveness of AI enabled Fintech services on corporate ESG performance. In *2025 international conference on recent innovation in science engineering and technology (ICRISET)* (pp. 1–6). IEEE. [CrossRef]

Cini, F., & Ferrari, A. (2025). Towards the estimation of ESG ratings: A machine learning approach using balance sheet ratios. *Research in International Business and Finance*, *73*, 102653. [CrossRef]

Clement, A., Robinot, É., & Trespeuch, L. (2025). Beyond resource use: XGBoost reveals unexpected drivers of corporate. CO_2 emissions. *Contemporary Economics*, *19*(2), 163–185. [CrossRef]

Climent, F., Momparler, A., & Carmona, P. (2025). Predicting mutual fund performance with machine learning: Are ESG pillar scores relevant predictors of fund return? *Borsa Istanbul Review*, *25*(6), 1403–1419. [CrossRef]

Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of Management*, *37*(1), 39–67. [CrossRef]

D'Amato, V., D'Ecclesia, R., & Levantesi, S. (2024). Firms' profitability and ESG score: A machine learning approach. *Applied Stochastic Models in Business and Industry*, *40*(2), 243–261. [CrossRef]

David, L. K., Wang, J., Angel, V., & Amjad, N. (2025). ESG scoring and forecasting in China: Advancing sustainable business with multidimensional modeling. *Journal of the Asia Pacific Economy*, *30*(1), 128–157. [CrossRef]

Del Vitto, A., Marazzina, D., & Stocco, D. (2023). *ESG ratings explainability through machine learning techniques*. Annals of Operations Research. [CrossRef]

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics (ACL).

Dincă, M., Ciotlăuși, V., & Akomeah, F. (2025). Estimating the impact of ESG on financial forecast predictability using machine learning models. *International Journal of Financial Studies*, *13*(3), 166. [CrossRef]

Ding, H., & Xu, Y. (2025). Rating of financing ability of listed companies based on ESG performance. *Sustainability*, *17*(18), 8512. [CrossRef]

Di Ventura, E., Gotti, G., Morrone, C., & Savio, R. (2026). How to mitigate the impact of ESG on beta: The hidden power of internal controls. *Corporate Social Responsibility and Environmental Management*, *33*(3), 3737–3761. [CrossRef]

Dorfleitner, G., Halbritter, G., & Nguyen, M. (2015). Measuring the level and risk of corporate responsibility—An empirical comparison of different ESG rating approaches. *Journal of Asset Management*, *16*(7), 450–466. [CrossRef]

Dossa, J., Ukwuoma, C., Thomas, D., Dossa, J., & Gopang, A. (2025). Prediction of nexus among ESG disclosure and firm performance: Applicability, explainability and implications. *Innovation and Green Development*, *4*(4), 100261. [CrossRef]

dos Santos, I., de Arau'jo Moxoto', A., & de Souza Mota, E. (2025). Greenwashing in Brazilian corporations: A machine learning approach to unmask the discourse-practice gap. *Corporate Social Responsibility and Environmental Management*, *32*(6), 8214–8224. [CrossRef]

Dwivedi, D., Batra, S., & Pathak, Y. (2023). A machine learning based approach to identify key drivers for improving corporate's ESG ratings. *Journal of Law and Sustainable Development*, *11*(1), e0242. [CrossRef]

Eckberg, J., Dorfleitner, G., Kathan, M. C., & Utz, S. (2026). Determinants and forecasting of corporate greenwashing behavior. *Journal of Economic Behavior & Organization*, *241*, 107354. [CrossRef]

Ergenç, C., & Aktaş, R. (2025). Clustering S&P 500 companies by machine learning for sustainable decision-making. *Economics and Business Review*, *11*(3), 91–117. [CrossRef]

Fabijańska, A., Wołczek, P., & Sikacz, H. (2025). Can machine learning bring ESG ratings closer to small and medium-sized enterprises? *Oeconomia Copernicana*, *16*(4), 1615–1663. [CrossRef]

Feng, H., Chen, S., Lyu, J., & Wu, J. (2025). Epidemic risk and corporate financing decisions: Machine learning based prediction and ESG factor analysis. In *Proceedings of the 2nd Guangdong-Hong Kong-Macao greater bay area international conference on digital economy and artificial intelligence* (pp. 951–957). Association for Computing Machinery. [CrossRef]

Ferrari, A., Cini, F., & Castellano, R. (2025). Beyond traditional metrics: ESG and financial performance in innovation-driven sectors. *Journal of Financial Management, Markets and Institutions*, *13*(1), 2550001. [CrossRef]

Ferro, A., Marazzina, D., & Stocco, D. (2025). Uncovering ESG ratings: The (Im)balance of aspirational and performance features. *Corporate Social Responsibility and Environmental Management*, *32*(5), 5895–5917. [CrossRef]

Fischbach, J., Adam, M., Fischbach, V., Mendez, D., Frattini, J., Kosenkov, O., & Elahidoost, P. (2023). Automatic ESG assessment of companies by mining and evaluating media coverage data: NLP approach and tool. In J. He, T. Palpanas, X. Hu, A. Cuzzocrea, D. Dou, D. Slezak, W. Wang, A. Gruca, J. C.-W. Lin, & R. Agrawal (Eds.), *Proceedings—2023 IEEE international conference on big data, BigData 2023* (pp. 2823–2830). Institute of Electrical Electronics Engineers Inc. [CrossRef]

Friede, G. (2019). Why don't we see more action? A metasynthesis of the investor impediments to integrate environmental, social, and governance factors. *Business Strategy and the Environment*, *28*(6), 1260–1282. [CrossRef]

Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, *5*(4), 210–233. [CrossRef]

Gangi, F., Meles, A., D'Angelo, E., & Daniele, L. M. (2019). Sustainable development and corporate governance in the financial system: Are environmentally friendly banks less risky? *Corporate Social Responsibility and Environmental Management*, *26*(3), 529–547. [CrossRef]

Gaoli, H., Shan, W., Yang, Q., & Lv, X. (2025). ESG rating disagreement, takeover activities and firm performance: Double/debiased machine learning approach. *Applied Economics Letters*, 1–9. [CrossRef]

Gaurav, A., Baishnab, K., & Singh, P. (2025). Intelligent ESG portfolio optimization: A multi-objective AI-driven framework for sustainable investments in the Indian stock market. *Sustainable Futures*, *9*, 100832. [CrossRef]

Giudici, P., & Wu, L. (2025). Sustainable artificial intelligence in finance: Impact of ESG factors. *Frontiers in Artificial Intelligence*, *8*, 1566197. [CrossRef]

Guerrero, S., & Viteri, J. (2025). What are environmental, social, and governance scores measuring? The role of outcome and impact indicators in ESG scores. *Finance Research Letters*, *72*, 106529. [CrossRef]

Gyönyörová, L., Stachoň, M., & Stašek, D. (2023). ESG ratings: Relevant information or misleading clue? Evidence from the S&P global 1200. *Journal of Sustainable Finance and Investment*, *13*(2), 1075–1109. [CrossRef]

Hassan Nassar, O., Jafari, F., & Jain, C. (2025). From news to knowledge: Leveraging AI and knowledge graphs for real-time ESG insights. *Sustainability*, *17*(24), 11128. [CrossRef]

Herman, A., Oplotnik, Ž., & Jagrič, T. (2025). The impact of ESG on business performance: An empirical analysis of NASDAQ–NYSE-listed companies. *Sustainability*, *17*(21), 9683. [CrossRef]

Hoffmann, S., & D'Ecclesia, R. L. (2026). Measuring national sustainability: ESG scores from corporate data. *Socio-Economic Planning Sciences*, *104*, 102408. [CrossRef]

Hong, M., & Kim, J. (2025). AI-driven responsible supply chain management and ethical issue detection in the tourism industry. *Sustainability*, *17*(21), 9622. [CrossRef]

Hsu, W.-L., Lin, Y.-L., Lai, J.-P., Liu, Y.-H., & Pai, P.-F. (2025). Forecasting corporate financial performance using deep learning with environmental, social, and governance data. *Electronics*, *14*(3), 417. [CrossRef]

Hu, T. (2025). Impact mechanism of ESG ratings on corporate financing costs: A hybrid machine learning analysis using marginal effect and ESG rating effect in developed and developing countries. *Systems and Soft Computing*, *7*, 200318. [CrossRef]

Jain, A., Agarwal, D., Jain, D., Kumar, M., & Gupta, N. (2025). AI-driven environmental, social, and governance score predictions: A machine learning approach using SDG indicators. In *2025 international conference on sustainability, innovation & technology (ICSIT)*. IEEE. [CrossRef]

Jain, V., & Sahay, R. (2025). Leveraging ESG rating and advanced analytics for portfolio optimization. *International Journal of System Assurance Engineering and Management*. [CrossRef]

Jlifi, B., Diaye, M.-A., Guesmi, K., Ferjani, S., & Hamdi, N. (2026). *Reconciling divergence among ESG scores: A bi-level artificial intelligence based methodology for corporate governance controversies prediction*. Computational Economics. [CrossRef]

Joglekar, P., Gaikwad, V., Jadhav, P., Jain, A., & Kulkarni, O. (2025). Semi-automated platform for ethical investing. In *2025 international conference on information, implementation, and innovation in technology (I2ITCON)* (pp. 1–5). IEEE. [CrossRef]

Joshi, H., & Chauhan, R. (2020). Determinants and prediction accuracy of price multiples for South East Asia: Conventional and machine learning analysis. *Indonesian Capital Market Review*, *12*(1), 42–54. [CrossRef]

Kaleem, M., bin Jusoh, H., Raza, H., Sadiq, M., & bin Hamzah, A. (2024). A machine learning approach to predict bankruptcy in Chinese companies with ESG integration. *Pakistan Journal of Commerce and Social Sciences*, *18*(2), 335–357. [CrossRef]

Kantos, C., Joldzic, D., Mitra, G., & Hoang, K. T. (2023). Comparative analysis of NLP approaches for earnings calls. In *Handbook of alternative data in finance, volume I* (pp. 215–227). CRC Press.

Katterbauer, K., Syed, H., Genç, S., & Cleenewerck, L. (2022). Environmental compliance and financial performance of shariah-compliant enterprises—A data-driven analysis. *Revista de Gestao Social e Ambiental*, *16*(2), e03043. [CrossRef]

Khaw, T. Y., Amran, A., & Teoh, A. P. (2024). Factors influencing ESG performance: A bibliometric analysis, systematic literature review, and future research directions. *Journal of Cleaner Production*, *448*, 141430. [CrossRef]

Kim, H., & Lee, M. (2025). Unraveling the drivers of ESG performance in Chinese firms: An explainable machine-learning approach. *Systems*, *13*(7), 578. [CrossRef]

Lee, H., Hwang, J., Ryu, D., & Kim, J. (2025a). Open-source data-driven prediction of environmental, social, and governance (ESG) ratings using deep learning techniques. *Intelligent Systems in Accounting, Finance and Management*, *32*(1), e70003. [CrossRef]

Lee, H., Jung, H., Park, H., & Kim, J. (2024a). CORRECT? CORECT!: Classification of ESG Ratings with earnings call transcript. *KSII Transactions on Internet and Information Systems*, *18*(4), 1090–1100. [CrossRef]

Lee, H., Kim, J., & Jung, H. (2025b). ESG-KIBERT: A new paradigm in ESG evaluation using NLP and industry-specific customization. *Decision Support Systems*, *193*, 114440. [CrossRef]

Lee, H., Kim, J. H., & Jung, H. S. (2024b). Deep-learning-based stock market prediction incorporating ESG sentiment and technical indicators. *Scientific Reports*, *14*(1), 10262. [CrossRef]

Lee, H., Kim, J. H., & Jung, H. S. (2025c). From corporate earnings calls to social impact: Exploring ESG signals in S&P 500 ESG index companies through transformer-based models. *Journal of Cleaner Production*, *501*, 145320. [CrossRef]

Lee, H., Lee, S. H., Park, H., Kim, J. H., & Jung, H. S. (2024c). ESG2PreEM: Automated ESG grade assessment framework using pre-trained ensemble models. *Heliyon*, *10*(4), e26404. [CrossRef]

Lee, K. (2026). Bridging financial disclosures and ESG ratings: A data-driven predictive framework. *Quantitative Finance and Economics*, *10*(1), 86–107. [CrossRef]

Li, D., & Adriaens, P. (2024). Deconstruction of ESG impacts on US corporate bond pricing: The cost of capital benefits across industry sectors. *Journal of Management in Engineering*, *40*(1), 04023052. [CrossRef]

Li, J., Lian, G., & Xu, A. (2023). How do ESG affect the spillover of green innovation among peer firms? Mechanism discussion and performance study. *Journal of Business Research*, *158*, 113648. [CrossRef]

Li, Z., & Yang, Z. (2025). ESG rating disagreement and corporate total factor productivity: Inference and prediction. *Finance Research Letters*, *78*, 107127. [CrossRef]

Liang, J. (2025). Deep Multi-layer perceptron regression for ESG-driven stock return prediction in China's energy industry. In *Proceedings of the 2025 international conference on big data, artificial intelligence and digital economy* (pp. 75–80). Association for Computing Machinery. [CrossRef]

Lin, H.-Y., & Hsu, B.-W. (2023). Empirical study of ESG score prediction through machine learning—A case of non-financial companies in Taiwan. *Sustainability*, *15*(19), 14106. [CrossRef]

Lin, S. L., Lin, Y. R., & Jin, X. (2025). How can media attention reveal ESG improvement opportunities? A multi-algorithm machine learning-based approach for Taiwan's electronics industry. *The North American Journal of Economics and Finance*, *78*, 102431. [CrossRef]

Lin, S.-W., & Lin, Y.-R. (2025). Evaluating CEO hubris effects on sustainable performance in the IC design industry: An integrated dynamic network DEA framework with machine learning. *Applied Soft Computing*, *185*, 113986. [CrossRef]

Liu, F., Wang, R., & Fang, M. (2024). Mapping green innovation with machine learning: Evidence from China. *Technological Forecasting and Social Change*, *200*, 123107. [CrossRef]

Liu, S., & Yu, B. (2025). The impact of synergistic effects of ESG practices and digital innovation on enhancing organizational resilience in the transportation industry. In *Proceedings of the 2025 international conference on digital economy and information systems* (pp. 510–515). Association for Computing Machinery. [CrossRef]

Lv, W., Meng, Q., Gan, Z., & Sun, S. (2025). ESG and manager integrity: Evidence from China. *International Review of Economics and Finance*, *101*, 104133. [CrossRef]

Ma, Y., Shi, Y., Yan, Y., & Yu, Y. (2026). *Investigating corporate default risk from carbon risk and ESG perspectives using explainable artificial intelligence*. Energy and Environment. [CrossRef]

Macpherson, M., Gasperini, A., & Bosco, M. (2021, June). *Implications for artificial intelligence and ESG data*. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3863599 (accessed on 11 December 2025).

Mann, M. (2025, April). *How some sustainable investors are divesting from Tesla*. Corporate Knights. Available online: https://www.corporateknights.com/category-finance/how-some-sustainable-investors-are-getting-tesla-out-of-their-portfolios/ (accessed on 11 December 2025).

Martin-Melero, I., Gomez-Martinez, R., Medrano-Garcia, M., & Hernandez-Perlines, F. (2025). Comparison of sectorial and financial data for ESG scoring of mutual funds with machine learning. *Financial Innovation*, *11*(1), 84. [CrossRef]

Martín-Cervantes, P., & Valls Martínez, M. (2023). Unraveling the relationship between betas and ESG scores through the random forests methodology. *Risk Management*, *25*(3), 18. [CrossRef]

Martínez-Barbero, X., Cervelló-Royo, R., Jordán, J., & Ribal, J. (2026). Constructing ESG-driven portfolios using genetic algorithms based on LSTM-predicted data. *International Journal of Data Science and Analytics*, *22*(1), 69. [CrossRef]

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, *83*(2), 340–363. [CrossRef]

Momparler, A., Carmona, P., & Climent, F. (2025). Catalyzing sustainable investment: Revealing ESG power in predicting fund performance with machine learning. *Computational Economics*, *65*(3), 1617–1642. [CrossRef]

Motz, M., Uzun, S., Hariharan, A., & Weinhardt, C. (2025). Unveiling green facades: Detecting greenwashing tendencies in corporate sustainability reports. In T. X. Bui (Ed.), *Proceedings of the 58th Hawaii international conference on system sciences* (pp. 951–960). IEEE Computer Society. [CrossRef]

Nagesh, C., Jayudu, T. V. N., Rao, N. S., Hariprasad, E., Naresh, A., & Balaji, C. (2025). Predicting market volatility and risk analysis with ESG factors: A hybrid approach using LSTM and ARIMA. In *2025 9th international conference on inventive systems and control (ICISC)* (pp. 102–107). IEEE. [CrossRef]

Nampoothiri, M. (2022, June). *ESG ratings and corporate sustainabilty performance—A primer*. Available online: https://nordesg.de/en/esg-ratings-a-primer/ (accessed on 11 December 2025).

Ngee, H., Ganesh, A., Azmi, M., Tang, T., Mukred, M., Mohammed, F., & Ahmad, A. (2024). Environmental, social and governance (ESG) scores automation in global reporting initiative (GRI) with natural language processing. In *2024 7th international conference on internet applications, protocols, and services (NETAPPS)*. IEEE. [CrossRef]

Nie, R., & Wang, X. (2025). Environmental, social and governance features and financial performance using random search based extreme gradient boosting. In *2025 3rd international conference on data science and network security (ICDSNS)* (pp. 1–6). IEEE. [CrossRef]

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. [CrossRef]

Park, M. (2025). Enhancing ESG risk assessment with litigation signals: A legal-Ai hybrid approach for detecting latent risks. *Systems*, *13*(9), 783. [CrossRef]

Ragazou, K., Passas, I., Garefalakis, A., & Dimou, I. (2022). Investigating the research trends on strategic ambidexterity, agility, and open innovation in SMEs: Perceptions from bibliometric analysis. *Journal of Open Innovation: Technology, Market, and Complexity*, *8*(3), 118. [CrossRef]

Read, C. (2023). *Understanding sustainability principles and ESG policies: A multidisciplinary approach to public and corporate responses to climate change*. Springer International Publishing. [CrossRef]

Ren, X., Zeng, G., & Sun, X. (2023). The peer effect of digital transformation and corporate environmental performance: Empirical evidence from listed companies in China. *Economic Modelling*, *128*, 106515. [CrossRef]

Rosinus, M., & Lansky, J. (2025). Predictive power of ESG factors for DAX ESG 50 index forecasting using multivariate LSTM. *International Journal of Financial Studies*, *13*(3), 167. [CrossRef]

Sajadi, S., Fereydooni, A., Athari, S., & Farhadi, S. (2025). A sustainable portfolio construction model based on ESG and deep learning algorithms: Evidence from the U.S. market. *Computational Economics*, *67*, 2927–2959. [CrossRef]

Sariyer, G., Kumar Mangla, S., Chowdhury, S., Erkan Sozen, M., & Kazancoglu, Y. (2024). Predictive and prescriptive analytics for ESG performance evaluation: A case of Fortune 500 companies. *Journal of Business Research*, *181*, 114742. [CrossRef]

Sarkheil, H., Salahjou, T., & Hashemi, A. (2026). A robust hybrid machine learning framework for multidimensional ESG assessment in critical mineral supply chains: The case of cobalt mining. *Results in Engineering*, *29*, 109501. [CrossRef]

Saxena, A., Santhanavijayan, A., Shakya, H. K., Kumar, G., Balusamy, B., & Benedetto, F. (2024). Nested sentiment analysis for ESG impact: Leveraging FinBERT to predict market dynamics based on eco-friendly and non-eco-friendly product perceptions with explainable AI. *Mathematics*, *12*(21), 3332. [CrossRef]

Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M., & Leippold, M. (2024). Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, *61*, 104979. [CrossRef]

Sedghiani, N., & Doğan, B. (2026). A survey on comparative analysis of ESG performance: FinTech vs. traditional financial institutions. *Industrial Policy*, *6*(1), 15–25.

Seow, R. Y. C. (2025). Transforming ESG analytics with machine learning: A systematic literature review using TCCM framework. *Corporate Social Responsibility and Environmental Management*, *32*(6), 7358–7389. [CrossRef]

Serafeim, G. (2020). Social-impact efforts that create real value they must be woven into your strategy and differentiate your company. *Harvard Business Review*, *98*, 38–48.

Shapley, L. S. (2016). 17. A value for n-person games. In H. W. Kuhn, & A. W. Tucker (Eds.), *Contributions to the theory of games, Volume II* (pp. 307–318). Princeton University Press. [CrossRef]

Shi, J., Zhang, C., Wen, J., Zhang, Z., & Wu, T. (2024). Research on improving the key indicators of enterprise ESG rating. *Procedia Computer Science*, *242*, 912–919. [CrossRef]

Sikacz, H., Teneta, D., & Fabijańska, A. (2025). The use of feature selection methods in the materiality analysis of banks' ESG data. *Procedia Computer Science*, *270*, 2115–2124. [CrossRef]

Singla, C. (2025, May). *Why tesla got dropped from an ESG index*. OneStop ESG. Available online: https://onestopesg.com/esg-news/why-tesla-got-dropped-from-an-esg-index-1748105587456?share#article (accessed on 11 December 2025).

Sokolov, A., Caverly, K., Mostovoy, J., Fahoum, T., & Seco, L. (2021). Weak supervision and black-litterman for automated ESG portfolio construction. *Journal of Financial Data Science*, *3*(3), 129–138. [CrossRef]

Spence, M. (1978). Job market signaling. In P. Diamond, & M. Rothschild (Eds.), *Uncertainty in economics* (pp. 281–306). Academic Press. [CrossRef]

Storchevoy, M., Rogova, E., & Kochetkova, D. (2025). Quality of corporate governance assessment: A methodic approach to a regional ESG rating. *Journal of Corporate Finance Research*, *19*(4), 105–125. [CrossRef]

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review*, *20*(3), 571–610. [CrossRef]

Sultana, N., & Zeya, F. (2025). Advancing firm-specific ESG sentiment analysis: A machine learning approach to clustering, prediction, and financial performance implications. *Business Strategy & Development*, *8*(3), e70150. [CrossRef]

Svanberg, J., Ardeshiri, T., Samsten, I., O¨ hman, P., Neidermeyer, P., Rana, T., Semenova, N., & Danielson, M. (2022). Corporate governance performance ratings with machine learning. *Intelligent Systems in Accounting, Finance and Management*, *29*(1), 50–68. [CrossRef]

Taskin, D., Sariyer, G., Acar, E., & Cagli, E. (2025). Do past ESG scores efficiently predict future ESG performance? *Research in International Business and Finance*, *74*, 102706. [CrossRef]

Telukdarie, A., Mahure, H., & Sishi, M. (2024). The digitization of ESG matrix. *Procedia Computer Science*, *239*, 808–815. [CrossRef]

Terzioğlu, M., Ersoy Bozcuk, A., U¨ nal Uyar, G., Kaya, N., Tutcu, B., & Dursun, G. (2026). Machine learning classification of return on equity from sustainability reporting and corporate governance metrics: A SHAP-based explanation. *Sustainability*, *18*(1), 194. [CrossRef]

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, *14*(3), 207–222. [CrossRef]

Truyols-Pont, M., Bilbao-Terol, A., & Arenas-Parra, M. (2024). Machine learning for sustainable portfolio optimization applied to a water market. *Mathematics*, *12*(24), 3975. [CrossRef]

Tsalis, T. A., Terzaki, M., Koulouriotis, D., Tsagarakis, K. P., & Nikolaou, I. E. (2023). The nexus of United Nations' 2030 Agenda and corporate sustainability reports. *Sustainable Development*, *31*(2), 784–796. [CrossRef]

Tseng, Y.-M., Chen, C.-C., Huang, H.-H., & Chen, H.-H. (2023). DynamicESG: A dataset for dynamically unearthing ESG Ratings from news articles. In *32nd ACM international conference on information and knowledge management* (pp. 5412–5416). Association for Computing Machinery. [CrossRef]

United Nations, The Global Compact. (2004). *Who cares wins connecting the financial markets to a changing world United Nations*. Scientific Research Publishing. Available online: https://www.scirp.org/reference/referencespapers?referenceid=3033722 (accessed on 11 December 2025).

van der Heever, W., Satapathy, R., Park, J. M., & Cambria, E. (2024). Understanding public opinion towards ESG and green finance with the use of explainable artificial intelligence. *Mathematics*, *12*(19), 3119. [CrossRef]

van Duuren, E., Plantinga, A., & Scholtens, B. (2016). ESG integration and the investment management process: Fundamental investing reinvented. *Journal of Business Ethics*, *138*(3), 525–533. [CrossRef]

Vannoni, V., & Ciotti, E. (2020). ESG or not ESG? A benchmarking analysis. *International Journal of Business and Management*, *15*(8), 152. [CrossRef]

Wan, J., Wang, Y., & Wang, Y. (2025). Promoting or hindering: The impact of ESG rating differences on energy enterprises' green transformation—A causal test from double machine-learning algorithms. *Energies*, *18*(3), 464. [CrossRef]

Wang, C., Yang, S., & Zhang, Y. (2025). A semi-permeable attention network for ESG score prediction. *PeerJ Computer Science*, *11*, e3333. [CrossRef]

Wang, J., Ma, M., Dong, T., & Zhang, Z. (2023). Do ESG ratings promote corporate green innovation? A quasi-natural experiment based on SynTao Green Finance's ESG ratings. *International Review of Financial Analysis*, *87*, 102623. [CrossRef]

Wang, X. (2025). Construction and optimization of enterprise ESG performance evaluation model based on support vector machine. In S. Agaian (Ed.), *Second international conference on big data, computational intelligence, and applications (BDCIA 2024)* (Vol. 13550). SPIE. [CrossRef]

Wang, Y., Wang, Y., Chen, Z., & Shen, H. (2025). The flip side of ESG: Do ESG rating discrepancies undermine corporate energy efficiency?—-A comprehensive assessment using double machine learning. *Applied Economics Letters*, *33*(10), 1683–1693. [CrossRef]

Wang, Z., Wang, X., Liu, X., Zhang, J., Xu, J., & Ma, J. (2024). A novel stacked generalization ensemble-based hybrid SGM-BRR model for ESG score prediction. *Sustainability*, *16*(16), 6979. [CrossRef]

Wei, D., & Zeng, Y. (2025). Enhancing risk management in inclusive ESG investment portfolios in financial markets through big data analysis. *Journal of Computational Methods in Sciences and Engineering*, *25*(3), 2275–2287. [CrossRef]

Wu, L., Tang, Y., Meng, L., Zhu, Q., & Zhou, D. (2025). Navigating ESG rating divergence: Implications for labor investment efficiency and firm adaptation strategy. *Global Finance Journal*, *67*, 101141. [CrossRef]

Wu, Z., Yang, L., Fei, Y., & Wang, X. (2023). Regularization methods for sparse ESG-valued multi-period portfolio optimization with return prediction using machine learning. *Expert Systems with Applications*, *232*, 120850. [CrossRef]

Xu, J. (2024). AI in ESG for financial institutions: An industrial survey. *arXiv*. [CrossRef]

Xue, Q., Jin, Y., & Zhang, C. (2024). ESG rating results and corporate total factor productivity. *International Review of Financial Analysis*, *95*, 103381. [CrossRef]

Yang, B., Huang, J., & Chen, Y. (2025). The relationship between ESG ratings and digital technological innovation in manufacturing: Insights via dual machine learning models. *Finance Research Letters*, *71*, 106362. [CrossRef]

Yang, S.-L. (2025). Predicting corporate carbon disclosure ratings with deep learning: Evidence from GRU-LSTM vs. traditional models. *Applied Economics Letters*, 1–5. [CrossRef]

Ying, H., Zhang, X., He, T., Feng, Q., Wang, R., Yang, L., & Duan, J. (2023). A bibliometric analysis of research on heart failure comorbid with depression from 2002 to 2021. *Heliyon*, *9*(2), e13054. [CrossRef]

Zhan, S. (2023). ESG and corporate performance: A review. *SHS Web of Conferences*, *169*, 01064. [CrossRef]

Zhang, H., Lai, J., & Jie, S. (2024). Quantity and quality: The impact of environmental, social, and governance (ESG) performance on corporate green innovation. *Journal of Environmental Management*, *354*, 120272. [CrossRef]

Zhang, J., & Wang, D. (2025). The impact of controversial events on corporate resilience: The chain-mediating role of ESG and value-at-risk. *Sustainability*, *17*(24), 11032. [CrossRef]

Zhang, J., & Zhao, Z. (2026). Corporate ESG rating prediction based on XGBoost-SHAP interpretable machine learning model. *Expert Systems with Applications*, *295*, 128809. [CrossRef]

Zhang, M., Shen, Q., Zhao, Z., Wang, S., & Huang, G. Q. (2025). Optimizing ESG reporting: Innovating with E-BERT models in nature language processing. *Expert Systems with Applications*, *265*, 125931. [CrossRef]

Zhang, Y. (2026). Industry heterogeneity of environmental, social, and governance performance, impact mechanisms and trend forecasts—Evidence from big data and machine learning analysis. *Journal of Cleaner Production*, *538*, 147305. [CrossRef]

Zhang, Y., Huam, H. T., & Zhang, Z. R. (2025). Sustainable development in the tourism sector: The impact of environmental, social, and governance performance on operational efficiency—A multilevel analytical approach. *Sustainable Futures*, *10*, 100923. [CrossRef]

Zhu, J., Zhao, M., Zhang, T., & Li, R. (2025). The role of ESG score in forecasting china corporate green bond issuance: Evidence from tree-based learning models. *Sustainable Futures*, *10*, 101324. [CrossRef]

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