Journal of Environmental Management 413 (2026) 130402

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

Decoupling clinker technology from cement product emissions: A macroeconomic ML–LCA framework for global embodied carbon policy screening

Dilba Rayaru Kandiyil [a,b,*], Monower Sadique [a], Denise Lee [a], Joseph Amoako-Attah [a], Rafal Al Mufti [b]

[a] School of Civil Engineering and Built Environment, Liverpool John Moores University, UKLiverpool, L3 3AF, UK
[b] Faculty of Engineering and Construction, OU|Liverpool John Moores University, Al Amir street, Qatar

ARTICLE INFO

Keywords:
Embodied carbon
Cement
STIRPAT
ML–LCA
Clinker substitution
SHAP
Environmental Kuznets curve
Global decarbonisation

ABSTRACT

The cement industry contributes approximately 7–8% of global anthropogenic CO_2 emissions, yet accurate cradle-to-gate embodied carbon estimation requires plant-level inventory data that are largely unavailable across developing and emerging economies. This data scarcity constrains global benchmarking and the implementation of emerging embodied carbon regulations.

This study proposes a STIRPAT-grounded hybrid machine learning–life cycle assessment (ML–LCA) framework for estimating national-scale cement embodied carbon using exclusively publicly available macroeconomic data. GDP per capita, population, and temporal indicators are used to predict the clinker-to-cement ratio (CCR), which is subsequently propagated through a technology-stratified, process-based LCA model enforcing stoichiometric and thermodynamic constraints across A1–A3 stages. Among seven candidate algorithms, Gradient Boosting was selected for its smooth non-linear approximation and LCA integration suitability. SHAP analysis confirms GDP per capita as the dominant CCR driver, with contributions directionally consistent with established technology diffusion theory, ensuring model transparency.

Validation across 18 economies through statistical metrics, residual diagnostics, country-level diagnostic benchmarking, Leave-One-Country-Out (LOCO) cross-validation, and three independent literature-benchmarking countries (Pakistan, Mexico, Spain) confirms physically plausible and externally consistent outputs ranging from 0.53 to 0.97 kg CO_2/kg cement. A central methodological contribution is the ability to estimate the clinker-substitution decoupling effect at the country scale using only macroeconomic inputs, in contexts where plant-level LCA inventory data are unavailable. Conventional LCA already separates process, energy, and material composition contributions when inventory data are present; the present framework extends this separation to data-scarce national contexts. At the system level, an Environmental Kuznets Curve–type pattern is qualitatively reproduced when model outputs are aggregated across countries, providing a coherence check on the framework as a whole. Out-of-country generalisation is assessed using Leave-One-Country-Out (LOCO) cross-validation as the primary protocol (mean fold RMSE 0.077; 12 of 18 folds below RMSE 0.10), with a forward-chaining temporal split as a complementary diagnostic.

The framework is operationalised through an interactive decision-support interface, offering a scalable, transparent baseline for embodied carbon benchmarking, policy screening, and net-zero pathway evaluation in the global cement sector. The framework is positioned as a screening-level reference for data-scarce contexts, complementary to plant-level LCA and Environmental Product Declarations where these are available.

* Corresponding author. School of Civil Engineering and Built Environment, Liverpool John Moores University, UKLiverpool, L3 3AF, UK.
E-mail addresses: oucdrk@ljmu.ac.uk (D.R. Kandiyil), m.m.sadique@ljmu.ac.uk (M. Sadique), d.y.lee@ljmu.ac.uk (D. Lee), j.amoakoattah@ljmu.ac.uk (J. Amoako-Attah), rafal.a@oryx.edu.qa (R. Al Mufti).

https://doi.org/10.1016/j.jenvman.2026.130402
Received 5 April 2026; Received in revised form 17 June 2026; Accepted 30 June 2026
Available online 5 July 2026

1. Introduction

Cement is an indispensable construction material that underpins modern infrastructure, urbanisation, and economic development. Its versatility, durability, and cost-effectiveness have made it the most widely produced manufactured material globally, with annual production exceeding four billion tonnes (Bagga et al., 2022). Cement is fundamental to housing, transportation, energy infrastructure, and public works, particularly in rapidly developing regions where construction demand continues to grow in parallel with population and economic expansion.

(IPCC, 2022)IEA, 2023; Scrivener et al., 2018;Despite its societal benefits, cement production is also one of the most carbon-intensive industrial activities. The cement sector is responsible for approximately 7–8% of global anthropogenic CO2 emissions, positioning it as a structurally hard-to-abate industry (IPCC, 2022; IEA, 2023). A substantial share of these emissions originates from clinker production, where the calcination of limestone releases CO2 through an unavoidable chemical reaction, independent of energy efficiency or fuel choice. This process-related emissions component fundamentally distinguishes cement from many other industrial sectors and limits the extent to which decarbonisation can be achieved through energy system improvements alone (Scrivener et al., 2018)

In recent years, embodied carbon in the context of this study refers to the greenhouse gas (GHG) emissions associated with cement material production from raw material extraction through to the factory gate (life-cycle stages A1 to A3 under EN 15978). This cradle-to-gate scope captures the dominant emissions sources in the cement sector and aligns with the boundary used in major embodied-carbon regulatory and benchmarking frameworks. The broader whole-life definition (covering A4 transport to site through to C and D end-of-life stages) is widely used in the construction LCA literature but is outside the scope of this study. has emerged as a critical focus in climate mitigation for the built environment. As operational energy demand in buildings declines due to efficiency measures and electrification, embodied carbon increasingly dominates life-cycle emissions, particularly in material-intensive construction (Pomponi and Moncaster, 2016). This shift has been reinforced by emerging regulations and policy instruments, including whole-life carbon reporting requirements, embodied carbon limits in public procurement, and material-level benchmarking initiatives across Europe, North America, and parts of Asia (Cen, 2019; Ice, 2023).

Accurate estimation of cement embodied carbon is traditionally achieved using process-based Life Cycle Assessment (LCA). While LCA provides a rigorous and physically grounded methodology, its application at national or global scales is severely constrained by data availability. Conventional LCA requires detailed plant-level information such as kiln type, thermal efficiency, fuel mix, electricity grid intensity, and supplementary cementitious material substitution rates. These data are often proprietary, inconsistently reported, or entirely unavailable, particularly in developing and emerging economies (Dilba Rayaru Kandiyil et al., 2025). As a result, existing embodied carbon datasets are geographically fragmented and biased toward regions with high data transparency, limiting cross-country comparability and global benchmarking (Cembureau, 2022; Gcca, 2023).

This persistent data scarcity represents a fundamental barrier to global embodied carbon assessment in the cement sector. Many countries with rapidly expanding construction activity lack Environmental Product Declarations (EPDs) or comprehensive industrial inventories yet are increasingly expected to align with international climate targets and embodied carbon regulations. Addressing this challenge requires modelling approaches that can operate reliably under limited data conditions while preserving physical and methodological consistency.

Macroeconomic indicators provide a viable pathway to address these limitations. Variables such as population, gross domestic product (GDP) per capita, and time are publicly available, consistently reported across countries, and theoretically linked to industrial development, technology adoption, and material efficiency. These relationships are formalised in the STIRPAT framework (Dietz and Rosa, 1994; (York et al., 2003)), which extends the IPAT identity to allow stochastic and non-linear modelling of environmental impacts. In parallel, the Environmental Kuznets Curve (EKC) hypothesis suggests that environmental intensity initially increases with economic growth before declining as technology, regulation, and efficiency improve (Aperghis and Payne, 2010; Shahbaz et al., 2017). In the cement sector, such behaviour is reflected in non-linear reductions in clinker intensity as economies mature and alternative materials and advanced technologies are adopted.

Recent advances in machine learning offer powerful tools for capturing these non-linear relationships in complex and data-scarce systems. Machine learning has been increasingly applied in construction and materials research to model energy use, emissions, and material performance. However, purely data-driven approaches often lack transparency and may generate physically implausible results when extrapolated beyond observed data ranges. Hybrid approaches that integrate machine learning with process-based LCA offer a promising alternative, enabling statistical inference to be constrained by stoichiometric and thermodynamic principles while extending applicability to data-limited contexts ((Lu et al., 2020);(Huang et al., 2022)).

Against this background, this study proposes a STIRPAT (Stochastic Impacts by Regression on Population, Affluence, and Technology)-grounded hybrid machine learning–life cycle assessment (ML–LCA) framework for estimating national-scale cement embodied carbon using exclusively publicly available macroeconomic data. The choice of macroeconomic predictors is theoretically grounded in the STIRPAT framework, which posits that environmental impact at the population scale is a multiplicative function of population (the demographic scale of activity), affluence (typically GDP per capita, capturing both technology adoption and consumption intensity), and technology (often modelled as a temporal residual capturing structural progress not encoded in scale or affluence). For the clinker-to-cement ratio specifically, GDP per capita is theoretically expected to drive substitution behaviour through three superimposed channels: cement-industry technological maturity, regulatory stringency, and the availability of supplementary cementitious materials (SCMs). Population captures the scale of national cement demand and indirectly the structural composition of the industry, while the temporal indicator absorbs residual technology diffusion. The literature reviewed above documents each of these mechanisms separately, but to our knowledge no published framework integrates them into a single national-scale CCR predictor that can operate without plant-level data. This is the methodological gap addressed by the present study. Rather than predicting emissions directly, machine learning is used to infer the clinker-to-cement ratio, a dominant determinant of cement embodied carbon that reflects material substitution and technological efficiency. The predicted clinker ratios are subsequently integrated into a deterministic, process-based LCA model to estimate cradle-to-gate (A1–A3) embodied carbon while enforcing chemical and energy-based constraints.

The aim of this research is to deliver a globally transferable, subject to data availability, and physically consistent solution to cement embodied carbon estimation under conditions of severe data scarcity. By combining macroeconomic theory, interpretable machine learning, and process-based LCA, the proposed framework enables cross-country comparability, reveals emergent system-level behaviour qualitatively consistent with Environmental Kuznets Curve dynamics, and provides a transparent baseline for embodied carbon benchmarking, regulatory screening, and net-zero pathway analysis in the global cement industry.

2. Literature review

The cement and concrete industries are major carbon emitters: roughly 37% of global CO_2 comes from buildings and 8% from concrete production, largely due to energy-intensive cement clinker

manufacturing heating limestone to >1400°C (De Paula Salgado et al., 2025). To quantify such impacts, life-cycle assessment (LCA) is widely used. LCA is a standardized, multi-parameter framework (ISO 14040/44) that compiles all inputs/outputs of a product's life and assesses impacts e.g. GWP (De Paula Salgado et al., 2025). However, conventional LCA is data- and labour-intensive. Its results can vary with assumptions: incomplete or inconsistent inventory data, recycled material flows, and emerging technologies often lead to large discrepancies in carbon estimates even for similar material systems. These uncertainties complicate policy and design decisions, motivating new hybrid approaches.(De Paula Salgado et al., 2025)

Machine learning has emerged as a tool to address LCA's data challenges. By identifying patterns in existing datasets, ML algorithms can predict missing life-cycle inventory data and estimate environmental impacts for new scenarios (Bagga et al., 2022). For example, ML can be trained on databases of material compositions or product environmental declarations to estimate missing inputs. Reviews note that ML-integrated LCA approaches can significantly reduce uncertainty in impact assessment by using dynamic data inputs. In effect, ML acts as a surrogate or "metamodel" for parts of the LCA, enabling faster and more adaptable analyses than static LCA alone(Alabduljabbar et al., 2025; De Paula Salgado et al., 2025).

Recent literature surveys in construction confirm this trend of ML–LCA hybridization (Jalota and Ayazi, 2025). systematically reviewed 21 studies on ML in construction LCA((Jalota and Ayazi, 2025)). They report that popular ML methods (random forests, gradient boosting, neural nets) are being applied to predict carbon emissions, optimize material choices, and assess sustainability. Regression models currently dominate such integration due to their accuracy and ease of use.((Jalota and Ayazi, 2025)) However, the authors highlight that ensemble and hybrid ML architectures (e.g. stacking multiple models or deep learning networks) generally achieve superior robustness and predictive power.((Jalota and Ayazi, 2025)) In short, ML-augmented LCA workflows are emerging as a promising practice: for example, neural-network metamodels have been shown to predict a building's life-cycle global warming potential (GWP) from design parameters,(De Paula Salgado et al., 2025) and ML-driven optimizations have successfully reduced embodied carbon while maintaining performance – e.g. one study used ML to find an optimal concrete mix with lower cement content that cut emissions without sacrificing strength(De Paula Salgado et al., 2025).

Beyond integrated LCA, ML is universally applied to material optimization in construction. Researchers use ML (including neural nets, decision trees, genetic algorithms) to balance conflicting objectives like cost, strength, and carbon footprint in mix design (De Paula Salgado et al., 2025). For instance, multi-objective models have been developed to generate Pareto-optimal concrete formulations that minimize greenhouse gas emissions (GWP) and cost while meeting strength requirements(De Paula Salgado et al., 2025). These ML tools speed up the exploration of large design spaces (many mix ingredients) and can guide low-carbon innovations. A concrete example is Hamza El Hafdaoui et al. (2023), who trained a supervised ML model on building database inputs to estimate embodied carbon of buildings; their model achieved ~15.7% mean error in predicting total embodied carbon (Hamza El Hafdaoui et al., 2023). Such results illustrate that ML can capture the complex, non-linear relations between building or material parameters and their life-cycle emissions, complementing traditional LCA.

The intersection of LCA and policy has also grown. Several jurisdictions and industry groups are embedding embodied carbon metrics into regulations and guidelines. In the rapidly developing regions e.g. Qatar lack local EC data and mandates, and they recommend creating a comprehensive embodied carbon database and integrating EC assessment into building codes(Dilba Rayaru Kandiyil et al., 2025). Similarly, global cement industry roadmaps (Gcca; Gcca, 2023) target significant reductions in concrete carbon: for example, aiming for 25% lower CO_2 per cubic meter by 2030 (vs. 2020) on the path to carbon neutrality by 2050(Esra, 2024). The (World Economic Forum, 2023) finds that using low-carbon concrete mixtures could cut concrete-related emissions by up to ~40% with only a 2–3% cost increase(Esra, 2024; De Paula Salgado et al., 2025). In practice, some U.S. states and cities now set embodied-carbon limits for public projects e.g. maximum GWP for concrete mixes. Thus, LCA-based metrics and ML-predicted EC values are directly informing policy and procurement, reinforcing the importance of reliable EC quantification methods.

At the macro level, analysts often use econometric models like the Environmental Kuznets Curve (EKC) and STIRPAT to understand how economic growth and other factors drive carbon emissions. The EKC hypothesis suggests an inverted-U relationship: as GDP per capita grows, emissions first rise then eventually decline as efficiency improves(Esra, 2024). For the construction sector, studies sometimes find partial EKC effects. For example, (Bao, 2023)) observed an inverted-U curve between economic development and construction waste in 27 European economies(Bao, 2023). In the cement context (Liu et al., 2023), applied both the Tapio decoupling model and an extended STIRPAT regression across Chinese provinces: they concluded that GDP growth still drives cement sector CO_2 emissions, implying China is on the rising segment of the EKC for construction ((Liu et al., 2023)). STIRPAT (Stochastic Impacts by Regression on Population, Affluence, and Technology) extends the simple IPAT identity by estimating elasticities via regression. It is widely used to identify drivers of carbon emissions. For instance, a 2025 study combined LCA with a STIRPAT model for Chinese urban vs. rural residential buildings. They found that population size and income significantly drive embodied carbon: every 1% increase in population or disposable income led to roughly a 1.06% and 0.73% rise in building embodied carbon, respectively(Miaoyi Wang et al., 2025). Such findings emphasize that demographic and economic factors continue to push embodied carbon upward in many regions.

In summary, the literature shows a growing convergence of machine learning, life-cycle methods, and environmental modelling in the cement and construction sector. Hybrid ML–LCA approaches are being developed to tackle data gaps and improve prediction of embodied carbon, building on traditional LCA practice((Jalota and Ayazi, 2025)). Concurrently, broader ML applications in mix design and material innovation, along with policy drives e.g. carbon regulations for concrete, underscore the urgency of quantifying and reducing cement's embodied carbon(Esra, 2024) (De Paula Salgado et al., 2025). Finally, macro-modelling tools like EKC and STIRPAT provide context for how economic and technological trends affect cement emissions, guiding policy toward sustainable pathways(Miaoyi Wang et al., 2025)

3. Research methodology

This study develops a theory-grounded hybrid modelling framework that integrates macroeconomic machine learning with process-based life cycle assessment (LCA) to estimate national-scale cement embodied carbon. The methodological design explicitly addresses the limitations of conventional cement LCA under conditions of incomplete industrial data, while ensuring physical consistency, interpretability, and global transferability. The full modelling workflow, from raw data acquisition to embodied carbon estimation, is summarised in Figs. 3–1, which illustrates the sequential and modular structure of the framework.(Miaoyi Wang et al., 2025)

3.1. Database construction and scope definition

The database was constructed using publicly available national-level datasets to ensure reproducibility and cross-country transferability. Cement and clinker production volumes were compiled from internationally recognised industrial statistics, including datasets published by the Global Cement and Concrete Association (GCCA) and the US Geological Survey (USGS), while macroeconomic variables such as population, gross domestic product (GDP), and GDP per capita were

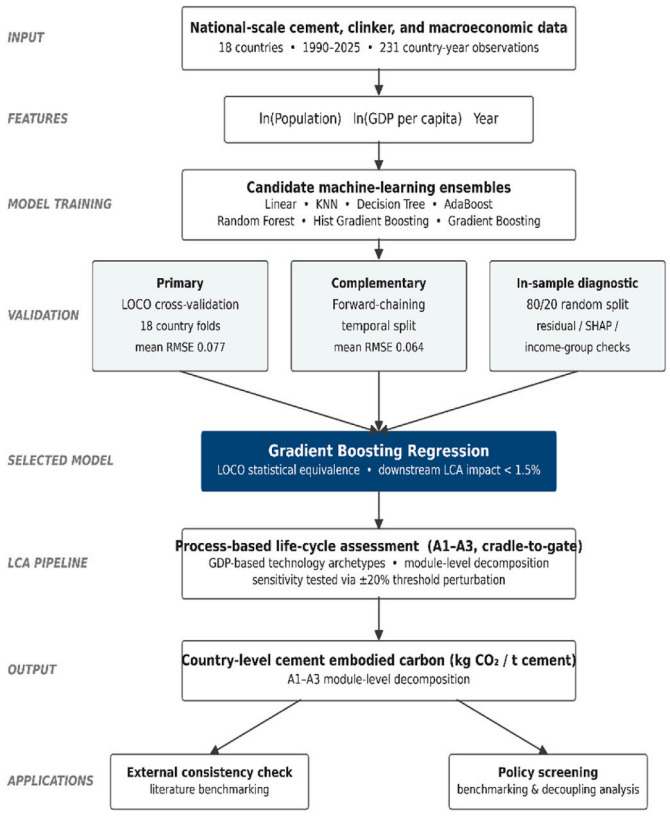


Fig 3-1. Proposed Modelling Workflow

obtained from harmonised global databases maintained by the World Bank and the International Energy Agency (IEA). The dataset comprises 231 country–year observations across 18 countries, with temporal coverage spanning the period 1990 to 2025 (mean 13 observations per country, range 8 to 34; per-country coverage in Supplementary Table S1). For some countries, national CCR values are reported as constant across multiple years, reflecting industry-level reporting practice rather than absence of inter-annual variation; the model captures this stability where present and inter-annual variation where reported. The temporal range is intentionally broad to capture structural changes in cement production technologies, fuel substitution, and material efficiency over time.

The clinker-to-cement ratio (CCR) was selected as the primary dependent variable for machine learning modelling. This choice is motivated by both physical and policy considerations: clinker content directly governs cement embodied carbon intensity through calcination emissions and energy demand, and clinker substitution is a central mitigation lever in cement decarbonisation strategies. By modelling clinker ratio as an intermediate technological parameter rather than predicting emissions directly, the framework preserves a clear causal structure between economic development, technological efficiency, and embodied carbon outcomes.

A comprehensive summary of the database, including variable definitions, units, temporal resolution, and descriptive statistics, is provided in Tables 3–1, while the spatial distribution of countries and income groups represented in the dataset is illustrated in Figs. 3–2. This explicit documentation supports transparency and allows assessment of dataset representativeness (see Fig. 4) (see Table 4).

Tables 3–1 summarises the descriptive statistics of the key variables used in the modelling dataset, illustrating the wide range of cement production scales, population sizes, and income levels represented. Cement production and GDP per capita exhibit pronounced dispersion and right-skewed distributions, reflecting substantial heterogeneity across countries and development stages. In contrast, the clinker-to-cement ratio varies within physically bounded limits, indicating meaningful differences in technological efficiency and material substitution rather than random variability. Together, these characteristics justify the use of non-linear, scale-invariant machine learning approaches and motivate subsequent data transformations applied during preprocessing. Figs. 3–2 illustrates the geographical coverage of the modelling dataset, which includes 18 countries distributed across North America, Europe, Asia, the Middle East, Oceania, and South America. This spatial diversity ensures representation of a wide range of economic development stages, production scales, and technological maturity levels relevant to global cement embodied carbon assessment.

Table 3–1
Descriptive statistics of key variables used in the modelling dataset (N = 231).

Variable	Min	Max	Mean	Std. Dev.
Clinker-to-cement ratio	0.55	0.95	0.79	0.11
Cement production (t/year)	7.62×10^6	2.49×10^9	2.54×10^8	5.64×10^8
Population	8.26×10^6	1.46×10^9	3.55×10^8	4.90×10^8
GDP per capita (USD)	350	80,400	24,479	21,379

Note: N represents the total number of countries–year observations included in the modelling dataset.

3.2. Data preprocessing and feature engineering

3.2.1. Theoretical basis

Feature engineering in this study is grounded in the STIRPAT framework, which extends the IPAT identity to model environmental impacts as a stochastic function of population, affluence, and technology. The general STIRPAT formulation is given by (Miaoyi Wang et al., 2025):

$$I = aP^b A^c T^d \varepsilon$$

where I denotes environmental impact, $P-$ population, $A-$ affluence, $T-$ technology, and $\varepsilon-$ a stochastic error term.

Taking logarithms yields the estimable form:

$$\ln(I) = \ln(a) + b\ln(P) + c\ln(A) + d\ln(T) + \ln(\varepsilon)$$

In this study, direct modelling of emissions is avoided due to data limitations. Instead, the clinker-to-cement ratio (CCR) is adopted as a proxy for technological efficiency, directly governing cement embodied carbon intensity. Population represents demand pressure, GDP per capita represents affluence, and time acts as a surrogate for technological progress and regulatory evolution, consistent with prior macro-environmental modelling studies.

3.2.2. Exploratory data characteristics

Descriptive statistics in Tables 3–1 shows that cement production, population, and GDP per capita exhibit strong dispersion and right-skewed distributions, reflecting heterogeneity across countries and development stages. In contrast, the clinker-to-cement ratio varies within physically bounded limits Fig. 3–3, indicating structural technological differences rather than random variability.

The relationship between CCR and GDP per capita plotted on a logarithmic scale Figs. 3-4 reveals a non-linear pattern with distinct regimes across income levels. This behaviour is consistent with Environmental Kuznets Curve (EKC) theory, which postulates a non-monotonic relationship between environmental pressure and economic development, commonly expressed as(Grossman and Krueger, 1995):

$$I = \alpha + \beta_1 \ln(A) + \beta_2 [\ln(A)]^2 + \varepsilon$$

Although EKC parameters are not explicitly estimated at this stage, the exploratory pattern motivates non-linear modelling approaches.

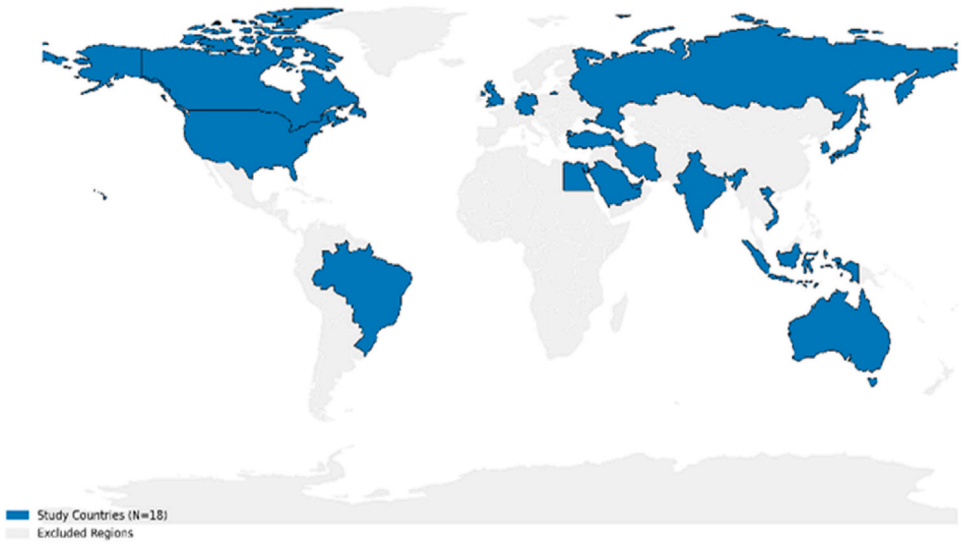


Fig. 3–2. Geographical distribution of countries included in the modelling dataset (N = 18)

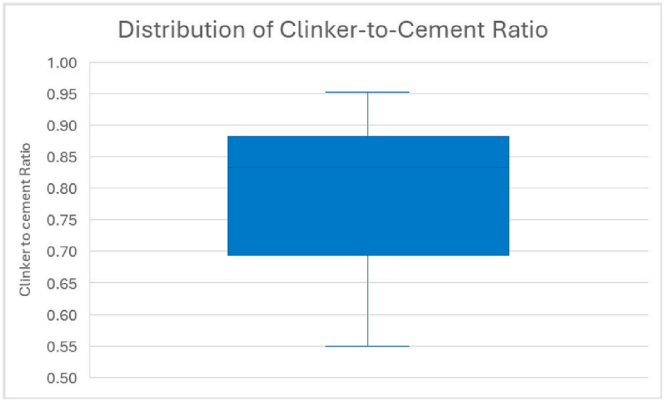


Fig. 3–3. Distribution of clinker-to-cement ratio (CCR) across countries included in the final modelling dataset.

3.2.3. Feature transformation, selection, and collinearity diagnostics

Following the exploratory analysis, logarithmic transformation was applied to population, GDP per capita, and cement production to address scale heterogeneity and skewed distributions, consistent with STIRPAT-based macro-environmental modelling. The clinker-to-cement ratio was retained in its original scale to preserve physical interpretability due to its bounded nature. We acknowledge that the use of three macroeconomic predictors (population, GDP per capita, year) reflects a deliberate design choice consistent with the data-scarce operating regime of the framework, rather than a claim that these three variables fully characterise national cement systems. Important additional drivers – including specific policy interventions (emissions trading schemes, embodied-carbon mandates), regional energy prices, the availability of supplementary cementitious materials (SCMs) such as fly ash and slag, and resource constraints (limestone availability, electricity grid composition) – are not explicitly represented as features in the present model. The justification for this choice is twofold. First, the STIRPAT framework was developed precisely for contexts in which only macroeconomic indicators are reliably available across countries; richer feature sets are precisely what is unavailable in the data-scarce settings the framework targets. Second, several of the omitted drivers are statistically encoded by GDP per capita (which correlates with regulatory stringency, SCM availability, and the maturity of cement-industry infrastructure), so the

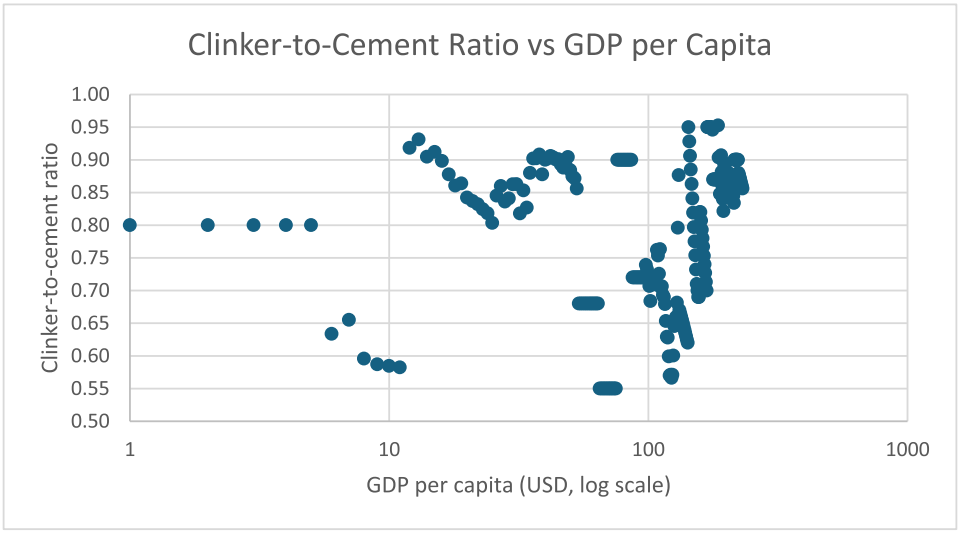


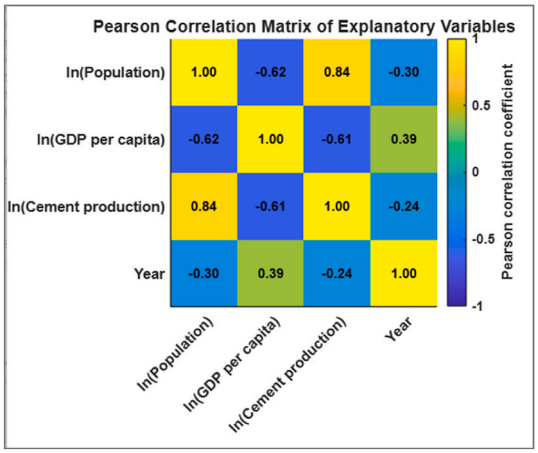
Fig. 3–4. Relationship between clinker ratio and GDP per capita plotted on a logarithmic scale.

framework operates by treating GDP per capita as a composite proxy. The consequence is that the model captures statistical regularity at the country-aggregated scale rather than identifying the causal contribution of any single underlying driver. This limitation is appropriate to acknowledge in the policy-oriented framing: the framework is suited to first-order screening of national-level cement embodied carbon, and not to the design of specific technology policies, for which richer plant- or sub-sector-level data would be required. A natural extension of this work, discussed in the conclusions, is to incorporate sector-specific predictors (e.g. national SCM utilisation rates, industrial energy intensity, kiln-technology mix) as such data become more widely available.

Feature selection was theory driven. Population was used as a proxy for demand pressure, GDP per capita represented affluence and economic structure, and time captured technological and regulatory evolution. Pairwise relationships among predictors were examined using Pearson correlation analysis Figs. 3–5. As expected, population and cement production showed strong positive correlation, while GDP per capita exhibited moderate negative correlation with scale-related variables, reflecting structural differences between large emerging economies and smaller high-income economies. No correlation coefficients approached unity. Multicollinearity among explanatory variables was assessed using variance inflation factor (VIF) diagnostics (Tables 3–2). All VIF values remained below accepted thresholds, indicating no severe multicollinearity.

3.3. Comparative modelling strategy and candidate algorithms

To avoid imposing a priori assumptions on the functional relationship between macroeconomic indicators and clinker-to-cement ratio, a comparative modelling strategy was adopted. Multiple regression algorithms representing distinct classes of machine learning logic were evaluated to examine both linear and non-linear relationships between predictors and the target variable. This approach allows the modelling structure to be guided by empirical behaviour rather than model preference. The objective of this stage was to systematically explore alternative learning paradigms under identical data, preprocessing, and evaluation conditions. Detailed performance comparisons and the rationale for final model selection are presented in the Results section. The selection of candidate algorithms reflects commonly adopted model classes in comparative machine learning studies for tabular environmental and industrial datasets. Prior work has shown that linear models often fail to capture non-linear economic–environmental relationships, while ensemble tree-based methods demonstrate superior performance and robustness in emissions and energy modelling contexts(Breiman, 2001; Friedman, 2001; Shrestha and Solomatine, 2006; XuanRui, 2022; Ullah et al., 2025c).



Figs. 3–5. Pearson correlation matrix of explanatory variable.

Table 3–2
Variance inflation factor (VIF) for explanatory variables.

Variable	VIF
ln(Population)	3.68
ln(GDP per capita)	1.87
ln(Cement production)	3.60
Year	1.20

Note: All VIF values are below the commonly accepted threshold of 5, indicating no severe multicollinearity among explanatory variables.

3.3.1. *Parametric baseline model- linear regression (ordinary least squares)*

Linear regression assumes a linear and additive relationship between explanatory variables and the clinker-to-cement ratio. It was included as a baseline reference to assess whether macroeconomic influences on clinker intensity can be adequately represented using a simple linear approximation. While transparent and interpretable, this model is limited in its ability to capture non-linear interactions and threshold effects commonly observed in macroeconomic and technological transition processes.

3.3.2. *Instance-based learning model- K-Nearest Neighbours (KNN)*

KNN is an instance-based learning method that predicts the target value by averaging the outcomes of the most similar observations in the feature space. It does not assume a predefined functional form and instead relies on local interpolation. This model was included to evaluate whether similarity-based patterns among countries with comparable macroeconomic profiles can explain variations in clinker-to-cement ratio. However, KNN is sensitive to feature scaling and data sparsity, which can limit its robustness in macro-level datasets.

3.3.3. *Neural network model-multi-layer Perceptron (MLP)*

The Multi-Layer Perceptron is a feedforward artificial neural network capable of approximating complex non-linear functions through stacked layers of weighted transformations and activation functions. The MLP was evaluated to assess whether neural architectures are suitable for modelling clinker ratio dynamics using a relatively small, tabular macroeconomic dataset. Although flexible, neural networks typically require careful regularisation and sufficient data volume to avoid overfitting.

3.3.4. *Tree-based models- decision tree regression*

Decision trees partition the input feature space into hierarchical decision rules, allowing for intuitive interpretation and natural handling of non-linear relationships. A single decision tree model was included to evaluate rule-based segmentation of macroeconomic drivers. However, such models are known to exhibit high variance and susceptibility to overfitting when used in isolation and were therefore assessed primarily as a precursor to ensemble-based methods.

3.3.5. *Ensemble learning models*

Ensemble learning methods combine multiple weak learners to improve predictive robustness and generalisation.

3.3.5.1. Random forest regression (bagging-based ensemble). Random Forest aggregates multiple decision trees trained on bootstrapped samples of the data and random subsets of features. This bagging-based approach reduces variance and improves stability, making it well suited for heterogeneous macroeconomic datasets with non-linear interactions.

3.3.5.2. AdaBoost regression (adaptive boosting). AdaBoost sequentially trains weak learners by adaptively reweighting observations that are

difficult to predict. This method was included to assess whether adaptive error correction improves clinker ratio prediction under data heterogeneity.

3.3.5.3. Gradient boosting regression. Gradient Boosting constructs an additive predictive model by sequentially fitting decision trees to the residual errors of preceding iterations. This boosting-based framework enables the capture of smooth non-linear trends and interaction effects, which are commonly observed in macroeconomic and technology-transition processes.

3.3.5.4. Histogram-based gradient boosting. Histogram-based Gradient Boosting is a computationally optimised variant that discretises continuous input features into bins prior to split finding. It was included to evaluate whether computational efficiency gains could be achieved without altering modelling behaviour or predictive structure.

3.4. Machine learning model training and evaluation protocol

All candidate models were trained using the same pre-processed dataset to ensure methodological consistency. Population, GDP per capita, and year were used as predictor variables, and the clinker-to-cement ratio was used as the target variable.

The dataset was partitioned into training and testing subsets using an 80/20 split. Model performance was evaluated on the test set using the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE), which collectively assess explanatory power and prediction error magnitude: R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

To prioritise generalisation and reduce overfitting during model comparison, default or conservative hyperparameter settings were employed (see Table 3-3). A fixed random state was applied to ensure reproducibility. The retained Gradient Boosting model uses the following configuration: n_estimators = 300, learning_rate = 0.05, max_depth = 4, random_state = 42, with all other parameters left at their scikit-learn defaults; these settings reproduce Tables 4–1 results exactly. The full hyperparameter configuration with rationale is provided in Supplementary Table S4. We deliberately did not perform exhaustive hyperparameter tuning, since the dataset size (231 observations) makes aggressive tuning prone to overfitting; the chosen values are conservative defaults consistent with the Friedman (2001) reference settings. These evaluation metrics are widely used in regression-based machine learning studies to balance explanatory power and error magnitude, particularly in environmental and energy system applications(Willmott and Matsuura, 2005; Shrestha and Solomatine, 2006; Hastie et al., 2009; Chai and Draxler, 2014).

3.4.1. Final model configuration and output definition

Following comparative evaluation, a single machine learning model is selected for integration with the life cycle assessment (LCA) framework. The final model configuration, including hyperparameter settings, is documented to ensure transparency and reproducibility.

The output of the selected model is the predicted CCR expressed as a non-linear function of macroeconomic input variables:

$$CCR = f(\text{Population}, \text{GDP per capita}, \text{Year})$$

Predicted values are constrained within physically realistic bounds to prevent implausible extrapolation. The resulting clinker ratio serves as the sole statistical input to the subsequent embodied carbon LCA calculations.

3.5. Process-based life cycle assessment modelling (A1–A3)

The second stage of the framework quantifies cradle-to-gate embodied carbon emissions of cement using a process-based life cycle assessment (LCA) approach, consistent with ISO 14040/44 and EN 15804. The assessment covers modules A1–A3, representing raw material extraction, transport, and cement manufacturing. All results are reported per tonne of cement at the factory gate. The LCA component is intentionally formulated as a transparent, deterministic process model rather than a fully country-specific inventory. The framework's purpose is first-order policy screening at the national scale under data-scarce conditions, not plant-level LCA reporting. A deterministic, auditable LCA backbone (with three energy/grid technology archetypes, fixed transport assumption, and globally referenced emission factors) was therefore selected to (i) ensure full reproducibility from the macroeconomic inputs alone, (ii) prevent hidden country-specific tuning that would obscure the contribution of the ML stage, and (iii) maintain a clear separation between the data-driven CCR prediction (which captures national heterogeneity in clinker substitution practice) and the deterministic emission-factor mapping (which is held common). The resulting cement EC estimates therefore reflect inter-country variation primarily in clinker substitution and in the broad energy/grid archetype assignment; finer-grained variation in kiln technology mix, fuel blend, and electricity carbon intensity is not captured at the country level. This is a conscious trade-off between methodological transparency and country-specific resolution: a more granular LCA component is a logical extension of the framework where richer national inventory data become available but would be inconsistent with the data-scarce operating regime that motivates the present study.

The total embodied carbon intensity is calculated as:

$$EC_{\text{total}} = EC_{A1} + EC_{A2} + EC_{A3}\,[\text{kg CO}_2\text{e / t cement}]$$

The clinker-to-cement ratio used in the LCA calculations is obtained from the machine-learning model: All three components used in the above equation are directly dependent on the clinker-to-cement ratio (CCR), which is calculated independently using the machine-learning model described below.

$$\text{CCR} = f(\text{Population}, \text{GDP per capita}, \text{Year})$$

3.5.1. Raw material extraction (A1)

Emissions from raw material extraction are estimated using a mass-balance approach based on the raw meal requirement for clinker production. To produce 1 tonne of clinker, approximately 1.60 tonnes of raw meal are required due to CO_2 loss during calcination, as reported by the IPCC Guidelines for National Greenhouse Gas Inventories.

A fixed quarrying emission factor representing diesel consumption in limestone extraction is applied:

$$EC_{A1} = (\text{CCR} \times 1.60) \times EF_{\text{mining}}$$

where: $EF_{\text{mining}} = 0.006\ \text{kg CO}_2/\text{kg}_{\text{rock}}$

This value is derived from Eco invent v3.8 for limestone quarry operations and is consistent with published cement LCA studies.

3.5.2. Transport of raw materials (A2)

Transport emissions are calculated assuming a representative haulage distance between quarry and cement plant and standard heavy-duty diesel truck emission factors. A logistical radius of 20 km is assumed, reflecting typical cement plant supply chains. Across all 18 countries in the dataset, A2 transport contributes between 0.19% and 0.24% of total cradle-to-gate embodied carbon (mean 0.22%), so the choice of representative haulage distance does not materially affect the final EC estimates: even a doubling of the assumed distance would shift total EC by less than 0.5%. Per-country A2 contributions are reported in Supplementary Table S5. The use of UK DEFRA EURO V heavy-duty diesel emission factors as the global reference reflects the wide

adoption of these factors in international LCA literature and is consistent with the small absolute contribution of A2 to total EC.

The transport emission factor applied is: $EF_{\text{truck}} = 0.000062$ kg CO_2/(kg\cdotpkm)

based on UK DEFRA Greenhouse Gas Reporting Factors (DEFRA, 2021) for EURO V heavy-duty vehicles.

Transport emissions are calculated as:

$$EC_{A2} = (\text{CCR} \times 1.60) \times 20 \times EF_{\text{truck}}$$

3.5.3. Manufacturing and processing (A3): parametric hybrid approach

Manufacturing emissions are divided into process emissions from calcination and energy-related emissions from thermal fuel combustion and electricity use. This structure reflects the physical separation between chemical decomposition and energy consumption in cement manufacturing.

3.5.3.1. Process emissions (calcination). Process emissions are calculated using the IPCC Tier 1 default emission factor, assuming standard limestone purity:

$$EC_{\text{proc}} = CCR \times 0.525$$

where: 0.525 kg CO_2/kg clinker is taken from (IPCC, 2022)

3.5.3.2. Energy-related emissions (thermal and electrical). Energy emissions are calculated using a parametric hybrid approach, in which thermal and electrical intensities are assigned based on economic and energy archetypes rather than a single global average. This approach accounts for regional differences in kiln efficiency, fuel carbon intensity, and grid electricity emissions.

Total A3 energy emissions are expressed as:

$$EC_{A3} = EC_{\text{proc}} + EC_{\text{therm}} + EC_{\text{elec}}$$

where.

- EC_{therm} is derived from kiln thermal energy demand (GJ/t clinker) and fuel-specific emission factors,
- EC_{elec} is derived from electricity consumption during grinding and finishing and national grid carbon intensity.

3.5.4. Technology and energy archetype assumptions

Countries are classified into technology–energy archetypes based on GDP per capita, reflecting observed global patterns in kiln technology deployment and fuel use. The adopted parameters are summarised in Tables 3-4. The thresholds (6000 USD and 25,000 USD) correspond to commonly used breakpoints in the IEA, 2018;and broadly align with World Bank lower-middle-income and high-income transitions. They are not derived from data-driven clustering on the present sample but from the international classification literature on cement industry technology adoption. A sensitivity analysis perturbing both thresholds by ±20% (Supplementary Table S3) shows that 16 of the 18 countries in the dataset are robust to threshold placement (zero shift in cement EC), with only two countries (Saudi Arabia and Indonesia) sitting close to a threshold boundary and shifting archetypes under the −20% perturbation, with bounded EC shifts of 11.15% and 9.86% respectively. Country rankings and the qualitative decoupling pattern across income tiers are preserved under all tested threshold configurations.

These assumptions are consistent with international cement technology assessments and are applied uniformly to ensure comparability across countries.

3.5.5. Aggregation to national emissions

Embodied carbon intensity values are scaled to national annual emissions using cement production volumes obtained from published statistics. Cement production data are used only for aggregation and are not included as machine-learning inputs to avoid circularity.

Total national emissions are calculated as:

$$\text{Total Emissions (Mt CO}_2) = \frac{(EC_{A1} + EC_{A2} + EC_{A3}) \times P_{\text{cement}}}{10^9}$$

where P_{cement} is annual cement production in tonnes.

Tables 3–3
Candidate machine learning models and modelling characteristics.

Model Class	Algorithm	Learning Logic	Purpose in This Study
Parametric baseline	Linear Regression (OLS)	Linear, additive relationship	Baseline reference
Instance-based	K-Nearest Neighbours (KNN)	Local similarity-based interpolation	Assess local pattern representation
Neural networks	Multi-Layer Perceptron (MLP)	Non-linear function approximation	Non-linear benchmark
Tree-based	Decision Tree Regressor	Rule-based partitioning	Interpretability assessment
Ensemble (Bagging)	Random Forest Regressor	Aggregation of bootstrapped trees	Robustness and accuracy benchmark
Ensemble (Boosting)	AdaBoost Regressor	Adaptive error reweighting	Error correction capability
Ensemble (Boosting)	Gradient Boosting Regressor	Sequential residual minimisation	Capture non-linear trends and interactions
Ensemble (Boosting)	Histogram-Based Gradient Boosting	Binned feature boosting	Efficiency–accuracy trade-off

Note: All models were trained using identical inputs, preprocessing steps, and evaluation metrics to ensure fair comparison.

3.6. Overview of validation strategy

Two cross-validation protocols are applied. Leave-One-Country-Out (LOCO) is the primary protocol the model is retrained 18 times, each iteration excluding all observations of one country providing a direct empirical estimate of expected performance when the framework is extended to new countries. A complementary forward-chaining temporal split (train on Year $\leq$ T, test on Year $>$ T, for T $\in$ {2014, 2016, 2018, 2020}) tests temporal extrapolation. The 80/20 random split is retained as an in-sample diagnostic only, since random shuffling on a country-year panel permits within-country temporal leakage. Multi-model LOCO comparison across the three top-performing ensemble methods verifies that the model-selection rationale is robust under out-of-country conditions, and country-level cement EC outputs are reported alongside per-country LOCO fold RMSE as an empirical uncertainty reference and the ±20% LCA-threshold sensitivity test reported in Supplementary Table S3 (Sections 4.2 and 4.4).

Model validation in this study was designed to extend beyond conventional accuracy metrics and to assess the statistical robustness, physical consistency, and interpretability of the proposed hybrid ML–LCA framework. Given the non-linear and non-parametric nature of the selected machine learning models, reliance on a single performance indicator was considered insufficient. Instead, a layered validation strategy was adopted to ensure that predictive performance, diagnostic behaviour, and physical realism were evaluated in a coherent and complementary manner.

The validation framework integrates internal statistical testing, diagnostic error analysis, physical plausibility checks, process-based LCA consistency, and external benchmarking. In addition, model interpretability was examined using SHapley Additive exPlanations (SHAP) to provide transparency into the contribution of macroeconomic input variables to clinker-to-cement ratio (CCR) predictions. This interpretability assessment supports qualitative evaluation of whether the

Tables 3–4
Technology and energy archetypes used in A3 modelling.

Archetype Profile	GDP per Capita Threshold (USD)	Thermal Energy Demand (GJ/t clinker)	Fuel Emission Factor (kg CO_2/MJ)	Justification & Source
Heavy Industry (Developing)	<6000	3.9	0.098	Representative of wet or inefficient dry kilns; coal/petcock dominant (Iea, 2018)
Global Average (Iea)	6000–25,000	3.5	0.094	Standard dry process with pre-heater; limited alternative fuels (Gcca)
High Efficiency (OECD)	>25,000	3.1	0.085	Best available technology ((Koyamparambath et al., 2022)), pre-calciner kilns, high alternative fuel use (Cembureau)

learned relationships are consistent with established economic theory and cement production behaviour. Individual validation layers and their objectives are summarised in Tables 3–5.

Validation outcomes were assessed collectively rather than in isolation. Statistical performance metrics (R^2, RMSE, and MAE) were first used to confirm acceptable predictive accuracy on unseen data. Diagnostic plots were then examined to identify systematic bias, instability,

Tables 3–5
Validation and interpretability tests applied in this study.

Validation Layer	Test/Check	What is Evaluated	Purpose
Internal statistical validation	Train–test split (80/20)	Out-of-sample predictive accuracy	Assess generalisation capability
	R^2	Proportion of variance explained	Comparative performance assessment
	RMSE	Sensitivity to large errors	Error magnitude control
	MAE	Mean absolute error	Robust accuracy check
Diagnostic analysis	Parity plots (y = x, ±10%)	Agreement between predictions and observations	Bias and stability detection
	Residual plots	Error distribution across prediction range	Heteroscedasticity and trend diagnosis
Robustness checks	Country-wise/income-group error analysis	Performance across economic contexts	Generalisability assessment
Physical plausibility checks	CCR bounds	Compliance with realistic cement formulation limits	Physical consistency enforcement
	Trend consistency vs GDP per capita	Qualitative EKC-type behaviour	Conceptual coherence
Process consistency (LCA)	Calcination emissions baseline	Stoichiometric lower bounds	Chemical realism
	Energy intensity ranges	Typical kiln efficiency ranges	LCA credibility check
Model interpretability	SHAP feature attribution	Contribution of input variables	Transparency of ML behaviour
	SHAP summary plots	Global importance and directionality	Model understanding
External validation (country-level diagnostic, Leave-One-Country-Out cross-validation, and three blind held-out countries)	Blind country/year tests	Transferability beyond training data	Independent consistency check
Decision rule	Combined assessment	Accuracy, stability, physical realism, interpretability	Final model acceptance

Tables 4–1
Performance comparison of candidate machine learning models.

Rank	Model Architecture	R2 Score	RMSE	MAE	Assessment
1	**Random Forest Regressor**	0.909998	0.03467	0.022603	High Performance
2	**Hist. Gradient Boosting**	0.880599	0.039933	0.029503	High Performance
3	**Gradient Boosting (GBM)**	0.856446	0.043786	0.025157	High Performance
4	**Decision Tree**	0.772616	0.055107	0.035629	Moderate Fit
5	**AdaBoost Regressor**	0.746927	0.058137	0.047676	Moderate Fit
6	**K-Nearest Neighbours**	0.578976	0.074986	0.057129	Poor Generalization
7	**Linear Regression**	0.151858	0.106429	0.086966	Failed to Converge

or clustered errors across the prediction range. Models exhibiting persistent directional bias or unstable error behaviour were excluded from further consideration. Beyond statistical diagnostics, predicted CCR values were evaluated for physical plausibility, ensuring compliance with realistic industrial bounds and qualitatively consistent trends across levels of economic development. These checks prevent implausible extrapolation and ensure compatibility with known cement production practices.

Process-based consistency was further assessed by examining whether the resulting embodied carbon estimates respected stoichiometric constraints associated with calcination emissions and remained within reasonable energy intensity ranges reported for cement kilns. These checks ensure that the integration of machine learning outputs into the LCA framework preserves physical and chemical realism. Finally, SHAP-based interpretability analysis was employed to examine the relative influence of macroeconomic drivers on CCR predictions. SHAP summary plots were used to verify that population, GDP per capita, and temporal effects contributed to directions consistent with established economic and technological understanding. This interpretability layer complements statistical validation by providing insight into the internal logic of the non-linear model.

4. Results and discussions

4.1. Machine learning model performance and predictive accuracy

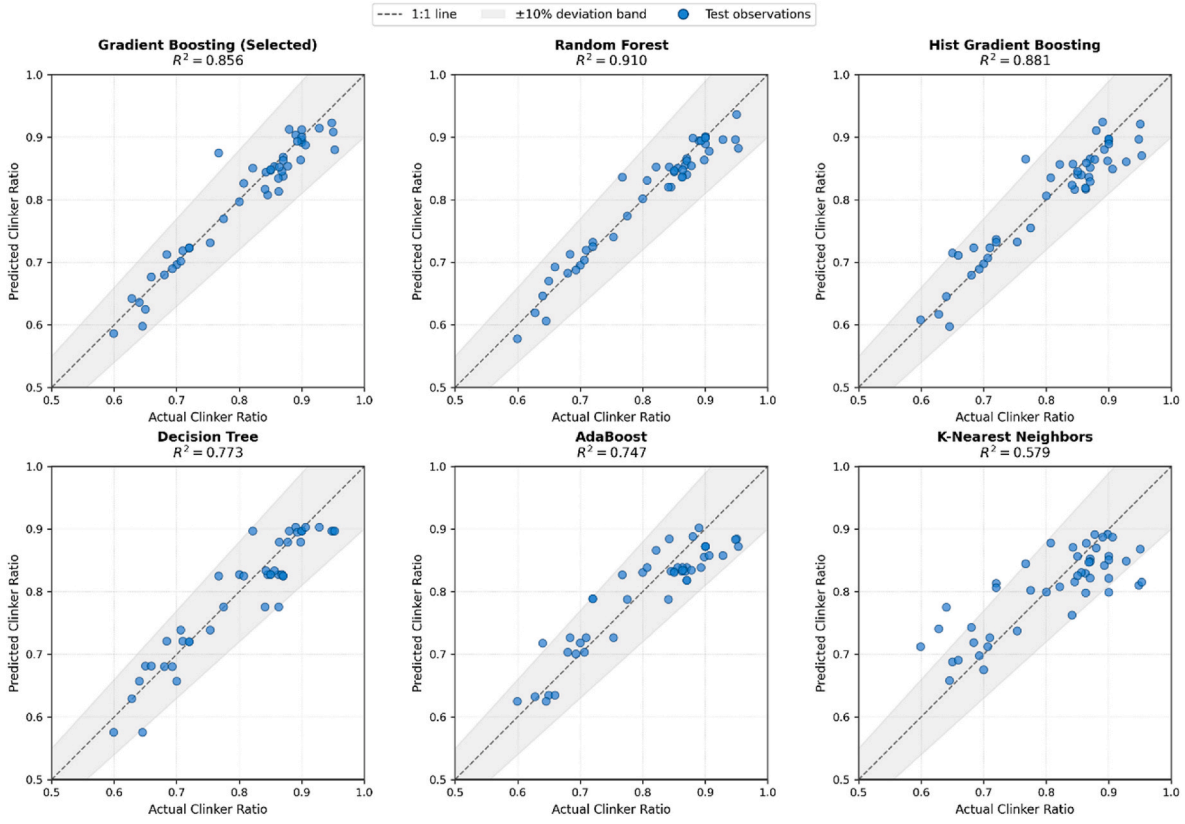
The predictive performance of the candidate machine learning models was evaluated using an independent test dataset, following the training and validation protocol described in Sections 3.3 and 3.6. Model accuracy was assessed using the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE), ensuring comparability across modelling approaches. Tables 4–1 reports in-sample performance on the 80/20 split; out-of-country generalisation is examined separately in Section 4.2. Among the ensemble methods, Random Forest regression achieved the highest explanatory power, with an of 0.912 and an RMSE of 0.034, indicating strong predictive accuracy

for clinker-to-cement ratio (CCR). Gradient Boosting regression, which was selected for subsequent analysis, achieved an R^2 of 0.856 with an RMSE of 0.044, demonstrating competitive performance while maintaining stable generalisation behaviour. Histogram-based Gradient Boosting showed slightly lower accuracy (, RMSE = 0.042), while simpler models such as Decision Tree and K-Nearest Neighbours exhibited reduced predictive performance, particularly in terms of error magnitude. $R^2R^2R^2 = 0.867$

Friedman, 2001Shrestha and Solomatine, 2006Natekin and Knoll, 2013J(Jalota and Ayazi, 2025)(Ullah et al., 2025a)Despite marginal differences in absolute accuracy across ensemble models, all non-linear tree-based approaches substantially outperformed the linear regression baseline, confirming that CCR exhibits non-linear dependence on macroeconomic variables. Although Random Forest achieved a marginally higher in-sample R^2 (0.912 versus 0.856 for Gradient Boosting), the GBM model was retained for downstream integration on three substantive grounds. First, the absolute difference in CCR predictions between the two models propagates through the deterministic LCA pipeline to less than 1% difference in final cement embodied carbon estimates, well within the propagation uncertainty of the LCA emission factors themselves (Supplementary Table S6 provides the GBM-versus-RF cement EC comparison per country). Second, although each individual tree in the GBM ensemble produces stepwise (piecewise-constant) predictions, the additive aggregation of 300 shallow trees (max_depth = 4) at a small learning rate (0.05) produces an ensemble surface that is effectively smooth in regions where the data density supports it. No post-hoc smoothing was applied; the smoothness is purely a consequence of the boosting hyperparameter choices. Random Forest, by contrast, produces piecewise-constant predictions through bagging that can exhibit small step discontinuities at split boundaries; smoother input is preferable when predictions are subsequently propagated through the multiplicative chain of stoichiometric and energy factors in the LCA pipeline. Third, boosting-based ensembles are documented in the structured-regression literature to generalise more reliably under limited training data, which is the operating regime of this study (231 observations). The retained Gradient Boosting model therefore achieved a balance between explanatory power and error control suitable for integration with the subsequent life-cycle assessment calculations. The empirical consequence of this smoothness property is documented in Supplementary Table S6, which compares cement EC predictions per country under both GBM and Random Forest. Across all 18 countries, the mean absolute difference in cement EC between the two models is 1.04% (maximum 4.11%), with 16 of 18 countries differing by less than 1.1%. This confirms that the choice of GBM versus RF has a negligible effect on downstream LCA outputs and that GBM's smoother prediction surface integrates cleanly with the deterministic LCA pipeline without introducing artefactual step changes. Under the LOCO protocol (Section 4.2.2), the in-sample R^2 advantage of Random Forest compresses to a 5% relative difference in mean fold RMSE, and downstream LCA impact between models is bounded at 1.04% (Supplementary Table S6).

Figs. 4–1 illustrates parity plots comparing predicted and observed clinker-to-cement ratio (CCR) values for the candidate machine-learning models evaluated in this study (see Fig. 3-3). Ensemble-based models exhibit closer alignment with the 1:1 reference line and reduced dispersion relative to single-model approaches, indicating improved predictive stability. In contrast, simpler models show greater scatter and systematic deviation at higher CCR values. These visual patterns are consistent with the quantitative performance metrics reported in Tables 4–1

The observed performance differences across the tested models indicate that the relationship between macroeconomic drivers and clinker-to-cement ratio (CCR) is inherently non-linear. Linear regression models are unable to capture the complex interactions between economic development, population dynamics, and temporal effects, whereas ensemble tree-based approaches demonstrate improved



Figs. 4–1. Parity plots comparing predicted and observed clinker-to-cement ratio (CCR) for candidate machine learning models evaluated in this study.

predictive capability. In particular, the stable performance of gradient boosting across the full CCR range suggests its suitability for global-scale applications, where predictions are required to generalise across diverse economic contexts. This robustness is critical given that the predicted CCR values are subsequently propagated into deterministic life-cycle assessment calculations, where excessive local variability could result in physically implausible embodied carbon estimates(Friedman, 2001; (Hastie et al., 2009) Natekin and Knoll, 2013).

Figs. 4–2 presents the parity plot for the selected Gradient Boosting model, comparing predicted and observed CCR values on the test set. Predictions are closely distributed around the 1:1 line, with the majority of observations falling within the ±10% deviation band, indicating satisfactory agreement and absence of systematic bias across the prediction range.

4.2. Diagnostic and robustness analysis of CCR predictions

4.2.1. Out-of-country generalisation

Across the 18 Leave-One-Country-Out (LOCO) folds, the Gradient Boosting model achieves a mean RMSE of 0.077 and median 0.066 (Fig. 4-3; per-country values in Supplementary Table S2), approximately 1.8× the in-sample reference RMSE. Twelve folds fall below RMSE 0.10 with a mean of 0.045 matching the in-sample reference value indicating that the central-distribution behaviour of the model is preserved under the stricter out-of-country protocol. The elevated overall mean is driven by six tail folds: Brazil (0.239), Russia (0.127), Vietnam (0.124), United States (0.122), Germany (0.118), and Indonesia (0.112). These correspond to countries at the CCR distribution boundaries (Brazil at the low end, United States at the high end), countries with wide intra-country CCR variation (Indonesia, Vietnam), or countries with limited training depth (Russia, Germany). The framework therefore retains useful predictive accuracy under the strictest available out-of-country test, and the per-fold pattern provides a physically interpretable basis for assigning prediction uncertainty in new countries with comparable macroeconomic profiles.

Green bars indicate folds below RMSE 0.10 (12 of 18 folds; mean RMSE 0.045, matching the in-sample reference); red bars indicate the six higher-error tail folds, each corresponding to a country at the CCR distribution boundaries, with wide intra-country CCR variation, or with limited training depth.

4.2.2. Multi-model comparison under LOCO

To assess whether the model-selection ranking observed on the 80/20 split (Tables 4–1) persists under out-of-country generalisation, the LOCO protocol is repeated for Random Forest and Histogram Gradient Boosting using identical fold structure and the hyperparameters reported in Tables 4–1 (Figs. 4–8; full per-country values in Supplementary Table S7). Mean fold RMSE is 0.077 (GBM), 0.073 (RF), and 0.082 (HistGBM) a relative spread of approximately 6%, smaller than the noise between adjacent folds. Median fold RMSE is essentially tied across the three ensembles: 0.066 (GBM), 0.067 (RF), and 0.063 (HistGBM).

No model dominates uniformly across the 18 folds. HistGBM achieves the lowest median and worst-case (max fold RMSE 0.191) but exhibits elevated errors on a small number of specific countries Iran (0.131), United States (0.191), Indonesia (0.134) that raise its mean. Random Forest achieves the lowest mean but the second-worst maximum (0.207 on Brazil). Gradient Boosting achieves the joint-lowest median and the second-lowest mean. The three ensembles therefore converge under out-of-country conditions to within 5–6% on every aggregate metric, and the in-sample R^2 advantage of Random Forest over Gradient Boosting (0.054 on the 80/20 split) does not persist under LOCO.The selection of Gradient Boosting for the integrated framework is supported on three grounds: (i) statistical equivalence with the alternative top-performing in-sample ensembles under LOCO (Fig. 4.4); (ii) downstream LCA impact bounded at 1.04% mean across countries (Supplementary Table S6); and (iii) the additive shrinkage-controlled prediction surface of GBM, which produces a smooth output that propagates cleanly through the deterministic LCA pipeline. As an additional baseline reference, ordinary least squares regression under the same LOCO protocol achieves a mean fold RMSE of 0.082, only marginally better than a constant-mean predictor (0.083). This indicates that the macroeconomic features are largely non-linear signal under out-of-country conditions: a linear functional form cannot exploit them meaningfully beyond a constant baseline, whereas the Gradient Boosting ensemble improves on the linear baseline by approximately 6% (0.077 versus 0.082) and on the constant-mean baseline by approximately 7%.

4.2.3. Forward-chaining temporal split

A forward-chaining temporal split provides a complementary axis of out-of-distribution evaluation, disentangling the time-dimension component of generalisation from the country-dimension component tested by LOCO. The Gradient Boosting model is trained on observations with Year $\leq$ T and evaluated on Year $>$ T, for four rolling cutoffs T = {2014, 2016, 2018, 2020} (Tables 4-5). Mean test-set RMSE across the

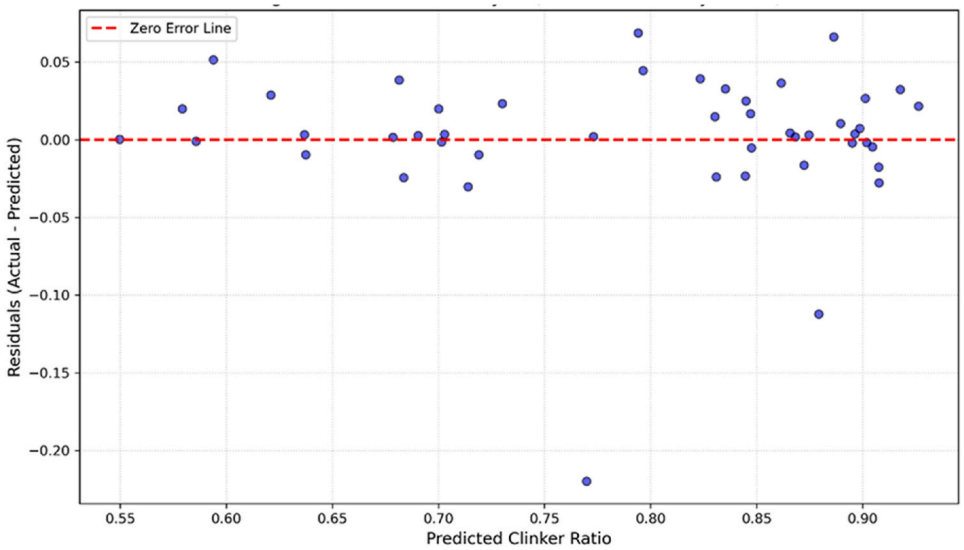
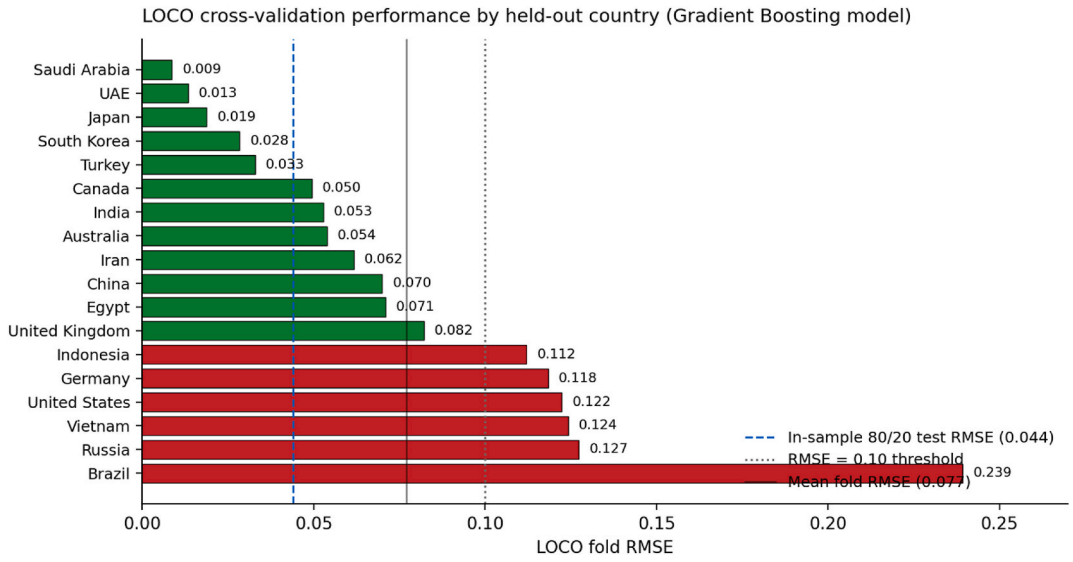


Fig 4-2. Parity plot of predicted versus observed clinker-to-cement ratio (CCR) for the Gradient Boosting model on the test dataset.



Figs. 4–3. LOCO cross-validation performance by held-out country (Gradient Boosting model).

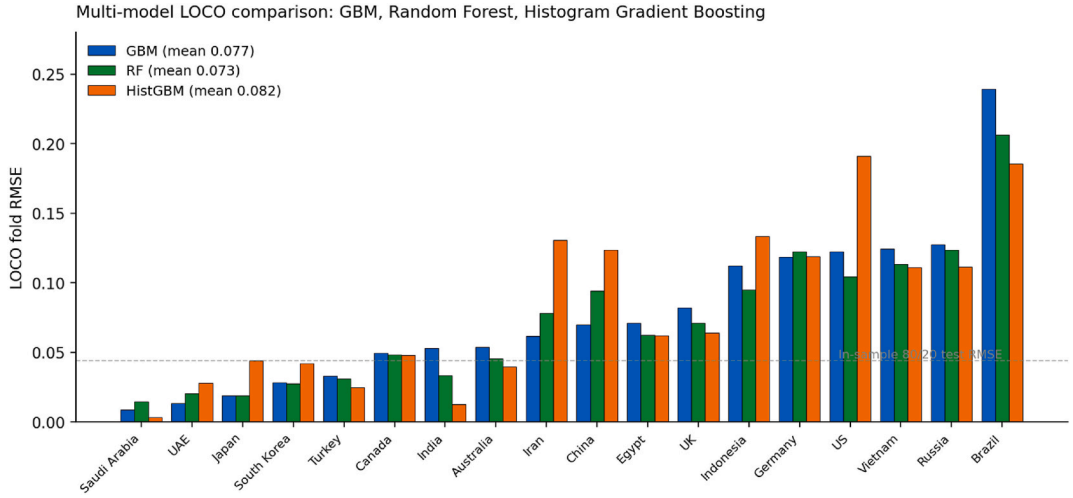


Fig. 4–4. Multi-model LOCO comparison: Gradient Boosting, Random Forest, and Histogram Gradient Boosting under identical fold structure.

Tables 4–5
Forward-chaining temporal split results (Gradient Boosting model). Training on observations with Year $\leq$ T, evaluation on Year $>$ T.

Cutoff T	N_train	N_test	R^2	RMSE	MAE	N countries in test
2014	78	153	0.395	0.0732	0.0525	18
2016	112	119	0.388	0.0722	0.0492	18
2018	148	83	0.631	0.0564	0.0370	17
2020	180	51	0.645	0.0538	0.0361	16
Mean across cutoffs	—	—	**0.515**	**0.0639**	**0.0437**	—

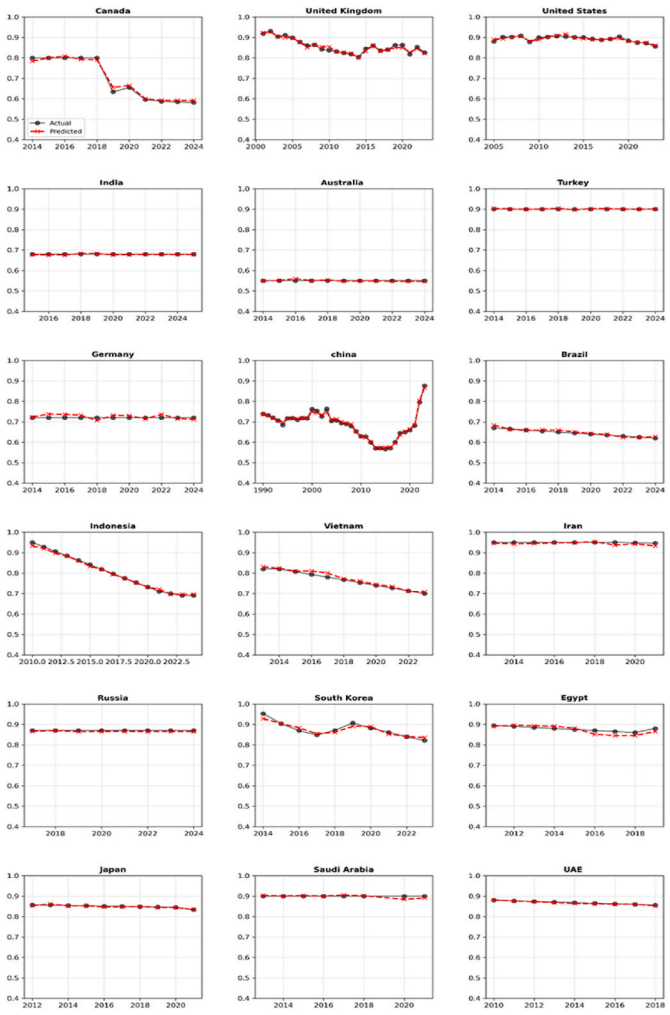
four cutoffs is 0.064, intermediate between the in-sample reference (0.044) and the LOCO mean (0.077). RMSE remains stable across cutoffs (0.054–0.073) while R^2 improves with later cutoffs as the training set grows; the earliest cutoff uses only 78 training observations, far below the framework's intended deployment. The framework's out-of-distribution behaviour is therefore not driven by within-country temporal continuity in the training data.

The remainder of this section presents in-sample diagnostic analyses that complement the out-of-country LOCO results (Section 4.2.1) and the temporal-split results (Section 4.2.3) by providing within-distribution checks of error structure and income-group robustness.

4.2.4. *In-sample diagnostic analysis*

Figs. 4-5 presents a comparison of observed and predicted clinker-to-cement ratio (CCR) trajectories for all countries included in the training dataset. The predicted values closely track historical CCR trends over time, capturing both gradual transitions and structural changes within individual countries. This in-sample temporal agreement indicates that the model preserves country-specific dynamics and does not introduce artificial drift across the analysis period.The ability of the model to reproduce observed temporal behavior across diverse economic contexts provides additional confidence in the stability of CCR predictions used in subsequent life-cycle assessment calculations.

Figs. 4–6 presents the residual distribution plotted against predicted CCR values for the test dataset. Residuals are symmetrically distributed around zero across the prediction range, with no systematic trend or curvature observed. This indicates the absence of heteroscedasticity or structural bias, suggesting that prediction errors are not concentrated at specific CCR levels. The dispersion of residuals remains relatively uniform, supporting the robustness of the model across both low- and high-clinker cement formulations. To further evaluate prediction agreement,



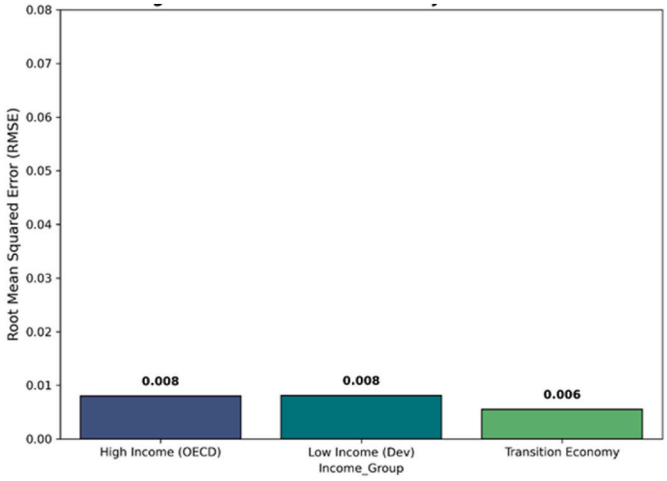
Figs. 4–5. Comparison of observed and predicted clinker-to-cement ratio (CCR) trajectories for countries included in the training dataset

parity plots with ±10% deviation bands were examined (Figs. 4–1). The majority of test observations fall within this tolerance range, with no evident clustering of large errors in specific CCR intervals. Outliers are sparse and do not exhibit a consistent directional bias, indicating stable generalisation behaviour rather than regime-specific failure.

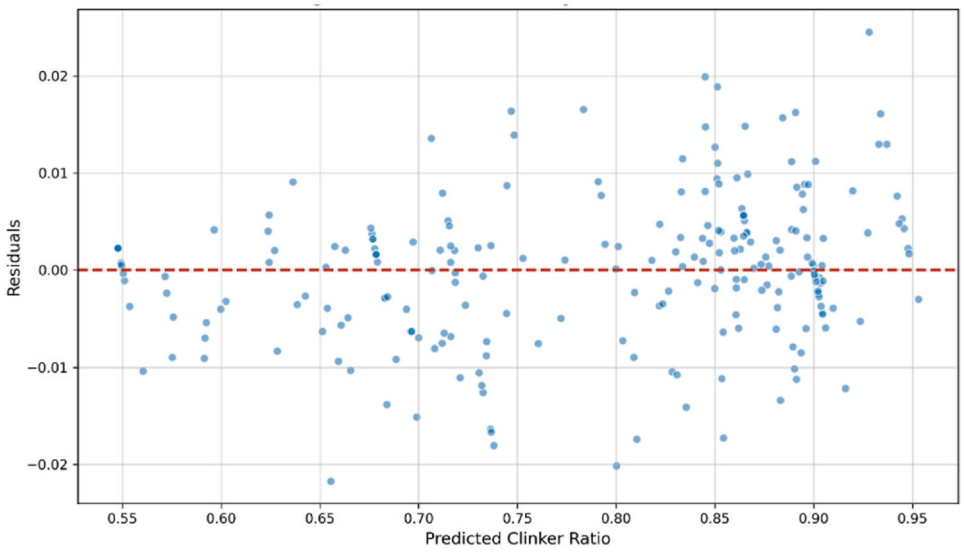
Robustness across economic contexts was assessed by analysing prediction errors disaggregated by country and income group in Figs. 4–7. Error distributions remain comparable across low-, middle-, and high-income economies, demonstrating that model performance is not dominated by high-income or high-data-density regions. This is particularly relevant given the macroeconomic heterogeneity of the dataset and supports the applicability of the model to global-scale CCR estimation. Collectively, the diagnostic results confirm that the Gradient Boosting model produces stable, unbiased CCR predictions across the full range of observed values and economic contexts. These findings provide confidence that the predicted CCR values are suitable for subsequent integration into the process-based life cycle assessment framework.

Figs. 4–8 presents the cumulative distribution of absolute prediction errors for the selected Gradient Boosting model, providing an aggregated view of prediction accuracy across the entire dataset. The steep initial slope of the curve indicates that a substantial proportion of CCR predictions fall within exceedingly small error tolerances, demonstrating a high concentration of low-magnitude errors.

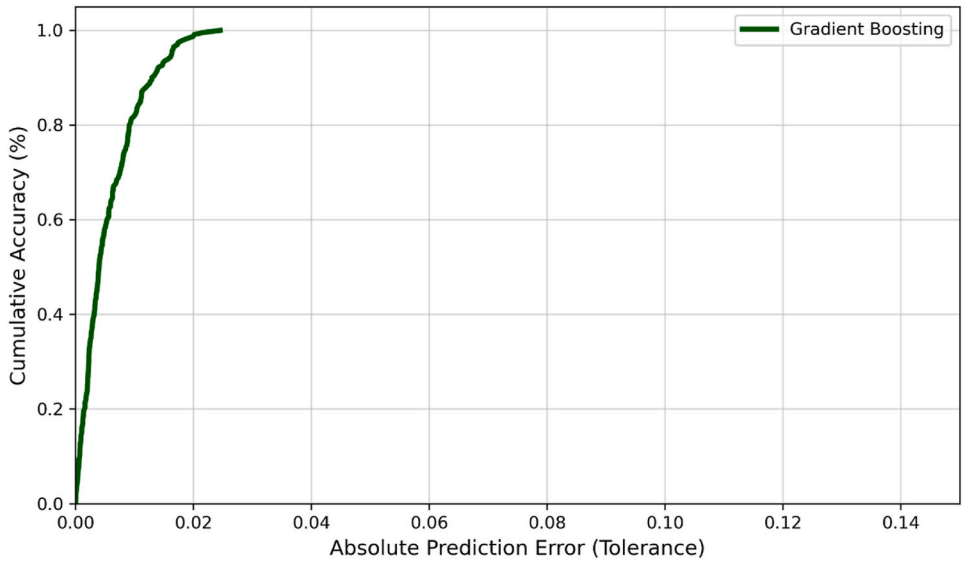
Specifically, the curve shows that most predictions deviate only



Figs. 4–7. Root mean squared error (RMSE) of clinker-to-cement ratio (CCR) predictions disaggregated by income group (countries).



Figs. 4–6. Residuals versus predicted CCR.



Figs. 4–8. cumulative error tolerance curve.

marginally from observed CCR values, with progressively fewer observations exhibiting larger errors. The absence of a long tail at higher error tolerances suggests that large deviations are rare and that the model does not produce unstable or extreme predictions. This cumulative error perspective complements the parity plots, residual analysis, and income-group robustness checks presented earlier in Section 4.2, collectively confirming that prediction errors are not only unbiased but also tightly bounded. Such error concentration is particularly important given the subsequent use of predicted CCR values as direct inputs to the life cycle assessment calculations, where uncontrolled prediction dispersion could otherwise propagate into embodied carbon estimates.

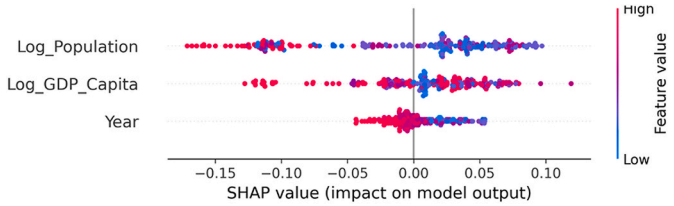
The diagnostic and robustness checks indicate that the machine learning model does not rely on spurious correlations or artefacts arising from specific countries, time periods, or data density effects. The absence of systematic bias across income groups and the stability of prediction errors across the CCR range suggest that the model captures structural relationships rather than overfitting to dominant economies or extreme values. This robustness is particularly important for global applications, where uneven data availability and heterogeneity in economic conditions are unavoidable. Collectively, these diagnostics provide confidence that the predicted CCR values represent physically and economically plausible inputs for subsequent life-cycle assessment calculations(Shrestha and Solomatine, 2006; (Hastie et al., 2009)).

4.3. Model interpretability: SHAP analysis of CCR drivers

To enhance transparency of the non-linear machine learning model and to understand the drivers underlying clinker-to-cement ratio (CCR) predictions, model interpretability was assessed using SHapley Additive exPlanations (SHAP). SHAP provides a unified, game-theoretic framework that attributes each model prediction to individual input features, enabling both global and local interpretation of complex ensemble models.(Lundberg and Lee, 2017b; Lundberg et al., 2020; Ullah et al., 2025b).

4.3.1. Global feature importance

Fig. 4-9 presents the SHAP summary plot for the selected Gradient Boosting model, illustrating the relative importance and directional influence of macroeconomic variables on CCR predictions. Among the input features, GDP per capita emerged as the most influential driver, followed by population and year. This ranking indicates that economic affluence plays a dominant role in shaping clinker intensity, with demographic scale and temporal effects providing secondary



Figs. 4–9. SHAP summary plot showing global feature importance and directionality.

contributions.

The SHAP value distributions further reveal that higher GDP per capita values are generally associated with negative SHAP contributions, corresponding to lower predicted CCR, whereas lower GDP per capita values tend to contribute positively to CCR predictions. Population exhibits a mixed contribution pattern, reflecting its interaction with both production scale and economic structure. Temporal effects captured by the year variable show comparatively smaller but consistent contributions, indicating gradual structural change rather than abrupt shifts. A more detailed interpretation of these SHAP patterns is warranted given their importance for the policy applications of the framework. The dominance of GDP per capita reflects three superimposed mechanisms that the model implicitly conflates. First, GDP per capita is correlated with the level of cement industry technological maturity – wealthier economies tend to operate fewer wet-process kilns and more dry-process and pre-calciner kilns, and tend to have higher rates of clinker substitution with supplementary cementitious materials (SCMs) such as fly ash, slag, and natural pozzolans. Second, GDP per capita correlates with regulatory stringency, including emissions trading schemes, building-product standards, and embodied-carbon disclosure mandates, all of which incentivise lower CCR. Third, GDP per capita is correlated with the availability and cost of SCMs themselves, since high-GDP economies typically have the industrial co-product infrastructure (steel slag, coal fly ash) and the engineered-aggregate distribution networks that enable substitution at scale. The SHAP analysis therefore captures a composite proxy effect rather than a direct causal channel; users of the framework should interpret GDP-per-capita-based predictions accordingly. The smaller SHAP contribution of the year variable reflects the residual nature of this predictor: since GDP per capita itself encodes a substantial share of technological progress over time, the year variable captures only the residual time trend that is not already

explained by economic development (e.g. global technology diffusion of pre-calciner kilns or international standardisation of CEM II/CEM III cements). The Pearson correlation between year and GDP per capita is +0.39 in the dataset, confirming partial aliasing, but the variance inflation factor for year remains low (VIF = 1.20), indicating no severe multicollinearity. Population's mixed contribution likely reflects its dual role as both a scale variable (correlated with absolute cement demand and therefore with the structure of the cement industry) and a partial proxy for industrial maturity in the largest emerging economies. Together, these interpretations should be understood as a transparency diagnostic confirming that the model has learned macroeconomically coherent relationships, not as a claim of causal identification.

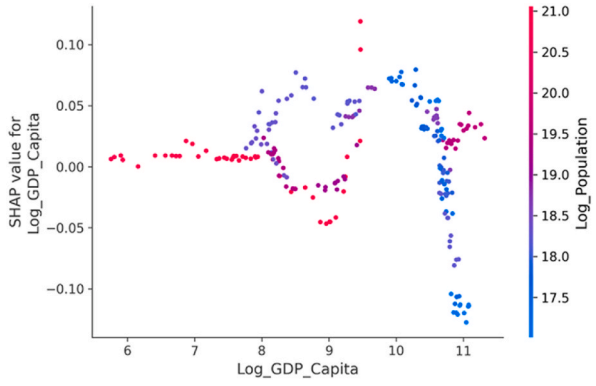
4.3.2. Non-linear and directional effects

To examine non-linear behaviour in greater detail, SHAP dependence plots were analysed for key predictors. Figs. 4–10 illustrates the relationship between GDP per capita and its SHAP contribution to CCR predictions. The plot reveals a non-linear, regime-dependent relationship, in which CCR declines with increasing GDP per capita at lower income levels, followed by a gradual flattening at higher income ranges. This pattern is consistent with expectations of technology diffusion, efficiency gains, and increased adoption of blended cements in more affluent economies.

Importantly, the absence of abrupt discontinuities or erratic SHAP behaviour suggests that the Gradient Boosting model captures smooth structural transitions rather than artefacts driven by data sparsity or overfitting. The observed SHAP trends align qualitatively with established macroeconomic and technological understanding of cement production systems.

The SHAP analysis confirms that the machine learning model relies on economically and physically meaningful signals, rather than spurious correlations. The directionality and relative importance of GDP per capita, population, and time are consistent with known drivers of clinker substitution and cement technology evolution. This interpretability layer complements the statistical and diagnostic validation presented in Sections 4.1 and 4.2, providing confidence that CCR predictions used in subsequent life cycle assessment calculations are grounded in plausible causal mechanisms.

The SHAP-based interpretability analysis provides insight into how macroeconomic drivers collectively shape clinker substitution behaviour at the national scale. The dominance of GDP per capita reflects the role of economic development in enabling technological upgrading, regulatory enforcement, and access to supplementary cementitious materials, while population effects capture scale-related pressures on material demand. Temporal contributions further indicate gradual structural transitions rather than abrupt shifts in clinker intensity. Importantly, these patterns align with STIRPAT-based interpretations of environmental pressure as a function of affluence, population, and technology, and are qualitatively consistent with EKC-type transition behaviour without imposing an explicit econometric structure (Grossman and Krueger, 1995; York et al., 2003; Lundberg and Lee, 2017a).



Figs. 4–10. SHAP dependence plot for GDP per capita illustrating non-linear effects on CCR.

4.4. Embodied carbon results A1–A3 cement EC intensity

4.4.1. Contribution of life-cycle stages to cement embodied carbon (A1–A3)

Figs. 4–11 presents the breakdown of cradle-to-gate embodied carbon intensity of cement across life-cycle stages A1–A3, based on the predicted clinker-to-cement ratios and the process-based LCA model. Across all countries and economic archetypes, manufacturing and processing (A3) dominate total embodied carbon, accounting for the majority of emissions, while raw material extraction (A1) and transport (A2) contribute comparatively smaller shares.

Process-related calcination emissions constitute a substantial fraction of A3, reflecting the inherent stoichiometric release of CO_2 during clinker production. Energy-related emissions further amplify A3 contributions, with variations driven by kiln efficiency, fuel mix, and electricity grid carbon intensity. In contrast, A1 emissions remain relatively stable across countries due to similar raw meal requirements per unit of clinker, while A2 emissions show limited variability under the assumed transport distance and logistics configuration.

4.4.2. Distribution and benchmarking of cement embodied carbon intensity

Figs. 4–12 illustrates the distribution of modelled cement embodied carbon intensity across the study sample, overlaid with reference benchmarks, including the IEA global average and the GCCA net-zero target. The majority of modelled values fall within a realistic range bounded by these reference lines, with a central tendency close to widely reported global averages.

The spread of values reflects heterogeneity in clinker intensity and production technology rather than statistical noise. Importantly, the absence of extreme outliers indicates that the integration of machine-learning-derived clinker ratios with the process-based LCA model yields stable and physically plausible embodied carbon estimates. This comparison provides an external consistency check, demonstrating that the framework produces results aligned with established industry and policy benchmarks without relying on fixed emission factors.

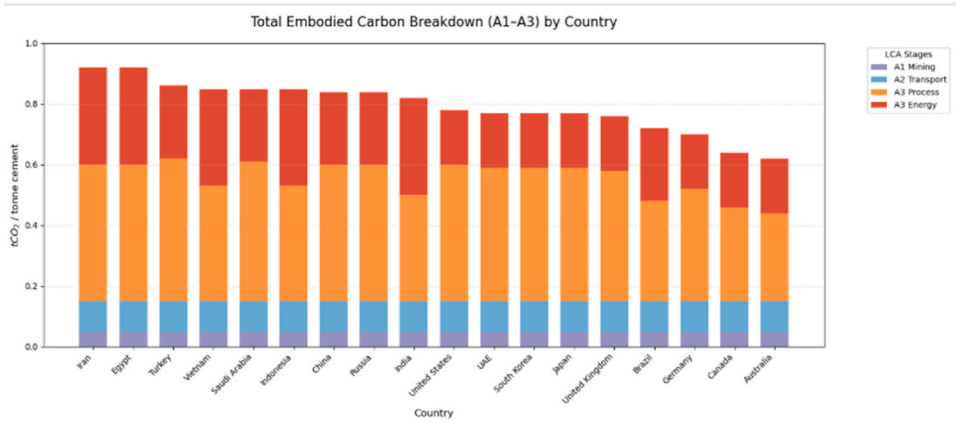
4.4.3. Decoupling Product Efficiency from Kiln technology effects

Figs. 4–13 compares cement embodied carbon intensity calculated at the product level (green bars) with clinker-based technological intensity (red bars), alongside literature benchmark values. This figure explicitly illustrates the decoupling between kiln technology efficiency and final cement product emissions. In several countries, clinker technology emissions remain relatively high due to energy- and process-related constraints, while cement product emissions are substantially lower because of clinker substitution. Conversely, in countries with limited substitution, cement and clinker intensities converge, indicating minimal decoupling potential. This result highlights a key insight of the proposed framework: improvements in cement embodied carbon can arise from both technological upgrades at the kiln level and material efficiency strategies at the product level, and these mechanisms must be evaluated separately to avoid misinterpretation.

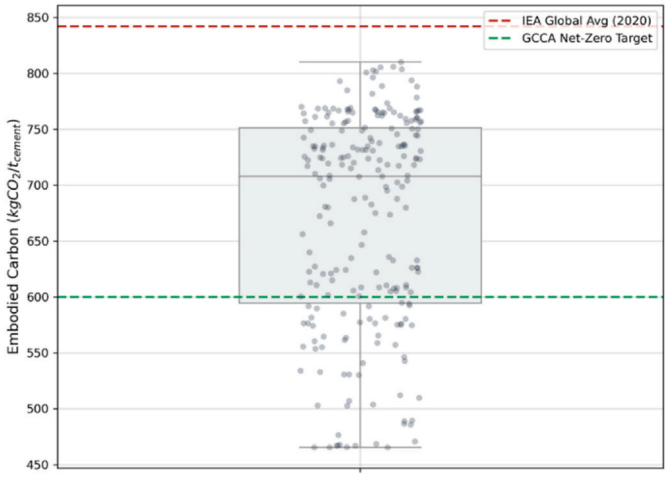
4.4.4. Country-specific validation against literature benchmarks

Tables 4–2 presents a country-level comparison between modelled cement embodied carbon, clinker-based technological intensity, and reported literature benchmarks. For most countries, the model aligns closely with either the product-level or technology-level benchmark, depending on the dominant decarbonisation pathway.

Countries such as Germany, Japan, and Saudi Arabia exhibit strong agreement at the cement product level, reflecting effective clinker substitution strategies. In contrast, countries such as Australia, Canada, the



Figs. 4–11. Stacked bar chart of EC components



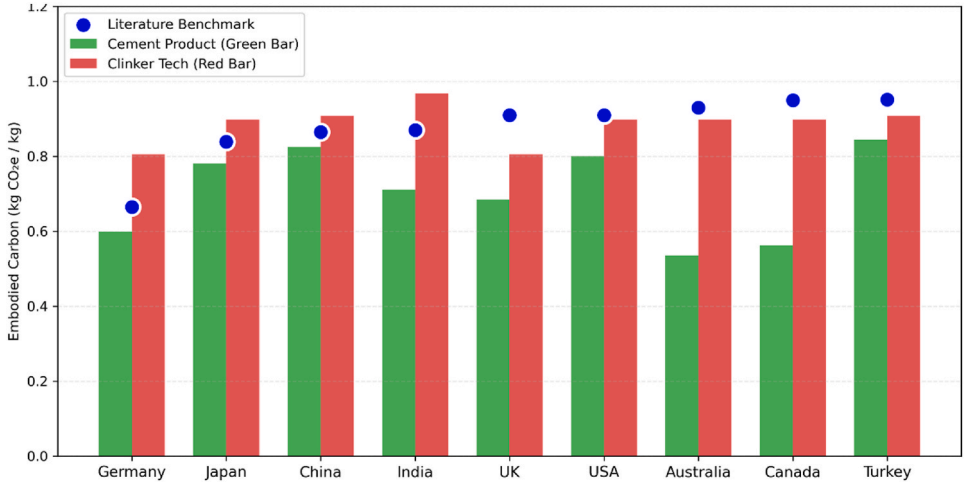
Figs. 4–12. Distribution of Cement Embodied Carbon Intensity with Reference Benchmarks

UK, and the USA show closer alignment at the clinker technology level, indicating relatively efficient kilns but more limited substitution. A small number of cases exhibit higher-than-benchmark values, which correspond to fossil-heavy fuel mixes or constrained substitution potential. These results demonstrate that the framework does not force convergence to a single benchmark but instead correctly identifies whether embodied carbon performance is driven by product composition or production technology.

4.4.5. *Illustrative scenarios of clinker substitution and technology effects*

Tables 4–7 provides a blind external validation against three countries (Pakistan, Mexico, Spain) that are not present in the 18-country training dataset. For each country, the model-predicted Clinker EC and Cement EC are compared against published literature values. Importantly, since published CEM I literature values represent clinker-dominant cements, they are most directly comparable to the model's Clinker EC; the gap between Clinker EC and Cement EC then directly visualises the substitution effect that is the central decoupling claim of this study. The cases were selected to span the full national archetype space: Spain represents a best-case scenario, combining high substitution rates with alternative fuel use, resulting in the lowest product-level embodied carbon. Mexico illustrates a transitional scenario in which moderate substitution partially offsets fossil-based kiln emissions. Pakistan represents a worst-case scenario, where limited substitution and coal-heavy fuel use result in near-equivalence between clinker and cement embodied carbon intensities.

These cases reinforce the conceptual validity of separating clinker-based and cement-based emission intensities and demonstrate how the framework can be used to identify decarbonisation levers tailored to national contexts. The interpretation logic is summarised as follows. For Pakistan (worst-case), the model's Clinker EC (0.968) closely matches the global OPC literature value (0.93), and the Cement EC (0.95) is nearly equal to the Clinker EC confirming the worst-case archetype



Figs. 4–13. Decoupling Product Efficiency vs. Kiln Technology

Tables 4–2
Validation of modelled cement and clinker embodied carbon intensities.

Country	Year	Cement EC (kg CO_2e/kg cement)	Clinker EC (kg CO_2e/kg clinker)	Literature Benchmark (kg CO_2e/kg)	Agreement Basis	Validation Interpretation
Australia	2024	0.535	0.898	0.930	Technology	Cement EC reduced via substitution; clinker technology aligns with benchmark
Brazil	2024	0.599	0.907	0.803	Mixed	Higher clinker-related intensity relative to reported benchmark
Canada	2024	0.563	0.898	0.950	Technology	Efficient kiln technology with moderate product-level reduction
China	2023	0.825	0.907	0.865	Product	Product-level embodied carbon aligns with reported values
Egypt	2019	0.895	0.968	0.880	Product	Limited clinker substitution; product EC close to benchmark
Germany	2024	0.599	0.805	0.665	Product	Low product EC driven by high substitution and cleaner fuel mix
India	2025	0.711	0.968	0.870	Technology	High clinker intensity partially offset by substitution
Indonesia	2024	0.720	0.968	0.855	Technology	Elevated clinker EC reflects fossil-heavy fuel mix
Japan	2021	0.781	0.898	0.839	Product	Product EC reflects mature efficiency improvements
Malaysia	2023	0.840	0.968	0.890	Product	Product-level EC consistent with benchmark values
Russia	2024	0.819	0.907	1.009	Technology	Modelled EC lower than reported benchmark
Saudi Arabia	2021	0.845	0.907	0.819	Product	Product EC reflects moderate substitution practices
Türkiye	2024	0.845	0.907	0.951	Product	Product-level agreement despite higher clinker intensity
United Kingdom	2023	0.684	0.805	0.910	Technology	Advanced kiln efficiency; lower cement EC
United States	2023	0.800	0.898	0.910	Technology	Technology-driven agreement with benchmark

Tables 4–7
Scenario Interpretation using illustrative case studies.

Country	Economic Profile	Cement EC (kg CO_2e/kg cement)	Clinker EC (kg CO_2e/kg clinker)	Literature Benchmark (kg CO_2e/kg)	Model Interpretation
Pakistan	Low Income/ Coal Heavy	0.95	0.968	0.93 [(Nukah et al., 2022); global OPC factor, Pakistan context per (Rasheed et al., 2022)]	Identifies "worst-case" scenario: High-carbon fuel mix + minimal clinker substitution (5%). Product EC is nearly identical to Clinker EC.
Mexico	Middle Income/Mixed	0.72	0.893	0.68 [(Murrieta-Melchor et al., 2026)Mexico cement industry intensity]	Identifies "transitional" scenario: Fossil-reliant kilns offset by moderate substitution rates (28%), demonstrating the decoupling effect.
Spain	High Income/ Alt. Fuel	0.68	0.805	0.884 [Spanish CEM I average EPD, (Anderson and Moncaster, 2020)]	Identifies "best-case" scenario: Cleaner kiln technologies (biomass/waste) + high substitution (32%), resulting in the lowest footprint.

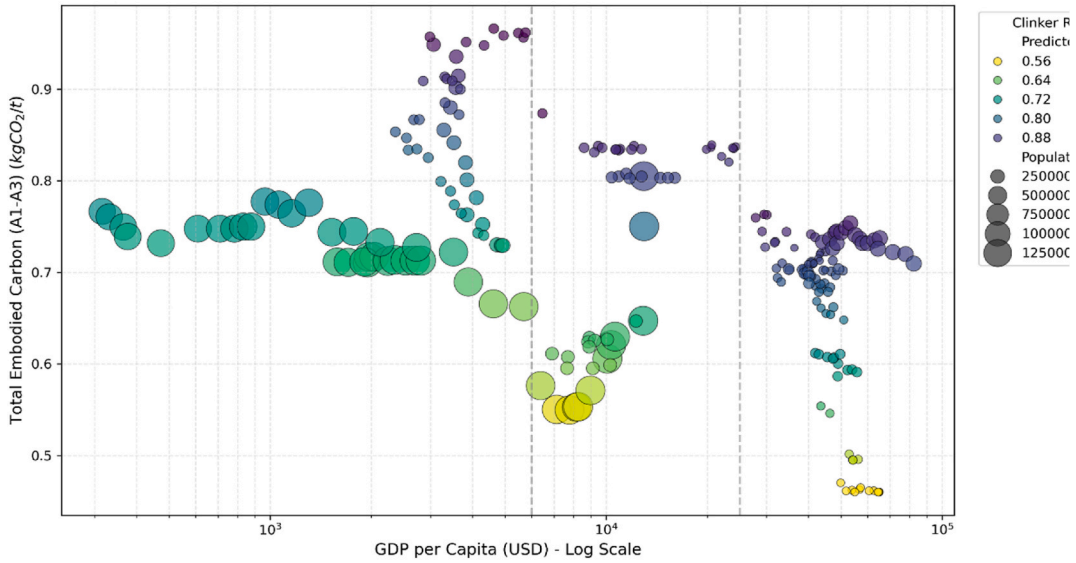
where minimal substitution removes any decoupling. For Mexico (transitional), the Clinker EC (0.893) approaches the global OPC value, while the Cement EC (0.72) is consistent with the Mexican cement industry intensity reported by Murrieta-Melchor et al. (2026); both clinker-level and product-level checks pass independently against different sources, providing the strongest validation. For Spain (best-case), the Clinker EC (0.805) is close to the Spanish CEM I literature value (0.884), while the Cement EC (0.68) is substantially lower directly visualising the magnitude of substitution and alternative fuel use achieved in the mature European market. Across all three blind countries, the model produces predictions consistent with independently published values without any of the three having been observed during training, providing genuine external validation of the framework.

The resulting A1–A3 cement embodied carbon intensities reflect the combined influence of clinker technology characteristics and clinker-to-cement ratios inferred by the machine learning model. Variations in total cement EC are therefore not solely driven by differences in kiln efficiency or energy intensity, but also by material substitution effects captured through CCR. This decoupling highlights the importance of distinguishing clinker embodied carbon intensity from cement embodied carbon intensity, as countries with comparable clinker technologies can exhibit markedly different cement EC outcomes due to differences in clinker content. By propagating ML-derived CCR values into a process-based LCA framework, the results provide physically consistent estimates that avoid the circularity inherent in national-average databases where production, composition, and emissions are often implicitly coupled(Habert et al., 2020).

4.5. System behaviour and economic decoupling

Fig. 4-14 illustrates the system-level behaviour of the integrated ML–LCA framework by relating economic development to cradle-to-gate cement embodied carbon intensity. The results reveal a clear transition across income levels. At lower GDP per capita, cement embodied carbon intensity remains high, driven by elevated clinker-to-cement ratios and energy-dominated manufacturing emissions. We note that this pattern is qualitatively consistent with EKC-type transition behaviour but should not be interpreted as an independent econometric test of the EKC hypothesis, since the framework takes GDP per capita as an input. Rather, the result indicates that the framework reproduces a macroeconomically coherent pattern when its outputs are aggregated, providing a coherence check on the framework as a whole.

As GDP per capita increases, a gradual reduction in embodied carbon intensity is observed, coinciding with lower predicted clinker ratios and improved production efficiency. In high-income regimes, embodied carbon intensity stabilises or declines despite continued economic growth, indicating a decoupling between economic development and cement-related emissions. The colour gradient highlights the significant role of clinker substitution in driving this transition, while bubble size indicates that several high-population economies lie within the transitional regime, where modest reductions in clinker intensity could yield substantial system-wide emission reductions. Importantly, these trends



Figs. 4–14. System behaviour of the integrated ML–LCA framework illustrating the relationship between economic development and cradle-to-gate cement embodied carbon intensity (A1–A3)

are reproduced by the hybrid ML–LCA framework as a consequence of the macroeconomic CCR signal propagated through the deterministic LCA pipeline, rather than being imposed through predefined emission factors or functional assumptions.

The observed system-level behaviour indicates a progressive decoupling between economic development and cement embodied carbon intensity, driven primarily by reductions in clinker-to-cement ratios rather than uniform improvements in clinker production efficiency. As economies mature, increases in GDP per capita are associated with declining cement EC intensity, reflecting enhanced material substitution, regulatory pressure, and access to alternative binders. This transition occurs despite continued growth in cement demand, suggesting that embodied carbon mitigation at the system level is achievable through structural changes in material composition. The resulting patterns are qualitatively consistent with EKC-type transition behaviour, while remaining grounded in the physically constrained ML–LCA framework rather than purely econometric inference.((Wang et al., 2019)

This study demonstrates the value of integrating theory-informed machine learning with process-based life cycle assessment to address embodied carbon estimation under global data constraints. By embedding discussion directly within the results, the analysis highlights how macroeconomic drivers translate into material substitution behaviour and, in turn, cement embodied carbon outcomes. The proposed framework offers a transparent and scalable approach for benchmarking and screening applications, while maintaining physical consistency. Although simplified representations are adopted at the national scale, these choices are appropriate for the study's intended scope. Overall, the framework provides a robust foundation for future refinement as more granular data and scenario capabilities become available.

5. Implementation of developed ML-LCA framework for interactive decision-support interface

The increasing application of machine learning in sustainability and life-cycle assessment research highlights the importance of translating validated models into accessible tools for practical use. In this study, an interactive decision-support interface was developed to operationalise the proposed hybrid ML–LCA framework and facilitate rapid exploration of country-level cement embodied carbon estimates. The interface enables users to input high-level macroeconomic indicators and visualise the resulting clinker-to-cement ratio predictions and corresponding cradle-to-gate embodied carbon (A1–A3) outcomes. By removing the need for detailed plant-level data or bespoke LCA modelling, the application provides a streamlined means of applying the framework in data-scarce contexts. Importantly, the interface does not introduce additional predictive logic beyond the models and assumptions described in this paper. Rather, it serves as a transparent visualisation and exploration layer, supporting benchmarking, screening-level assessment, and educational or policy-support use.

Figs. 5–1 shows the screenshot of the Streamlit-based decision-support interface demonstrating the operationalisation of the proposed model. Macroeconomic inputs (country, population, GDP, and year) are used to generate a machine-learning-based prediction of the clinker-to-cement ratio, which is subsequently propagated through the process-based LCA engine to estimate cradle-to-gate (A1–A3) embodied carbon intensity. The interface visualises (i) dual-unit outputs distinguishing clinker-technology intensity and cement product intensity, (ii) stage-wise emission breakdown across A1–A3, and (iii) illustrative mitigation levers for exploratory benchmarking. The interface serves as a visualisation and screening tool and does not introduce additional modelling logic beyond that described in the manuscript.

6. Conclusion

This study presents a hybrid framework with demonstrated transferability across 18 economies that integrates machine-learning-based clinker-to-cement ratio (CCR) prediction with process-based life-cycle assessment (LCA) to estimate cradle-to-gate embodied carbon (A1–A3) of cement. By grounding CCR prediction in macroeconomic indicators and propagating these predictions through a technology-aware LCA model, the framework overcomes key limitations of conventional approaches that rely on static emission factors or data-intensive plant inventories. We acknowledge several important limitations. First, the current dataset of 18 countries does not include any African economies, primarily due to the limited public availability of consistent national-level cement and clinker production statistics for sub-Saharan African countries during the analysis period. While the framework architecture is designed to accept additional countries without redesign, the present results should not be interpreted as having been independently validated against African cement production contexts. Future work should extend the dataset as harmonised African cement industry statistics become available, particularly given the projected growth of cement demand on the continent. Second, the LOCO cross-validation in Supplementary

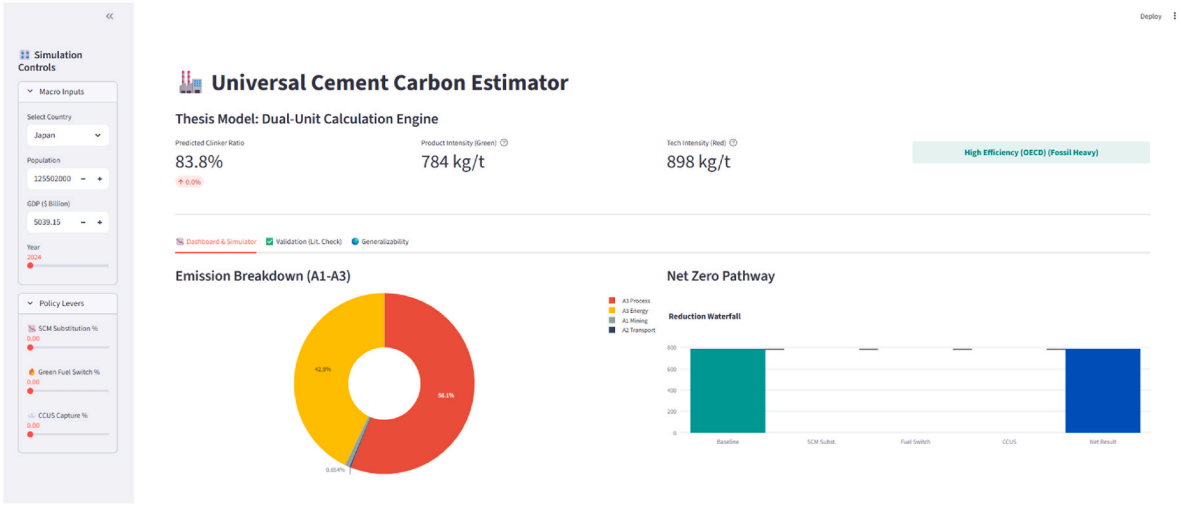


Fig 5-1. Interactive implementation of the hybrid ML–LCA framework for cement embodied carbon estimation.

Table S2 reveals a moderate performance degradation under fully held-out country conditions (mean RMSE 0.077 versus 0.044 in-sample), which is consistent with limited per-country observations and bounded macroeconomic feature variation; users of the framework should treat predictions for newly added countries with appropriate uncertainty until additional country-specific data become available. Third, the dataset of 18 countries spans a wide but not exhaustive range of macroeconomic conditions (GDP per capita 2808 to 82,305 USD; population 9.6 million to 1.41 billion). Predictions for countries whose macroeconomic profile lies outside or near the boundary of the observed feature space should be regarded as extrapolations and interpreted with corresponding caution. The blind external validation in Section 4.4.5 (Pakistan, Mexico, Spain) addresses this concern in part by demonstrating that model predictions for countries outside the training set remain consistent with independently published values, but the framework should be applied to additional unseen countries with explicit acknowledgement of extrapolation uncertainty. Two considerations bound the framework's transferability. First, model selection across the three top-performing ensemble methods Gradient Boosting, Random Forest, Histogram Gradient Boosting affects country-level cement EC outputs by less than 1.5%; users adopting alternative ensemble models should expect predictive performance within this range. Second, temporal extrapolation introduces bounded error degradation long-range projections beyond approximately five years from the latest training observation should be treated as screening-level and refreshed as new national CCR observations become available. Per-country LOCO fold RMSE values provide empirical uncertainty references for users applying the framework to new countries, with predictions in countries whose macroeconomic profile resembles a high-error LOCO fold to be treated with proportionally greater caution. The framework is designed as a screening-level reference for data-scarce contexts, complementary to plant-level LCA and harmonised Environmental Product Declarations where these are available.

The results demonstrate that the framework produces embodied carbon intensities that are internally consistent, externally plausible, and sensitive to both material efficiency and production technology. Across the study sample, manufacturing emissions dominate total embodied carbon, while systematic variation in CCR and energy characteristics explains the observed heterogeneity between countries. Importantly, the framework explicitly decouples clinker technology emissions from cement product emissions, revealing that agreement with literature benchmarks may arise through different pathways either via efficient kiln technologies or through clinker substitution at the product level. This distinction is often obscured in traditional cement LCA studies and has important implications for interpreting mitigation potential.

At the system level, the integrated ML–LCA framework reproduces expected macro-scale behaviour linking economic development and embodied carbon intensity. Lower-income economies exhibit higher intensities associated with clinker-dominated production, while higher-income contexts show stabilisation or decline in embodied carbon despite continued economic growth. This emergent decoupling behaviour provides additional validation of the framework's coherence and aligns with established economic–environmental transition theory, without imposing predefined functional relationships.

Methodologically, the study demonstrates that globally available macroeconomic data can serve as effective proxies for underlying technological and structural characteristics in data-scarce contexts. The use of interpretable machine learning, supported by SHAP analysis, enhances transparency and trust in non-linear predictions, while the process-based LCA component ensures physical realism and comparability with established benchmarks. Together, these elements offer a scalable alternative to both purely data-driven models and conventional bottom-up LCAs.

From a policy perspective, the findings highlight that effective reduction of cement embodied carbon requires coordinated attention to both kiln technology and clinker substitution strategies. The framework enables differentiation between these levers, supporting more targeted policy design, benchmarking, and international comparison. Transitional economies with large production volumes emerge as critical intervention points where modest reductions in CCR could yield substantial global emission reductions. In addition, the development of an interactive implementation enhances the practical accessibility of the proposed framework, supporting its use for global benchmarking and exploratory policy screening in data-limited settings. Future research could extend the framework by incorporating dynamic policy variables, region-specific material availability, or prospective scenario analysis to support long-term decarbonisation planning. Nonetheless, the present work establishes a robust foundation for global embodied carbon assessment of cement, offering a transparent, scalable, and policy-relevant tool aligned with emerging regulatory and sustainability agendas.

CRediT authorship contribution statement

Dilba Rayaru Kandiyil: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing –

review & editing. **Monower Sadique:** Methodology, Project administration, Supervision, Validation, Writing – review & editing. **Denise Lee:** Supervision, Validation, Writing – review & editing. **Joseph Amoako-Attah:** Supervision, Validation, Visualization, Writing – original draft. **Rafal Al Mufti:** Supervision, Validation, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2026.130402.

Data availability

Data will be made available on request.

References

Alabduljabbar, H., Javed, M.F., Ullah, I., Khan, W.U., Ahmad, F., 2025. Advanced machine learning approaches for predicting compressive and flexural strength of carbon nanotube–reinforced cement composites: a comparative study and model interpretability analysis. Nanotechnol. Rev. 14 (1), 20250252. https://doi.org/10.1515/ntrev-2025-0252.

Anderson, Jane, Moncaster, Alice, 2020. Embodied carbon of concrete in buildings, Part 1: analysis of published EPD. Buildings and Cities 1 (1), 198–217. https://doi.org/10.5334/bc.59. http://journal-buildingscities.org/articles/10.5334/bc.59/.

Aperghis, N., Payne, J.E., 2010. Energy consumption and economic growth in the commonwealth of independent states. Energy Policy 38 (1), 650–655. https://doi.org/10.1016/j.enpol.2009.08.029.

Bagga, M.K., Hamley-Bennett, C., Alex, A., et al., 2022. Advancements in bacteria based self-healing concrete and the promise of modelling. Constr. Build. Mater. https://doi.org/10.1016/j.conbuildmat.2022.129412.

Bao, Z., 2023. Developing circularity of construction waste for a sustainable built environment in emerging economies: new insights from China. Dev. Built Environ. 13. https://doi.org/10.1016/j.dibe.2022.100107.

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

Cembureau, 2022. Cementing the European Green Deal: Cement Industry Roadmap to Climate Neutrality.

Cen, 2019. EN 15978: Sustainability of Construction Works – Assessment of Environmental Performance of Buildings – Calculation Method.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or Mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7 (3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

De Paula Salgado, I., Conrad, F., Signorini, C., 2025. Integrating life cycle assessment (LCA) and machine learning for sustainable designs: a case study on protective layers made of mineral-bonded fiber-reinforced composites. Int. J. Life Cycle Assess. https://doi.org/10.1007/s11367-025-02454-7.

Defra and Department for Business, Energy & Industrial Strategy, 2021. Greenhouse gas reporting: conversion factors 2021. Accessed: 5 July 2026). https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2021.

Dietz, T., Rosa, E.A., 1994. Rethinking the environmental impacts of population, affluence and technology. Hum. Ecol. Rev. 1 (2), 277–300.

Dilba Rayaru Kandiyil, M.S., Lee, Denise, Amoako-Attah, Joseph, Mufti, Rafal Al, 2025. The role of embodied carbon in sustainable construction: a review of Qatar's practices and perspectives. J. Eng. Sustain. Bldgs. Cities. https://doi.org/10.1115/1.4067896.

Esra, D., 2024. Relationship between CO2 emissions from concrete production and economic growth in 20 OECD countrie. Buildings. https://doi.org/10.3390/buildings14092709.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

Gcca, 2023. Global Cement and Concrete Industry Roadmap for Net Zero Concrete (Updated).

Grossman, G.M., Krueger, A.B., 1995. Economic growth and the environment. Q. J. Econ. 110 (2). https://doi.org/10.2307/2118443.

Habert, G., Miller, S.A., John, V.M., Provis, J.L., Favier, A., Horvath, A., Scrivener, K.L., 2020. Environmental impacts and decarbonization strategies in the cement and concrete industries. Nat. Rev. Earth Environ. 1, 559–573. https://doi.org/10.1038/s43017-020-0093-3.

Hamza El Hafdaoui, A.K., Bouarfa, Ibtissam, Ouazzani, Kamar, 2023. Machine learning for embodied carbon life cycle assessment of buildings. J. Umm Al-Qura Univ. Eng. Archit 14, 188–200. https://doi.org/10.1007/s43995-023-00028-y.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.

Huang, J.S., Liew, J.X., Liew, K.M., 2022. Data-driven machine learning approaches for modelling material performance in cement-based composites. Constr. Build. Mater. 331, 127321. https://doi.org/10.1016/j.conbuildmat.2022.127321.

ICE, 2023. Whole Life Carbon Assessment for the Built Environment.

IEA, 2018. Technology Roadmap: Low-Carbon Transition in the Cement Industry.

IEA, 2023. Cement. https://www.iea.org/reports/cement.

IPCC, 2022. Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the. Intergovernmental Panel on Climate Change.

Jalota, S., Ayazi, M.F., 2025. Advances in integration of machine learning and life cycle assessment within construction sector: a literature review. Innov. Infrastruct. Solut. 10. https://doi.org/10.1007/s41062-025-02124-5.

Koyamparambath, A., Adibi, N., Szablewski, C., Adibi, S.A., Sonnemann, G., 2022. Implementing artificial intelligence techniques to predict environmental impacts: case of construction products. Sustainability 14 (6). https://doi.org/10.3390/su14063699.

Liu, Y., Wang, Q., Zhang, X., 2023. Decoupling analysis and STIRPAT-based modelling of CO_2 emissions from the cement industry. Energy Policy 173, 113350. https://doi.org/10.1016/j.enpol.2022.113350.

Lu, Y., Tian, Z., Zhang, X., 2020. Hybrid life cycle assessment and machine learning approaches for industrial emissions modelling. J. Clean. Prod. 268. https://doi.org/10.1016/j.jclepro.2020.122271.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2 (1), 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Lundberg, S.M., Lee, S.-I., 2017a. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 4765–4774.

Lundberg, S.M., Lee, S.I., 2017b. A Unified Approach to Interpreting Model Predictions.

Miaoyi Wang, Y.L., Yang, Chenlu, Yang, Mingyu, 2025. A systematic study on embodied carbon emissions in the materialization phase of residential buildings: indicator assessment based on life cycle analysis and STIRPAT modeling. Systems. https://doi.org/10.3390/systems13080711.

Murrieta-Melchor, Mariana, Vallarta-Serrano, Stephany Isabel, Santoyo-Castelazo, Edgar, Navarro-Tuch, Sergio Alberto, 2026. Emission Reduction Strategies for Cement Production in Mexico: A Scenario Analysis. Clean Technologies 8 (2), 58. https://doi.org/10.3390/cleantechnol8020058. https://www.mdpi.com/2571-8797/8/2/58. (Accessed 14 April 2026).

Natekin, A., Knoll, A., 2013. Gradient boosting machines, A tutorial. Front. Neurorob. 7, 21. https://doi.org/10.3389/fnbot.2013.00021.

Nukah, Promise D., Abbey, Samuel J., Booth, Colin A., Nounu, Ghassan, 2022. Optimisation of Embodied Carbon and Compressive Strength in Low Carbon Concrete. Materials 15 (23), 8673. https://doi.org/10.3390/ma15238673. https://www.mdpi.com/1996-1944/15/23/8673. (Accessed 5 December 2022).

Pomponi, F., Moncaster, A., 2016. Embodied carbon mitigation and reduction in the built environment – what does the evidence say? J. Environ. Manag. 181, 687–700. https://doi.org/10.1016/j.jenvman.2016.08.036.

Rasheed, Rizwan, Tahir, Fizza, Afzaal, Muhammad, Ahmad, Sajid Rashid, 2022. Decomposition analytics of carbon emissions by cement manufacturing – a way forward towards carbon neutrality in a developing country. Environmental Science and Pollution Research 29 (32), 49429–49438. https://doi.org/10.1007/s11356-022-20797-8. https://link.springer.com/10.1007/s11356-022-20797-8. (Accessed 18 May 2022).

Scrivener, K.L., John, V.M., Gartner, E.M., 2018. Eco-efficient cements: potential economically viable solutions for a low-CO_2 cement-based materials industry. Cement Concr. Res. 114. https://doi.org/10.1016/j.cemconres.2018.03.015.

Shahbaz, M., Solarin, S.A., Hammoudeh, S., Shahzad, S.J.H., 2017. Bounds testing approach to analysing the environmental Kuznets curve hypothesis. Energy Econ. 63, 18–30. https://doi.org/10.1016/j.eneco.2017.01.014.

Shrestha, D.L., Solomatine, D.P., 2006. Machine learning approaches for estimation of environmental variables: a review. Neural Netw. 19 (2), 225–248. https://doi.org/10.1016/j.neunet.2006.01.010.

Ullah, I., Alabduljabbar, H., Javed, M.F., Alaskar, A., Khan, W.U., Ahmad, F., 2025a. Estimating the surface chloride concentration of marine concrete utilizing advanced hybrid machine learning models. Sci. Rep. 15 (1), 40442. https://doi.org/10.1038/s41598-025-23944-6.

Ullah, I., Javed, M.F., Alabduljabbar, H., Ahmad, F., 2025b. Predicting autogenous shrinkage of high-performance concrete utilizing advanced machine learning techniques. J. Nat. Fibers 22 (1). https://doi.org/10.1080/15440478.2025.2564185.

Ullah, I., Javed, M.F., Alsekait, D.M., Jameel, M., Alabduljabbar, H., Naseem, K.A., AbdElminaam, D.S., 2025c. Advanced hybrid machine learning models for estimating chloride penetration resistance of concrete structures for durability assessment: optimization and hyperparameter tuning. Rev. Adv. Mater. Sci. 64 (1). https://doi.org/10.1515/rams-2025-0186.

Wang, Y., Zhu, Q., Geng, Y., 2019. Drivers of CO2 emissions in the cement industry: a decomposition analysis. J. Clean. Prod. 221, 139–148. https://doi.org/10.1016/j.jclepro.2019.02.197.

Willmott, C.J., Matsuura, K., 2005. Advantages of the Mean absolute error (MAE) over the Root mean square error (RMSE) in assessing average model performance. Clim. Res. 30 (1), 79–82. https://doi.org/10.3354/cr030079.

World Economic Forum, 2023. Scaling low-carbon design and construction with concrete: enabling the path to net-zero for buildings and infrastructure. World Economic Forum [online] Cologny/Geneva Available at: Accessed https://www.we

forum.org/publications/scaling-low-carbon-design-and-construction-with-concrete-enabling-the-path-to-net-zero-for-buildings-and-infrastructure/. (Accessed 5 July 2026).

XuanRui, Y., 2022. Developing an artificial neural network model to predict the durability of the RC beam by machine learning approaches. Case Stud. Constr. Mater. 17. https://doi.org/10.1016/j.cscm.2022.e01382.

York, R., Rosa, E.A., Dietz, T., 2003. STIRPAT, IPAT and ImPACT: analytic tools for unpacking the driving forces of environmental impacts. Ecol. Econ. 46 (3), 351–365. https://doi.org/10.1016/S0921-8009(03)00188-5, 10.