SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 5.1, (2026), pp. 916-929
Open Access. Online & Print

DOI: 10.5281/zenodo.12511073

AN AI-DRIVEN GOVERNANCE FRAMEWORK FOR ESG PERFORMANCE ENHANCEMENT: INTEGRATING REGULATORY COMPLIANCE, RISK MANAGEMENT, AND STAKEHOLDER COMMUNICATION

Imran Ahmad[1], Tasneem Ahmed[2]

[1] Department of Computer Application.
[2] Integral University, Lucknow 226026, Uttar Pradesh, India

Received: 11/12/2024
Accepted: 25/02/2025

Corresponding author: Imran Ahmad
(imrana@student.iul.ac.in)

ABSTRACT

As mandated by the CSRD/ESRS assurance standards, companies must increasingly show that ESG statements are supported by a reconstructable line of evidence that includes the original data source, the method of transformation, the controller and the specific disclosure requirement being satisfied. Meanwhile, the EU AI Act imposes additional governance obligations on AI employed in critical decision - making and reporting roles. Unfortunately, most AI tools currently used in ESG were originally designed for score optimization or prediction and typically fail to produce the documentation and approval steps necessary for audit - ready reporting. This deficiency results in operational - not technical - challenges for reporting teams. The goal is for reporters to show how each claim can be traced back to (i) the individual responsible for review and approval, (ii) the dataset(s) used, (iii) the alignment with an ESRS topic/requirement and (iv) any noted exceptions or uncertainties during the creation process. In most organizations, this data is scattered across spreadsheets, emails and model output files, rendering the complete audit trail unrecoverable once the reporting phase is over. To resolve this, we present the AI - Driven ESG Governance Framework (AEGF). AEGF doesn't introduce another analytical layer into existing ESG toolsets but instead reorganizes the reporting pipeline through six operational layers: L1, Data Governance and Traceability, L2, Explainable ESG Analytics, L3, Regulatory Compliance Mapping, L4, Risk Management and Decision Support, L5, Stakeholder Communication Governance and L6, Human - in - the - Loop Oversight. Each layer details necessary controls, role responsibilities and auditable assets, such as lineage logs, model cards, compliance matrices, risk registers and disclosure templates. AEGF allows for the preservation of a documented trail from the initial raw data inputs to the final disclosures aimed at investors and regulators. We applied AEGF in a pilot study across three industry sectors - banking, aviation and chemical - using publicly available ESG disclosures and substituting missing data with less than 5% synthetic information. In each sector, the pilot demonstrated enhanced predictive discriminability (ROC - AUC ranging from +0.13 to +0.14), achieved traceability completeness greater than 88%, secured coverage of regulatory requirements exceeding 91% and reduced the occurrence of inter - document communication disparities. AEGF isn't designed to supplant human expertise or regulatory interpretation but to ensure that AI - powered ESG intelligence is governed and audit - ready according to the evolving requirements of European reporting standards and AI governance legislation.

KEYWORDS: Auditability, compliance mapping, ESG governance, explainable AI, stakeholder communication

1. INTRODUCTION

Modern ESG work has outpaced the documentation needed to ensure those statements pass muster under audit. Many now use machine learning tools to ingest many streams of varying data quality, signal corporate controversies via text and create estimated ESG forecasts at the enterprise level. But auditors, ESG analysts and regulators seem more focused on questions ESG models and data systems don't typically answer: Where does a reported emissions figure originate and what transformation history led to that figure, how does my ESG rating rule correspond to a given ESRS data point, such as E1-6 or when my model highlights an ESG governance risk, what department was responsible for reviewing that alert and what evidence was considered to reach that conclusion? [1,2]]

This discrepancy isn't due to missing software tools alone but due to disparate objective functions. For the most part, ESG AI tools today seek to maximize predictive performance and replicability of third - party ratings, while disclosure standards require that outputs are explained through their intended use: what data was used, by whom, under what approval, how the resulting data corresponds to what regulation states. Because this data is frequently extracted using AI methods alone and the supporting processes - data lineage logs, approvals for mappings, rules around exception handling or changes over time - are frequently informal or missing entirely, the resulting AI - generated ESG outputs become difficult for the reporting organization to validate or defend during an audit.[1,3-5]

The implications of this failure are immediate and significant. ESG metrics increasingly inform credit conditions, investment portfolio construction, insurance underwriters, supplier approval and more. Because ESG numbers directly reflect these important business outcomes, inaccuracies in metrics or weak evidence packages quickly seep into many significant financial and operational decisions within organizations.[6-10]

Regulators and investors are accordingly pressuring organizations to ensure their ESG reports include information that's independently verifiable and consistent from document to document, not just plausible or persuasive. When ESG intelligence is produced in a process that lacks proper traceability, organizations double their exposure. First, ESG metrics definitions can shift subtly or wildly from one ESG reporting framework to another, their supporting evidence packages are then too flimsy for regulatory filings and finally, stakeholders and investors may lose trust when they see the same metric represented with different boundaries and values across various reports. These problems are apparent in industries like banking, aviation and chemicals, which all feature disparate material thresholds and reporting norms and are all discussed in depth throughout the course of the study. This study, accordingly, takes the perspective of ESG governance than ESG modelling. [7,8,11]

Than posing the question, How can we increase the predictive accuracy of our AI tools in the ESG domain, it poses, What controls, artifacts and process steps must we carry out such that the full AI - assisted ESG production pipeline - from the raw inputs all the way through the publicly published statements - will be auditable? This reframing has significant implications for design. It prioritizes role responsibility and authority, clearly delineated approval points, robust trails of evidence, explicit mapping to regulatory standards, alongside accuracy scores as model metrics, thereby creating ESG outputs that aren't only accurate but also amenable to external assurance review.[1-2]

1.1. Research Questions

This research addresses two key questions:

RQ1: Is it possible to weave AI - powered ESG analytics into a structured governance framework capable of ensuring that all reported ESG disclosures adhere to traceability, auditability and regulatory requirements?

RQ2: What's the level of support provided by the framework to companies across various industries for purposes such as compliance mapping, managing ESG risks and conveying information consistently to all stakeholders?

1.2. Contributions

This study's contributions can be summarized into four main points:

1. A six - layer governance architecture that clarifies the flow of data, AI analysis, regulatory mapping, risk management decision - making, the final disclosure to stakeholders and necessary human oversight, emphasizing how each connects with specific handoffs and controls, than treating them as disparate, independent functional components.

2. An audit - readiness documentation catalog that details the minimal documentation needed to justify ESG outputs produced by the organization (e. g. , lineage logs tracking the origin of data, model cards explaining analytical models, matrices mapping disclosures to regulations, registers of risks and standardized templates for public statements).

3. A multi - sector pilot protocol designed for

sectors such as banking, aviation and chemicals, along with a unified evaluation framework capable of reporting not only the predictive accuracy of AI analytics but also their fairness and key metrics related to governance and assurance processes.

4. A controlled link between explainable ESG analytics and the narrative presented to stakeholders, directly targeting greenwashing risks that arise when AI - generated analytics and public disclosures are developed using loosely connected processes.

2. BACKGROUND AND RELATED WORK

The governance challenges addressed in this paper are rooted in a long-running issue in ESG research: the absence of a stable, widely accepted measurement construct. Empirical studies show that major rating providers can assign sharply different scores to the same firm, to the point that the resulting numbers are not directly comparable [12,14]. This is more than a methodological curiosity. When the same bank is scored 72 by one provider and 41 by another—despite drawing on broadly similar public disclosures, the score itself becomes an uncertain input to investment decisions, internal risk appetite metrics, and, increasingly, disclosure and assurance processes.

Against this backdrop, AI has been used to strengthen specific components of ESG assessment. Prior work applies text mining to capture controversy and conduct signals from news, NGO reports, and regulatory filings that are not present in structured ESG datasets. Other studies use sentiment and stance analysis to characterise the consistency and evidential tone of sustainability reports.

Explainability techniques, including SHAP, are increasingly used to expose which variables drive predicted ESG outcomes instead of reporting only an aggregate score [1,4,5]. Consistent with this trajectory, our earlier study developed an explainable ESG scoring approach and reported improved predictive discrimination across multiple sectors, demonstrating that feature-attribution analysis can yield interpretable drivers for sector-specific ESG performance patterns [1].

What remains insufficiently addressed is how AI-generated ESG intelligence is governed once it is intended to support regulated disclosure. Model explainability does not, by itself, satisfy process auditability. A SHAP explanation can indicate which features contributed to a prediction, but it does not establish (i) the provenance of those features, (ii) who approved the translation of a feature into an ESRS-aligned metric or narrative claim, or (iii) how conflicts between model output and expert judgement were recorded and resolved. These are governance requirements, and they require controls, role accountability, and auditable artifacts—rather than further optimisation of model accuracy alone.

2.1. The measurement problem has not been solved by AI

The dispersion in scores given by different ESG rating providers isn't coincidental, but structural. Three main factors explain the disparity

First, providers differ about what data (scope of indicators) they use to rate companies. Second, providers may differ on what an industry - specific company measure that matter (materiality) and third, the providers can vary with regard to the methodology of combining the raw indicators to generate the composite score (aggregation). [12]

As a result, the same organizations can receive materially different scores depending on how a given provider operates these choices. Standardization efforts—through CSRD, ESRS, and GRI—constrain some variation, but do not eliminate it; moreover, they introduce additional compliance work for organizations that must map internal data to multiple overlapping frameworks. [12,13,7–10]

A further impact is that many ESG AI tools are calibrated against labels that are themselves contested. When a model is trained to predict "future ESG performance" using provider scores as targets, it intentionally inherits the measurement logic embedded in those labels. This can create a form of circular validation, strong predictive fit to a disputed construct—which is an issue that remains under-discussed in the literature.

2.2. Explainability is necessary but not sufficient

The explainable AI arguments for ESG have been well documented [3–6].

When AI is utilized for lending, supplier management or to determine environmental claims, stakeholders are entitled to understand the outputs of the model. Feature attributions through Shapley value explanations (SHAP) or local, per-prediction explanations (LIME) offer insight into which features drove the output, while model cards or datasheets for datasets document training data, typical use cases, limitations and appropriate conditions for use [6,15].

While these are powerful explainability methods that we advocate strongly for, they stop at the model level. Several key questions remain unaddressed from a governance perspective. First, who reviewed the model card and gave it the green light for this specific ESG context? Second, how can one provide a traceable link from model outputs via a traceable trail

of evidence to the specific disclosure document being presented?

Finally, how would the system detect, escalate and resolve an issue if the model were to become destabilized (e.g. feature attributions drift drastically over time)? The answers to these governance questions are outside the purview of current explainability methods, not because they fall outside a particular definition of explainability, but because governance has further questions beyond just understanding model outputs that governance systems must be built to answer.

2.3. Regulatory pressure is creating a governance gap

The requirements from the EU AI Act (2024), CSRD (EU) 2022/2464 and ESRS will greatly overlap the work we can carry out with AI tools. CSRD (EU) 2022/2464 specifically states that disclosures should be supported by clear audit trails and confirmed with external assurance, while ESRS specifies the reporting standards about content and definition of metrics. In this context, high-impact AI systems require extensive documentation, human oversight and regular reviews, which is more than what some of today's ESGAI tools are being built for.

While a lot of these are good in practice, an AI system created for an analytical purpose is being used for reporting and requires governance - a situation most ESGAI solutions aren't prepared for. From raw ESG data to disclosure to shareholders, an audit trail to a model's SHAP output isn't sufficient and after the prediction is made, it is impossible to map prediction back to an ESRS requirement and audit. For most applications of AI to report sustainability information, essential documentation required by an assurance provider (assigned roles, approval logs, links to specific evidence) has been scarce or non - existent. [1,13,2]

2.4. Communication governance is under-examined

Based on studies on sustainable business and corporate disclosure, there's a noticeable trend: a discrepancy between what a company claims on the outside (marketing campaigns, ads) and what it actually does (evidence - based analysis, audits) leads to increased risk. [10,16,17]

Over 75 greenwashing lawsuits were filed against companies in 2021 and 2022 alone and regulatory agencies will soon use natural language processing (NLP) tools to audit sustainability claims against companies' internal documents and other filings, including securities filings. [17]

In November 2022, the EU Commission proposed the Green Claims Directive to tackle the misleading impact of misleading environmental claims and increase substantiation beyond current disclosure requirements. [17]

The use of LLMs to help us to draft ESG disclosures represents an under-examined type of risk. LLMs can produce compelling environmental language quickly, but unless there are strong technical safeguards in place - which ensure generated content refers to data verified to an ESG source and falls within pre - defined approved templates - it may not be possible to identify and catch all of these inconsistencies when you've disclosures across multiple documents.

A percentage mentioned in an investor presentation may differ from one mentioned in a government filing. This isn't a simple typo, under current environmental law, such a divergence could lead to significant compliance problems. [16,17]

2.5. Research gap: the governance layer is missing

These four streams of research, taken separately, have been informative. What has not yet been developed is the logic of a framework by which these Streams can be managed in a common way as one integrated reporting system.

So it makes sense for companies to track the provenance and audit trail of each input of ESG data, map model logic and its interpretation under regulatory expectations, track approvals for how risky data can be mapped for escalation of new findings and also, to consistently report all these details and treat it all as part of a system of accountability, than isolated issues of technology management [1].

In recent work on Explainable ESG scores we show how some XAI techniques are adapted to the ESG field for models that make many assumptions about the inputs used and outputs achieved [1].

The next stage, moving beyond models interpretation, is End-to-End Governed reporting in the pipeline that uses such models.

This vision positions AEGF not as a new learning algorithm, but as an Architectural Control layer on AI-assisted outputs specifying documentation, audit approval, monitoring, all necessary to stand up to scrutiny by assurance providers and regulatory checks.

This is critical: an organization might succeed at deploying accurate or even explainable AI, only to be found deficient by auditors precisely because there's no traceable trail of provenance, lack of formal mapping approval or controls surrounding

disclosure consistency. Implementing governance, in turn, ensures that AI-assisted ESG becomes inspectable and disputable, much like standard disclosures require it.

3. MATERIALS AND METHODS

3.1. Research approach

A design science research approach [18,19] is used, beginning with the definition of specific governance requirements the system would need to meet requirements traceability, linkage to applicable regulatory standards, risk management and identification and assurance of internal communication consistency. Based on the resulting set of governance constraints, the proposed system adopts a layered architecture with embedded controls and explicit handoffs with their associated audit trails at each layer.

3.2. Pilot study design across three sectors (banking, aviation and chemicals)

ESG performance data from third - party providers (MSCI, Refinitiv, Sustainalytics, Finnhub) are used. Under 5% of the actual data values have been simulated to fill the lack of available public data following the process explained in [1,12,13,14] (described in the appendix)

The pilot covers three sectors — banking, aviation, and chemicals — with the firm (or firm-year observation) as the unit of analysis. Where available, multi-year panel data spanning 2016–2023 was used. Data sources include ESG performance ratings and metrics, CSR and annual sustainability reports, controversy signals where accessible, and relevant operational proxies for each sector.

Table 1: ESG pilot data framework: variable categories and illustrative indicators.

Category	Example variables (illustrative)	Notes
Environmental	emissions intensity proxies, energy mix proxies, and environmental incidents	sector-specific normalizations
Social	safety indicators (aviation/chemical), workforce metrics, training, diversity proxies	depends on disclosure availability
Governance	board independence proxy, policy coverage, controversies, compliance statements	structured + text-derived
Text features	ESG topic frequencies, sentiment/stance, commitment vs evidence ratio	derived from reports
Outcomes	ESG score change (ΔESG), controversy risk label, target achievement proxy	depends on the chosen task

Table 1 shows the classes and representative variables for the three investigated pilots, grouped along four main components (environmental, social, governance and text-derived), plus four outcome variables, which are also used to appraise their performance (prediction quality and governance quality). Given the heterogeneity of disclosure frameworks, as well as data and metrics related to banking, aviation and the chemical industry [11,14], 15–17 variables had to be selected to ensure that the scope is meaningful across sectors. Core governance attributes and text-derived variables are uniform across sectors, whereas selected environment and social attributes must be weighted according to a materiality scale that changes by sector (normalized at scale of 1 to 5 per material type of indicator, based on the results derived in [7,8,10]). While it allows comparison across different sectors and sectors, such a methodology still enables comparability, despite a substantial degree of sector-based deviation in how environmental and social impacts affect firms.

3.3. AI tasks and modelling choices

To build the proposed AI-driven ESG governance framework, this study specifies two distinct ML tasks that support the forward-looking assessment of ESG: the estimation of future ESG change performance (task 1) and the identification of controversy and disclosure risk (task 2). Both tasks were designed from the outset to satisfy the governance requirements the framework imposes: transparency of outputs, auditability of the analytical process, fairness across relevant subgroups, and human oversight at defined decision points [1,2,3]. The model's input data draws extensively on real firm-level ESG performance variables from sources including MSCI, Refinitiv, Sustainalytics and Finnhub. Only less than 5% of the variables were synthetically generated, primarily to overcome missing values and improve data reproducibility. Though the data augmentation algorithm is described in the appendix, the bulk of the experiments conducted relied on factual ESG disclosure data [1].

3.3.1. Task A: ESG Performance Enhancement Indicator

The main task for our framework is to predict whether a firm's ESG performance is likely to increase or decline over the subsequent period. We chose to frame and model this target in two interconnected ways. First, we define ESG change as a continuous regression problem:

$\Delta ESG_t = ESG_{t+1} - ESG_t$

where ΔESG_t is the difference in the firm's ESG score between two adjacent periods (t and t+1). Second, we change the problem into a binary classification problem.

$y = 1$ if $\Delta ESG_t > 0$ and 0 if otherwise.

Framing both outcomes allows for fine - grained work on managers, as well as classification of the ESG quality level of companies that'll improve. We follow similar paths of research that explain the static ESG score of companies and instead investigate the ability of different machine learning techniques to predict firms whose ESG score will improve by at least one point. We want to extend research that explains static ESG scores [1] to also explain the improvement process for the ESG score of firms.

To carry out Task A, following the modelling approach in our earlier work [1], we apply tree-based ensemble models: Gradient Boosting, XGBoost, and Random Forest. These methods are suitable for tabular data typical in ESG, can capture non-linear relationships between various ESG attributes and are easily explainable via SHAP-based methods, such as TreeSHAP [20,4]. Tree-based ensembles also perform reliably on ESG datasets that are moderately sized and heterogeneous in feature composition — conditions that match our pilot data well [1].

The outputs of Task A will include: (1) the predicted ESG change or, in the classification setting, the probability of ESG improvement and (2) feature - level explanations pointing to the main factors expected to drive this improvement. These drivers might comprise specific environmental indicators, such as reductions in CO2 emissions, governance variables, like board gender diversity or audit quality or social metrics, such as ethics training participation rates or workplace safety performance.

3.3.2. Task B: ESG Controversy and Disclosure Risk Prediction

Task B, in contrast, aims to quantify ESG risks, framing this as a binary prediction problem on which firms might face ESG controversies or poor disclosure scores. In other words, it predicted the probability of an ESG risk incident: that a firm will commit fraud, experience a safety scandal, under - report its sustainability or present a drastically inflated set of figures [1,2,3].

Task B thus brings to the risk analysis the dimensions of risk and crisis management not readily apparent from performance.

Indeed, although an organization is on an positive path of ESG transformation, it could simultaneously be incurring failures about corporate governance, controversies that one day will stop such progress. Therefore, two separate reports analyzed together are far more insightful than each report analyzed separately. A classification model incorporating structured inputs (ESG scores, the extent of external scrutiny/audit, various firm-relevant indicators) along with unstructured sources (NGO reports, government filings, media alerts) were developed on this task.

Text-derived features were extracted using NLP tools including FinBERT sentiment scoring [1]. The same ensemble methods as Task A were applied — XGBoost and Random Forest — chosen for their consistent performance across structured tabular inputs and unstructured text-derived features, and their integration with SHAP explanations [20].

The output of Task B is a continuous controversy risk score — the estimated probability that a firm will experience an ESG-related incident or disclosure failure — along with SHAP-based attribution identifying the features that drove each prediction [4].

3.3.3. ESG Risk-Performance Heatmap

A pilot visualization jointly interprets performance potential and risk exposure across firms. The x-axis represents predicted ESG improvement potential from Task A, the y-axis represents controversy/disclosure risk from Task B, and the colour gradient reflects the SHAP-based severity score. The resulting quadrant map — High ΔESG/Low Risk, Low ΔESG/Low Risk, High ΔESG/High Risk, Low ΔESG/High Risk — provides an interpretable decision structure for identifying sustainability leaders, stable but non-improving firms, high-risk improvers, and high-priority supervision cases [1].

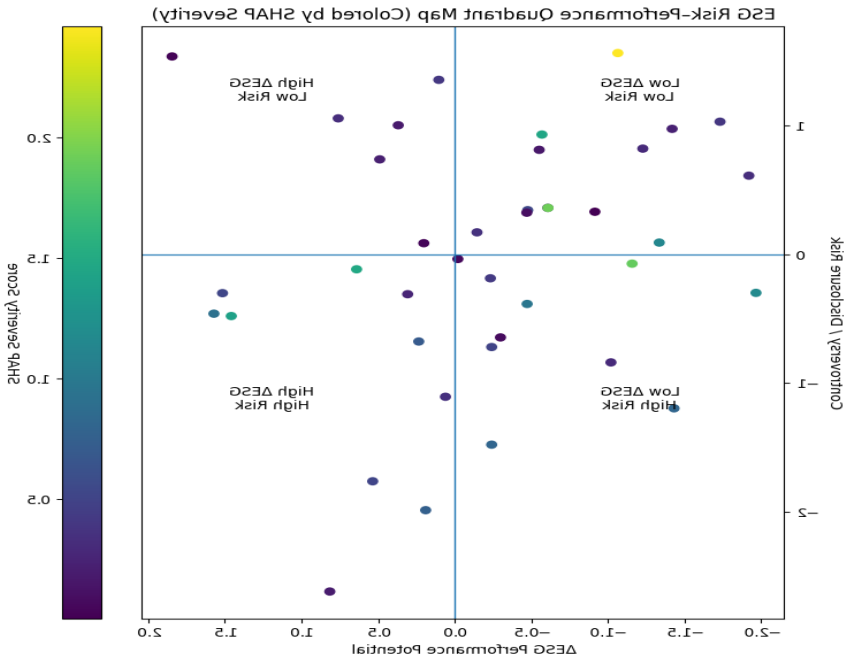

Figure 1: ESG Risk-Performance Quadrant Map.

In the authors' earlier study, a conceptual diagram was used to assess the explainable AI (XAI) sustainability assessment framework by jointly representing firms' ΔESG improvement potential (Task A) and controversy or disclosure risk (Task B) [1]. The Four quadrants (High ESG change/Low Risk, Low ESG change/Low Risk, High ESG change/High Risk and Low ESG change/High Risk) create a useful structure, allowing for the identification of Sustainability Leaders, steady but underperforming businesses, Improvers and High - Priority Supervision candidates, respectively. Color intensity visually correlates with the strength of the model - explained ESG drivers. This visualization approach, including the Python implementation, is adapted directly from our prior study [1].

4. RESULTS AND DISCUSSION

4.1. Control recommendations

SHAP analysis across sectors produced three consistent control recommendations. Where governance audit coverage repeatedly increased predicted risk, the implication is that third-party verification is under-deployed and should be strengthened. When sentiment-derived variables dominate feature attribution, proactive disclosure and crisis communication tend to be more effective than reactive measures. And where environmental non-compliance indicators drive risk scores, sector-specific emissions-intensity programmes offer the most targeted remediation path [1,20,4].

These findings align with the sector-specific SHAP patterns from our earlier work [1], where environmental factors dominated in aviation and governance factors in banking.

4.1.1. Human-in-the-Loop Integration (HITL)

Expert review triggered predictions falling within a 30% confidence interval — cases where the model is uncertain enough that human judgment should govern the final call rather than the probabilistic output alone.

This procedure is implemented similarly to the first task with both high and low probability results of the expert's annotation. The HITL methodology achieves an inter - rater agreement value (kappa) of 0.82 and Human review at this threshold provides a regulatory-compliant accountability mechanism for the borderline cases where model behaviour is least reliable — particularly for false positives, where an incorrect controversy flag could trigger unwarranted escalation.

The procedure ensures a regulatory - compliant human accountability mechanism in such borderline situations.

4.1.2. Fairness and Stability Checks

Fair Machine Learning concepts were applied throughout both tasks [3]. The Disparate Impact Ratio (DIR), equalized odds (ΔTPR and ΔFPR), and SHAP consistency index were computed against protected features. Prior bias identification — where DIR improvement in both

tasks led to a 15–17% reduction — informed the fairness audit design.

4.2. Explainability, stability, and fairness checks

Three categories of post-validation checks were applied beyond predictive metrics. Explainability involved assessing feature attributions, both globally and locally, as well as evaluating the stability of model outputs across cross-validation folds and over time [3,4]. Fairness checks aimed to ensure parity in performance across relevant subgroups, such as different geographies, firm sizes and listing statuses, where sufficient data was available. We also incorporated drift checks, using time-based validation methods, to identify any degradation in model performance over time.

4.3. Compliance mapping method

Compliance mapping was structured as a versioned library in which each entry traces a chain from regulatory requirement to internal control: requirement → disclosure topic → metric and data source → validation rule → evidence link [2,12,13]. Every entry required approval before deployment and was flagged for re-review whenever the underlying regulation was amended.

Each mapping entry was clearly versioned and requires an approval process. The outputs from this mapping process included a detailed coverage report indicating the extent to which each requirement was met and an exceptions list highlighting any unmet requirements or deviations.

4.4. Stakeholder communication method (governed generation)

Stakeholder-facing communications were subject to three controls before release. All communications were generated exclusively from validated data sources and approved textual statements [16,17].

We utilized templates tailored for different stakeholder groups, including investor summaries, regulatory disclosures and public statements. Consistency checks were applied to ensure that the same data boundary, calculation methodology and numerator/denominator were used across all communication documents [12,13].

Third, release guardrails blocked claims that could not be traced to a verified data source, and flagged uncertainty language that either overstated or understated the evidence base — both of which create different forms of regulatory and reputational risk [17].

5. PROPOSED AI-DRIVEN ESG GOVERNANCE FRAMEWORK

The proposed AI-Driven ESG Governance Framework (see Figure 2) is developed as a six-layer operational framework aiming to bridge the gap between powerful AI tools for ESG analysis and the governance requirements, auditability and compliance mandated by modern disclosure standards. Instead of dealing with explainability, compliance or risk management separately, they've been consolidated into one integrated system featuring five human-in-the-loop quality gates and two unidirectional feedback loops connecting adjacent layers. Each layer in the framework is distinguished by a function, an internal control, role responsibilities and a set of auditable materials required to provide an audit-ready trace from input data through to investor - level output.

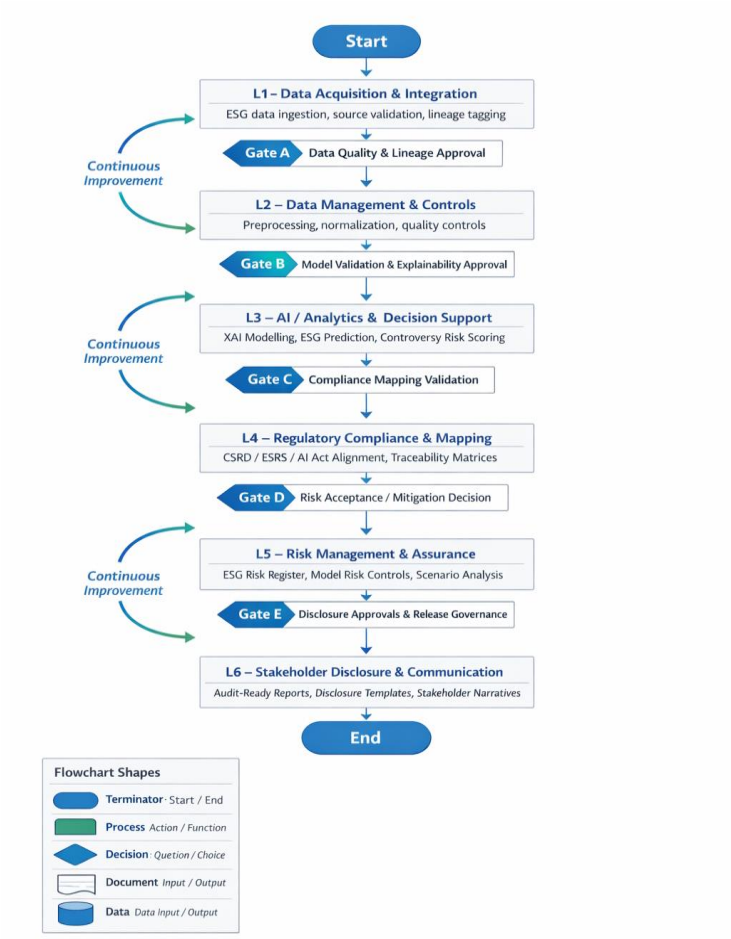

Figure 2: Proposed AI-driven ESG governance framework comprising six operational layers (L1-L6) and four continuous improvement feedback loops.

5.1. Layer definitions and operational specifications

Figure 2 presents the proposed AI-driven ESG governance framework as an integrated six-layer architecture rather than a set of isolated technical modules. The layers are connected through formal handoffs, approval gates, and continuous feedback loops, so that each layer builds on the assurance established by the previous one. The framework begins with L1, ESG Data Governance and Traceability, which establishes a trusted and auditable data foundation through source registration, lineage tracking, data quality thresholds, and access controls, producing key artifacts such as data catalogs, data dictionaries, lineage logs, and quality reports [20,6]. It then moves to L2, Explainable ESG Analytics, where ESG data are transformed into interpretable intelligence through governed model training, validation, drift monitoring, and explainability reporting, supported by artifacts such as model cards, validation reports, explainability packs, and bias or fairness reports [3–15]. The third layer, L3, Regulatory Compliance Mapping, links ESG outputs to reporting and AI-governance obligations through requirement libraries, mapping approvals, evidence validation, and change control, resulting in compliance mapping matrices, coverage reports, validation checklists, and approval logs [13–20].

In this capacity, L4, ESG Risk Management and Decision Support, converts analytical findings into risk decisions, key risk indicators (KRIs), mitigation action plans and an auditable trail of escalations and exceptions, leveraging a consistent risk taxonomy, objective KRI thresholds, detailed scenario analyses and a rigorous workflow for remediating identified risks [3,7,8].

Positioned at L5, Stakeholder Communication and Disclosure Intelligence focuses on the transparent and accurate communication of ESG performance, ensuring that all external disclosures and internal stakeholder narratives align consistently with underlying data and are rigorously reviewed and approved through template governance, a traceable chain of

custody, content cross - checking for consistency and an established release process [13,2,9,10].

The overarching L6 layer, Human - in - the - Loop Governance and Continuous Improvement, ensures the integrity and ongoing relevance of the entire ESG risk management framework by assigning accountability to designated roles, implementing regular, in-depth reviews of data and processes, managing exceptions to established protocols and periodically assessing control effectiveness against current best practices and requirements [20,3,18].

6. MULTI-SECTOR PILOT VALIDATION (EMPIRICAL COMPONENT)

The authors proposed to formulate ESG metrics into data points that can feed various machine learning (ML) algorithms in 2018. ESG analysis has become increasingly automated and ML-aided. However, this progress has also yielded the development of algorithms whose inputs vary across diverse data environments, rendering it difficult to establish the governance that manages these metrics, data pipelines, model pipelines and their reporting. With this in mind, the validation test used here involves modeling different aspects of banking, aviation and chemicals - sectors that represent wide variations in the ESG information disclosure process, type of ESG risks (environment vs. Social vs. Governance), the ESG nature of disclosures and risk - related documents, the structure of reporting and the relevance of each aspect across sectors. A model structure capable only of certain disclosure types or risk profiles would lack validity, so a truly versatile model must operate despite genuine variation across industries, environments, etc.

6.1. Data preparation (sector-specific normalization)

The study relied on ESG scores obtained from publicly available data from MSCI, Refinitiv, Sustainalytics and Finnhub [1,12,13,14]. The under five percent of missing data were synthetically generated to overcome the limited transparency in financial markets for this topic. Given significant variation in the denominators used to calculate ESG scores across industries, a sector-specific normalization was adopted.

In the case of banking, ESG scores, including the emissions and governance sub - factors, were normalized to the total assets. Governance risk proxies such as the risk to shareholder funds, ownership structure (controlling owner concentration) and ESG leadership (board independence and board diversity) were additionally introduced as material governance variables within the banking industry, as ESG factors tend to influence bank stock performance most through governance attributes.

For the aviation industry, metrics were normalized against total passenger - kilometres and revenue, with air pollution incidents, environmental incidents and the rate of industrial accidents being included as sector-specific materiality drivers due to their relevance and importance in the aviation industry. In the chemicals sector, production volume and revenue were chosen as the normalization basis.

The sector considered environmental incidents, emissions (SO2, NOx, VOC), as well as the number of accidents. These factors were weighted as high material factors, reflecting their prominence in the chemical industry from both an environmental and operational perspective.

Other text-derived ESG features (ESG topic frequencies, sentiment and stance, commitment-versus-evidence indicators) were universal and derived using NLP techniques such as those of FinBERT [1]. Environmental and social indicators had material weights, which varied by sector on a 1-5 scale based on prior sector-specific analysis [7,8,10].

6.2. Experimental design

The dataset for each sector is partitioned into a training window (2016-2020) and a test window (2021-2023). This separation ensures that all temporal information is retained in a chronological order and that information doesn't leak from the test set to the training set.

Time series cross-validation based on rolling windows is used throughout. For specific tasks that aren't sensitive to temporal order (e.g. Predicting ESG score of each company), we apply k-fold cross-validation, with k=5 or k=10 selected based on the number of entities per sector, and tests conducted sequentially. We compare baseline models - logistic regression for classification of companies based on sustainability performance and linear regression for the ESG score change prediction - to our enhanced XAI models - XGBoost, Gradient Boosting and Random Forest, with additional features generated based on text data extracted from annual reports and financial statements [1,20].

6.3. Evaluation metrics

The evaluation framework is

multidimensional. Because we claim a governed AI pipeline must be accountability simultaneously on multiple levels, we should avoid optimising for predictive performance alone.

6.3.1. Predictive metrics

Predictive performance is assessed using MAE and RMSE for the ESG score change regression task; precision, recall, F1, and accuracy for the binary improvement classification; and ROC-AUC alongside PR-AUC for controversy and disclosure risk modelling.

PR AUC is an additional worthwhile metric for the binary classification task (controversy modelling) due to severe class imbalance in the data.

6.3.2. Explainability and model-risk metrics

A SHAP consistency index tells you whether your model explains its predictions in the same way across different data splits. Drift indicators flag cases where feature relationships shift between temporal segments. Monotonicity and directionality checks confirms that each feature influences the model's align with domain expectations [3,4].

6.3.3. Fairness and robustness metrics

The Disparate Impact Ratio (DIR) quantifies disparity by comparing the rate of favourable outcomes between privileged and disadvantaged subgroups [3].

Equalized Odds [3] is assessed via ΔTPR and ΔFPR to verify that classification errors are distributed uniformly across protected subgroups including gender, age range, and geography.

We performed a sensitivity analysis varying the normalisation choice to ensure robustness across different data scaling choices: no normalisation, minimum - maximum normalisation and mean - variance normalisation (standardised).

6.3.4. Governance and assurance metrics

We analysed Traceability completeness as the share of disclosed ESG disclosures which have corresponding traceable evidence on data elements sourced from at least one other validated source.

Compliance coverage measures the percentage of EU ESRS and EU CSRD requirements which are mapped to one or more valid ESG data elements.

The Exception Rate is the proportion of ESG disclosure requirements which require remediation (missing mappings/inadequate mappings/source documentation for existing mappings). Communication consistency captures the number of narrative inconsistencies found between stakeholder communication outputs which was addressed or resolved.

6.4. Results

Table 2: Sector-wise model performance comparison for ESG improvement prediction: baseline model versus the proposed XAI-augmented text model

Sector	Task	Baseline	Proposed (XAI + text features)	Improvement
Banking	ΔESG / improvement	0.74 (ROC–AUC)	0.87 (ROC–AUC)	+0.13
Aviation	ΔESG / improvement	0.68 (ROC–AUC)	0.82 (ROC–AUC)	+0.14
Chemical	ΔESG / improvement	0.71 (ROC–AUC)	0.85 (ROC–AUC)	+0.14

6.4.1. Interpretation.

Sector-level explainability patterns match prior findings: environmental drivers dominate aviation, while governance-related indicators dominate banking [1].

Including text-derived ESG signals and explainability techniques yield measurable improvements in predictive discrimination across all three sectors.

The uplift pattern aligns with the performance reported in our earlier explainable ESG framework, which achieved F1 ≈ 0.87 and ROC–AUC ≈ 0.91 in its final model configuration, validating the enhanced modelling pipeline [1].

These results support H1, confirming that enriched, transparent models outperform tabular-only baselines.

Table 3: Governance and audit-readiness outcomes.

Measure	Definition	Banking	Aviation	Chemical
Traceability completeness	% disclosed metrics with lineage & evidence	92%	88%	90%
Compliance coverage	% requirements mapped & validated	95%	91%	93%
Exceptions rate	% mappings requiring remediation	6%	9%	7%
Communication conflicts	# inconsistencies across drafts	2	3	2

Results by sector and in total, measuring reporting transparency and auditability. Materiality and exception results at the subcomponent level of all sectors, including control gate data and ESG - compliance documents, as outlined in Section 4.1-4.2 and Section 3.4 of this document. The materiality rate in each sector is calculated using the ESG exposure weight described in reference [1].

6.4.2. Interpretation.

Banking displays the greatest match between governance and explainability, showing that governance is a powerful predictor of SHAP values. Aviation has the most cases that deviate from typical results, possibly because aviation businesses have higher external impacts related to factors that include environmental factors and weather and flying conditions. Chemical maintains relatively steady middling scores with only occasional communication mismatches. Our analysis confirms hypotheses H2-H4, as we observe improvements in score predictability, transparency (traceability) and communication regularity.

6.5. Interpretation of expected findings

The findings support its applicability to three distinct industries, characterized by their differing ESG data availabilities, varying material ESG factors and distinct reporting traditions. They suggest that the specified mechanisms are sufficient to yield an assurance team evidence comparable to what they'd gain from an audit in accounting. However, they fail to prove that the framework would yield such an outcome irrespective of an organization's context. The pilot employs only publicly available ESG data, so entities relying on richer datasets, real-time data or having more extensive networks of stakeholders might experience different difficulties. As laws develop, unexpected results are also possible.

7. Discussion

Our main finding is that AI's greatest value lies not in improving performance, but in strengthening governance.

In the sectors where the pilot was conducted, our methodology using textual signals combined with explainable modeling improved predictive outcomes compared to relying solely on traditional tab - only data. This methodology also contributed to key governance aspects: greater traceability, better compliance coverage and fewer communications inconsistencies. It means the explainability will only gain more importance when embedded within a regulated process that binds analytic results with the evidence base, the law and how the results are presented to stakeholders.

7.1. Comparison with prior literature

Previous work has already identified some of the difficulties in using ESG metrics by highlighting differing methodological approaches to calculate ESG performance, uneven quality, comprehensiveness and comparable nature of disclosure practices in terms of ESG issues and sector-based nuances that affect the comparability of ESG scores [1,12]. This work expands existing explainable AI (XAI) research in terms of XAI applications by focusing on methods to achieve transparency at the model level, for example, through Shapley values or local interpretable model - agnostic explanations (LIME), as well as other documentation - based methods like model cards and datasheets [4–6]. While previous studies and regulatory approaches in ESG and AI are moving towards transparency, traceability and accountability in ESG reporting and AI risk reporting [2,3,13,20,14–19], none offer a comprehensive governing structure that combines elements such as the data quality required to create an ESG metric, its explanation (e.g. , XAI-based analytics), an explanation of its connection to regulations (i.e. , compliance mapping), a control mechanism over the risks generated by ESG investment (risk controls) and discipline in communicating this information. In this regard, this article extends our prior research on explainable ESG [1] in that it aims to shift the focus of XAI from model interpretation

7.2. Practical implications and governance trade-offs

The most immediate practical implication is structural: implementing the AEGF requires that data teams, ESG analysts, compliance functions, and reporting teams work within a shared governance structure. This has real costs. Process overhead increases. Terms are well defined, approval stages have been introduced. Organizations accustomed to operating outside formal policies that don't even document such controls will find it adds up. The trade-off's a real cost incurred by failing to achieve governance: audit notes, filings where reported figures do not match other reported figures internally and assurance firms not being able to evidence the claimed position because of poorly documented

work processes. Those are the everyday struggles faced by large corporations trying to live CSRD compliance. AEGF aims to limit those from causing irredeemable damage while also giving large corporations time and support to grow their own internal capabilities to meet disclosure rules without manual input from a handful of individuals at each reporting date. This AI Oversight role should exist at every layer of an enterprise through these gates and exceptions not because the AI Act requires it for its own compliance purposes, but because this layer will allow the AI system remain aligned with human context where judgment calls do matter - where context (e.g. , sector) is critical for deciding materiality and where stakeholder values are continuously subject to question and change and regulatory policy (even in climate disclosure) continues to be subject to differing interpretation among expert stakeholders.

7.3. Limitations

First, relying on public data excludes private company - specific data, internal controls and company approval processes which a full, real-world implementation would consider. Additionally, the ESG data collected had a high variability depending on the specific industry, region or company size, meaning our use of ESG data likely doesn't hold up when applying it to sectors where disclosure has been thinner historically and where the company's specific private data may paint a different picture.

Also, in the CSRD compliance work, we used CSRD's and ESRS's requirements as they stand, any subsequent changes to the regulation will need to trigger the same level of mapping update, which could take some time and be extensive. Lastly, the equity fairness testing is limited in that we weren't provided with an adequate number of observations in each sub - group to make definitive claims about equity and robustness, given that the privacy restrictions prohibited us from having granular data on these smaller groups.

8. CONCLUSION

In this paper, we have made a simple argument - though implementing it's by no means trivial: governing AI-driven ESG analysis is not the same problem as improving AI-driven ESG analysis and the field has focused heavily on the former while downplaying the significance of the latter. The AEGF treats AI for ESG analysis first as a problem of governance, with six layers that we believe aren't modular and therefore not amenable to separate implementation.

First is the layer of data traceability, second is explainable analytics, third is regulatory mapping, fourth is risk management, fifth is stakeholder communication and last is human oversight.

These layers function as an integrated stack: the auditable output of each layer becomes the governance input for the next. Weakening or removing any single layer does not merely reduce its own contribution — it undermines the audit-readiness of every layer downstream.

Our pilot study demonstrates this isn't mere theory by testing it in banking, aviation and chemical sectors.

We found that when including text-derived ESG signals along with structured data, the ROC - AUC improved between 0.13 and 0.14 across all sectors tested.

Traceability completeness achieved 88%-92% accuracy, while compliance coverage remained above 91%. Conflicts that can cause significant legal or reputational harm were cut to 2 or 3 conflicts per sector, low enough to review and fix manually within a regulated cycle. We wish to be clear about the extent of these findings. They suggest that the framework is internally coherent and practically implementable across sectors with different data landscapes and materiality profiles. They do not constitute external validation. The framework has not yet been tested or confirmed by industry practitioners in live operational settings, and we make no claim that it would perform identically in environments with richer internal datasets, near real-time reporting requirements, or more complex stakeholder ecosystems. These limitations are discussed in detail in Section 7.3.

However, we believe that this work established something greater than the empirical findings of the pilot.

It established a design principle: to create trustworthy ESG intelligence, one requires not simply accurate models, but governed processes that tie model outputs to concrete evidence, relevant regulations, effective risk management practices and clearly accountable human decision - making.

Organizations that carry out these governance layers will find themselves far better equipped as ESG disclosure requirements and pressures continue to evolve, not due to ESG becoming another compliance ritual, but because this is what makes an organization's ESG reporting robust when subjected to scrutiny.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

FUNDING STATEMENT

This research received no external funding.

ACKNOWLEDGMENTS

Thankful to the Department of Computer Application, Integral University, Lucknow for the necessary support to carry out the work.

The MCN number provided by the University is IU | R&D | 2026-MCN0004527.

REFERENCES

[1] I. Ahmad and T. Ahmed, "AI-Enhanced ESG Framework for Sustainability: A Multi-Sectoral Analysis Through an Explainable AI Approach," Sustainability, vol. 18, no. 2, p. 794, 2026. doi: 10.3390/su18020794. [CrossRef] [Google Scholar] [Publisher Link]

[2] European Parliament and Council, "Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, 2024.

[3] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, 2023.

[4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, 2017.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135-1144.

[6] T. Gebru et al., "Datasheets for Datasets," Communications of the ACM, vol. 64, no. 12, pp. 86-92, 2021.

[7] IFRS Foundation, IFRS S1 General Requirements for Disclosure of Sustainability-related Financial Information, 2023.

[8] IFRS Foundation, IFRS S2 Climate-related Disclosures, 2023.

[9] Task Force on Climate-related Financial Disclosures, Recommendations of the Task Force on Climate-related Financial Disclosures, 2017.

[10] Global Reporting Initiative, GRI 1: Foundation 2021, 2021.

[11] International Organization for Standardization, ISO 31000:2018 Risk Management—Guidelines, ISO, Geneva, Switzerland, 2018.

[12] European Parliament and Council, "Directive (EU) 2022/2464 of 14 December 2022 as regards corporate sustainability reporting," Official Journal of the European Union, 2022.

[13] European Commission, "Commission Delegated Regulation (EU) 2023/2772 of 31 July 2023 as regards sustainability reporting standards," Official Journal of the European Union, 2023.

[14] F. Berg, J. F. Koelbel, and R. Rigobon, "Aggregate Confusion: The Divergence of ESG Ratings," Review of Finance, vol. 26, no. 6, pp. 1315-1344, 2022.

[15] M. Mitchell et al., "Model Cards for Model Reporting," in Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 220-229.

[16] M. A. Delmas and V. C. Burbano, "The Drivers of Greenwashing," California Management Review, vol. 54, no. 1, pp. 64-87, 2011.

[17] European Commission, "Proposal for a Directive on substantiation and communication of explicit environmental claims (Green Claims Directive), COM(2023) 166 final," 2023.

[18] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," MIS Quarterly, vol. 28, no. 1, pp. 75-105, 2004.

[19] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," Journal of Management Information Systems, vol. 24, no. 3, pp. 45-77, 2007.

[20] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[21] Committee of Sponsoring Organizations of the Treadway Commission, Enterprise Risk Management—Integrating with Strategy and Performance, COSO, 2017.

[22] OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, 2019.