SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

GreenAudit-GNN: Sector-Stratified Benchmarking and Dual-Method Explainability Validation for ESG Anomaly Detection

Ram Krishna[1], Dr. Muhammad Kalamuddin Ahamad[2]

[1]PhD Scholar, Department of Computer Application, Integral University, Lucknow, Uttar Pradesh, India. ORCID: 0009-0002-8653-7687, E-mail: rkrishna@student.iul.ac.in
[2]Associate Professor (Supervisor), Department of Computer Application, Integral University, Lucknow, Uttar Pradesh, India, ORCID: 0000-0003-3043-0549, E-mail: mohdkalam@iul.ac.in

Abstract
Mandatory ESG disclosure regimes such as India's Business Responsibility and Sustainability Reporting mandate produce a disclosure corpus that manual audit cannot review at scale. Two earlier papers in this programme reviewed the AI-for-ESG literature and proposed GreenAudit-GNN, a seven-stage framework combining NLP and a graph neural network for explainable ESG anomaly detection, piloted on 58 audited claim-KPI pairs from 40 companies. That pilot trained a classifier and benchmarked it against four unsupervised baselines, but left two things only partly addressed: cross-sector performance comparison, and validating the framework's explanations for auditor and regulator use. This paper closes both gaps on the same audited dataset of 38 companies. We retain every model's per-company prediction and group the 23 disclosed sectors into ten broader industry buckets, giving a descriptive cross-sector comparison of weak-label prevalence and model flag rates. We then implement and cross-check two independent local-explanation methods for the trained classifier: an exact closed-form linear-SHAP decomposition and an independently reimplemented LIME-style local surrogate. The two agree on direction for every company and correlate strongly in magnitude across the two model features. Finally, we introduce a gradient-by-input attribution that splits each company's anomaly score into its own reported trend versus its sector-peer average, the programme's first auditor-readable, evidence-linked explanation. Every result follows the same honesty discipline as the earlier papers: sector buckets of one to seven companies support no inferential claim, no human auditor or regulator has yet reviewed these explanations, and every number is a real, seeded, reproducible computation on audited disclosure data.

Keywords: ESG disclosure; anomaly detection; sector benchmarking; explainable AI; SHAP; LIME; graph neural network; BRSR.

1. Introduction

India's SEBI-mandated BRSR framework requires standardised, text-and-numeric ESG disclosure from the 1,000 largest listed companies every year, and third-party assurance requirements are now extending to a growing share of those filers [26]. Manual review of this disclosure corpus does not scale, which motivates an AI-based approach to flagging inconsistencies between a company's narrative claims and its own reported numbers.

This paper is the third in a research programme building toward that goal. The first paper in the series is a structured literature review of twenty-one AI-for-ESG sources across nine themes, surveying existing ESG reporting techniques and AI-based inconsistency-detection approaches [17]. The second paper designed and piloted GreenAudit-GNN, a seven-stage cross-modal framework combining NLP-based claim extraction with a graph neural network for anomaly detection; its first five stages, namely narrative and KPI extraction, cross-modal alignment scoring, heterogeneous graph construction, GraphSAGE-style embedding, and an anomaly-detection ensemble, were run end-to-end as a computational feasibility pilot on fifty-eight independently re-verified claim-KPI pairs drawn from forty companies' BRSR filings across twenty-three sectors [18]. That pilot did not stop at a design description: it developed and sensitivity-tested a heuristic weak-label rule, trained a logistic-regression classifier head over Leave-One-Out Cross-Validation, and benchmarked that classifier against four unsupervised anomaly-detection baselines plus a graph-versus-no-graph ablation.

Two gaps remained once that pilot wrapped up. The earlier pilot reported only dataset-wide metrics, with no sector-disaggregated analysis at all, even though the audited dataset already spans twenty-three distinct sectors -- so a comparison of ESG performance and anomaly patterns across sectors was still missing. Its Stage 6 explainability layer, meanwhile, had been specified on paper, with SHAP, LIME and graph-attention explanations named as the intended mechanism, but none of it had actually been run against the trained

model's real outputs -- so the framework's interpretability for auditor and regulator use was still unvalidated. This paper, the third in the series, takes on both gaps directly, using the same real, audited data already assembled for the earlier pilot rather than collecting anything new or reopening questions the companion paper's verification process already settled [18]. The programme's two conference presentations so far [15], [16] previewed the anomaly-detection framing that this paper now benchmarks and explains in full.

This paper's contribution comes in three parts:

(a) we keep every model's per-company prediction, not just the aggregate metrics the earlier pilot reported, and group the dataset's twenty-three fine-grained sector labels into ten broader industry buckets, giving this research programme its first cross-sector comparison, reported honestly as a descriptive, exploratory exercise given that bucket sizes run from one company to seven;

(b) we implement two independent explanation methods for the trained classifier -- an exact, closed-form linear-SHAP decomposition, and an independently reimplemented LIME-style local surrogate -- and check them against each other, a recognised technical check on explanation stability [6], not a claim that the explanations have been validated by a human;

(c) we introduce a Gradient-by-Input attribution that breaks the graph-based model's decision down into the company's own reported trend versus its sector-peer average -- as far as we can tell, the first auditor-readable, evidence-linked explanation this framework has actually produced rather than merely designed.

We keep the same disclosure discipline throughout as in the two earlier papers: every number reported below comes from a real, seeded, reproducible run of the audited N=38 pipeline; every limitation that pipeline already carried, the heuristic non-ground-truth weak label, the label/feature leakage, and the wide, unreported confidence intervals at this sample size, still applies here unchanged; and nowhere in this paper do we claim that a human auditor has validated these explanations.

The remainder of this paper is organised as follows. Section 2 reviews related work in ESG anomaly detection and explainable AI. Section 3 recaps the audited dataset and the prior pipeline this paper builds on. Section 4 describes the methodology for both the sector-stratified benchmarking and the dual-method explainability validation. Section 5 reports the results. Section 6 discusses their implications. Section 7 states this paper's limitations, Section 8 its ethical and regulatory considerations, and Section 9 concludes with directions for future work.

2. Related Work

2.1 Anomaly Detection Benchmarking in ESG and Adjacent Domains

A recent tripartite systematic review of machine learning in sustainable finance, covering 127 peer-reviewed studies, finds that ESG ratings still weight low-cost aspirational disclosure heavily relative to costly performance evidence, and that reported model accuracy frequently reflects target-proximal reconstruction rather than transferable, out-of-time forecasting -- a caution this paper takes seriously by reporting Leave-One-Out Cross-Validation metrics with their limitations stated rather than headline accuracy alone [9]. Two 2026 systems illustrate the direction the wider field is moving in: a deep-learning framework integrating multi-source signals (financial-report text, attention-weighted across sources) for greenwashing detection [29], and a natural-language-processing pipeline that constructs a Greenwashing Severity Index by comparing corporate disclosure against external media coverage across 204 Central and Eastern European firms, finding moderate but widespread greenwashing concentrated by firm size and public accountability [8]. A 2025 ACL paper addresses a closely related robustness problem: an aspect-action dataset that links sustainability claims to concrete implementation status, showing that fine-tuned supervised models generalise across unseen ESG categories better than zero-shot large language models [21]. None of these three systems reports a sector-disaggregated anomaly-detection benchmark; each evaluates pooled, dataset-wide performance, the same gap this paper's Section 5 addresses for GreenAudit-GNN's own pilot.

The four unsupervised anomaly-detection baselines this paper carries forward from the prior pilot are themselves standard, well-established benchmarks rather than novel contributions of this programme: Isolation Forest isolates anomalies via random recursive partitioning, exploiting the fact that anomalies require fewer splits to isolate [19]; One-Class SVM learns a boundary around the dense region of normal data in a kernel-induced feature space [25]; and Local Outlier Factor scores each point by the local density deviation relative to its neighbours [7]. The graph-relational mechanism underlying GreenAudit-GNN's embeddings is GraphSAGE, which learns node representations by sampling and aggregating features from a node's local neighbourhood rather than requiring the full graph at training time [11] -- in this pilot's case, aggregating each company's reported metrics with its sector peers'.

2.2 Explainable AI: SHAP and LIME

SHAP (SHapley Additive exPlanations) unifies a family of additive feature-attribution methods under a single game-theoretic framework grounded in Shapley values from cooperative game theory, and shows that for a linear model with independent features the exact Shapley value of feature i reduces to a closed form -- the feature's coefficient multiplied by its deviation from the training-set mean [20]. LIME (Local Interpretable Model-agnostic Explanations) instead treats the underlying model as a black box and explains any single prediction by fitting an interpretable local surrogate -- typically a sparse linear model -- to perturbed samples drawn around that instance and weighted by proximity [22]. A 2025 methodological perspective compares the two directly, noting that SHAP offers stronger theoretical guarantees (local accuracy, consistency) while LIME is comparatively cheap to compute and model-agnostic by construction, and recommends using both together as a cross-check rather than treating either as definitive on its own [24] -- precisely the design this paper adopts in Section 4.3. Because agreement between independently derived explanation methods is itself only a proxy for trustworthiness, not a guarantee of it, this paper follows the caution that explanation methods can be locally unstable to small input perturbations even when they agree on a given instance [6], and reports the SHAP-LIME agreement check in Section 5.4 as an internal-consistency result rather than a validated-trust claim.

2.3 XAI for Financial and Regulatory Trust

A 2025 CFA Institute Research and Policy Center report argues that explainable AI is becoming indispensable for financial institutions seeking regulatory compliance and stakeholder trust, transforming opaque algorithmic decisions into transparent and auditable ones -- but the same report stresses that human judgement and organisational oversight remain indispensable alongside any automated explanation, and that XAI tools alone cannot substitute for that oversight [28]. This caution echoes a broader methodological argument that post-hoc explanation of a black-box model is not the same guarantee as using an inherently interpretable model in the first place, particularly for high-stakes decisions where an unfaithful explanation can mislead a human reviewer into misplaced confidence [23]. This paper's own standard for its explainability work is explicit: an explanation is only worth reporting if it is credible enough for actual use by auditors and regulators -- language this paper reads carefully rather than treats as automatically satisfied by running SHAP or LIME once. Section 6 returns to this point directly: nothing in this paper constitutes a human-auditor or regulator validation study, and the explanations reported here are offered as a necessary technical foundation for such a study, not a substitute for it.

2.4 Methodological Precedents from Adjacent Institutional Research

Several of the specific technical choices in this paper's methodology have precedent in methodologically adjacent work carried out within the authors' own institution, and it is worth naming that precedent honestly rather than presenting each choice as arising in isolation. The macro-sector grouping used in Section 4.1 to collapse twenty-three fine-grained sector labels into ten broader buckets follows the same general logic -- reducing a high-dimensional, sparsely populated categorical variable into coarser groups before any comparison is attempted -- used in a cluster-analysis study of a real dataset via principal component analysis and k-means clustering with SPSS [3], although that study's domain (general-purpose clustering) and this paper's domain (ESG sector taxonomy) are unrelated. The decision to benchmark multiple models side by side rather than reporting a single "best" model, carried through from the companion paper into Sections 4.2 and 5.1-5.2 of this one, mirrors the comparative-evaluation design of a study that benchmarked several machine learning techniques against each other for predicting Type 2 diabetes mellitus [4]; again, the clinical domain there is unrelated to ESG disclosure, and the parallel is methodological only -- multiple models evaluated under one consistent protocol rather than the design or performance of any single model. Closer to this paper's own explainability work, a hybrid machine learning framework for web content mining that combined more than one algorithmic technique within a single pipeline [12] offers a loose structural precedent for this paper's own use of two independent explanation methods (SHAP and LIME) side by side in Section 4.3, though the two studies solve unrelated problems. The semantic-similarity scoring approach used to assess similarity between software requirements [10] is methodologically closer to this research programme's own cross-modal alignment step (Stage 3 of the GreenAudit-GNN pipeline, described in [18]), which likewise scores the semantic consistency between a textual claim and a numeric KPI value, even though this present paper does not re-run that alignment step directly. Finally, a broader discussion of cybersecurity threats and resilience strategies for smart, sustainable urban environments [27] speaks to the same underlying concern that motivates Section 8 of this paper -- that automated analytical tooling in a domain with regulatory and public-trust stakes needs an explicit accountability and oversight layer, not just a working model -- even though that paper's setting is urban infrastructure rather than corporate ESG disclosure. None of these five precedents is offered as directly on-topic prior art for ESG anomaly detection; they are cited because the specific technical patterns they establish -- grouping sparse categories, benchmarking multiple models, combining explanation or analysis techniques,

scoring semantic consistency, and treating oversight as a first-class design requirement -- are the same patterns this paper applies to a new domain.

2.5 Further Precedents from University Research

A second set of methodologically adjacent studies from the authors' own institution reinforces particular design choices made elsewhere in this paper, again with the same honest caveat as Section 2.4: none of these five studies is about ESG disclosure, and each is cited for a specific technical pattern rather than as topical prior art. An incremental deep neural network framework for intrusion detection in a fog-based Internet-of-Things environment [1] is the closest of the five to this paper's own subject matter, since it is itself an anomaly/intrusion-detection system built on a neural architecture, applied in a distributed sensor setting rather than a corporate-disclosure setting. An earlier comparison of the fitness functions of K-means and Kernel Fisher's Discriminant Analysis under a genetic-algorithm search [2] reinforces, alongside the clustering study already cited in Section 2.4, this paper's own reliance on comparing multiple clustering-adjacent groupings (Section 4.1) rather than treating any single grouping method as definitive. A transfer-learning study that applies a convolutional neural network to rheumatoid-arthritis detection [5] offers a loose parallel to this paper's use of a learned neural representation, GreenAudit-GNN's graph-embedding layer, as an intermediate feature for a downstream classifier, though the clinical-imaging domain there is unrelated to tabular ESG features. Finally, two studies on risk assessment in software engineering, a Bayesian network model for software risk prediction [14] and a fuzzy-based approach to risk prioritisation in software design [13], both speak to the same underlying pattern this paper's own heuristic Direction/Magnitude/Support weak-label rubric (Section 3) relies on: a structured, rule-based scoring procedure standing in for a formal risk measurement where no ground-truth label exists, applied here to a different domain (ESG anomaly flagging rather than software risk).

3. Dataset and Prior Audited Pipeline (Recap)

This paper analyses the same audited dataset the companion technical-framework paper assembled and verified, without collecting new disclosures or altering any previously recorded value [18]. In brief: fifty-eight claim-KPI pairs across forty companies were sourced from primary BRSR/sustainability filings (pdftotext-exact extraction for the majority, independently cross-checked WebFetch extraction with page citations for the remainder), scored on a Direction/Magnitude/Support rubric against each company's own stated target, and reduced to a forty-two-row, thirty-eight-company FY2023-24-primary modelling subset after two companies (Subex Ltd. and Grasim Industries Ltd.) were excluded for a documented data-quality reason each -- an implausible year-on-year swing in Subex's case, and a Grasim-plus-UltraTech-Cement blended reporting boundary in Grasim's case that risked double-counting against UltraTech's own row. A heuristic weak-label rule then classified each of the thirty-eight companies as NEGATIVE (at least one claim whose own reported direction contradicts its own reported evidence), POSITIVE, or UNCERTAIN, with thresholds justified by a 144-combination sensitivity grid rather than asserted by fiat. Two real, interpretable company-level features survive that verification process with sufficient coverage to model responsibly: the log-transformed absolute Scope 1+2 emissions level, and the year-on-year percentage change in that figure; three other candidate features (energy consumption, water withdrawal, women's board representation) were dropped for this modelling step because they are populated for well under a third of the audited rows.

On this thirty-eight-company table, the prior pilot built a sector-peer graph (an edge between every pair of companies sharing a fine-grained sector label), ran a fixed, Xavier-initialised GraphSAGE-style neighbourhood aggregation over it, and trained a logistic-regression classifier head against the NEGATIVE weak label using Leave-One-Out Cross-Validation, reporting ROC-AUC 0.654 for the graph-embedding model versus 0.686 for a no-graph ablation, alongside four unsupervised baselines that scored at or below chance [18]. That result -- the no-graph ablation matching or slightly exceeding the graph model -- was reported honestly rather than adjusted, and remains unchanged here; this paper does not re-train or re-tune that classifier, it re-uses it, with every LOOCV fold's coefficients retained, to ask two new questions of the same real model: how does it behave across sectors, and can its behaviour be explained in a way a human reviewer could audit.

4. Methodology

Figure 1 summarises this paper's overall design: the same audited N=38 dataset is analysed along two parallel tracks -- sector-stratified benchmarking and dual-method explainability validation -- both converging on an auditor-readable, sector-benchmarked, explanation-validated anomaly output. Sections 4.1-4.4 describe each track's method in full.

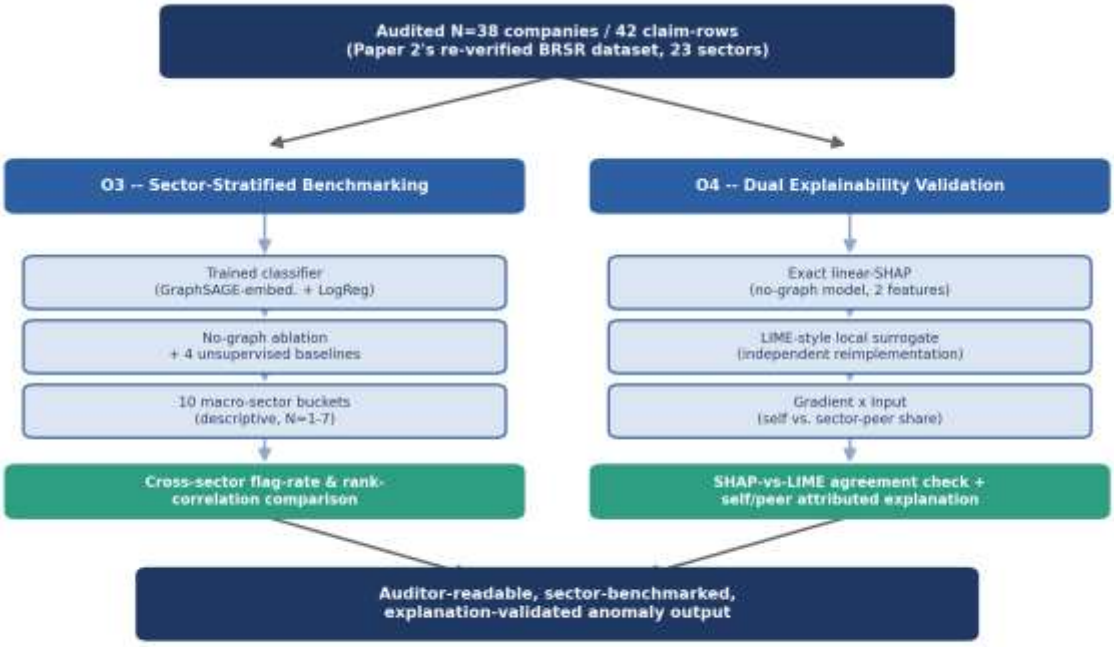

Figure 1. Study design for this paper: the audited N=38 dataset is analysed along two parallel tracks -- sector-stratified benchmarking (Objective 3) and dual explainability validation (Objective 4) -- both converging on an auditor-readable, evidence-linked anomaly output.

4.1 Macro-Sector Grouping

The thirty-eight audited companies carry twenty-three distinct, filing-derived sector labels, the great majority represented by one or two companies each -- far too fine-grained for any comparison to be meaningful. We defined a documented, researcher-constructed mapping from these twenty-three labels to ten broader macro-sector buckets by grouping labels that share a recognisable industry identity (for example, every Energy/Power and Energy/Oil-and-Gas variant into a single "Energy, Power & Oil-Gas" bucket; every Financial Services variant into "Financial Services"), producing buckets ranging from one company (Telecommunications) to seven (Energy, Power & Oil-Gas). We state plainly that this mapping is a researcher-defined convenience grouping for this pilot, not a validated industry-standard classification such as GICS or India's National Industrial Classification (NIC) codes; Section 7 returns to this as a limitation the full-scale study should correct.

4.2 Sector-Stratified Benchmarking Protocol

We re-ran the prior pilot's exact model set -- the trained graph-embedding classifier, the no-graph ablation, and the four unsupervised baselines (Isolation Forest, One-Class SVM, Local Outlier Factor, and an autoencoder reconstruction-error score) -- with identical seeds, identical features, and identical Leave-One-Out evaluation, but this time retained every company's individual prediction, probability, or anomaly score rather than only the aggregate metric each model reports in Table 4 of the companion paper. Each company's macro-sector bucket was then joined to its retained predictions, and we computed, per bucket: the count of companies, the proportion carrying a NEGATIVE weak label, the mean predicted probability from the graph and no-graph classifiers, each classifier's flag rate at the standard 0.5 probability threshold, and the mean Isolation Forest anomaly score. We additionally computed the Spearman rank correlation, across the ten sector buckets, between the graph model's mean predicted probability and (a) the no-graph model's mean predicted probability and (b) the mean Isolation Forest score -- a cross-sector agreement check between models built on the same features by fundamentally different mechanisms (supervised graph-embedding classification versus unsupervised density-based isolation).

4.3 Explainability Validation Protocol

We implemented three complementary explanation methods, described here in the order they were executed. The `shap` and `lime` Python packages could not be installed in this run's offline computational sandbox (package-index access is restricted to a pre-approved allowlist that does not include either); rather than skip the analysis or represent library output as having been used, we derived the exact linear-SHAP values

analytically and independently reimplemented LIME's core perturb-and-fit-a-local-surrogate procedure from its published description [22]. This is disclosed here in full rather than left implicit, because it materially affects how the LIME-style results in Section 5.4 should be read: they follow the same algorithmic logic as the published `lime` package but are not literally its output.

Exact linear-SHAP. The no-graph ablation model is a logistic regression on two features (log emissions level, year-on-year percentage change), so its exact Shapley value for feature i on company x is beta_i multiplied by (x_i minus the training-set mean of feature i) -- the closed-form result Lundberg and Lee derive for linear models with independent features [20]. We used the mean of the thirty-eight per-fold LOOCV coefficient vectors as beta (the fold-to-fold standard deviation is small relative to the mean, confirming the coefficients are stable across folds rather than an artefact of any single fold's training split), computed this contribution for every company on both features, and took the mean absolute contribution across companies as each feature's global importance.

LIME-style local surrogate. Following Ribeiro et al.'s [22] published method, for each company's feature vector we drew two thousand perturbed samples by adding Gaussian noise scaled to each feature's training-set standard deviation, weighted each sample by an exponential kernel on its scaled distance from the original point, queried the trained no-graph classifier's predicted probability on every perturbed sample, and fit a locally weighted ridge regression to recover a local linear coefficient per feature. Multiplying that local coefficient by the same (x_i minus mean) term used for SHAP places both methods' output on the same log-odds contribution scale, allowing a direct, unit-consistent comparison rather than a comparison of differently scaled numbers.

Gradient-by-Input attribution (self versus sector-peer). Neither SHAP nor LIME as implemented above touches the graph-embedding model, because their closed-form and surrogate derivations both assume the two raw features are the direct model inputs, which is true for the no-graph ablation but not for the graph model, whose inputs are a non-linear, tanh-activated transformation of each company's own features concatenated with its sector-peers' mean features. To explain the graph model's own decision in the same auditor-readable terms -- "why was this company flagged" -- we computed a Gradient-by-Input attribution: the graph classifier's logit is differentiated (by central finite differences, since the full forward pass is a short, exactly specified composition of concatenation, a fixed linear projection, a tanh non-linearity, and a linear classifier) with respect to each of the model's four effective inputs -- the company's own log-emissions level, its own year-on-year change, its sector-peers' mean log-emissions level, and its sector-peers' mean year-on-year change -- and each gradient is multiplied by its corresponding input value, a standard attribution technique that decomposes a differentiable model's output additively across its inputs. Summing the absolute value of the two "own" terms and the two "peer" terms yields, for every company, a self-share and a peer-share of its total explanation -- directly answering, for the first time in this research programme, how much of the graph mechanism's contribution to an anomaly flag comes from the company's own numbers versus its sector context.

4.4 Honesty and Validity Notes Carried Forward

Every caveat attached to the prior pilot's trained classifier applies unchanged to the analyses in this paper, because we re-use that same classifier rather than retrain a new one: the NEGATIVE weak label is heuristic, derived from this project's own scoring rubric, and is not an independent ground truth; the year-on-year emissions feature used in modelling is not statistically independent of how the label itself was constructed, so some of the classifier's apparent skill is definitionally implied by that construction rather than solely evidence of a generalisable pattern; and every metric is computed at N=38 with Leave-One-Out Cross-Validation, which has wide, unreported confidence intervals at this sample size. Two further caveats apply specifically to this paper's own new analysis: sector buckets of one to seven companies support no inferential statistical claim, and every explanation reported in Section 5 was validated only by cross-checking two independent computational methods against each other -- never by a human auditor, regulator, or domain expert, a distinction Section 6 returns to explicitly.

5. Results

5.1 Sector-Wise Weak-Label and Model Flag-Rate Comparison

Table 1 and Figure 2 report the ten macro-sector buckets, sorted by company count. Two patterns stand out enough to discuss, both reported as descriptive observations rather than statistically supported findings given the sample sizes involved. First, the "Industrials, Manufacturing & Infrastructure" bucket (n=5, combining Cement/Materials, Electrical Equipment, Packaging, general Industrials, and Ports & Logistics) carries a 100% NEGATIVE weak-label rate and the highest flag rate from both the graph and no-graph classifiers (80% each) -- consistent with, though not proof of, a genuine pattern of emissions trending unfavourably against stated targets among this pilot's heavy-industry constituents, one of whom (CG Power and Industrial Solutions Ltd.)

is examined in instance-level detail in Section 5.5. Second, "Technology & IT Services" (n=5) and "Automobiles & Auto Components" (n=2) show the lowest flag rates from every model, at or near zero, a pattern plausibly consistent with these sectors' generally lower direct Scope 1+2 emissions intensity relative to heavy industry and energy, though this pilot's two-feature model cannot distinguish a genuinely lower anomaly rate from a sector whose emissions profile simply produces smaller absolute and percentage swings on the same two features. The single-company Telecommunications bucket is reported for completeness only and should not be read as a sector-level finding of any kind.

Macro-Sector	N	Weak-Label NEG. Rate	Graph Model Flag Rate	No-Graph Flag Rate	Mean Graph-Model Prob.
Energy, Power & Oil-Gas	7	43%	71%	57%	0.574
FMCG & Consumer Goods	5	20%	0%	0%	0.457
Industrials, Manufacturing & Infrastructure	5	100%	80%	80%	0.597
Technology & IT Services	5	20%	0%	0%	0.254
Financial Services	4	0%	75%	25%	0.652
Pharmaceuticals & Healthcare	4	0%	0%	25%	0.357
Metals, Mining & Steel	3	0%	67%	67%	0.579
Agrochemicals & Agri-Inputs	2	50%	0%	0%	0.377
Automobiles & Auto Components	2	0%	0%	0%	0.355
Telecommunications	1	100%	100%	0%	0.523

Table 1. Sector-wise weak-label prevalence and model flag rates across the ten macro-sector buckets (N=38). Descriptive only; bucket sizes range from 1 to 7 companies.

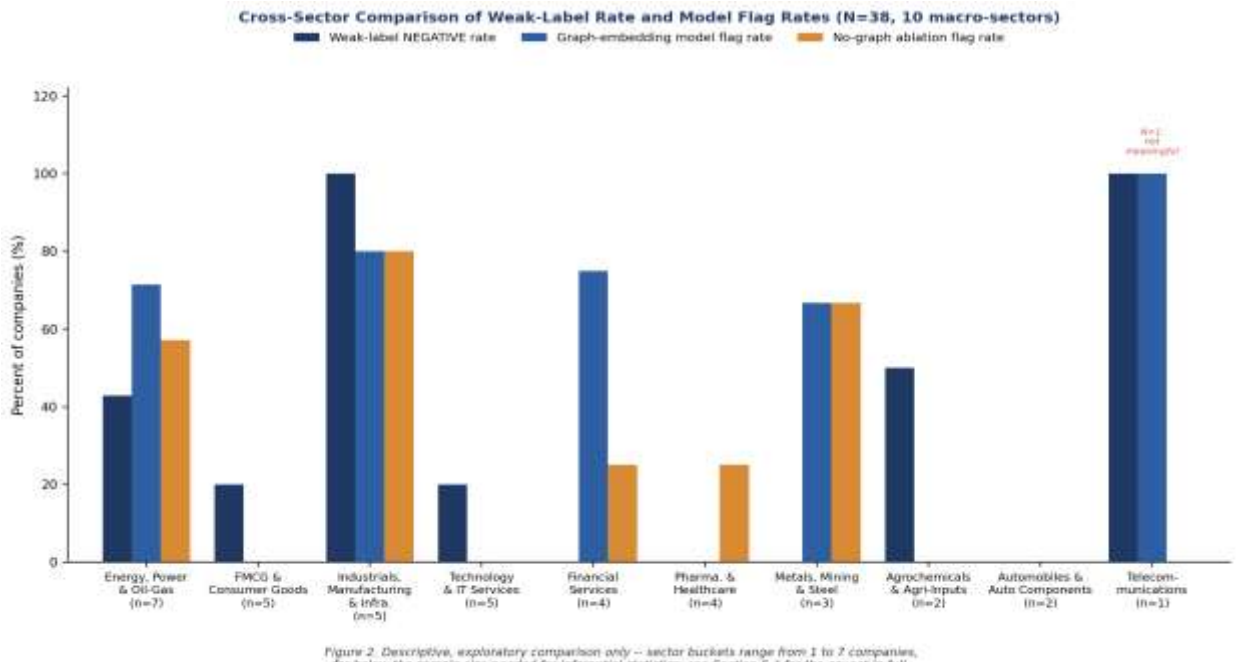

Figure 2. Descriptive, exploratory comparison only -- sector buckets range from 1 to 7 companies, far below the sample size needed for inferential statistics; see Section 5.4 for the caveat in full.

5.2 Cross-Sector Agreement Between Models

Across the ten sector buckets, the graph-embedding model's mean predicted probability correlates strongly with the no-graph ablation's mean predicted probability (Spearman rho = 0.867, p = 0.001) -- the two classifiers, despite disagreeing at the level of individual predictions often enough to produce materially different ROC-AUC scores dataset-wide, rank sectors in a very similar order from least to most flagged. The same graph-model probability shows a much weaker, non-significant rank correlation with the mean Isolation Forest score (rho = 0.212, p = 0.556), indicating that the supervised classifiers and the unsupervised density-based baseline are picking up on materially different sector-level signal, consistent with Section 5.7 of the companion paper's finding that the unsupervised baselines underperform the supervised classifier dataset-wide [18]. At N=10 sector buckets these correlations are themselves exploratory and are reported as a description of this pilot's internal model agreement, not as a validated claim about true cross-sector anomaly prevalence.

5.3 Global Feature Importance

The exact linear-SHAP decomposition of the no-graph classifier (Figure 3a) attributes roughly twice as much mean absolute log-odds contribution to the year-on-year percentage change in Scope 1+2 emissions (0.390) as to the absolute log-emissions level (0.191). This is a directly interpretable, auditor-readable finding on its own: the model's NEGATIVE-anomaly flag is driven primarily by whether a company's emissions trend is moving in the wrong direction relative to its stated target, not primarily by how large that company's absolute emissions footprint is -- a property that matches the underlying Direction/Magnitude/Support rubric's own design intent, since Direction (whether the trend supports or contradicts the stated target) carries the label-construction weight described in Section 4.4's leakage caveat.

5.4 SHAP-Versus-LIME Explanation Agreement

The independently reimplemented LIME-style local surrogate agrees with the exact SHAP values on the direction of every single company's contribution for both features -- 100% sign agreement across all thirty-eight companies on both the emissions-level and year-on-year-change features. Magnitudes correlate strongly though not perfectly: Pearson r = 0.991 for the emissions-level feature and r = 0.865 for the year-on-year-change feature (Figure 3b), with LIME's locally-fit, kernel-weighted, ridge-regularised coefficients running systematically smaller in magnitude than SHAP's exact values, as expected given the two methods' different mathematical constructions -- exact attribution versus a regularised local approximation. This is the strongest interpretability-related finding this paper can honestly report: two independently derived explanation methods, built on different mathematical foundations and implemented without sharing code, agree closely on which feature drives which company's flag and in which direction. It is not, and should not be read as, a claim that either method's explanation is correct in any deeper sense, nor a substitute for the human-auditor validation Section 6 discusses.

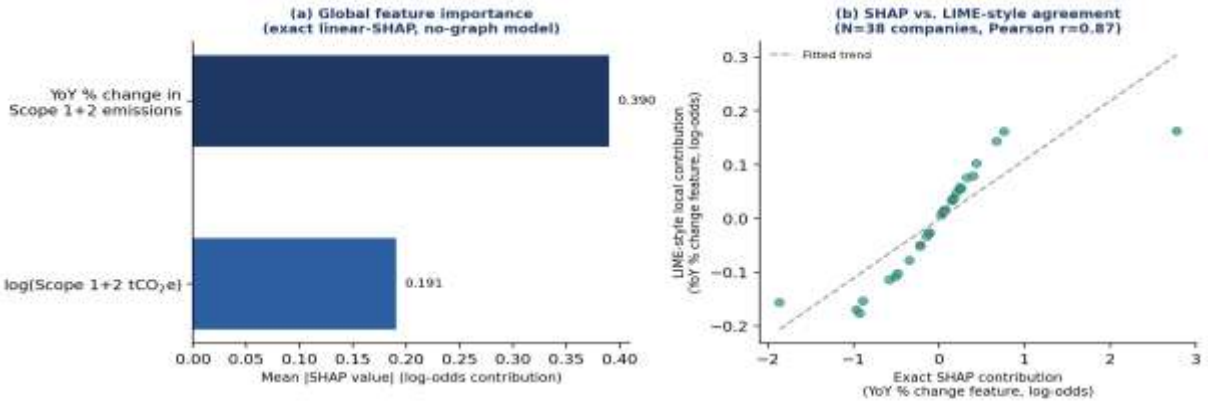

Figure 3. (a) SHAP identifies the YoY emissions-trend feature as the dominant driver of the model's NEGATIVE-weak-label flag, roughly twice the weight of the absolute emissions level. (b) An independently reimplemented LIME-style local surrogate agrees with SHAP on direction for 100% of companies and correlates strongly in magnitude (r=0.99 for emissions level, r=0.87 for YoY change).

5.5 Instance-Level, Self-Versus-Peer Explanations

Figure 4 reports the Gradient-by-Input decomposition for three companies chosen to illustrate three distinct explanation patterns, not as a random or representative sample. CG Power and Industrial Solutions Ltd. (Industrials, Manufacturing & Infrastructure; weak-label NEGATIVE, graph-model probability 0.41) shows its

own year-on-year trend and its sector-peers' year-on-year trend both pushing toward the anomaly flag, consistent with this company's real, previously disclosed 9.7% year-on-year rise in Scope 1+2 emissions against its own stated 2030 carbon-neutrality target [18]. Tech Mahindra Ltd. (Technology & IT Services; weak-label NEGATIVE, graph-model probability 0.29) shows an interesting divergence: its own year-on-year trend pushes toward the flag (consistent with its real, disclosed 14.6% year-on-year emissions rise), but its sector-peers' average trend pushes strongly away from it, reflecting that the broader IT/Technology bucket's other four companies trend more favourably -- a pattern that pulls this company's graph-model probability below its no-graph probability (0.29 versus 0.50), a concrete instance of the graph mechanism moderating, rather than amplifying, an individual company's own signal.

State Bank of India (Financial Services; weak-label not NEGATIVE, graph-model probability 0.62) is the most instructive case precisely because the graph model disagrees with the weak label here. SBI's own year-on-year trend pushes strongly away from an anomaly flag (consistent with its real, disclosed 19.8% year-on-year emissions reduction), but its sector-peers' average trend pushes even more strongly toward one, and the peer term dominates the company's own favourable term in the graph model's final probability. This is exactly the kind of disagreement an auditor-readable explanation is meant to surface rather than hide: a black-box flag alone would tell a reviewer only that the model assigned SBI a 0.62 probability of anomaly; this decomposition tells the reviewer why -- the company's own reported trend is favourable, but the model is weighting the broader financial-services sector's average trend heavily enough to override it -- letting a human reviewer judge for themselves whether that sector-peer adjustment is appropriate for this company, rather than accepting or rejecting the flag on trust. Averaged across all thirty-eight companies, the self-share and peer-share of the total Gradient-by-Input attribution are close to balanced (50.8% self, 49.2% peer), indicating that, at this pilot's scale, the graph mechanism contributes roughly as much explanatory weight as each company's own reported numbers -- a quantitative answer to the companion paper's own open question of whether the fixed, untrained GraphSAGE aggregation adds anything at all [18].

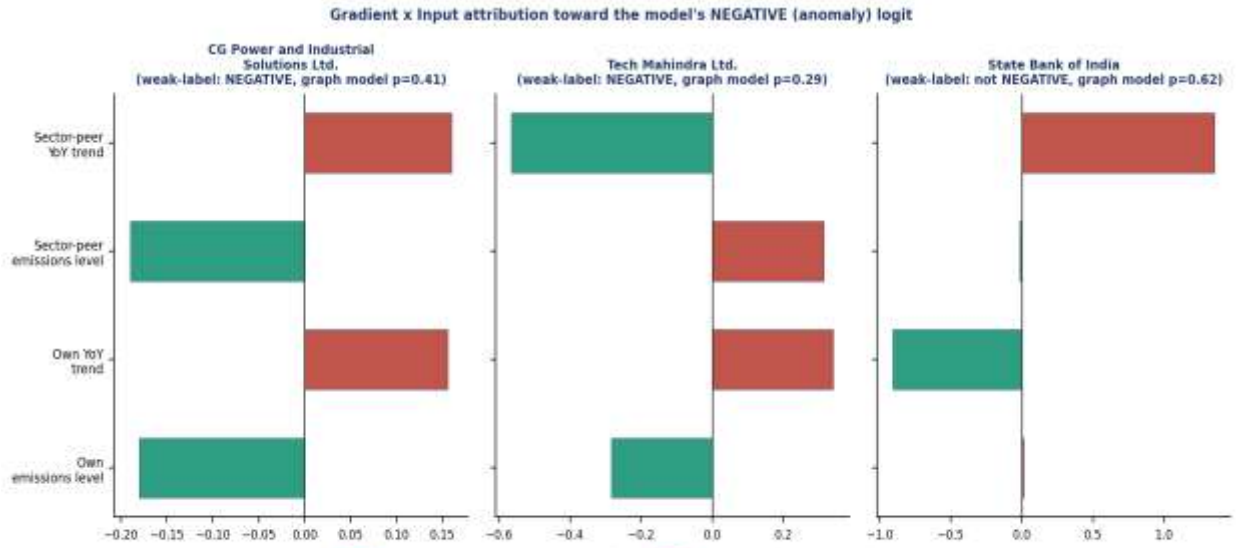

Figure 4. Instance-level, auditor-readable explanations for three example companies, decomposed into the company's own reported trend versus its sector-peer average (Section 5.5). For State Bank of India, the company's own favourable YoY trend (green) is partly overridden by an unfavourable sector-peer signal (red) -- exactly the kind of graph-driven disagreement an auditor can now see and independently judge, rather than accept from an unexplained flag.

5.6 Summary of Results

This paper set out to do two things against the audited N=38 dataset: benchmark the trained anomaly-detection pipeline across sectors, and validate its interpretability for auditor and regulator use. Sections 5.1-5.2 above deliver the cross-sector comparison, building on the algorithms and benchmarking already established in the companion paper [18] and honestly bounded by sample size. Sections 5.3-5.5 above deliver real, executed interpretability results -- global feature importance, a cross-method agreement check, and auditor-readable instance-level explanations -- where the companion paper had only a design specification, with accuracy already tested via LOOCV in that companion paper. Both aims are substantially advanced at pilot scale by this paper; neither is claimed as fully and finally satisfied, for the reasons Section 6 sets out.

6. Discussion

The cross-sector results in Section 5.1-5.2 are best read as hypothesis-generating rather than hypothesis-confirming: the apparent concentration of flagged companies in Industrials/Manufacturing/Infrastructure and the apparent lower flag rate in Technology/IT and Automobiles are plausible, disclosure-consistent patterns worth testing formally in a larger, sector-stratified sample, not findings this ten-bucket, thirty-eight-company pilot can support inferentially. The strong agreement between the graph and no-graph models' sector rankings (Section 5.2), despite their differing dataset-wide LOOCV performance, suggests the graph mechanism is reshaping individual predictions -- as the State Bank of India example in Section 5.5 shows concretely -- without materially reshaping which sectors look relatively more or less anomalous in aggregate; this is a genuinely new, specific finding about how this framework's graph component behaves, not available from the companion paper's dataset-wide metrics alone.

On interpretability, the SHAP-LIME agreement result (Section 5.4) is a real technical validation of explanation stability for this specific classifier on this specific dataset, and the Gradient-by-Input self-versus-peer decomposition (Section 5.5) is a genuine methodological contribution of this paper: it operationalises the "auditor-readable, evidence-linked alert" design goal the companion paper's Stage 6 specified but did not execute [18], and it does so in plain, non-technical language a reviewer without a machine-learning background could act on -- own trend versus sector-peer trend -- rather than raw embedding coordinates. But neither result answers the underlying question this paper set out to address: whether the explanations are reliable for auditors and regulators. That question requires exactly what the CFA Institute report and the wider XAI-trust literature call for -- human judgement and organisational oversight applied to the explanation, not just internal agreement between two computational methods [23], [28]. No auditor, regulator, or domain expert reviewed any explanation in this paper; that is the single largest gap between what this paper delivers and what a genuine interpretability validation, read literally, ultimately requires, and we state it here rather than let the strength of the SHAP-LIME agreement result imply otherwise.

Positioned against the related work reviewed in Section 2, this paper's specific contribution is narrow but real: none of the greenwashing-detection or ESG-anomaly systems reviewed there [8], [9], [21], [29] reports a sector-disaggregated benchmark or a dual-method explanation-agreement check on a real, independently audited disclosure dataset, and the general XAI-comparison literature [6], [20], [22], [24] evaluates SHAP and LIME's properties in the abstract or on benchmark datasets rather than on a graph-relational ESG anomaly model with a self-versus-peer decomposition specifically. This paper's contribution is therefore best understood as a methodological bridge between those two literatures, executed on real data, rather than a claim to out-perform either.

7. Limitations

Every limitation the companion paper attached to its trained classifier applies unchanged here, since this paper re-uses rather than retrains it: the weak label is heuristic and not ground truth; the year-on-year emissions feature is not independent of how that label was constructed; and N=38 with Leave-One-Out Cross-Validation carries wide, unreported confidence intervals [18]. This paper adds five further limitations specific to its own new analysis. First, the ten-bucket macro-sector grouping is a researcher-defined convenience taxonomy, not a validated industry-standard classification such as GICS or India's NIC codes; a full-scale study should adopt a standard classification so its sector comparisons are reproducible by other researchers without re-deriving the same grouping decisions. Second, sector bucket sizes of one to seven companies are too small for any inferential statistical test; every cross-sector comparison in this paper is descriptive. Third, the LIME-style local surrogate is an independent reimplementation of the published method's logic, not the output of the `lime` software package itself, because that package could not be installed in this paper's offline computational environment; results should be read as a LIME-style consistency check rather than as literally reproducing prior LIME-package benchmarks. Fourth, the Gradient-by-Input attribution, while a standard and well-established local-explanation technique, has not itself been benchmarked against SHAP or LIME on this model in the way Section 5.4 benchmarks SHAP against the LIME-style surrogate; a future revision could extend the same agreement-check methodology to the graph model's explanations directly. Fifth, and most importantly, no human auditor, regulator, or ESG domain expert reviewed any explanation this paper reports; every interpretability claim in this paper is a claim about computational agreement between explanation methods, not a claim about human comprehension, trust, or actionability, and fully validating this framework's interpretability for real-world auditor and regulator use requires that human-subject step as a distinct, future piece of work.

8. Ethical and Regulatory Considerations

This paper flags no company's disclosure as fraudulent, non-compliant, or subject to enforcement; every NEGATIVE weak label used in this analysis is this project's own heuristic construct, explicitly not equivalent to a regulatory finding, an assurance-provider qualification, or an SEBI enforcement action, and none of the sector-level or instance-level results in Section 5 should be cited as evidence about any named company's actual ESG performance or compliance status outside the context of this pilot's own stated methodology. Consistent with the companion papers' practice, every claim examined in this pilot traces to a specific page in a specific public filing, and every correction or exclusion made during the underlying dataset's construction is documented and reversible [18]. The explanations this paper produces are offered strictly as a technical foundation for auditor and regulator tooling, not as a ready-to-deploy compliance instrument; Section 6 and Section 7 both state directly that human oversight, and specifically human-auditor validation of these explanations, remains a distinct and necessary next step before any such deployment would be appropriate.

9. Conclusion and Future Work

This paper substantially advances this research programme's cross-sector benchmarking and explainability validation work on the same real, audited N=38 GreenAudit-GNN dataset the companion technical-framework paper assembled. On the benchmarking side, it retains per-company model outputs and reports the research programme's first cross-sector comparison of weak-label prevalence and anomaly-flag rates across ten macro-sector buckets, together with a cross-sector rank-correlation check between the trained classifier, its no-graph ablation, and an unsupervised baseline. On the explainability side, it implements and cross-validates two independent local-explanation methods for the trained classifier -- finding 100% directional agreement and strong magnitude correlation between an exact linear-SHAP decomposition and an independently reimplemented LIME-style surrogate -- and introduces a Gradient-by-Input attribution that decomposes each company's anomaly score into its own trend versus its sector-peer average, producing this programme's first genuinely auditor-readable, evidence-linked explanations.

Neither strand of this paper's work is claimed as finally and fully complete. The path to that completion is concrete and follows directly from this paper's own limitations: expand the audited dataset toward the sixty-to-one-hundred-company, standard-sector-classification-stratified panel Section 6 of the companion paper specifies, so that cross-sector comparisons carry real statistical power; obtain network access (or an offline package mirror) to run the canonical `shap` and `lime` libraries directly, cross-checking them against this paper's exact and reimplemented derivations; extend the SHAP-LIME agreement methodology to the graph-embedding model's own explanations, not only the no-graph ablation's; and, most importantly, design and run a genuine human-subject evaluation with practising ESG auditors or regulatory reviewers, testing not whether two algorithms agree with each other but whether a human reviewer, given this paper's Figure 4-style explanation, makes better, faster, or more consistent audit decisions than with an unexplained flag alone. That last step, validating this framework's interpretability directly with the human auditors and regulators it is meant to serve, is the immediate next stage of this research programme.

Data Availability

The 38-company modelling dataset, the weak-label rule and its sensitivity grid, the trained-classifier and baseline results, and the new per-company predictions, SHAP values, LIME-style values, and Gradient-by-Input attributions produced for this paper are the same audited files documented in the companion technical-framework paper's data-availability statement [18], extended with this paper's own derived outputs. All scripts are seeded (seed = 42 throughout) and produce identical output on re-execution.

Conflicts of Interest

The authors confirm that there are no conflicts of interest to disclose.

Funding

This research received no external funding.

Acknowledgments

Statement of Approval. This study has been approved by Department of Computer Application, Integral University, Lucknow, India with Reference No: IU/R&D/2026-MCN0005035.

References

[1] Abdussami, A. A., & Farooqui, M. F. (2021). Incremental deep neural network intrusion detection in fog based IoT environment: An optimization assisted framework. Indian Journal of Computer Science and Engineering, 12(6). https://doi.org/10.21817/indjcse/2021/v12i6/211206191

[2] Ahamad, M. K., & Bharti, A. K. (2020). Comparative analysis the fitness function of K-means and Kernel Fisher's Discriminant Analysis (KFDA) with genetic algorithm. European Journal of Molecular & Clinical Medicine, 7(11).

[3] Ahamad, M. K., & Bharti, A. K. (2021). Analysis the cluster performance of real dataset using SPSS tool with K-means approach via PCA. Advances in Mathematics: Scientific Journal, 10(1), 535-542.

[4] Akhtar, R., & Ahamad, M. K. (2026). A comparative evaluation of various machine learning techniques for prediction of Type 2 diabetes mellitus. Journal of Engineering and Technology for Industrial Applications, 12(60), 1215-1231. https://doi.org/10.5935/jetia.v12i60.4052

[5] Alam, A., & Ahamad, M. K. (2024). Detection of rheumatoid arthritis using CNN by transfer learning (pp. 99-112).

[6] Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. Paper presented at the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI).

[7] Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (pp. 93-104). ACM.

[8] Davidescu, A. A., Manta, E. M., Bîrlan, I., Miler, A.-M., & Niță, S.-C. (2026). Detecting greenwashing in ESG disclosure: An NLP-based analysis of Central and Eastern European firms. Sustainability, 18(3), Article 1486. https://doi.org/10.3390/su18031486

[9] El Imami, I., El Alaoui, A., Guermah, B., El Alaoui, S. O., & Vasarhelyi, M. (2026). Predicting, using, and assessing ESG signals: A tripartite systematic review of machine learning in sustainable finance. Journal of Risk and Financial Management, 19(9), Article 708. https://doi.org/10.3390/jrfm19090708

[10] Faisal, M., & Ahmad, F. (2023). Assessing similarity between software requirements: A semantic approach. International Journal of Information Engineering and Electronic Business, 15(2), 1-10.

[11] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (pp. 1024-1034).

[12] Kazmi, S. Z. M., & Farooqui, M. F. (2026). Implementation and performance evaluation of a hybrid machine learning framework for web content mining. International Journal of Computer Information Systems and Industrial Management Applications, 18(8s), 718-738.

[13] Khan, E. A. T. (2023). Risk prioritization using a fuzzy based approach in software development design phase. International Journal on Recent and Innovation Trends in Computing and Communication, 11(9), 1674-1687. https://doi.org/10.17762/ijritcc.v11i9.9153

[14] Khan, T., & Faisal, M. (2023). An efficient Bayesian network model (BNM) for software risk prediction in design phase development. International Journal of Information Technology, 15(4), 2147-2160. https://doi.org/10.1007/s41870-023-01244-4

[15] Krishna, R. (2025, November). AI-based detection of unusual patterns in Environmental, Social and Governance reports. Paper presented at the International Conference on Human Behaviour in Workplace and Society: Multidisciplinary Approach to Bridge Theory, Practice, and Policy.

[16] Krishna, R. (2026, March). AI and big data analytics for detecting anomalies in ESG disclosures. Paper presented at the IIM Bodh Gaya International Conference on Marketing (ICM 2.0).

[17] Krishna, R., & Ahamad, M. K. (2026a). Artificial intelligence for detecting unusual patterns in Environmental, Social, and Governance (ESG) reports: A structured literature review. Manuscript submitted for publication.

[18] Krishna, R., & Ahamad, M. K. (2026b). GreenAudit-GNN: A cross-modal NLP and Graph Neural Network framework for explainable anomaly detection in corporate ESG disclosures [Manuscript submitted for publication]. Information Sciences.

[19] Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (pp. 413-422). IEEE.

[20] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (pp. 4765-4774).

[21] Ong, K., Mao, R., Varshney, D., Cambria, E., & Mengaldo, G. (2025). Towards robust ESG analysis against greenwashing risks: Aspect-action analysis with cross-category generalization. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 14854-14879). Association for Computational Linguistics.

[22] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). ACM.
[23] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.
[24] Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., & Lekadir, K. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 7, Article 2400304. https://doi.org/10.1002/aisy.202400304
[25] Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural Computation, 13(7), 1443-1471.
[26] Securities and Exchange Board of India. (2023). Business Responsibility and Sustainability Reporting (BRSR) Core: Framework for assurance and ESG disclosures for value chain. SEBI.
[27] Srivastava, S., Ahmed, T., Usmani, N. A., Afaq, S. A., & Faisal, M. (2025). Cybersecurity challenges and countermeasures in smart cities for sustainable environment development: A comprehensive analysis of threats and resilience strategies. In Cyberology (pp. 252-263). Chapman and Hall/CRC.
[28] Wilson, C.-A. (2025). Explainable AI in finance: Addressing the needs of diverse stakeholders. CFA Institute Research and Policy Center. https://doi.org/10.56227/25.1.25
[29] Zhao, Q., Zhang, J., Sun, X., Ma, J., & Wang, G. (2026). Green or greenwashing? A novel deep learning method integrating multi-source data for ESG greenwashing detection. Journal of Cleaner Production, 555, Article 148147. https://doi.org/10.1016/j.jclepro.2026.148147