

Review

Mapping the Methodological Landscape of Green Finance and Digital Technology Integration: A Scoping Review

Jiacheng Liu 

School of Economics and Management, Southeast University, Nanjing 211189, China; 220242825@seu.edu.cn

Abstract

The integration of green finance and digital technology exhibits considerable methodological diversity; however, existing scholarship lacks a systematic account of how empirical evidence is generated in this field. This scoping (PRISMA-ScR) review develops a literature classification framework centered on methodological approaches and applies it to 423 English-language articles on the green finance–digital technology nexus from the Web of Science Core Collection (2018 to April 2026). Research practices are organized along a two-dimensional framework spanning the epistemological objectives (explanatory vs. predictive) and the technical toolkit (traditional econometrics vs. machine learning), with causal machine learning integrating both paradigms. Each study is assigned a primary methodological label via a lexicon-based multi-label classification engine employing a 2×2 weight matrix. Weight sensitivity analysis confirms that classification outcomes are robust to parameter perturbation. Inter-coder reliability was near-perfect (Cohen's $\kappa = 0.93$); algorithm–gold standard agreement was substantial ($\kappa = 0.76$, $F1 = 0.79$). The classification reveals that traditional econometrics remains the most prevalent method (26.95%), followed by predictive machine learning (17.26%) and technology application (16.31%), while causal machine learning (9.69%) remains underutilized and behavioral research is virtually absent (0.95%). Over a quarter of studies (26.24%) employ multiple methodologies. These findings provide a methodological map for researchers and policymakers navigating the green finance–digital technology landscape.

Keywords: green finance; digital technology; methodology; machine learning; causal inference; econometrics; multi-label classification; lexicon-based classification

1. Introduction

As the global community accelerates toward carbon neutrality, financing the low-carbon transition has become one of the most pressing economic challenges of this decade [1]. Green finance, including instruments such as green credit, green bonds, and transition financing, has become established as a core policy tool for directing capital flows toward environmental objectives [2]. At the same time, the rapid diffusion of digital technologies has transformed the pricing, monitoring, and disclosure of environmental risks: AI-driven ESG analytics have enhanced transparency and enabled real-time monitoring of environmental performance [3], while digital platforms have reduced transaction costs, optimized information disclosure mechanisms, and facilitated green capital flows [4]. The deep integration of green finance and digital technology has given rise to a rapidly expanding body of empirical research supported by an increasingly diverse array of methodological tools, spanning traditional econometrics, machine learning, and causal machine learning [5].



Academic Editor: Cristina Raluca Gh. Popescu

Received: 8 July 2026

Revised: 31 July 2026

Accepted: 5 August 2026

Published: 7 August 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

This methodological diversification raises a critical concern: different analytical strategies may produce fundamentally different research findings and policy implications. Blakely et al. [5] theoretically demonstrate the close linkage between methodological choice and resource allocation outcomes; Gur Ali [6] empirically confirms this effect. This concern is further grounded in the broader methodological literature: Athey and Imbens [7] systematically catalog the expanding toolkit of machine learning methods that economists should know about, while Angrist and Pischke [8] argue that research design, not methodological sophistication alone, is the foundation of credible empirical evidence. To understand which methodological tools researchers in this field actually employ, their relative prominence, and the overall configuration of methodological choices is therefore of fundamental importance for assessing the reliability of the field's knowledge base.

In recent years, several scholars have reviewed the green finance–digital technology nexus from diverse perspectives, making important contributions to the understanding of this rapidly advancing research frontier (Table 1). Kwong et al. [9] were the first to systematically examine the synergistic relationship between green finance and financial technology, clarifying the theoretical interaction mechanisms and identifying structural barriers to deeper integration. Rojas-Valdez and Fajardo-Piñan [10] provided a panoramic mapping of Industry 4.0 and green finance research by industry and operational domain, revealing the enabling pathways of digital technologies in clean production and the role of financial innovation in inclusive development. Saeedi and Ashraf [11] constructed an analytical framework for analyzing barriers to green finance practice, organized around core challenges of risk management, product innovation, and compliance complexity. Geetha and Biju [12] systematically delineated the conceptual boundaries of Green FinTech, providing a clear terminological foundation for this subfield. Debrah et al. [13] employed a mixed bibliometric–qualitative approach to comprehensively identify knowledge gaps and future directions in green finance research. Teixeira [14] outlined the technology-driven pathways of green digital finance across key technological dimensions including blockchain, artificial intelligence, big data, and digital platforms. These works have deepened scholarly understanding of the field from the perspectives of topic identification, conceptual delineation, and trend assessment, forming an important intellectual foundation for this study.

Table 1. Comparison of Existing Reviews on the Green Finance–Digital Technology Nexus: Methodological Coverage.

Study	Research Focus	Review Method	Key Contribution	Methodological Dimension
Kwong et al. [9] (2023)	Synergy between green finance and financial technology	Bibliometric analysis	Identified theoretical interaction mechanisms and structural barriers to integration	Not covered
Rojas-Valdez & Fajardo-Pinan [10] (2026)	Intersection of Industry 4.0 and green finance	Cluster analysis and literature review	Panoramic mapping of digital technology applications in clean production by industry and operational domain	Not covered
Saeedi & Ashraf [11] (2024)	Role of technology in promoting green finance	Systematic literature review	Constructed a framework for technology-enabled green finance development	Not covered
Geetha & Biju [12] (2024)	Green FinTech reshaping the financial sphere	PRISMA systematic review	Proposed a Green FinTech conceptual framework and research agenda	Not covered
Debrah et al. [13] (2023)	Knowledge gaps and directions in green finance research	Mixed bibliometric–qualitative review	Identified six research hotspots and future directions	Not covered

As Table 1 demonstrates, existing reviews share a common feature: they adopt a topic-oriented organizing logic; that is, they organize the literature around “what is studied” (core themes, technological dimensions, or practical challenges). This approach is efficient and intuitive for quickly locating research progress on specific topics, and the aforementioned reviews have all produced high-quality knowledge syntheses within their respective focus areas. At the same time, it is evident that existing reviews have not incorporated methodology itself, that is, “how research is conducted,” as an analytical dimension within their frameworks. In other words, these reviews have effectively answered “what questions has the field studied” but have not yet provided a systematic answer to “what methods has the field employed to study these questions.”

When methodology remains a blind spot, a series of critical questions become unanswerable. What does the field’s methodological toolkit actually look like—is it dominated by a narrow set of conventional techniques, or has it diversified in step with the growing complexity of green finance problems? Can this toolkit be mapped at a scale large enough to reveal systematic patterns rather than anecdotal impressions? And what does the resulting methodological landscape imply for the credibility of causal claims, the rigor of evidence standards, and the overall quality of knowledge production? These questions point to a direction that lies beyond the scope of existing reviews yet is essential for the healthy development of the field.

Reviews that adopt methodology as the primary classification dimension are generally scarce in the financial economics literature, although several scholars have made important contributions in this direction. Hoang and Wiegatz [15], in reviewing machine learning applications in finance, proposed a function-based “three-prototype taxonomy” that distinguishes machine learning applications by three functional roles: constructing new measurement indicators, reducing prediction error, and extending the standard econometric toolkit, one of the few examples of methodology-function-oriented classification, whose innovation lies in moving beyond the conventional algorithm-type-based taxonomy to a classification grounded in the epistemic needs of economists. Mullainathan and Spiess [16] provided an influential methodological account of the complementary relationship between machine learning and traditional methods from an applied econometrics perspective, offering a bridge between the two paradigms. Bahammou et al. [17] constructed a four-layer methodological classification framework for financial prediction, integrating data preprocessing, model selection, optimization strategies, and evaluation metrics into a unified system. D’Amico et al. [18] proposed a multi-dimensional modal fusion taxonomy that demonstrates strong organizational capacity in multi-modal analysis scenarios. These works provide valuable references for this study; they demonstrate the feasibility and value of methodological classification within their respective method families or application contexts, but they also suggest that migrating such frameworks to a comprehensive mapping of the methodological landscape in the green finance–digital technology domain requires additional design considerations.

This study aims to provide a literature analytical framework complementary to existing reviews from a methodological perspective. Specifically, this paper poses the following research questions: (1) What is the distributional pattern of methodological categories—traditional econometrics, predictive machine learning, causal machine learning, financial econometrics, operations research/game theory, interpretability methods, behavioral research, and technology application—across the green finance–digital technology nexus? (2) Can a lexicon-based multi-label automated classification scheme reliably annotate the methodological approaches of the literature in this field? (3) What implications does this methodological landscape hold for researchers’ method selection practices, policymakers’ evidence-use standards, and the overall quality of knowledge production in the field?

To address these questions, this study designs a literature classification framework centered on methodology, situating methodological practices within a two-dimensional framework spanning traditional econometrics to causal machine learning, and constructs a lexicon-based multi-label automated classification system. A gold standard is established through dual independent coding and adjudication; classification reliability is validated through algorithm–gold standard agreement testing and confusion matrix analysis; parameter robustness is assessed through weight sensitivity analysis; and the scoping review (PRISMA-ScR [19]) is completed for 423 articles from the Web of Science Core Collection.

2. Conceptual Framework

2.1. Methodology

The integration of green finance and digital technology has long been characterized by a clear methodological distinction: explanatory modeling aims to test theoretically motivated causal hypotheses, while predictive modeling strives to maximize out-of-sample prediction accuracy. Traditional econometrics and financial econometrics belong to the explanatory domain, centered on unbiased estimation and hypothesis testing. The distinction between the two is not merely an operational division based on terminology assignment but rests on methodological paradigm foundations: traditional econometrics takes causal inference as its core objective, with a toolkit including instrumental variables, difference-in-differences, and panel fixed effects as quasi-experimental methods; financial econometrics has developed a specialized methodological system oriented toward the structural properties of financial data: Generalized Autoregressive Conditional Heteroskedasticity (GARCH)-family volatility models, asset pricing factor tests, high-frequency data analysis, and risk measurement including Value at Risk (VaR) and Expected Shortfall (ES), constituting a coherent specialized paradigm in terms of modeling assumptions (e.g., time-varying volatility, no-arbitrage conditions) and inferential targets (risk premium estimation, volatility forecasting) [16]. In contrast, machine learning originates from the predictive domain, emphasizing predictive accuracy and generalization capability, often at the expense of interpretability [7,16].

In recent years, the rise of causal machine learning and interpretability methods has complicated this binary distinction. Causal machine learning employs flexible computational tools to estimate treatment effects, while interpretability methods aim to enhance the transparency of predictive models [20,21]. We argue that the methodological approaches in this field can no longer be viewed as belonging to two fundamentally opposed camps; rather, they can be situated within a two-dimensional framework defined by two orthogonal axes (Figure 1): the epistemological objectives (explanatory vs. predictive) and the technical toolkit (traditional econometrics vs. machine learning). Based on this two-dimensional framework, we define eight methodological categories (Table 2). Under this two-dimensional framework, B1, B2, and B4 each occupy a distinct position. B1 represents explanatory $\times$ traditional econometrics, while B2 represents predictive $\times$ machine learning. B4 (Causal ML) bridges the explanatory and predictive poles of the first axis while remaining anchored to machine learning on the second axis, thereby integrating both epistemological objectives within a single technical toolkit [22]. B5 is situated in the explanatory domain, resting on the specialized paradigm of financial data structures, while B6 occupies a normative/optimization position, resting on the specialized paradigm of optimization–equilibrium analysis. B3 (interpretability methods) and B7 (behavioral research) do not occupy a position within this two-dimensional plane; rather, they constitute two independent cross-cutting dimensions that can be overlaid on any position within the plane, including B1, B2, B4, B5, or B6. B3 functions as a second-order diagnostic tool pertaining to the model itself, applicable regardless of the underlying method's position on

the two axes; B7, by contrast, is grounded in an entirely different epistemological paradigm external to the explanatory–predictive and econometrics–machine learning distinctions. C (Technical/Engineering Applications) likewise falls outside the two-dimensional plane, but for a different reason than B3 and B7: rather than being a cross-cutting dimension overlaid on the plane, C is excluded from the explanatory–predictive axis altogether, since its primary contribution is the technical artifact itself (e.g., a system, platform, or engineering pipeline) rather than an explanatory or predictive claim about an economic phenomenon.

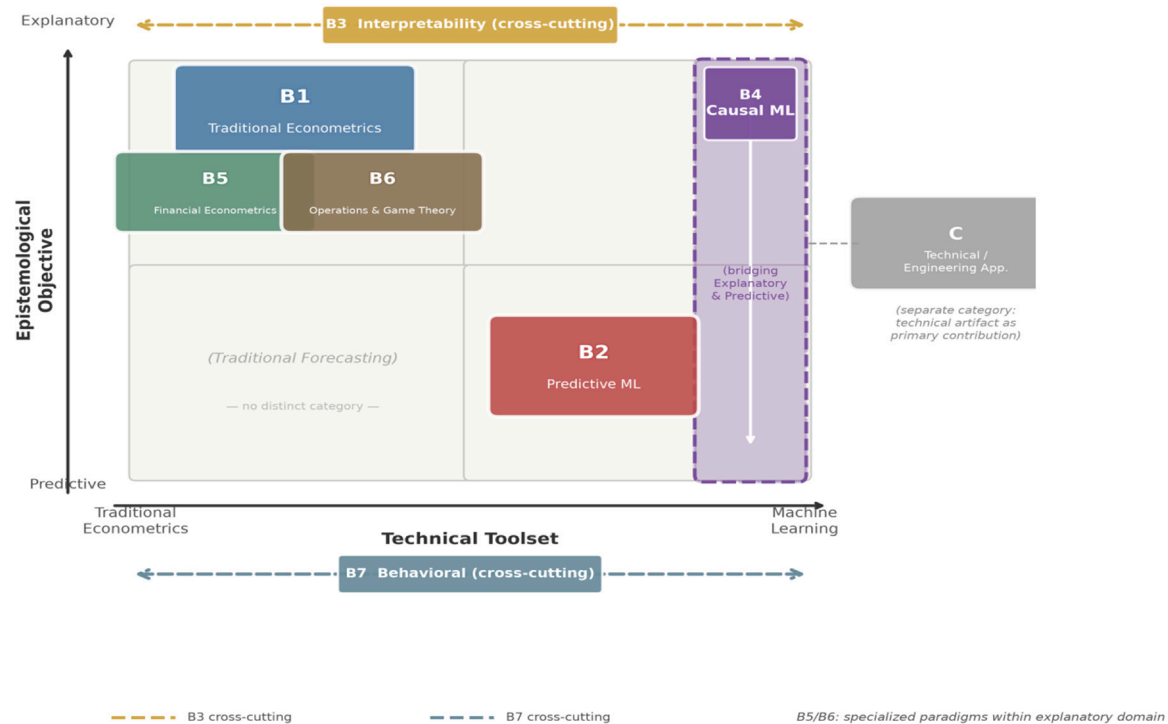


Figure 1. Two-Dimensional Methodological Framework.

Table 2. Methodological Taxonomy and Its Position in the Two-Dimensional Framework.

Code	Category	Position in the Two-Dimensional Framework
B1	Traditional Econometrics	Explanatory (Anchor)
B2	Predictive ML	Predictive
B3	Interpretability Methods	Explanatory–Predictive Bridge
B4	Causal ML	Explanatory–Predictive Bridge
B5	Financial Econometrics	Explanatory (dual-purpose)
B6	Operations & Game-Theoretic Modeling	Normative/Optimization
B7	Behavioral Approaches	Cross-cutting
C	Technical/Engineering Application	Predictive/Applied

Note: Abbreviations used in this table are defined in the Abbreviations section at the end of this article.

Conceptualizing methodological practices within a two-dimensional framework does not imply that the boundaries between methodological approaches are sharply delineated. Indeed, multiple positions within this cross-structure exhibit inherent ambiguity: for instance, GARCH-family models in financial econometrics serve a dual function, explanatory (parameter estimation and hypothesis testing) and predictive (volatility forecasting), and Double/Debiased ML straddles predictive modeling (flexible first-stage estimation) and causal inference (second-stage debiasing) at the epistemological level, with its positioning dependent on the researcher’s interpretation of its intended use rather than the method’s

technical characteristics per se. This inherent ambiguity implies that any empirical classification based on the two-dimensional framework must confront a fundamental methodological tension: the framework acknowledges that boundaries are gradual, while an operable classification system must convert gradual boundaries into discrete breakpoints to ensure inter-coder reliability. This study addresses this tension through two design choices. First, discrete categories are positioned according to the epistemological distinguishability of methodological functions, with breakpoints corresponding to locations of significant functional difference (Table 3). Second, the multi-label framework allows a single study to carry multiple category labels. These design choices endow the classification system with the operational clarity required by discrete categories while maximizing fidelity to the methodological landscape, though this fidelity is limited, as discussed in Section 5.4.

Table 3. Operational Definitions of the Methodological Taxonomy.

Code (Category)	Inclusion Criteria (with Representative Methods)	Boundary Rules
B1 (Traditional Econometrics)	Classical regression-based inference focuses on unbiased coefficient estimation and hypothesis testing (OLS, fixed/random-effects panels, IV, standard time series analysis).	Quasi-experimental designs are included here only if they use conventional parametric estimation; ML-based effect identification → B4.
B2 (Predictive ML)	Out-of-sample predictive performance is the principal focus (e.g., tree ensembles, SVM, neural networks, clustering).	It differs from B4 in that it lacks a causal estimand and from B3 by not prioritizing interpretation as the central objective.
B3 (Interpretability Methods)	Explanation/transparency of predictive models as the central aim (SHAP, LIME, feature importance, partial dependence).	Applies only when model interpretation is the primary goal; if interpretability serves as a supplementary diagnostic, → B2.
B4 (Causal ML)	ML estimators are explicitly designed for predictive causal identification—treatment, heterogeneous, or counterfactual effects (double/debiased ML, causal forests, TMLE, ML-augmented designs).	Defined by an explicit causal estimand; however, when identical algorithms are used solely for accuracy → B2.
B5 (Financial Econometrics)	Techniques specifically tailored for financial-data properties (GARCH-family volatility, cointegration/ECM, event studies, spillover/connectedness).	Distinguished from B1 by methods designed for volatility clustering, co-movement, and contagion, rather than general-purpose regression techniques.
B6 (Operations & Game-theoretic Modeling)	Formal modeling of optimal or strategic behavior (game theory, optimization/mathematical programming, evolutionary games, mechanism design, simulation).	Defined by a normative or equilibrium orientation, deriving optimal decisions from assumed structures rather than estimating them from observed data.
B7(Behavioral Approaches)	Designs that elicit and measure decision-making (lab/field experiments, survey-based behavioral measurement, behavioral-economics constructs).	Reserved for studies where the behavioral mechanism is the explicit object of measurement; behavioral interpretation of regressions stays in B1/B5.
C (Technical/Engineering Application)	Design or deployment of a technical artifact or system as the primary contribution (blockchain/DLT, IoT monitoring, platform/algorithm design, applied engineering).	Classified here when the primary technical system is the core contribution; if the central result is a predictive model built on its data → B2.
Unclassified	Studies not confidently assignable: (i) purely conceptual/policy papers lacking an identifiable method; (ii) insufficient methodological reporting; (iii) multiple methods without a dominant contribution.	Retained as a distinct residual category rather than being forced into substantive categories, thereby preserving classification integrity and reproducibility.

Note: Abbreviations used in this table are defined in the Abbreviations section at the end of this article.

2.2. Theoretical Foundations of Automatic Classification Design

To ensure sufficient transparency and reproducibility, this study designs a literature automated classification system based on methodological key-term lexicon matching. The

system's design is grounded in the differential classification utility of bibliographic meta-data fields.

Sjögårde et al. [23] systematically evaluated algorithmic labeling methods across different levels of classification hierarchies and found that extracting terms from titles and keywords yielded higher labeling accuracy at the thematic level than methods incorporating abstract text; the latter introduces noise from contextual mentions. This finding is supported by Mateen et al. [24], who demonstrated that while adding abstracts to title-based screening marginally improves precision, using title information alone achieves high screening efficacy. Bramer et al. [25] further noted that restricting search to title and abstract fields in systematic reviews may miss 22–33% of relevant articles, underscoring the trade-off between precision and recall across fields. At the term level, Spärck Jones [26] established the foundational principle of term specificity (IDF): terms appearing in fewer documents carry greater discriminative power and should receive higher weights; Chen and Xiao [27] confirmed that TF-IDF-based keyword selection significantly outperforms pure frequency selection in bibliometric domain analysis; Sjögårde et al. [23] extended this theory by demonstrating that the term frequency–specificity (TFS) ratio outperforms either term frequency or specificity alone. Collectively, these studies provide empirical justification for differentiating high-specificity “core” terms (e.g., “causal forest,” “XGBoost”) from low-specificity “generic” terms (e.g., “machine learning,” “endogeneity”) in automated classification, and for applying differentiated weights that reflect the higher signal-to-noise ratio of titles and keywords relative to abstracts.

In designing the automated classification system, each study's core method is identified through a multi-label classification framework, assigning one primary label; when multiple substantive methods are present, the study may receive secondary labels. The primary label is determined by term evidence weights and the priority bonus mechanism. Term evidence weights are generated by a 2×2 weight matrix that differentiates weights along two dimensions: the term specificity dimension (core vs. generic terms) and the field-level dimension (title/keywords vs. abstract). Studies for which no methodology is identified are retained as an explicit residual category.

The rationale for employing a predefined 2×2 weight matrix rather than TF-IDF- or TFS-based dynamic weight computation is as follows. First, the term lexicon used in this classification system is a small, expert-curated lexicon (approximately 8–94 terms per category; see Supplementary Table S1). The statistical validity of TF-IDF and TFS depends on frequency distribution characteristics across a large number of terms in a large corpus; in a small curated lexicon, the selection of most terms reflects deliberate dictionary construction strategy rather than natural term distribution, causing IDF values to be systematically inflated and lacking in discriminative power. Second, for small, domain-specific lexicons, expert binary classification of terms (e.g., classifying “causal forest” as core and “machine learning” as generic) is more precise than IDF-based statistical estimation; IDF is an indirect statistical proxy for specificity, while expert classification is more direct and accurate when vocabulary size is limited. Third, the predefined weight matrix ensures complete reproducibility of classification results: any researcher using the same lexicon and weight parameters will obtain identical output, while TF-IDF/TFS weights depend on corpus-specific frequency statistics that change with corpus size, temporal scope, or search strategy, hindering cross-study comparability and iterative improvement. Therefore, this study's weighting scheme draws on the design principles of the information retrieval literature (the differentiated treatment of term specificity and field signal-to-noise ratio) without directly adopting its algorithms.

A related boundary issue concerns methodological approaches that are prominent in the broader green finance literature but do not receive independent categories in this

taxonomy: carbon accounting and life cycle assessment (LCA), ESG rating methodology, climate scenario analysis (NGFS/IPCC scenarios), and green bond impact reporting. These approaches are absorbed into B5 (financial econometrics) when they involve quantitative modeling of financial data, or into C (technology application) when they emphasize technical implementation. This design choice reflects the framework's focus on the epistemological and technical dimensions of methodology rather than on domain-specific procedural standards. While LCA and ESG rating methods have distinct procedural traditions, their analytical logic (quantitative data processing for evaluation or assessment) aligns with the broader methodological categories defined here. This is consistent with the boundary-handling principle discussed above: rather than creating additional categories for every domain-specific procedure, the framework treats such procedures as instances of the underlying epistemological and technical dimensions already captured by B1–C. Future research could extend the classification to include domain-specific methodological subcategories where finer-grained regulatory or disciplinary distinctions are needed.

We note that this boundary decision concerns studies whose core methodological contribution is the construction, validation, or comparison of these domain-specific procedures (e.g., proposing a new ESG rating construction method or LCA protocol), rather than the downstream measurement-error implications of using existing ESG ratings or Scope 3 estimates as explanatory variables in studies classified elsewhere in this taxonomy. The latter is addressed separately in Section 5.4 as a data-quality constraint that cuts across, rather than falls within, this classification framework.

3. Materials and Methods

This scoping review follows the PRISMA Extension for Scoping Reviews (PRISMA-ScR) guideline [19,28]. The complete PRISMA-ScR checklist is provided in the Supplementary Materials (Table S2). A review protocol containing the detailed search strategy and classification scheme has been publicly archived on the Mendeley Data platform (Elsevier, Amsterdam, The Netherlands; DOI: 10.17632/rbdph2w7bz.1). As the PROSPERO database does not accept scoping reviews, no formal prospective registration was conducted. All custom code and extended analysis results are archived on the Mendeley Data platform (DOI:10.17632/nt4vsvc3z.1).

3.1. Corpus Construction

The literature data for this study are sourced from the Web of Science (WoS) Core Collection (Clarivate Analytics, Philadelphia, PA, USA), a database widely used in bibliometric and methodological analyses of sustainability research owing to its carefully curated coverage and structured metadata. A single search query was specifically designed to precisely capture the intersection of digital technology and green finance.

Search query: TS = (("big data" OR "artificial intelligence" OR "AI" OR "machine learning" OR "blockchain") AND ("green finance" OR "green credit" OR "green bond*" OR "carbon finance" OR "carbon trading")).

The search was limited to English-language articles, reviews, and early-access publications from 2018 through 30 April 2026. Data retrieval was completed on 30 April 2026. Following duplicate screening (which identified no duplicate records), the final dataset comprised $n = 423$ records (Figure 2). For each record, we extracted the title, abstract, author keywords, Keywords Plus field, and all other bibliographic metadata.

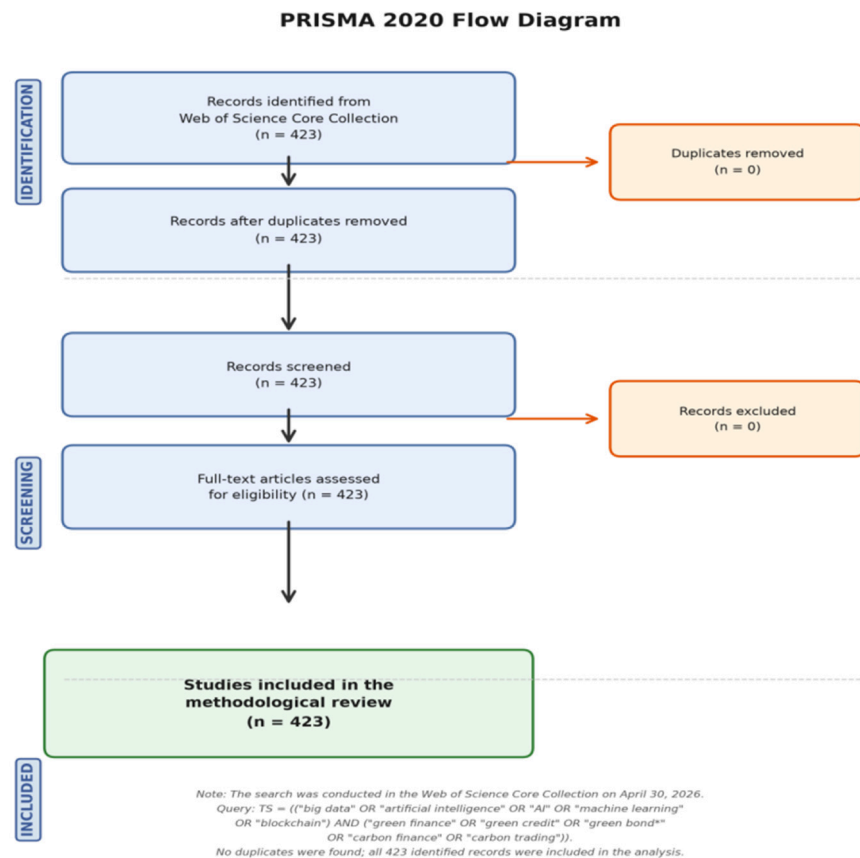


Figure 2. PRISMA 2020 literature screening flow diagram. Note: * indicates a truncation wildcard used in the Web of Science search query (e.g., ‘green bond*’ matches ‘green bond’ and ‘green bonds’).

Across the three stages of the PRISMA workflow (identification, screening, and inclusion), the literature sample remained at $n = 423$. Since this study utilized only the WoS Core Collection as a single database, there were no cross-database duplicates in the retrieval results. The indexing mechanism of the WoS Core Collection itself has already de-duplicated duplicate records within the database; therefore, the number of records after deduplication remained 423. This study’s search strategy employed highly restrictive Boolean logic, requiring the literature to focus on the intersection of two themes: “green finance” and “digital technology.” Additionally, with relatively broad inclusion criteria and an “Unclassified” category within the classification framework to accommodate the marginal interdisciplinary literature, the exclusion of potentially relevant articles with ambiguous thematic affiliation was avoided. After independent dual screening with arbitration, all 423 articles were retained.

3.2. Literature Classification and Coding

3.2.1. Lexicon-Based Multi-Label Classification with a 2×2 Weight Matrix

We employ a lexicon-based multi-label classification scheme to identify the core methodological approach of each article in the corpus. The classification framework determines the primary and secondary methodological categories for each article based on the cumulative weight of methodological terms matched from the article’s metadata against the term lexicon. Each article is counted under its primary methodological category in the distributional analysis. All custom scripts for automated classification, sensitivity analysis, automated sampling, and reliability analysis were written in Python 3.13 (Python Software Foundation, Wilmington, DE, USA).

(1) Match Text Construction and Term Matching

To construct the match text for each article, we extract two text fields from the WoS Core Collection metadata: (1) Title + Keywords, comprising the combination of title (TI), author keywords (DE), and Keywords Plus (ID); and (2) Abstract (AB). Both text fields are converted to lowercase, with one exception: to distinguish 'DID' (Difference-in-Differences) from the English verb 'did', the uppercase form of 'DID' is preserved.

Term matching employs a word-boundary matching mechanism, requiring that text tokens match lexicon terms not only in spelling but also with non-alphanumeric characters on both sides, to prevent substring false matches.

(2) Weight Matrix and Scoring Rules

Each matched term receives a weight according to a 2×2 weight matrix that crosses term level (core vs. generic) with field type (title/keywords vs. abstract): $W1 = 0.70$ (core term in title/keywords), $W2 = 0.30$ (core term in abstract), $W3 = 0.15$ (generic term in title/keywords), $W4 = 0.05$ (generic term in abstract). This ordering ensures that specific and prominently positioned terms dominate article classification, while terms merely mentioned in abstract context exert only a minor influence. When the same term appears in multiple fields, only its highest weight is retained (i.e., term deduplication). Methodological categories with a total cumulative weight below MIN_LABEL_WEIGHT (default: 0.10) are not recorded as method labels for the article. If all categories fall below this threshold, the article is classified as "Unclassifiable" (Unc).

After computing each term's weight, the term weights for the same methodological category are summed to obtain the weight of each methodological category employed in the article, and a priority bonus is then added to each category's weight according to the category's priority. The category with the highest sum of term weight and priority bonus is designated as the article's primary methodology. When a category's weight proportion is $\geq 15\%$ among all categories in the same article, that category is automatically marked as a secondary methodology.

(3) Priority Bonus Mechanism

The priority bonus is a small additive value applied only during the ranking stage, designed to resolve ties when the cumulative term weights of different methodological categories are equal. The priority ordering follows the principle that more specific categories should take precedence over generic ones, reflecting the term specificity principle (IDF) in information retrieval [25].

We evaluated this ex ante using three independent criteria: (a) rarity in general-purpose corpora (few non-field papers use these terms), (b) unambiguous denotation (each term maps to a single technique), and (c) absence from adjacent methodological fields. This ex ante assessment, conducted before running the classification engine on the corpus, ensures that the bonus structure is not an artifact of observed classification results. Causal machine learning (B4) receives the highest bonus (0.40) because its lexicon terms (e.g., "causal forest," "double machine learning") exhibit the highest methodological specificity. The remaining categories receive progressively lower weight adjustments ($B1 + 0.05$, $B2 + 0.02$, $B6 + 0.015$, $B3 + 0.01$, $B5 + 0.005$, $C + 0.00$), reflecting the overlap of their terms with adjacent fields.

The priority bonus affects only the ranking of methodological categories within the same article (i.e., it determines which category is designated as the primary methodology) but does not alter the cumulative term weight of each category or its percentage of the total term weight across all categories.

The term lexicon, weight matrix, priority bonus values, and decision rules are documented in Supplementary Table S1.

(4) Co-occurrence Constraint Mechanism

By default, matched generic terms (low-specificity terms) are assigned weight only when at least one core term of the same methodological category is also matched. This co-occurrence constraint prevents review articles that mention broad terms like “machine learning” from being classified as employing B2 methodology. Exceptions exist: generic terms in B1 (traditional econometrics) such as “endogeneity” and “panel data,” and in C (“blockchain”) possess sufficient methodological specificity to be weighted independently upon matching. For C, the generic term “blockchain” can trigger classification independently, while other C classifications require specific construction or application terms (e.g., “smart contract,” “digital twin,” “fraud detection”).

3.2.2. Regarding B7

Behavioral research (B7) belongs to a different epistemological paradigm from econometrics and machine learning [29]. Given the fundamental differences between paradigms, the automated classification system is designed such that: if specific, high-specificity behavioral terms (e.g., “theory of planned behavior,” “perceived behavioral control,” “intent to invest,” “PLS-SEM,” etc.; see Supplementary Table S1) appear in the abstract, the article is labeled as employing B7 as its sole methodology. This rule is based on two considerations: first, when an article’s abstract uses behavioral terms as its methodological identifier, its research design framework most likely belongs to the behavioral paradigm. Even when econometric tools are simultaneously employed, they typically serve as auxiliary processing of behavioral data rather than independent inference; second, the B7 trigger threshold requires high-specificity terms rather than generic expressions such as “survey” or “questionnaire,” so articles centered on econometric methods with only ancillary questionnaire-based data collection will not trigger this rule. The cost of this rule is that for articles employing B7 methodology, other methodologies are no longer recorded. However, given B7’s extremely low prevalence (see Sections 4.1 and 4.3) and its high-specificity trigger conditions, this trade-off has negligible impact on overall classification diversity.

3.3. Weight Sensitivity Analysis

To verify that the classification results are not artifacts of specific weight parameter choices, we conducted a systematic sensitivity analysis comprising three dimensions. First, a one-factor-at-a-time (OFAT) grid scan varied each of the four weights (W1 to W4) individually across 4 to 5 values (including the baseline), generating 19 configurations. For each configuration, three stability metrics were computed: the proportion of articles with changed primary category (Prim_Changed), changed complete label set (Labels_Changed), and switched primary category among multi-label articles (PrimFlip_MultiLabel). This analysis focuses on the four weight-matrix parameters (W1 to W4) and the label-weight threshold (MinW), and the priority reward value was not included in the OFAT scan range. However, the numerical design of the priority bonus values ensures a certain degree of robustness: since the sum of term weights is always an integer multiple of 0.05, while the bonus values for B2 (+0.02), B6 (+0.015) and B5 (+0.005) are not integer multiples of 0.05, these three types of bonuses mathematically cannot exactly compensate the original weight differences to zero—i.e., they cannot create new ties between originally unequal weights. Only between B4 (+0.40) and B1 (+0.05) does there theoretically exist a possibility that the bonus value difference (0.35) exactly equals the weight difference. However, this scenario requires both categories to have strong evidence and the weight difference to be exactly 0.35, which is extremely rare in practice.

We acknowledge this as a limitation of the current sensitivity analysis design.

Second, six configurations were designed specifically to test the impact of the “W2 > W3” ordering assumption on classification results. A dimensional conflict exists between W2 and W3: the term precision dimension requires W2 to be higher, while the field position dimension requires W3 to be higher. The six configurations span three comparison strategies: order reversal, equal-weight settings, and extreme ratios. Additionally, for the baseline and reversed configurations, an article-by-article comparative analysis was conducted to identify all studies whose primary classification or label set changed when the W2/W3 ordering was reversed.

Third, eight comprehensive multi-parameter scenarios (including the baseline configuration) were constructed to simultaneously perturb multiple weight parameters, simulating different classification strategy orientations, including core-term high-weight dominance, core-term low-weight, title/keyword dominance, abstract dominance, four-weight equality, W2/W3 reversal, and an elevated minimum label weight threshold (MinW = 0.20).

All evaluation metrics are computed by internally comparing each perturbed configuration against the baseline configuration, requiring no external reference standard.

The specific steps of sensitivity analysis are provided in the Supplementary Material (README_Methodology_Classification_and_Sensitivity_Analysis(Multilabel).txt).

3.4. Reliability and Algorithm Validation

Automated literature classification is performed by a lexicon-matching engine based on the classification parameters listed in Supplementary Table S1 (key-term lexicon, 2×2 weight matrix, priority bonuses, and decision rules). To assess reliability, two researchers independently coded a random sample of 85 studies (approximately 20% of the 423-article sample set); coding disagreements were resolved through third-party adjudication to form the gold standard [29]. Two comparisons were then evaluated: (1) Coder A vs. Coder B (inter-coder reliability); (2) automated algorithm vs. gold standard (algorithm validity). For each comparison, we computed raw agreement, Cohen’s κ with 95% bootstrap confidence intervals ($B = 5000$ percentile resamples), per-category precision/recall/F1, and the complete confusion matrix.

The classification parameters (lexicon, weight matrix, priority bonuses, and decision rules) were finalized prior to the reliability assessment. The 85-article gold standard was drawn independently from the same 423-article corpus after parameter fixation, and no iterative parameter adjustment was performed based on gold standard results. This design ensures that the reliability metrics reflect genuine out-of-sample agreement rather than in-sample fit.

4. Results

4.1. Methodological Distribution of the Research Field

The 423 articles in the corpus are automatically classified by primary research method as summarized in Table 4.

The automated classification results indicate that traditional econometrics remains the most widely employed primary method, accounting for 26.95% of the entire dataset. Predictive machine learning (B2, 17.26%) is the second-largest independent category, reflecting the growing adoption of data-driven approaches. Causal machine learning accounts for 9.69%, indicating that this conceptual frontier has been preliminarily established but still has significant room for application growth.

Interpretability methods (B3) cover 20 studies in the corpus (4.73%), of which 6 have B3 as the primary label (1.42%) and 14 as a secondary label. Among studies with B3 as a secondary label, the primary labels predominantly belong to B2 (10 articles), B6 (2 articles), B1 (1 article), or B5 (1 article), indicating that interpretability analysis in most cases appears

as an auxiliary dimension of predictive modeling. Upon manual review, the 6 articles with B3 as the primary label all employ interpretability techniques as ancillary components attached to established predictive or causal models, serving as model validation or results presentation, rather than constituting the core contribution of the interpretability method itself.

Table 4. Distribution of Methodological Approaches ($n = 423$).

Rank	Code	Category	n	Share (%)
1	B1	Traditional Econometrics	114	26.95
2	B2	Predictive ML	73	17.26
3	C	Technical Application	69	16.31
4	B4	Causal ML	41	9.69
5	B6	Operations & Game-theoretic Modeling	35	8.27
6	B5	Financial Econometrics	33	7.80
7	B3	Interpretability	6	1.42
8	B7	Behavioral Approaches	4	0.95
9	Unc	Unclassified	48	11.35
Total:			423	100.00

Note: The classification distribution in Table 4 is based on automated assignment of the 423 articles. An 85-article gold standard (dual-coder manual coding) achieved an 80.0% agreement rate with the automated classification algorithm ($\kappa = 0.76$, 95% CI [0.65, 0.85]; see Section 4.3). Readers should therefore interpret the percentages above as estimates with an approximate 20.0% misclassification margin at the category level, rather than as exact values. Abbreviations used in this table are defined in the Abbreviations section at the end of this article.

Operations research/game theory (B6, 8.27%) slightly exceeds financial econometrics (B5, 7.80%), indicating that mathematical optimization and game-theoretic analysis have some application in this field, while financial econometrics appears relatively underutilized given the finance-oriented nature of the domain. Behavioral analysis (B7, 0.95%, $n = 4$) is exceedingly rare. Technology application (C) accounts for 16.31%.

Articles that could not be assigned any methodological label (Unclassified) account for 11.35%. To assess the potential impact of these articles on the methodological distribution, we conducted a structural-quantitative analysis of all 48 unclassified studies. In terms of publication type, approximately 54% ($n \approx 26$) are reviews, policy analyses, or perspective pieces whose methodological contribution consists primarily of thematic synthesis rather than empirical analysis; excluding such articles introduces no bias into the empirical methodological distribution. The remaining approximately 46% ($n \approx 22$) contain no methodological terms from the lexicon, using only extremely generic expressions such as “analysis,” “approach,” and “data” to describe their methods, a characteristic reflecting a severe simplification of methodological description. This composition indicates that unclassified articles are not randomly missing along the methodological dimension but are systematically concentrated in two types: non-empirical reviews and policy discussions, and articles whose methodological descriptions are insufficiently specific for term-evidence-based identification. This finding is revisited in the discussion of methodological transparency in the field (Section 5.4).

In addition to the aforementioned primary findings, it was also observed: among the 423 studies, 111 (26.24%) were annotated with at least one secondary methodological label in the Secondary_Methods field, indicating that the simultaneous adoption of multiple methodologies is a non-trivial phenomenon. The most common methodological co-occurrence patterns include: B4 + B1 (causal machine learning combined with traditional econometrics, 28 papers), B6 + C (operations research/game theory combined with technology application, 22 papers), and B2 + B3 (predictive machine learning combined with interpretability methods, 15 papers). Among the 41 studies with B4 as the primary method,

20 (48.8%) listed B1 as a secondary method, reflecting the field’s trend of integrating causal inquiry with traditional econometric analysis.

Complete multi-label assignment results (including primary labels, secondary labels, and weight proportions) are provided in the Supplementary Materials (literature_classification(Multilabel).xlsx).

4.2. Sensitivity Analysis Results

The sensitivity analysis confirms the robustness of the classification results to weight parameter perturbation. In the OFAT grid scan, the maximum primary-category change rate was 7.6% (at $W3 = 0.05$), with most changes attributable to C-class articles shifting to Unclassified due to excessively low generic-term weights; excluding this extreme setting, the maximum change rate was 4.7% (at $W4 = 0.10$ or 0.15), and most parameter changes yielded change rates below 2.0%. The $W2$ and $W3$ ordering test showed a maximum change rate of 5.4% (when $W3 \gg W2$), indicating that the $W2 > W3$ ordering assumption has some impact on results but remains overall robust. Among the eight combined scenarios, the most extreme test was the elevated label threshold ($MinW = 0.20$), producing a primary-category change rate of 10.6%; the four-weight equality scenario ($W1 = W2 = W3 = W4 = 0.25$) yielded a change rate of 9.7%. Given the extreme nature of these configurations, these values remain at acceptably low levels. Parameter impact ranking shows $W3$ (range = 7.6%) as the most influential parameter, followed by $W4$ (1.2%), $W2$ (1.2%), and $W1$ (0.9%). Complete sensitivity analysis results are provided in the Supplementary Materials (weight_sensitivity(Multilabel).xlsx).

4.3. Classification Reliability and Algorithm Validation

The reliability analysis (Table 5) demonstrates near-perfect inter-coder reliability: Cohen’s κ between Coder A and Coder B was 0.93 (95% CI [0.86, 0.99], agreement = 94.1%). The automated algorithm vs. gold standard comparison yielded $\kappa = 0.76$ (95% CI [0.65, 0.85], agreement = 80.0%, weighted F1 = 0.79), indicating substantial agreement [30]. Categorical analysis indicates that the algorithm performs best in category B5 (precision = 1.00, recall = 1.00) and also performs well in categories B4 (precision = 0.89, recall = 0.89) and B1 (precision = 0.88, recall = 0.85). Category-level detection issues and their root causes are summarized in Table 6.

Table 5. Reliability Analysis Results.

Confusion Matrix Reliability Analysis—Summary								
Comparison	<i>n</i>	Agreement %	Cohen’s κ	κ 95% CI (Lo)	κ 95% CI (Hi)	κ Interpretation	Accuracy %	Weighted F1
Coder A vs. Coder B	85	94.1%	0.929	0.858	0.986	Almost Perfect	94.1%	0.940
Automatic script vs. Gold	85	80.0%	0.759	0.654	0.855	Substantial	80.0%	0.792

Note: κ = Cohen’s kappa, a measure of inter-rater agreement (0.81–1.00: almost perfect; 0.61–0.80: substantial; 0.41–0.60: moderate; 0.21–0.40: fair; <0.20: poor).

Table 6. Analysis of Category-Level Detection Issues in the Automated Classification.

Category	Performance Issue	Error Pattern	Root Cause	Supporting Evidence
B3	Precision = 0, Recall = 0	No true positives in the 85-article sample	Sample limitation, not algorithmic deficiency: the random sample included no article with B3 as the primary label. B3 is assigned a low priority bonus (+0.01, below B1's + 0.05 and B2's + 0.02) based on three theoretical considerations—epistemological second-order status, methodological hierarchical dependency, and logical non-independence—causing it to typically occupy a subordinate position in weight competition (Section 2.1).	Manual review of all 20 algorithm-assigned B3 articles (6 primary + 14 secondary): 18 papers (90%) were confirmed to employ post-hoc interpretability methods (e.g., SHAP, LIME, and feature importance analysis); the remaining 2 adopted inherently interpretable models (e.g., decision trees, generalized additive models), which also fall within the B3 category but achieve interpretability through model transparency rather than post-hoc explanation.
B7	Insufficient sample ($n = 2$)	Category-level metrics unreliable	Category rarity: only 4/423 articles (0.95%) in the full corpus, yielding 2 articles in the 85-article adjudication sample—too few for reliable metric computation.	Manual inspection confirms all B7 articles are studies with behavioral methods as the primary design, with no instances where the core method is a different approach but was forcibly overridden as B7.
B6	Recall = 0.57 (false negatives)	Genuine B6 papers undetected	Horizontal term overlap: B6's core and generic terms exhibit cross-category polysemy—"optimization" spans B6 (operations research), B5 (portfolio optimization), and B2 (hyperparameter optimization); "equilibrium" spans B6 (Nash equilibrium) and B1/B5 (market equilibrium). Given B6's low priority bonus (+0.015, below B1's + 0.05 and B2's + 0.02), genuine B6 papers are outcompeted in weight ranking by homologous terms from B5 or B1 (Section 3.2.1).	Game-theoretic papers using generic terms in their title or abstract have their weight diluted by homologous terms from B5 or B1, resulting in some genuine B6 papers going undetected.
C	Precision = 0.70 (false positives)	Non-methodological articles assigned to C	Blurred vertical classification boundary: the C lexicon contains only specific technology terms (e.g., "smart contract," "digital twin") and no generic term list (Section 3.2.1). Technology-themed review or policy articles contain these terms but their core contribution is thematic synthesis rather than methodological practice; the algorithm detects the technology terms and assigns C, while gold-standard coders judge them as not belonging to any methodological category.	Lack of methodologically distinguishing features between empirical technology-application studies and technology-themed review articles.
B2	Precision = 0.57 (lowest), Recall = 1.00 (perfect)	Over-classification (6 false positives)	ML terminology (e.g., "machine learning," "neural network," "deep learning") has become ubiquitous across methodological traditions in finance research: articles whose core contribution is statistical inference (B1), technology application (C), or mathematical modeling (B6) frequently invoke these terms as auxiliary tools or future research directions. The B1–B2 boundary is the most blurred interface in the classification system, as both traditions share foundational vocabulary ("regression," "prediction," "model selection") while differing in epistemological orientation.	6 false positives from B1 (2), C (2), B6 (1), Unc (1); all genuine B2 articles correctly identified.

Note: Abbreviations used in this table are defined in the Abbreviations section at the end of this article.

Complete category-level metrics and the full confusion matrix are provided in the Supplementary Materials (reliability_analysis(Multilabel).xlsx).

5. Discussion

5.1. Implications of the Methodological Landscape

We analyze this methodological landscape along four themes: the data environment and methodological choice, the causal–predictive fusion paradigm, the structural absence of behavioral finance, and evidence governance. These implications are based on the frequency distribution of methodological labels, which reflects how often each approach appears in the literature but does not directly assess the internal identification credibility of individual studies (e.g., whether B1 studies report parallel trend tests or exclusion restriction tests). The following implications should therefore be read as frequency-level inferences rather than causal claims about evidence quality.

5.1.1. Data Environment and Methodological Choice

The digitalization of the financial sector has generated high-frequency, high-dimensional data (including IoT and sensor data streams, ESG disclosure information, and text sentiment analysis results), whose characteristics fundamentally challenge the applicability of traditional estimation methods. Sensor data exhibits high frequency and low signal-to-noise ratio, violating the stable, low-dimensional error structures upon which standard inference relies [16]; ESG disclosure information constitutes unstructured text that cannot be naturally mapped to a fixed regression matrix as assumed by structural econometrics [31,32]. The rise of machine learning methods (predictive ML at 17.26% plus causal ML at 9.69%, totaling 26.95%) should be understood as a technological imperative driven by changes in the data environment.

5.1.2. The Causal–Predictive Fusion Paradigm

Green finance research simultaneously faces demands for causal evidence (policy evaluation) and predictive accuracy (risk management). As the Network for Greening the Financial System (NGFS) framework explicitly states, climate stress testing and scenario analysis favor models capable of predicting system-wide behavior under unobserved states, prioritizing out-of-sample generalization over individual structural coefficients [33]. In China, the climate risk stress testing conducted by the People’s Bank of China for financial institutions vividly exemplifies this forward-looking, system-level regulatory orientation [34]. By contrast, green bond certification and green credit eligibility determination are confirmatory tasks whose decision rules require far greater transparency than predictive accuracy [35]. Approximately 48.8% of causal ML primary studies include a traditional econometrics secondary label; B4 + B1 and B2 + B3 are the most common co-occurrence patterns; 26.24% of studies employ auxiliary methods, indicating that the field is adopting a “predictive + causal” dual-track fusion approach in response to these dual demands.

5.1.3. Structural Absence of Behavioral Finance

The core issues of green finance (green premium formation mechanisms, investor ESG preference heterogeneity, green credit demand elasticity, and carbon trading market participant behavior), are fundamentally behavioral finance questions. Yet behavioral research accounts for only 0.95% ($n = 4$), suggesting that behavioral experiments and survey designs are infrequent in this field. In the broader green finance literature, these behavioral questions are typically investigated through stated-preference surveys (e.g., eliciting investors’ willingness to pay for green assets) and laboratory experiments examining green investment decisions under controlled conditions. These methods fall within the B7 (behavioral approaches) paradigm as defined in this framework, yet their near-absence in the current corpus suggests that the digital technology dimension of green finance research has

not yet integrated the behavioral paradigm that animates much of the broader sustainable finance field [7,36].

5.1.4. Evidence Governance

Green finance regulation, faced with diverse regulatory tasks, requires different types of empirical evidence: climate stress testing under NGFS scenarios demands forward-looking predictive capability [33], EU Taxonomy alignment assessment requires confirmatory classification methods [37], TCFD-aligned disclosure evaluation requires standardized reporting frameworks [38], and green bond certification requires confirmatory methods [35]. The “green finance–emission reduction nexus” identified through ML cannot serve directly as causal evidence for policy intervention. The relatively low proportions of causal ML (9.69%) and interpretability methods (4.73%) suggest that the current methodological landscape may not be fully aligned with the diverse evidence demands of institutional regulation [36,39,40].

5.2. Implications for Practice

For researchers, first, when selecting a methodology, one should begin by assessing the characteristics of the data environment (structured/unstructured, dimensionality, frequency), allowing data characteristics to drive method selection, rather than choosing a method first and then fitting data to it. When confronting unstructured ESG text or high-frequency sensor data, researchers should proactively employ methods capable of handling high-dimensional nonlinearity rather than forcing the assumptions of traditional econometric frameworks. Predictive performance alone reflects only association and cannot answer causal questions of policy relevance; researchers are advised to integrate quasi-experimental or causal machine learning designs into predictive studies; for example, combining high-dimensional learning models with double machine learning or causal forests. This approach both leverages the flexibility afforded by massive digital finance data [41,42] and enables precise estimation of heterogeneous treatment effects. We recommend applying behavioral experimental designs (e.g., survey experiments, natural experiments, field experiments) to green finance research to fill the behavioral foundation gap in topics such as green premium formation mechanisms and investor ESG preference heterogeneity. Greater attention to interpretability methods themselves may be warranted; although this direction is currently dominated by dependent applications, it may still represent a relatively underexplored avenue for future research. Researchers should also be alert to how domain-specific measurement issues interact with method choice: for example, B1 regression studies using ESG scores as explanatory variables face attenuation bias from cross-provider rating divergence, while B2 predictive ML studies may inadvertently learn rating-agency heuristics rather than underlying green performance (see Section 5.4 for a fuller discussion of these data-quality constraints). Second, our analytical approach’s inability to classify certain studies is itself an important finding. This partly reflects overly abbreviated methodological descriptions that cannot be identified through titles, keywords, and abstracts alone. As a constructive improvement, we recommend that researchers place concise, standardized method descriptions in the title, keywords, or abstract rather than only in the body text. Such concise reporting would make core research methods easily retrievable and integrable in the literature. Transparent reporting is essential for ensuring research reproducibility and for promoting knowledge accumulation in this rapidly developing field [43].

For policymakers, first, the methodological source of research conclusions should determine how they are applied. Predictive models provide evidence capable of identifying patterns and supporting early warning and screening functions, but such evidence alone

cannot prove the causal effects of green finance interventions. Conclusions intended to justify policy adjustments should be based on research methods specifically designed for causal identification (e.g., quasi-experimental methods or causal machine learning) rather than relying solely on predictive accuracy [44,45]. Distinguishing between predictive and causal evidence helps avoid attributing mere correlation to policy effectiveness. Second, when algorithms inform consequential decisions (e.g., green credit scoring, climate risk classification, or transition finance allocation), interpretability should be a regulatory requirement rather than an optional add-on. Models deployed in such contexts should reveal the key features influencing their outputs through interpretable design or post-hoc explanation, thereby providing stakeholders with a basis for decision-making, enabling regulators to check for bias, and ensuring that decisions are contestable [46]. Incorporating interpretability into the green finance governance system can both enhance the efficiency of data-driven decisions and satisfy the accountability mechanisms required for public-interest decision-making. When designing green finance incentive policies, the actual behavioral patterns of investors and enterprises (e.g., inertia, herding, myopia) should be considered rather than relying solely on model-derived optimal policy pathways. When constructing green finance data infrastructure (e.g., ESG databases, carbon trading data platforms), the implications of data format and granularity for subsequent research methods should be considered: structured, standardized data disclosure formats can lower the methodological threshold for researchers.

5.3. A Case-Based Comparison of Identification Credibility

The technology application category (C) encompasses internally heterogeneous approaches whose regulatory implications differ substantially: blockchain smart contracts (automating compliance and verification), IoT monitoring (real-time environmental data collection), satellite remote sensing (verifying green project outcomes), and algorithmic credit scoring (assessing green credit eligibility). While the current framework groups these under a single category because they share the commonality of applying digital technology without advancing a distinct methodological paradigm, their policy implications diverge. For instance, the interpretability requirement for algorithmic credit scoring is far more stringent than for satellite remote sensing, given the former's direct impact on credit access. Future research could subdivide C into monitoring technologies, transaction technologies, and analytical technologies to better capture these differential regulatory implications.

This section illustrates how the automated classification system supports a second analytical dimension: the examination of identification credibility, rather than validating the classification system itself. The classification framework identifies the methodological paradigm employed by a study, from which one can further ask: Do studies under different paradigms exhibit systematic differences in evidence quality and robustness? To address this question, this study draws on the concept of bias risk assessment from evidence-based medicine and employs a four-dimensional robustness testing framework to conduct a structured assessment of four representative studies selected through the classification system, including placebo tests [P], alternative specification tests [A], subsample sensitivity tests [S], and endogeneity/selection tests [E] (Table 7). The four selected studies are drawn from the B1 (traditional econometrics) and B4 (causal machine learning) categories to represent the robustness of research designs under both explanatory and predictive paradigms. It should be clarified that components such as placebo tests and endogeneity tests are integral parts of the specific studies' identification strategies; this section assesses the completeness of the "specific study samples filtered by the classification system" on these identification strategies, not the validity of the classification system itself. The findings are illustrative rather than inferential: based on only four studies, they serve as concrete

examples of how studies selected by the classification system present patterns along the evidence quality dimension, not as statistically generalizable conclusions about the field, nor as validation of the classification system’s effectiveness.

Table 7. Structural assessment of identification credibility across representative B1 and B4 studies (methodological efficacy probe).

Study	[39]	[40]	[47]	[48]
Category	B1	B1	B4	B4
Research question	Impact of Carbon Information Disclosure Quality on Green Bond Financing Costs	Impact of Repeated Green Bond Issuance on Corporate Capital Costs (Reputation Effect)	Heterogeneous Causal Effects of Green Innovation on Tobin’s Q and ROA	Causal Effects of Green Finance Investment Intensity on Entrepreneurship (NBR), Employment (Em), and Energy Transition/Just Transition (ANE), and Cross-City Heterogeneity
Core stimand	ATE	ATT	CATE, HTE	CATE, ATE
Identification strategy	Fixed Effects Panel Regression (Two-way FE panel)	Difference-in-Differences	Causal Forest (GRF, Athey-Wager), Rubin Framework + Quasi-Natural Experiment (2012 Green Credit Guidelines)	Causal Forest (GRF), Selective Observability Framework + City/Year Fixed Effects
Key identifying assumption	Strict exogeneity	Parallel trends	Unconfoundedness (Conditional Independence) + Overlap (Figure 1 Propensity Score $\eta < e(x) < 1 - \eta$, Common Support Satisfied)	Unconfoundedness (Conditional Independence) + Overlap (Diagnostic Table 1 Confirms Common Support)
Confounding treatment	Firm FE, Industry FE, Low-dimensional Covariates (Leverage/Size, etc.)	Firm FE, Time FE, Low-dimensional Covariates (Bond Characteristics, Liquidity, etc.)	High-dimensional (15 covariates: Agency Cost/Operating Cost/State Ownership/Leverage/Size, Assets/Equity Concentration, etc.)	Low-dimensional covariates (~5: CRW, ATE, CDE/LCDE, ANE) + City FE + Year FE (Not high-dimensional)
Robustness	[P] $\times$	✓ (Placebo/Falsification Test)	$\times$	✓ (Placebo/Falsification Test)
	[A] ✓ (Alternative Variable/Model Specification)	✓ (Alternative Variable/Model Specification)	✓ (Alternative Variable/Model Specification)	✓ (Alternative Variable/Model Specification)
	[S] ✓ (Subsample/Outlier Exclusion)	✓ (Subsample/Outlier Exclusion)	✓ (Subsample/Outlier Exclusion)	✓ (Subsample/Outlier Exclusion)
	[E] $\times$	✓ Endogeneity/Selection Test: IV)	$\times$	$\times$
Heterogeneity	No	Yes	Yes	Yes
Score	2	5	2	4

Note: This four-study evaluation is illustrative, not inferential. With $n = 4$, Table 7 scores are worked examples, not summary statistics. The purpose is to show that the taxonomy can support a second axis (identification credibility) via the Cochrane checklist. This reinforces our central claim: credibility must be assessed study-by-study through design, not inferred from a methodological label. A full-scale credibility assessment across the corpus is underway as a separate study. Abbreviations used in this table are defined in the Abbreviations section at the end of this article.

The evidence quality characteristics shown by the selected cases (Table 7) reveal the following: First, the two B4 studies and two B1 studies display structural differences in

handling heterogeneity: B4 causal forest analyses incorporate high-dimensional covariates (e.g., one study includes fifteen firm-level characteristics and employs propensity score overlap diagnostics), while B1 studies rely on low-dimensional covariates and strong exogeneity or parallel trends assumptions. This case-level comparison suggests that, among the selected studies, causal machine learning demonstrates superior capacity for handling high-dimensional heterogeneity compared to traditional econometric methods, an observation consistent with the existing consensus in the causal inference literature regarding the advantages of high-dimensional controls [20,21]. Second, among the selected cases, the completeness of identification strategies does not increase monotonically with methodological novelty: the most comprehensive of the two B1 studies employed a difference-in-differences design (covering all four components, score = 5); while the two B4 studies, despite precisely capturing heterogeneity, neither implemented placebo tests nor endogeneity/selection tests, exhibiting a consistent pattern in this sample of strong diagnostic identification capability but weak self-falsification capability. These case observations further reinforce an important argument from the methodological literature: a method's capacity to address its core research challenge depends on whether it satisfies the relevant identification assumptions [20,49], not on its methodological label; identification credibility derives from the research design itself, not from technical novelty [45]. Accordingly, the method application distribution profile presented in this paper reflects the frequency and structural distribution of methodological paradigms in the corpus; it does not constitute a ranking of methodological quality, nor should it be used to infer that one methodological category is inherently superior to another; the differences presented in this section exist solely within the four-case sample, illustrating the analytical entry point that the classification system provides for subsequent identification credibility analysis, not serving as proof of the classification system's own validity.

5.4. Limitations

This study has several limitations, which can be grouped into four broad categories.

(1) Subjectivity and discretization inherent in the classification system.

The classification framework rests on several researcher-determined design choices. The priority bonus values and the weight matrix more broadly embed researcher judgment about which method should be treated as primary; alternative weighting schemes are examined in the sensitivity analysis (Section 4.2), and the weight matrix could benefit from future supervised-learning calibration rather than purely judgment-based design. Converting a continuous methodological landscape into discrete categories also produces forced boundary assignments for studies positioned at category breakpoints; this is an inherent cost of discretization, and its practical magnitude is bounded by the sensitivity analysis, where the maximum primary-category change rate across all perturbation scenarios was 10.6%. Relatedly, interpretability methods (B3) operate at a different epistemological level from the other categories—addressing “why did the model produce this output” rather than a substantive research question—yet are granted equal competitive standing to keep studies with interpretability as their core contribution visible. Manual review found that all six B3-primary articles in this corpus were in fact B1/B2 applications in which interpretability served only as a validation step, indicating that frequency- and co-occurrence-based matching cannot reliably separate “interpretability as validation” from “interpretability as core contribution” for hierarchically dependent categories such as B3.

(2) Lexicon coverage and corpus scope.

Several limitations stem from the boundaries of the lexicon and the corpus. 11.35% of studies ($n = 48$) could not be reliably classified: approximately 54% of these are non-

empirical reviews or policy pieces whose exclusion introduces no bias into the empirical distribution, while the remaining ~46% contain no lexicon-matched methodological terms in their abstracts, which may lead to mild underestimation of certain categories. This partly reflects the fact that classification relies solely on titles, keywords, and abstracts, without full-text information, and that the term lexicon requires continuous iteration as new methods emerge. The corpus is also restricted to the WoS Core Collection with a fixed set of search terms (2018–April 2026), a choice that prioritizes precision and reproducibility over exhaustive coverage; such omissions are expected to affect topic coverage more than the relative distribution of methodological categories.

(3) Reliability validation design.

The reliability assessment has two design limitations. B3 and B7 are weak-reliability categories: the 85-article simple random sample happened not to cover any article with B3 as the primary label, and B7 had only 2 articles in the adjudication sample, making category-level reliability metrics impossible to compute effectively—exposing a systematic inadequacy of simple random sampling for rare categories such as B3 (6 articles) and B7 (4 articles); future research should adopt stratified sampling by methodological category to ensure adequate coverage of rare categories. In addition, reliability was assessed on primary labels only; inter-coder agreement and algorithm validity for secondary labels were not separately evaluated. Because secondary labels are supplementary and do not drive the main distributional findings, this does not affect the core conclusions, but future research should extend reliability evaluation to the full multi-label assignment using metrics such as subset accuracy or Hamming loss.

(4) Domain-specific data-quality challenges in green finance, external to the classification framework.

Beyond the classification system itself, green finance research is subject to first-order data-quality constraints independent of method choice. At the conceptual-definitional level, no universally accepted definition of “green” exists—the EU Taxonomy, China’s Green Bond Endorsed Projects Catalogue, and the Climate Bonds Standard apply divergent eligibility criteria, limiting comparability across studies using different green-screening frameworks [35,36]; and greenwashing detection, which requires textual analysis of disclosure divergence between claimed and actual environmental performance, still lacks a validated lexicon for automated identification. At the measurement-error level, Scope 3 carbon emissions data—a key independent variable in many green finance studies—suffer from estimation-based measurement error that may bias regression coefficients toward attenuation, while ESG rating divergence across major providers (MSCI, Sustainalytics, Refinitiv), with pairwise correlations as low as 0.38–0.71, introduces upstream measurement error into any regression using ESG scores as explanatory variables [31]. These challenges cannot be resolved by refining the classification system itself; they call for domain-specific safeguards—multi-taxonomy robustness checks, greenwashing-aware textual controls, and errors-in-variables corrections for Scope 3 and ESG measures—that lie outside the scope of this study.

In conclusion, these limitations point to clear directions for future work: expanding the lexicon to include full-text terms, testing additional weight configurations, incorporating embedding- or LLM-assisted classification to reduce unclassified cases and refine boundary decisions, adopting stratified sampling to improve reliability assessment for rare categories, extending reliability evaluation to multi-label assignments, and developing domain-specific measurement corrections for the data-quality challenges noted above—while preserving the transparency and reproducibility that the current lexicon-based approach provides.

6. Conclusions

6.1. Research Background and Questions

The intersection of green finance and digital technology has given rise to a rapidly expanding body of empirical research, yet the methodological practices underpinning this knowledge base have long lacked systematic assessment. Existing reviews have adopted a topic-oriented organizing logic—synthesizing the literature around core themes, technological dimensions, or practical challenges—without incorporating methodology itself as an analytical dimension. Consequently, while the field can readily answer “what has been studied,” it cannot yet answer “how research has been conducted” or whether the methods employed are commensurate with the causal and explanatory claims being made.

This study addressed that gap by posing three research questions: What is the distributional pattern of methodological categories—traditional econometrics, predictive machine learning, causal machine learning, financial econometrics, operations research/game theory, interpretability methods, behavioral research, and technology application—across the green finance–digital technology nexus? Can a lexicon-based multi-label automated classification scheme reliably annotate the methodological approaches of the literature in this field? What implications does this methodological landscape hold for researchers’ method selection practices, policymakers’ evidence-use standards, and the overall quality of knowledge production in the field?

6.2. Research Methods

To address these gaps, this study designs a literature classification framework centered on methodology, situating the methodological practices of the green finance–digital technology domain within a two-dimensional framework defined by the epistemological objectives (explanatory vs. predictive) and the technical toolkit (traditional econometrics vs. machine learning), encompassing six methodological categories (B1–B6) plus two special categories: technology application (C) and behavioral research (B7). On this basis, a lexicon-based multi-label automated classification system is constructed, employing a 2×2 weight matrix to quantify methodological evidence strength, with a priority bonus mechanism to resolve ranking ties. Algorithmic validity is verified against a dual independent coding gold standard (inter-coder $\kappa = 0.93$; algorithm vs. gold standard $\kappa = 0.76$, $F1 = 0.79$), and result robustness is confirmed through weight sensitivity analysis (maximum OFAT change rate 7.6%; maximum combined extreme scenario 10.6%), culminating in a systematic scoping review of 423 articles.

6.3. Key Findings

The automated classification results reveal: (1) Traditional econometrics dominates at 26.95%, indicating that the field still rests on classical regression analysis as its foundation. (2) Predictive machine learning (17.26%) ranks second, reflecting the growing adoption of data-driven approaches. (3) Causal machine learning accounts for 9.69%, indicating preliminary establishment but with considerable room for application growth. (4) Interpretability methods (B3) cover 20 studies (4.73%), of which 6 have B3 as the primary label, but upon review all are ancillary components attached to predictive or causal models, indicating that this field severely lacks methodological research with interpretability algorithms as the core contribution. (5) Operations research/game theory (8.27%) slightly exceeds financial econometrics (7.80%), constituting two mid-sized branches. (6) Technology application (16.31%) has a relatively high share, reflecting that a considerable proportion of the literature in this field contributes through technology application rather than methodological innovation, with some articles using only extremely generic terms such as “analysis,” “approach,” and “data” to describe their methods, whose methodological ambiguity affects reproducibility

and evaluability to some extent. (7) Behavioral research (B7) accounts for only 0.95%, revealing that micro-level behavioral mechanism research is virtually absent. (8) Over one-quarter of studies (111 articles, 26.24%) employ at least one secondary methodological approach, confirming that cross-methodological combination has become a non-negligible practice pattern.

6.4. Theoretical Contributions

The contributions of this paper are as follows: First, it constructs the first methodology-centered literature classification system for the green finance–digital technology domain, proposing a two-dimensional framework, spanning the epistemological objectives and the technical toolkit, that provides a conceptual foundation for mapping methodological landscapes. Second, it designs and validates a lexicon-based multi-label automated classification scheme featuring a configurable weight matrix, transparent priority mechanism, and dual-validated reliability metrics, with reproducible, auditable, and extensible characteristics. Third, through systematic methodological landscape mapping, it reveals the data-environment-driven methodological transition, the rise of the causal–predictive fusion paradigm, the structural absence of behavioral research, and the imbalance between methodological supply and institutional regulatory demands, providing a baseline reference for tracking future methodological evolution.

6.5. Practical Implications

Based on the above findings, this paper offers recommendations at two levels: for researchers and for policymakers.

For researchers: First, methodological selection should follow the data-characteristics-driven principle: when facing unstructured ESG text or high-frequency sensor data, high-dimensional nonlinear methods should be employed rather than traditional econometric frameworks; second, quasi-experimental or causal machine learning designs should be integrated into predictive research to balance flexibility with causal inference needs; third, behavioral experimental designs should be introduced into green finance research to fill the behavioral foundation gap; and fourth, transparent reporting practices should be adopted, placing method descriptions in titles, keywords, or abstracts for easy identification and integration.

For policymakers: First, the methodological source of research conclusions should determine how they are applied, avoiding the conflation of correlation with policy effectiveness; second, when algorithms inform consequential decisions such as green credit scoring or climate risk classification, interpretability should be established as a regulatory requirement; third, incentive policy design should incorporate actual behavioral patterns of investors and enterprises; and fourth, data infrastructure construction should fully consider the implications of data format and granularity for research methodology.

6.6. Limitations and Future Directions

This study has four main limitations. First, the classification system embeds researcher judgment in its weight matrix rather than pure data-driven optimization, and discretization forces boundary assignments (up to 10.6% category shift in sensitivity tests); interpretability methods (B3) also sit at a different epistemological level yet compete equally with substantive categories, occasionally conflating validation use with core contribution. Second, 11.35% of studies remain unclassified due to lexicon gaps, and classification relies on titles, keywords, and abstracts rather than full text, within a corpus restricted to WoS Core Collection with fixed search terms. Third, random sampling failed to adequately cover rare categories (B3, B7) for reliability testing, and validation covered primary labels only, not multi-label assignments. Fourth, green finance research faces external data-quality chal-

allenges independent of this framework, including inconsistent “green” definitions across taxonomies, missing greenwashing-detection lexicons, and measurement error in Scope 3 emissions and divergent ESG ratings.

Future work should expand the lexicon, incorporate LLM-assisted classification, adopt stratified sampling, extend reliability testing to multi-label assignments, and address domain-specific measurement issues.

The sustainable development of green finance depends on high-quality empirical evidence, not only requiring causal credibility of evidence but also demanding that the methodological process generating such evidence be traceable and reproducible. The methodological classification system constructed in this paper provides the field with a framework tool for systematically organizing the literature, diagnosing the methodological landscape, and guiding research practice. As it undergoes continuous iteration and deep integration with human–machine collaborative intelligent annotation technologies, it can evolve into infrastructure for research methodology governance.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/su18168053/s1>, The raw bibliographic metadata analyzed in this study were retrieved from the Web of Science Core Collection and are not redistributable per the database’s licensing agreement. Accordingly, these data cannot be made publicly available as supplementary material. To ensure full reproducibility, the complete search query and all retrieval parameters (including database, collection, time span, and document types) are reported in Section 3.1 of the manuscript, enabling independent replication of the dataset. All custom code and extended analytical outputs that are not included in the main text are provided in the supplementary materials. All supplementary materials are located at Mendeley Data (DOI: 10.17632/nt4vsv3z.1). Supplementary material including: (1) Supplementary Table S1: The Multi-Label Classification Engine: Term Dictionaries, Weight Matrix, Priority Bonuses, and Decision Rules. (2) Supplementary Table S2: PRISMA-ScR checklist. (3) Coding_Manual_for_Human_Coders(Multilabel).docx. (4) Program code of (methodology_classification_and_sensitivity_analysis(Multilabel).py).txt. (5) README_Methodology_Classification_and_Sensitivity_Analysis(Multilabel).txt. (6) literature_classification(Multilabel)(Field Descriptions are manually added).xlsx. (7) weight_sensitivity(Multilabel).xlsx (Output file of methodology_classification_and_sensitivity_analysis(Multilabel).py). (8) Program code of (generate_coding_sheet(Multilabel).py).txt. (9) README_generate_coding_sheet(Multilabel).txt. (10) coding_sheet_coder-A(Multilabel).xlsx (The blank sheet is generated by generate_coding_sheet(Multilabel).py and filled manually.). (11) coding_sheet_coder-B(Multilabel).xlsx (The blank sheet is generated by generate_coding_sheet(Multilabel).py and filled manually.). (12) arbitration_results(Multilabel).xlsx (Gold standard from arbitration. Filled manually.). (13) Program code of (reliability_analysis(Multilabel).py).txt. (14) README_Reliability_Analysis(Multilabel).txt. (15) reliability_analysis(Multilabel).xlsx (Output file of reliability_analysis(Multilabel).py).

Funding: This research was funded by the National Social Science Foundation of China (Grant No. 22BGL029).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the article and its Supplementary Materials.

Acknowledgments: The author would like to thank the anonymous reviewers and the editor for their insightful comments and constructive suggestions, which have significantly improved the quality of this work. Any remaining errors or omissions are solely the responsibility of the author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AB	Abstract (WoS metadata field)
AI	Artificial Intelligence
ANE	Alternative and Nuclear Energy (energy-transition variable in study [38])
[A]	Alternative Specification Test (robustness check type, Table 7)
ATE	Average Treatment Effect
ATT	Average Treatment Effect on the Treated
CATE	Conditional Average Treatment Effect
CDE	CO ₂ Emissions (environmental-pressure variable in study [38])
CI	Confidence Interval
CRW	Combustible Renewables and Waste (circular-economy variable in study [38])
DE	Author Keywords (WoS metadata field)
DID	Difference-in-Differences
DLT	Distributed Ledger Technology
[E]	Endogeneity/Selection Test (robustness check type, Table 7)
ECM	Error Correction Model
Em	Employment (outcome variable in study [38])
ES	Expected Shortfall
ESG	Environmental, Social, and Governance
EU	European Union
FE	Fixed Effects
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GRF	Generalized Random Forest
HTE	Heterogeneous Treatment Effect
ID	Keywords Plus (WoS metadata field)
IDF	Inverse Document Frequency
IoT	Internet of Things
IPCC	Intergovernmental Panel on Climate Change
IV	Instrumental Variable
LCA	Life Cycle Assessment
LCDE	Logged CO ₂ Emissions (carbon-efficiency proxy in study [38])
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
ML	Machine Learning
NBR	New Businesses Registered (entrepreneurship proxy variable in study [38])
NGFS	Network for Greening the Financial System
OFAT	One-Factor-At-A-Time
OLS	Ordinary Least Squares
[P]	Placebo/Falsification Test (robustness check type, Table 7)
PRISMA-ScR	Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
ROA	Return on Assets
[S]	Subsample Sensitivity Test (robustness check type, Table 7)
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
TCFD	Task Force on Climate-related Financial Disclosures
TF-IDF	Term Frequency-Inverse Document Frequency
TFS	Term Frequency-Specificity
TI	Title (WoS metadata field)

TMLE	Targeted Maximum Likelihood Estimation
Tobin's Q	Tobin's Q Ratio
TS	Topic Search
Unc	Unclassified (residual category in the classification framework)
VaR	Value at Risk
WoS	Web of Science

References

1. COP28 UAE Presidency. The UAE Consensus. 2023. Available online: <https://www.cop28.com/en/the-uae-consensus-foreword> (accessed on 3 July 2026).
2. Ziolo, M.; Bak, I.; Spoz, A. The nexus between green bonds, green credits, and the energy transition toward renewable energy sources: State of the art. *Energies* **2025**, *18*, 4370. [[CrossRef](#)]
3. Iftikhar, H.; Luo, G.; Ullah, A.; Qian, N. Green development through disruptive financial technology and modern Belt and Road spillover: A pathway to the net-zero future in Asia, Europe, and Africa. *Financ. Innov.* **2026**, *12*, 121. [[CrossRef](#)]
4. Khan, I.; Shahid, M.S. The nexus between fintech and green finance in advancing sustainable performance across countries. *Discov. Sustain.* **2026**, *7*, 49. [[CrossRef](#)]
5. Blakely, T.; Lynch, J.; Simons, K.; Bentley, R.; Rose, S. Reflection on modern methods: When worlds collide—Prediction, machine learning and causal inference. *Int. J. Epidemiol.* **2020**, *49*, 2058–2064. [[CrossRef](#)] [[PubMed](#)]
6. Gur Ali, O. Targeting resources efficiently and justifiably by combining causal machine learning and theory. *Front. Artif. Intell.* **2022**, *5*, 1015604. [[CrossRef](#)] [[PubMed](#)]
7. Athey, S.; Imbens, G.W. Machine learning methods that economists should know about. *Annu. Rev. Econ.* **2019**, *11*, 685–725. [[CrossRef](#)]
8. Angrist, J.D.; Pischke, J.-S. The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *J. Econ. Perspect.* **2010**, *24*, 3–30. [[CrossRef](#)]
9. Kwong, R.; Kwok, M.L.J.; Wong, H.S.M. Green FinTech innovation as a future research direction: A bibliometric analysis on green finance and FinTech. *Sustainability* **2023**, *15*, 14683. [[CrossRef](#)]
10. Rojas-Valdez, K.; Fajardo-Piñan, V. Clustering green finance, fintech, and digital technologies for sustainable operations. *Discov. Sustain.* **2026**, *7*, 612. [[CrossRef](#)]
11. Saeedi, M.; Ashraf, B.N. The role of technology in promoting green finance: A systematic literature survey and the development of a framework. *J. Risk Financ. Manag.* **2024**, *17*, 472. [[CrossRef](#)]
12. Geetha, S.; Biju, A.V.N. Is green FinTech reshaping the finance sphere? Unravelling through a systematic literature review. *Environ. Sci. Pollut. Res.* **2024**, *31*, 1790–1810. [[CrossRef](#)] [[PubMed](#)]
13. Debrah, C.; Darko, A.; Chan, A.P.C. A bibliometric-qualitative literature review of green finance gaps and future research directions. *Clim. Dev.* **2023**, *15*, 432–455. [[CrossRef](#)]
14. Teixeira, N. Green digital finance and new quality productive forces: Bridging technological innovation and sustainable productivity. *China Financ. Rev. Int.* **2026**, 1–27. [[CrossRef](#)]
15. Hoang, D.; Wiegatz, K. Machine learning methods in finance: Recent applications and prospects. *Eur. Financ. Manag.* **2023**, *29*, 1657–1701. [[CrossRef](#)]
16. Mullainathan, S.; Spiess, J. Machine learning: An applied econometric approach. *J. Econ. Perspect.* **2017**, *31*, 87–106. [[CrossRef](#)]
17. Bahammou, I.; Ziti, S.; El Bouchti, K. Machine learning for financial market prediction: A systematic literature review. *Stat. Optim. Inf. Comput.* **2026**, *16*, 1697–1715. [[CrossRef](#)]
18. D'Amico, A.; Rossi, L.; Bianchi, F.M.; Lanzarini, L.C. Learning across modalities: A multi-perspective survey. *Inf. Fusion* **2025**, *114*, 102689. [[CrossRef](#)]
19. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [[CrossRef](#)] [[PubMed](#)]
20. Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; Robins, J. Double/debiased machine learning for treatment and structural parameters. *Econ. J.* **2018**, *21*, C1–C68. [[CrossRef](#)]
21. Wager, S.; Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* **2018**, *113*, 1228–1242. [[CrossRef](#)]
22. Kaddour, J.; Lynch, A.; Liu, Q.; Kusner, M.J.; Silva, R. Causal machine learning: A survey and open problems. *arXiv* **2022**, arXiv:2206.15475v3. [[CrossRef](#)]

23. Sjögarde, P.; Ahlgren, P.; Waltman, L. Algorithmic labeling in hierarchical classifications of publications: Evaluation of bibliographic fields and term weighting approaches. *J. Assoc. Inf. Sci. Technol.* **2021**, *72*, 853–869. [CrossRef]
24. Mateen, F.J.; Oh, J.; Tergas, A.I.; Bhayani, N.H.; Kamdar, B.B. Titles versus titles and abstracts for initial screening of articles for systematic reviews. *Clin. Epidemiol.* **2013**, *5*, 89–95. [CrossRef] [PubMed]
25. Bramer, W.M.; Giustini, D.; Kleijnen, J.; Franco, O.H. Searching Embase and MEDLINE by using only major descriptors or title and abstract fields: A prospective exploratory study. *Syst. Rev.* **2018**, *7*, 200. [CrossRef] [PubMed]
26. Spärck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *J. Doc.* **1972**, *28*, 11–21. [CrossRef]
27. Chen, G.; Xiao, L. Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods. *J. Informetr.* **2016**, *10*, 212–223. [CrossRef]
28. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
29. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*, 4th ed.; SAGE Publications: Thousand Oaks, CA, USA, 2018.
30. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. Available online: <https://www.ncbi.nlm.nih.gov/pubmed/843571> (accessed on 15 July 2026). [CrossRef]
31. Berg, F.; Kölbel, J.F.; Rigobon, R. Aggregate confusion: The divergence of ESG ratings. *Rev. Financ.* **2022**, *26*, 1315–1344. [CrossRef]
32. Zapata, H.O.; Mukhopadhyay, S. A Bibliometric Analysis of Machine Learning Econometrics in Asset Pricing. *J. Risk Financ. Manag.* **2022**, *15*, 535. [CrossRef]
33. Network for Greening the Financial System (NGFS). NGFS Climate Scenarios for Central Banks and Supervisors. 2024. Available online: <https://www.ngfs.net/en/ngfs-climate-scenarios-central-banks-and-supervisors> (accessed on 3 July 2026).
34. People's Bank of China. China Monetary Policy Implementation Report for the Fourth Quarter of 2021. 2022. Available online: <https://www.pbc.gov.cn/zhengcehuobisi/125207/125227/125957/4246985/75d90372c9174bde9287c2c30935b477/20221119311841777.pdf> (accessed on 3 July 2026).
35. International Capital Market Association (ICMA). Green Bond Principles: Voluntary Process Guidelines for Issuing Green Bonds. 2025. Available online: <https://www.icmagroup.org/assets/documents/Sustainable-finance/2025-updates/Green-Bond-Principles-GBP-June-2025.pdf> (accessed on 3 July 2026).
36. G20 Sustainable Finance Working Group. 2021 *Synthesis Report*; G20 Italian Presidency: Rome, Italy, 2021. Available online: https://g20sfwg.org/wp-content/uploads/2021/11/Synth_G20_Final.pdf (accessed on 3 July 2026).
37. European Union. Regulation (EU) 2020/852 of the European Parliament and of the Council on the Establishment of a Framework to Facilitate Sustainable Investment, and Amending Regulation (EU) 2019/2088 (EU Taxonomy Regulation). Official Journal of the European Union, L 150/13. 2020. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32020R0852> (accessed on 3 July 2026).
38. Task Force on Climate-Related Financial Disclosures (TCFD). Recommendations of the Task Force on Climate-Related Financial Disclosures (Final Report). 2017. Available online: <https://www.fsb-tcf.org/recommendations/> (accessed on 3 July 2026).
39. Wang, S.; Wijethilake, C. The Effect of Carbon Disclosure Quality on Green Bond Financing Cost. *Bus. Strategy Environ.* **2026**, *34*, e71126. [CrossRef]
40. Petreski, A.; Schäfer, D. The Reputation Effect of Repeated Green-Bond Issuance and Its Impact on the Cost of Capital. *Bus. Strategy Environ.* **2025**, *34*, 2436–2448. [CrossRef]
41. Wu, B.; Ding, Y.; Xie, B.; Zhang, Y. FinTech and Inclusive Green Growth: A Causal Inference Based on Double Machine Learning. *Sustainability* **2024**, *16*, 9989. [CrossRef]
42. Chen, J.-E.; Jing, A. Recent advances in causal machine learning and dynamic policy learning. *WIREs Comput. Stat.* **2025**, *17*, e70050. [CrossRef]
43. Christensen, G.; Miguel, E. Transparency, reproducibility, and the credibility of economics research. *J. Econ. Lit.* **2018**, *56*, 920–980. [CrossRef]
44. Nurmasari, N.D.; Bilad, M.R. Artificial intelligence applications for advancing sustainable green finance. *Discov. Sustain.* **2026**, *7*, 668. [CrossRef]
45. Kleinberg, J.; Ludwig, J.; Mullainathan, S.; Obermeyer, Z. Prediction policy problems. *Am. Econ. Rev.* **2015**, *105*, 491–495. [CrossRef] [PubMed]
46. Smith, J.A.; Todd, P.E. Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? *J. Econ.* **2005**, *125*, 305–353. [CrossRef]
47. Huang, X.; Wang, Y.; Chen, Y.; Hu, Z. Green Technology Innovation and Enterprise Performance: An Analysis Based on Causal Machine Learning Models. *Sustainability* **2024**, *16*, 2309. [CrossRef]

48. Abidi, I.; Alaraby, H.Y.; Rabaiah, G. A Machine Learning Approach to Green Finance: Pathways for Enhancing Entrepreneurship, Employment, and the Energy Pillar of the Just Transition. *Sustainability* **2026**, *18*, 2161. [[CrossRef](#)]
49. Creswell, J.W.; Creswell, J.D. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed.; SAGE Publications: Thousand Oaks, CA, USA, 2018.

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