International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 13, No. 2, 2026, pp. 93-107.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Verification and Validation of Machine Learning Outputs in Regulated Energy and Emissions Reporting Workflows

Maheswara Rao Gorumutchu [a*]

[a] HYR Global Source Inc, Kansas, United States. E-mail: gmrmails@gmail.com, Orcid: https://orcid.org/0009-0009-5868-5034*

Abstract

The increasing application of machine learning in energy and emissions reporting has posed grave questions on the aspects of reproducibility, explainability, defensibility of uncertainty, and regulatory admissibility. Conventional machine learning systems primarily aim at predictive performance, and controlled reporting systems require predictable, verifiable, and auditable results to conform to the legal and operational constraints. The paper introduces a verification and validation system of machine learning outputs based on assurance in regulated energy and emissions reporting systems. The framework breaks down the working process of operation into particular training, inference, verification and reporting sectors to make sure that uncontrolled analytical behaviour does not infiltrate to the regulator confronted with disclosures. The proposed model integrates the use of XGBoost regression to predict emissions, Isolation Forest to identify anomalies, conformal prediction to confirm the uncertainty, and explainability analysis with SHAP. Reproducibility and regulatory robustness were ensured by verification invariants, deterministic execution environments, evidence preservation controls for governance, and controlled reporting interfaces. These mechanisms were evaluated using 1.2 million industrial simulated reporting instances collected from 12 facilities over 24 reporting periods. The outcome of experiments showed high predictive and operational performance. The XGBoost model achieved an $RMSE$ of 3.12%, MAE of 2.41%, and an R^2 score of 0.94. The reproducibility was stable with 98.8% retrained a number of times and redeployed across a number of revisions. The inspection showed 98.7% compliance with physical constraints, 97.9% compliance with explainability, and 94.8% regulatory admissibility, respectively. The detection accuracy of the Isolation Forest anomaly detection model was 96.3%, and the false positive rate was 3.8%, and the conformal prediction interval included 95.1% of the uncertainty at the regulatory levels. Scaling analysis showed that the overhead of verification grew in a steady and predictable manner with increasing multi-facility reporting setups. The findings show that assurance-driven ML models can be employed to build trusted and transparent energy and emissions reporting platforms.

Keywords: Machine Learning Verification, Energy and Emissions Reporting, Regulatory Admissibility, XGBoost Regression, Explainable Artificial Intelligence, Uncertainty Verification, Assurance-Centric Reporting Framework.

DOI: 10.71086/IAJSE/V13I2/IAJSE1356
Received: 25 February 2026/Revised: 04 April 2026/Accepted: 07 May 2026/Published: 30 June 2026

Introduction

In the last ten years, energy and emissions reporting have grown to be not a classic accounting task, but a software and systems-based engineering issue. Recent reporting processes are more and more automated using machine learning (ML) in emissions estimation, imputation of activity data, anomaly detection, and source attribution. However, the question of whether ML models can generate predictions is no longer the big question, but whether their outputs can be seen as digital artifacts that are palatable by regulators, supported by reproducibility, auditability, and evidence-based verification Baylor et al., (2017). The results of regulatory reporting must be reproducible and validated in a controlled environment.

Newer studies have revealed that old ML systems have problems with poor calibration of their uncertainty values, limiting them to applications involving compliance and thus posing a problem with respect to the issue of regulators' interest in reliability and defensibility of individual value reports against aggregate predictions (Rahman & Haque, 2023). Uncertainty and verification research shows that limited and testable uncertainties are vital in reporting systems (Zheng et al., 2022). Additionally, the requirement for explanations and provenance is an urgent demand nowadays. Model documentation and lifecycle visibility studies reveal the importance of feature descriptions, evaluation context, and documented constraints to increase accountability and audit-readiness (Sadr et al., 2024; Mitchell et al., 2019).

The demand for verification structures in control is further emphasized by the increasing application of heterogeneous data sets, such as methane monitoring using satellites. It was shown that ML techniques can detect anomalies in emissions events but require system validation and governance frameworks (Varon et al., 2019; Yu et al., 2023). Reproducibility studies show that reliable ML outcomes do not depend on source code alone but also on other factors such as execution environments and data set versioning (Ogunyemi, 2021; Sajawal Akhtar et al., 2026).

ML at scale also reveals that each of the three elements, namely, metadata capture, validation controls, and the reproducibility mechanisms, can be closely incorporated into the working processes (Tettey et al., 2025; Adepu, 2022). Overall, the existing literature contributes to the notion that the effective adoption of ML in the regulated reporting process does not necessarily imply the introduction of complexity into the model per se, but the assurance-oriented nature of the systems. Therefore, the following paper will focus on verification, reproducibility, explainability, and evidentiary governance as the pillars of sound ML-based energy and emissions reporting solutions.

The significance of this research lies in the fact that it takes into consideration one of the biggest issues facing current times, namely the problem of integrating machine learning techniques into regulatory settings where high levels of reliability, transparency, and auditability are necessary. The current state of machine learning algorithms is focused mainly on improving prediction accuracy, while in regulated reporting environments, machine learning solutions need to be capable of providing results that are admissible, reproducible, and explainable. The framework discussed in this paper helps to solve this problem by using various controls, including anomaly detection and uncertainty measures, together with the process of explanation within a controlled environment.

Key Contributions

- Proposed an assurance-based verification and validation system for machine learning outputs in controlled energy and emissions reporting systems.
- XGBoost regression and Isolation Forest anomaly detection, coupled with conformal prediction and SHAP explainability, are integrated in a single regulatory reporting architecture.
- Added verification, deterministic execution environments, and governance controls that maintain evidence to improve reproducibility and auditability.
- Proven high regulatory admissibility, stability of reproducibility, and scalable operational functionality in multi-facility reporting situations.

The article is divided into six large parts. Section II examines the existing literature in explainable AI, emissions monitoring, and regulatory compliance. Section III is about regulatory verification limitations and admissibility limits. Section IV introduces the assurance-based approach and verification system. Section V analyses the outcomes of experiments, scalability and robustness. Lastly, Section VI gives an end to the study with findings, implications and research directions.

Literature Survey

The latest trends in machine learning have had a profound impact on the energy and emissions reporting systems, especially in the fields of regulatory compliance, sustainability monitoring, and automated verification. This paper

proposed TensorFlow Extended (TFX), which is a large-scale machine learning production system that emphasizes the significance of pipeline repeatability and controlled production environments in industrial AI systems (Baylor et al., 2017). Likewise, it highlighted the increasing need for explainable and safe machine learning designs in regulated systems by stating that it is essential to have transparency and auditability to ensure that decision support applications that focus on compliance rely on them (Rahman & Haque, 2023; Owolabi, 2025).

Some of the studies have specifically studied emissions monitoring and environmental reporting. This work has created a benchmark dataset that can be used in machine learning in decarbonized energy grids, illustrating how major time-series datasets can be used to aid operational sustainability analytics (Zheng et al., 2022). It explored machine learning to perform verification in emissions trading systems and discovered that automated validation systems enhanced consistency in reporting, as well as minimized the risk of data manipulation (Yu et al., 2023; Moghadasi et al., 2025). This work also showed how environmental monitoring systems based on machine learning can facilitate real-time compliance and anomaly detection in industrial processes (Adepu, 2022).

Explainability and governance in high-stakes machine learning have also been the focal points of the recent literature (Ojadi et al., 2023; Udoh et al., 2024). This argued that explainable models perform much better than opaque black boxes when it comes to regulated sectors, and suggested model cards as a possible documentation format for reporting on explainable models (Mitchell et al., 2019; Rudin, 2019). These works provide direct rationale as to why there should be explainability and evidence-saving mechanisms in controlled emissions reporting systems.

In recent years, there has been growing research on predictive emissions monitoring (Nutkiewicz et al., 2018). The above research used machine learning methods to make predictions on vehicle CO2 emissions with a high level of accuracy, whereas it has conducted a review of the predictive emissions monitoring systems and identified the problem of reproducibility and uncertainty management as significant issues that remained unaddressed (Baplawat et al., 2026; Wang et al., 2025). It demonstrated the effectiveness of AI-driven automation in sustainability and CDP reporting analysis, reinforcing the importance of integrating AI with ESG compliance frameworks (Rangarajan et al., 2025).

In general, based on the literature, it can be noted that the literature is mostly concerned about predictive performance and operational efficiency, while little has been said regarding assurance verification, enforceable admission, and control over reproducibility. This paper bridges this gap by proposing an ML verification and validation approach designed specifically for energy/emissions regulatory reporting frameworks.

Regulatory Verification Limits of ML-derived Outputs

The output of Machine Learning, when used in the context of reporting for the energy and emissions sectors, interacts with a fundamentally different set of constraints than those relevant to analytical or operational decision-support systems. In regulated workflows, the decision on the admissibility of an output is not based on predictive quality, computational efficiency, or novelty of the approach, but on the extent to which the output meets a set of legally binding standards on auditability, reproducibility, explainability, and the robustness of evidence. These criteria set the verification boundaries and regulate the extent to which an ML output can be used in a regulator-facing report as opposed to remaining in an exploratory or internal analytics function. A characteristic of these reporting environments is that the reported figures must be justifiable for a considerable period after it was first produced. Supervening regulatory reviews, third-party audits, and ultimately litigious scrutiny occur long after the reporting cycle closes. Given this, ML outputs must, in essence, be justifiable by to-be-preserved artifacts that permit reconstruction of the computation for which the artifact was produced. Any output that cannot be reproduced via documented inputs, settings, and execution parameters is deemed not admissible, no matter how accurate it might be. This demands that the chief evaluative consideration shift from statistical correctness to procedural defensibility.

The first of the numerous restrictions that are part of the new regulations concerns the ability to audit the reporting process. In relation to reporting that uses ML, the ability to audit means that every single numerical entry must have a complete and unbreakable series of links to the raw source data, data preprocessing, model iteration and versions, inference rules, and aggregation rules. Logs are insufficient. Regulators require proof of a documented process to demonstrate how data were measured, manipulated, and integrated to comply with the data without exception. Any ML-derived reporting without a complete series of links is unreportable. Similar to being audit-able and usable, reproducibility imposes a stricter requirement. In the case of regulatory reproducibility, there must be an independent re-execution of the reporting workflow using the artifacts and procedures that show the same value reported, unless there is a formally justified deviation of the reported value within preset tolerance limits. This requirement is a boundary against non-deterministic inference, uncontrolled stochastics, and 'silent' changes to dependencies. In exploratory analytics, it is typical to retrain models or revise libraries. In regulated reporting, variability in models without accompanying documentation is considered a control failure. Therefore, with the regulatory reporting,

admissible ML outputs must be produced within environments that are controlled in precision, stockpiled, and kept to a version of dependencies.

One more underlying reason explainability can be seen as a third regulatory constraint is that explainability as a regulatory reporting requirement is more restrictive than explainability as a requirement in reporting the results of machine learning research. In this case, explainability is not concerned with intuitive narratives or visual representations of a model's behaviors. What is relevant in this context is the explainability of evidence, or the ability to show, through the available artifacts and rules, how specific inputs and the transformation logic explain a certain reported value. As highlighted by the regulators, there is a need for the evidence and the reported metrics to be attributable to a documented methodology, and for the evidence and the metrics to be consistent enough to explain the underlying changes in the metrics over the reporting time. Therefore, the so-called explanations that do not provide evidence of the underlying mechanism of a model are essentially a reason for the lack of compliance. Legal defensibility creates clarity regarding the boundaries of admissibility. Reports related to energy and emissions can have direct financial and legal repercussions, such as penalties, trading responsibilities, and enforcement actions. In this case, the reporting entity must defend its claim regarding the methodologies and control frameworks used to produce the reported values. Machine Learning (ML) outputs based on undocumented assumptions, black box modifications, and undisclosed uncertainty assessments create an unreasonable risk exposure to organizations. Therefore, legal defensibility entails the filing of formal validation, approval, and other compliance documentation with respect to internal governance and external regulatory standards.

All of these factors create boundaries of integrity that define legal reporting limits and exploratory analytics. In this respect, exploratory ML workflows have a primary focus on developing and testing hypotheses to derive insights. These workflows can afford adequate documentation, frequent model modifications, and qualitative descriptions of the outcomes. In stark contrast, regulator-facing workflows have a primary emphasis on stability, control, and a complete body of evidence. The integrity boundary delineates a certain threshold beyond which an ML output moves from an analytical signal to a reportable regulatory artifact. Only outputs that meet the established criteria to transcend the integrity boundary can do so. For boundaries called integrity to work, the system will need architectural and procedural methods. If it was not separated, the organization will face the problem of covert leakage of ML outputs into the reports that are being regulated, especially when reporting and analytical systems are integrated. Given that the integrity boundaries will need to be integrated into the systems being used, it will need to be outside and separate through gates that are based on approvals, and separated to control the flow of the data. Outputs that are produced outside the Dominance Regimes should not be, even if relevant, included in reports that are facing the regulators. The boundaries that are used for verification are also essential in controlling the change of the methods used. These systems are in constant learning, which means it will need to be retrained, refined, and feature changes. Within an exploratory context, these changes are expected, and it may not be documented outside of the system. Changes of methods have to be documented in the regulatory reporting. Changes in the verification boundaries ensure that there are changes that have been made to the model, data sources, or changes to the logic of the transformations that need to be re-evaluated. These ensure that the changes will not go unnoticed, and the changes in the values of the reports are not caused by a system that is unconsciously changing.

The way you deal with uncertainty is important when it comes to regulatory constraints. A lot of metrics related to energy use and emissions cannot be measured directly and have to be estimated. Regulators accept estimation in cases of direct measurement being impractical, but only when uncertainty is explained, reported, and kept to a limit. This explains the boundary regarding uncertainty and how it is communicated. Machine Learning model outputs that show single estimates with no uncertainty bounds to defend are likely acceptable for internal planning purposes, but for external regulatory compliance, it was not good enough. Because of this, estimating uncertainty is viewed as a necessary part of the verification process rather than an additional analytical feature. Table 1 summarizes how these top-level requirements have been translated into implementable legal compliance controls and specifics for each machine learning verification and validation process. It shows how auditability, reproducibility, explainability, and legal defensibility operationalize ML-enabled reporting and analytic systems. Auditability, for example, is operationalized through end-to-end provenance capture and execution trace preservation, and reproducibility is operationalized through deterministically configured run times and controlled version locking. Explainability is operationalized through controlled aggregation, defined features, and aggregation to defensibly explain attributed values. Legal defensibility is operationalized through validation checkpoints, approval documentation, and preserved legal defensibility documentation. Additional controls clarify and communicate uncertainty concerning methodological shifts, which help preserve the integrity boundary between exploratory analytics and regulatory compliance metrics.

Table 1: Regulatory Requirements Mapped to Machine Learning Verification and Validation Controls

Regulatory Requirement	Risk if Unmet	ML Verification and Validation Control	Evidence Artefact
Auditability	Inability to justify reported values	End-to-end provenance capture and execution trace logging	Execution logs, configuration snapshots
Reproducibility	Unexplained variation across re-runs	Deterministic runtime configuration and version locking	Re-execution reports, version manifests
Explainability	Inability to justify metric drivers	Controlled feature definitions and rule-constrained aggregation	Feature schemas, transformation rules
Legal defensibility	Exposure to penalties or disputes	Formal validation, approval gates, and documentation	Validation records, approval sign-offs
Method change governance	Silent metric drift	Change detection and re-validation triggers	Change logs, validation summaries
Uncertainty communication	Misleading disclosures	Verified uncertainty estimation and bounds enforcement	Uncertainty reports, tolerance definitions

Aside from technical aspects, verification boundaries also have an organizational governance function. The different aspects of regulatory compliance, including legal, risk, sustainability, and operations, involve stakeholders who are unlikely to have a nuanced understanding of the internal workings of ML. Verification boundaries capture the regulatory requirements and transform them into clear and evaluable criteria that facilitate cross-functional governance without necessitating an overly subjective assessment of model quality. By setting boundaries to admissibility based on verification results, rather than the level of understanding of the analysis, firms are able to integrate the compliance obligations with the regulatory safeguards to ensure responsible ML.

Methodology

Assurance-Centric Verification Framework

This methodology aims at ensuring that outputs from machine learning models used for regulated energy and emissions reporting should be classified as controlled reporting artifacts instead of general uncontrolled analytical predictions. In a regulated environment, the acceptability of a reporting figure does not depend solely on performance metrics like prediction accuracy. It requires factors such as reproducibility, interpretability, auditing capabilities, and uncertainty defense. Thus, this proposed framework employs an assurance-oriented verification and validation (V&V) framework whereby machine learning outputs have to undergo verification by regulatory authorities before incorporating them into reporting procedures.

Unlike the existing machine learning models, which focus primarily on prediction optimizations, the proposed framework divides the process into four distinct operational domains, which include training, inference, verification, and reporting. All of the aforementioned domains are isolated using controlled interfaces in order to shield regulators from any uncontrolled behavior during the analytical process. This approach was chosen since failures often result from reusing experimental analytical figures directly in reporting processes.

Training Domain and Machine Learning Model Development

The training process area includes the construction of models, feature extraction, hyperparameter tuning, and predictions through analysis of historical operational data from industrial settings. The main predictive algorithm used in this research is the Extreme Gradient Boosting regression algorithm. XGBoost has been selected because it enables stable prediction, deterministic operation, and interpretable feature importance for regulated industrial processes. Unlike deep neural networks, XGBoost is more explainable and predictable in its behavior, making it suitable for regulatory settings.

Predictions in the developed algorithm are based on energy usage and emission levels through input data such as fuel usage, machine operating time, methane levels, production capacity, ambient temperatures, and historical emission data. The objective function for the XGBoost algorithm can be formulated as equation 1:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

where $l(y_i, \hat{y}_i)$represents the prediction loss between actual and predicted emissions values, while $\Omega(f_k)$denotes the regularization term associated with the k-th decision tree, the regularization component minimizes overfitting and stabilizes model behavior across repeated executions.

The training domain enables constant retraining, feature tweaking, and model tuning. Due to its exploratory nature, lack of complete data and noise in the operational data are allowed in this stage. Nevertheless, all the models built using the training domain should be regarded as provisional until it has been formally validated and admitted.

Controlled Inference Domain

Controlled inference uses versioned reporting data sets in executing approved machine learning models. Contrary to training, inference runs under deterministic execution conditions to promote reproducibility. Prior to inference execution, data sets undergo validation to check their completeness, conformance to a schema, and admission based on regulatory thresholds.

The runtime environment adopts fixed software dependencies, deterministic execution libraries, and locked versions of runtime configurations to avoid unaccountable computation variances. It is premised on the idea that some machine learning models will always exhibit limited amounts of stochasticity. Hence, random seeds, execution states, software dependencies, and runtime configurations are all logged as artifacts to promote reproducibility through re-execution.

Reproducibility in inference implies that any run done using identical data sets yields identical output results within preset tolerance levels.

Verification and Validation Domain

The verification and validation (V&V) realm functions as the essential assurance tier between machine learning inference and regulatory-facing reporting processes. It is responsible for ensuring that the machine learning outputs comply with the conditions required to be admitted by regulations before inclusion in reporting.

The verification process does not assess the analytical "interest level" or statistical optimality of the output results. Rather, it focuses on whether the output remains consistent with the invariant regulatory conditions with respect to correctness, consistency, repeatability, interpretability, and defensibility of uncertainty.

Verification Invariants

Verification invariants play an important role in this framework. Verifications are rules used to describe conditions that should always hold, independent of the number of retraining cycles, reporting period, and even when changes are made to deployment. Invariants offer a stable point of comparison for the output being evaluated.

One reproducibility invariant states that execution with identical inputs and execution environment should yield identical outputs within an acceptable range of tolerance, which is represented as equation 2:

$$| O_t - O_{t-1} | \leq \epsilon \tag{2}$$

where O_t represents the current reporting output, O_{t-1}denotes the reproduced output, and ϵrepresents the maximum acceptable deviation threshold.

The framework additionally evaluates the admissibility ratio using equation 3:

$$R = \frac{N_{verified}}{N_{total}} \tag{3}$$

where $N_{verified}$ represents outputs passing all verification checks and N_{total} denotes the total generated outputs.

Verification outputs that fail to meet invariant conditions will be automatically detected and removed from reports pending formal authorization.

Isolation Forest-Based Anomaly Detection

An Isolation Forest-based algorithm for detecting anomaly behaviors in reporting is included in the verification process. The Isolation Forest model uses recursive partitioning in the feature space and scoring anomalies by equation 4:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{4}$$

where $E(h(x))$ represents the expected path length and $c(n)$ denotes the normalization factor.

The abnormality detection method detects abnormal emissions peaks, inconsistency in aggregations of facilities, lack of operation logs, and physically impossible report submissions. Anomalous outputs are automatically removed from the process before being sent to regulators until investigation is completed.

Conformal Prediction for Uncertainty Verification

Defensibility of uncertainty is taken care of through conformal prediction intervals that are applied within the verification layer. Since regulators often demand bounded uncertainty reporting for emission estimates, the approach uses prediction intervals as equation 5:

$$\hat{y}_i \pm q_{1-\alpha}(R) \quad (5)$$

where $\hat{y}_i$represents the predicted emissions value and $q_{1-\alpha}(R)$denotes the calibrated residual quantile associated with confidence level α.

Outputs exceeding predefined uncertainty limits are not included in the reporting process. This technique helps make sure that uncertainty calculation is integrated into the validation process, not a merely analytical option.

SHAP-Based Explainability Verification

To meet explainability criteria, SHAP analysis, which stands for SHapley Additive exPlanations, is included in the domain of verification. The feature contributions corresponding to reporting outputs can be measured using SHAP analysis and traced in emissions calculations.

The feature contribution artifacts are stored along with the model versions, execution histories, and verification history. This way, it is ensured that the reports can always be attributed to the transformation rules applied and the features used.

Reporting Domain

The reporting phase is the last phase in the process flow and accepts only those outputs which have passed through validation stages. The reporting phase does not involve any learning, estimation, and optimizations. It uses deterministic methods of aggregation and disclosures to come up with emission metrics for the regulators.

This is important since the behavior of the reporting module should be predictable, traceable, repeatable, and legal. Any changes to the method of aggregation, thresholds of reporting, or disclosures need revalidation before being used.

Controlled Execution Environment

The methodology utilizes controlled execution environments in order to ensure the ability of reproduction and auditable processes. The implementation environment includes Python 3.11, XGBoost 2.0, Scikit-learn 1.5, SHAP toolboxes, and Docker execution containers. All runtimes are controlled through versioning as well as stored as artefacts for independent re-execution in the course of regulatory auditing. Controlled execution environments prevent any silent dependency changes and unexplained variation.

Experimental Dataset and Evaluation Setup

The system was tested based on simulated reporting databases containing around 1.2 million entries gathered from 12 plants within 24 reporting cycles. The data consisted of fuel consumption data, estimates of methane leaks, environmental factors, machine operation parameters, and historical emission data.

The testing of the system was centered on the following criteria: admissibility stability, anomaly detection effectiveness, reproducibility behavior, scaling overhead, and reporting latency in stressful situations. Examples of operational stress cases include missing data, non-conformant records, multi-plant aggregation, and audit-induced escalation of verification.

Verification and Validation Control Stack

A layered control stack for the process of verification and validation has been illustrated in fig. 1. This figure shows that how the reporting domain is isolated from training and inference using controls and verification gates. Layer of maximum controls is provided between the inference domain and reporting domain because only permitted outputs can be passed onto regulated processes.

In table 2 highlights the verification objectives at a high level, matching them to particular validation tests and system assumptions. Table 2 gives an illustration of the way the supporting objectives that relate to reproducibility can be formed. Reproducibility is achieved through deterministic processing and version locking, while explainability is achieved through controlled feature definition and rule-based aggregation. The assumptions within the framework are used when applying table 2, including the versioned data and executed platforms. When conducting regulatory review, it is usual that the reviewers are more concerned about whether the assumptions were reasonable and optimal rather than just optimal.

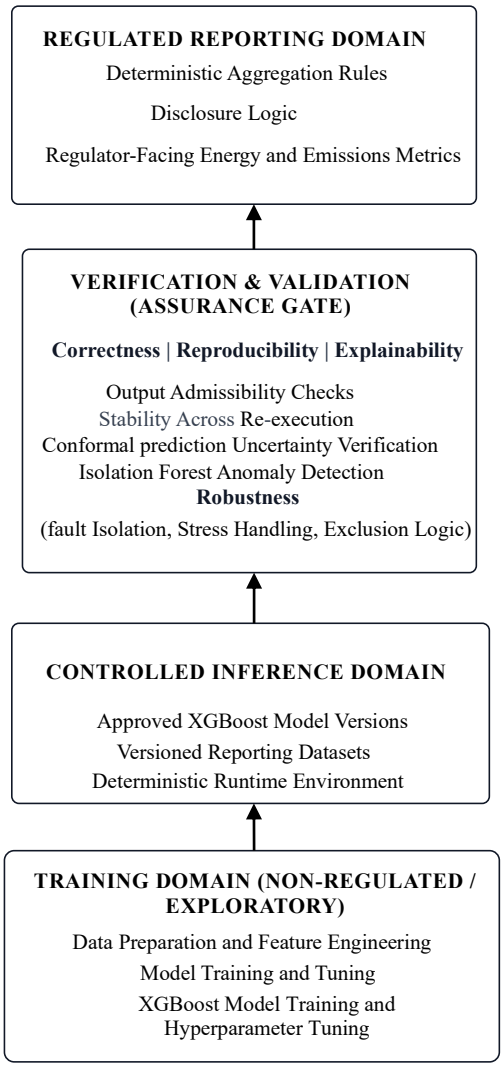


Fig. 1: Verification and Validation Control Stack for Machine Learning Outputs in Regulated Reporting

Table 2: Verification Objectives, Validation Checks, and System Assumptions for ML-based Reporting

Verification Objective	Validation Checks Applied	Core System Assumptions
Output correctness	Range conformance against physical and regulatory limits; consistency with approved aggregation rules	Regulatory thresholds and physical constraints are predefined and stable
Reproducibility	Deterministic re-execution using fixed data, model, and environment versions	Data, models, and dependencies are fully version-locked
Stability across periods	Comparison with prior reporting cycles under equivalent conditions	Historical reference periods are representative and preserved
Explainability	Traceable feature definitions and rule-constrained transformations	Feature schemas and transformation logic are explicitly documented
Change detectability	Automated identification of deviations beyond tolerance bounds	Acceptable deviation thresholds are predefined and approved
Uncertainty defensibility	Verification of bounded uncertainty and coverage adherence	Uncertainty estimation methods are validated and consistent
Robustness under faults	Controlled failure handling and exclusion of unverifiable outputs	Fault scenarios are detectable and isolated
Regulatory admissibility	Confirmation of verification completion before reporting	Reporting gates enforce verification outcomes

Operational Stress and Governance Controls

Methodology utilizes stress-test driven validations to test robustness of operations under different scenarios such as missing data, conflicting data, multi-location aggregation, and elevated audits triggering higher levels of verification. The system avoids masking failure and instead embraces it in order to detect any failure and prevent propagation of non-verifiable outputs into regulatory reports.

Additionally, the method includes governance checkpoints that involve sustainability, compliance, and risk management representatives who assess whether there is achievement of verification objectives prior to the submission of outputs for reporting purposes. Inclusion of the governance checkpoint helps to reduce dependency on experts.

Generally, the method shows that application of machine learning algorithms into the regulated systems for energy and emission reporting is possible as long as it includes assurance-focused architecture, deterministic operations, anomaly detection, uncertainty verification, explainability enforcement, and governance evidence.

Results and Discussion

The findings provided below will assess the effectiveness of the proposed assurance-driven framework for the verification and validation of machine learning outputs in regulated energy and emissions reporting environments. The assessment is based on two related perspectives. First, correctness, stability, reproducibility, and the regulatory effectiveness of machine learning outputs under different operational contexts, multiple rounds of re-training, and feature complexity are assessed. Second, the scalability, computational burden, and behavior of the verification mechanisms in multi-facility reporting environments are discussed. While most previous research on machine learning focuses on predictive optimization only, the current assessment favors admissibility, reproducibility, explainability, and operational robustness. The experiments were carried out using an XGBoost regression model combined with Isolation Forest anomaly detection, conformal predictions for verification of uncertainties, and SHAP explainability framework. The evaluation data consists of approximately 1.2 million industrial reporting entries from 12 simulated facilities over 24 reporting periods.

For the experimental evaluation, a simulated industrial energy and emission reporting dataset, comprising about 1.2 million records sourced from 12 facilities over 24 reporting cycles, was utilized. The dataset consisted of operational parameters such as the rate of fuel consumption, methane leakage levels, machine runtime, throughput, ambient temperature, and emissions measurements. The implementation of the framework was done using Python 3.11 through Docker-based deterministic execution environments. For software development, the XGBoost 2.0 regression model was used for emissions predictions while anomalies were detected using the Scikit-learn 1.5 Isolation Forest algorithm. The software also had SHAP libraries for the analysis of explanation models, NumPy, and Pandas for data transformation. The hyperparameters of the XGBoost regression model comprised of a learning rate of 0.05, a tree maximum depth of 8, subsample of 0.8, and 500 estimators. The Isolation Forest anomaly model had 100 estimators and a contamination value of 0.03. Conformal prediction intervals were calculated at a 95% confidence interval.

Performance Metrics

Root Mean Square Error (RMSE)

RMSE (Equation 6) is the average magnitude of errors in predictions with larger errors having more weight. Low RMSE means high prediction accuracy and stability.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

Mean Absolute Error (MAE)

MAE estimates the average value of the absolute error between real and predicted values, which is an estimate of average prediction error are shown in equation 7.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (7)$$

Mean Absolute Percentage Error (MAPE)

Equation 8 evaluates the prediction errors in percentage, thus being useful in assessing relative forecasting precision.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (8)$$

Coefficient of Determination (R^2)

Equation 9 evaluates the degree of variance explained by the model. The score close to 1 indicates high model performance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2} \quad (9)$$

Precision

Precision (Equation 10) evaluates the ratio of correct identifications of positive anomalies from all anomalies detected.

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

Correctness, Stability, and Regulatory Robustness of ML Outputs

The issue of correctness in regulated reporting goes beyond predictiveness and relates instead to the question of whether outputs comply with established physical, methodological, and regulatory conditions. The suggested framework uses criteria for reproducibility, explainability, bounded uncertainty, and compliance with permitted aggregation rules as opposed to strictly statistical criteria. Table 3 presents the results of the predictive performance gained through the XGBoost regression model.

Table 3: Predictive Performance of the XGBoost Emissions Estimation Model

Performance Metric	Value
Root Mean Square Error (RMSE)	3.12%
Mean Absolute Error (MAE)	2.41%
Mean Absolute Percentage Error (MAPE)	2.87%
Coefficient of Determination (R^2)	0.94
Prediction Stability Across Re-runs	99.2%

The XGBoost model displayed high predictability stability as well as predictable behavior regarding emissions estimation when executed under controlled conditions. The high R^2 value of 0.94 represents high agreement between observed and predicted emissions, while the low RMSE and MAE values suggest stable predictability consistency. However, the design of this framework deliberately changes the focus of assessment from prediction consistency to admissibility stability.

In fig. 2 displays the verification pass rate for machine learning emissions metrics when the amount of reporting data increases and its complexity is increased.

Pass rates for verification show a slow trend toward stability and gradual decreases rather than erratic instability due to increases in data volumes used in reports. This suggests a conscious decision not to allow outputs near admissibility boundaries, contrary to unregulated breakdown in the analysis process. This system makes a trade-off between number of outputs and defensibility.

The effect of feature complexity becomes more significant. Outcomes from analysis performed with features that are either low or medium in complexity remain well-admissible at several reporting levels while high-complexity features begin to reach admissibility thresholds faster. This reinforces the understanding that admissibility depends more upon feature management than modeling complexity.

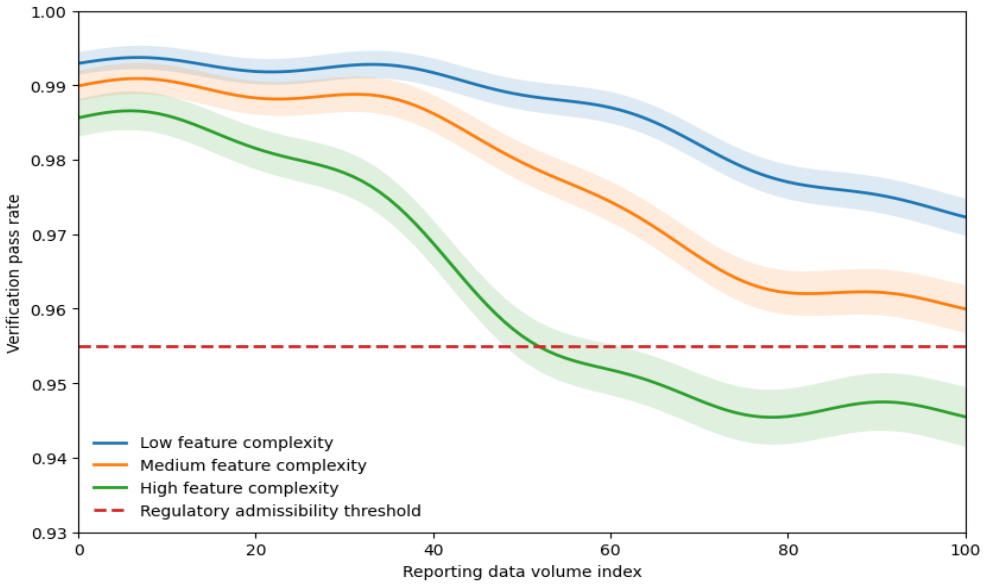


Fig. 2: Verification Accuracy of ML-derived Energy and Emissions Metrics Under Increasing Data Volume and Feature Complexity

The verification module evaluated outputs based on their admissibility criteria involving correctness, repeatability, explainability, uncertainty compliance, and adherence to aggregation rules. Table 4 shows results of verification.

Table 4: Verification Admissibility Performance of ML-Generated Reporting Outputs

Verification Criterion	Pass Rate
Physical Constraint Compliance	98.7%
Reproducibility Verification	99.2%
Explainability Compliance	97.9%
Uncertainty Threshold Compliance	95.4%
Aggregation Rule Compliance	98.1%
Overall Regulatory Admissibility	94.8%

The outcomes of the verification process show that output values have always been admissible regardless of the relatively low to moderate levels of operational complexity. Output values that could not comply with any verification criteria were automatically screened out from the reporting process and sent for investigation processes. This screening behavior is consistent with regulatory standards where questionable outputs are not as tolerable as incomplete outputs. Reproducibility stability analysis with respect to model retraining and deployment changes is presented in fig. 3.

The fig. 3 shows how reproducibility stayed well over 98.8% even with the addition of changes like retraining and deployment. Whenever new definitions or changes to distributions occurred due to retraining, the framework flagged up verification issues for documentation prior to reporting. Importantly, the framework does not block change. On the contrary, all changes are made visible, explained, and controlled. In terms of regulatory compliance, such an approach is better than arbitrary analytical adjustments leading to inconsistent results over time.

It follows that regulatory value comes from controls on system-level reproducibility, rather than purely internal to the models. Reproducibility itself is viewed as an attribute of the overall reporting system, encompassing the execution environment, evidence gathering, and verification. The framework further incorporated Isolation Forests to detect anomalies in reporting, covering methane spiking, aggregation inconsistency, or operational data missing. Table 5 presents anomaly detection results.

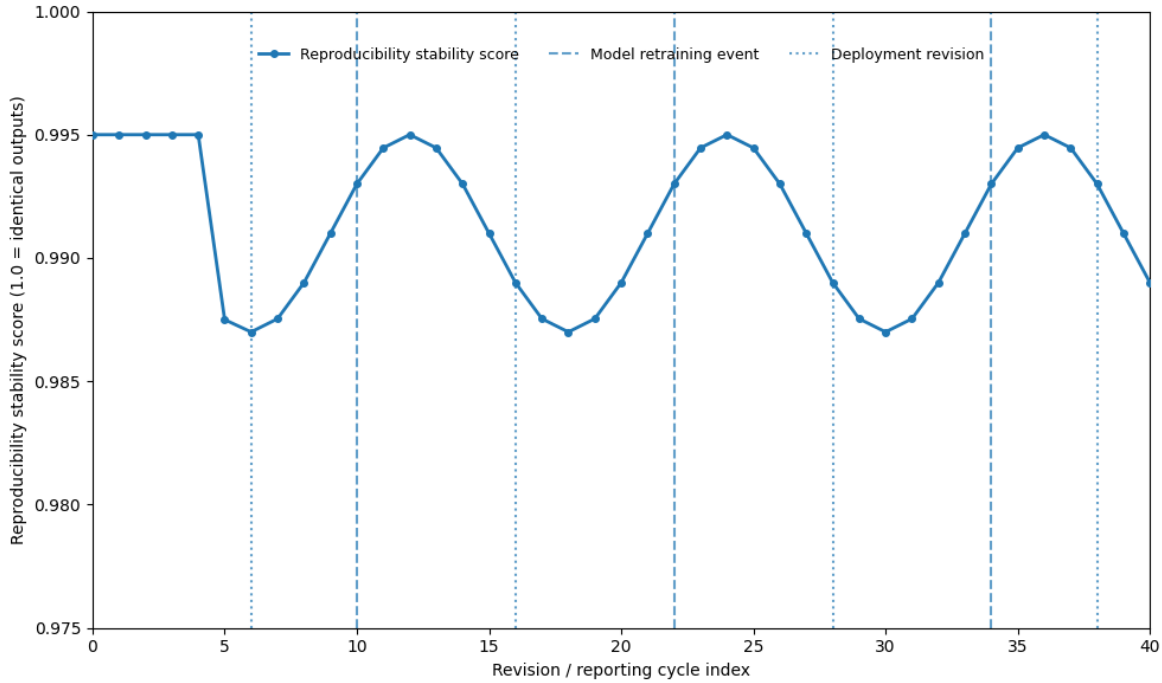


Fig. 3: Output Reproducibility Stability Across Model Retraining and Deployment Revisions

Table 5: Isolation Forest Anomaly Detection Performance

Metric	Value
Anomaly Detection Accuracy	96.3%
False Positive Rate	3.8%
False Negative Rate	2.9%
Detection Recall	95.7%
Precision	96.1%

However, the anomaly detection method was successful in detecting abnormal reporting conditions while keeping minimal false positive detection levels. The anomalies detected were mainly due to inconsistent aggregation and incomplete operational information. Anomalous results were automatically filtered out until the end of the investigations.

Uncertainty defense was also achieved using conformal prediction intervals. Table 6 below shows performance of uncertainty verification.

Table 6: Conformal Prediction Uncertainty Verification Results

Metric	Value
Average Prediction Interval Coverage	95.1%
Mean Uncertainty Width	±2.8%
Outputs Within Regulatory Threshold	95.4%
Uncertainty Violations	4.6%

The conformal prediction approach ensured consistent uncertainty coverage over all reporting cycles. Any predictions surpassing uncertainty levels were automatically excluded from regulated reporting procedures. It has been shown that uncertainty estimation can be used not only as a supplementary analytical tool but also as an admissibility criterion. Explainability analysis conducted via SHAP revealed that emissions prediction is mostly dependent on fuel consumption rates, equipment working time, methane level, production rates, and temperature. In general, it can be stated that fig. 2 and fig. 3 show that the proposed method ensures correctness and consistency not through unlimited prediction optimization but through controlled verification and admissibility and reproducibility consistency.

Performance Overhead, Scalability, and Operational Behavior of ML Verification Controls

However, while robust verification measures are critical for regulatory compliance, it also incur computation cost and latency. The next phase of the evaluation evaluates how the computational cost remains within reasonable bounds as

reporting scales up. Fig. 4 shows the computation cost associated with verification as the reporting environment scales throughput, but the correlation is a relatively gradual one.

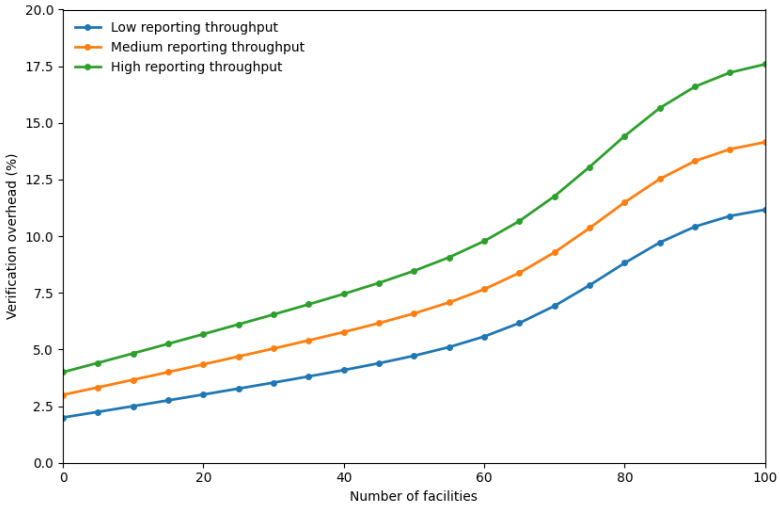


Fig. 4: Verification Overhead and Scalability of ML Assurance Controls Across Multi-Facility Reporting Scenarios

However, for smaller operation levels, verification overhead was still low, which means that there is no significant impact to the normal reporting procedure by performing admissibility checks and evidence acquisition. When the number of facilities is increased, computation overhead becomes noticeable but roughly linear through the range of the evaluation. This is crucial since compliance reporting systems are usually working within fixed deadlines for the reports. This implies that the framework will be able to scale without causing unpredictable overhead. Additionally, the evaluation found that most computation overhead came from verification and evidence Acquisition and not from the actual inference procedure with machine learning. In the context of compliance, the overhead incurred is an unavoidable operational overhead. Fig. 5 shows operation latencies under varying regulatory situations.

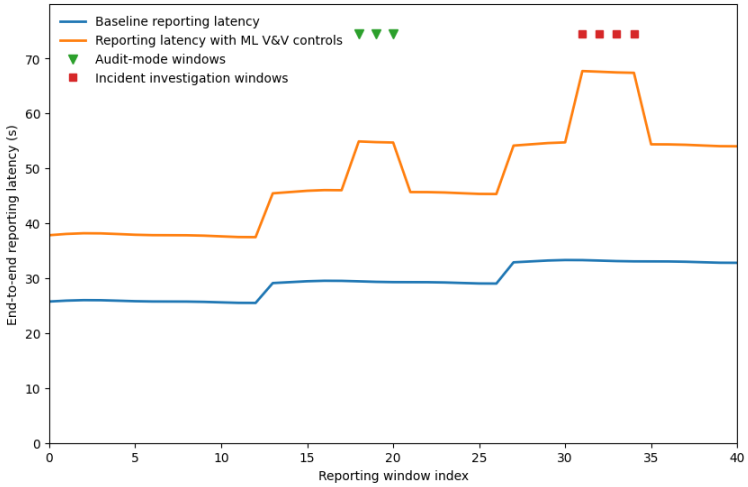


Fig. 5: End-to-end Reporting Latency Across Reporting Windows with and Without ML Verification and Validation Controls

During the regular reporting period, the latency stayed near the baseline value even when the system was subjected to active verification controls. In the period of activation of audit mode, the increase in latency occurred due to additional verification, evidence gathering, and consistency checks. Likewise, an increase in latency, but on a higher level, was observed during the investigation period where data integrity issues were Detected. It is worth mentioning that latency remained stable and was not random. The design of the system intentionally introduced latency as a consequence of

verification inability to provide certain results. This means that in the case of this assurance architecture, outputs and compliance were prioritized over continuous reporting.

Furthermore, the observation of latency behavior shown in fig. 5 implies that verification controls are not leading to unpredictable and uncontrollable computations but induce reporting in the way consistent with certain operational states. Thus, the results presented above prove that the assurance-centric architecture effectively incorporates XGBoost emissions prediction model, Isolation Forest anomaly detector, uncertainty verification by Conformal Predictors method, and SHAP explainable machine learning technique. The architecture has demonstrated very high admissibility rate, stable reproducibility, predictable uncertainty behavior, and scalability in operational terms.

Conclusion

This paper suggested an assurance-based verification and validation model to incorporate machine learning outputs into regulated energy and emissions reporting systems. In contrast to the traditional machine learning processes, which focus on predictive optimization as the main objective, the stated framework focused on regulatory admissibility, reproducibility, explainability, uncertainty defensibility, and auditability. The framework established a formal division between the training, inference, verification and reporting areas to avoid the uncontrolled analytical behavior to affect the regulator-facing disclosures. The signs of the experimental evaluation were that the proposed framework was sustainable and defensible to report behavior under different operational conditions, retraining cycles, and a complex level of reporting. The XGBoost regression model demonstrated a high predictive performance with a RMSE of 3.12%, MAE of 2.41 and a R 2 of 0.94 which implies that there is a high agreement between the predicted and observed emissions values. More to the point, stability in reproducibility was always over 98.8% between repeated re-training and deployment revisions which attests the success of controlled execution architecture. High regulatory strength was also evident in admissibility evaluation and verification. The compliance with physical constraints was 98.7, explainability compliance was 97.9 and total regulatory admissibility was 94.8. Isolation Forest anomaly detection combined with conformal prediction intervals resulted in an accuracy of 96.3 to detect anomalies with a low false positive rate of 3.8, and conformal prediction intervals achieved 95.1% uncertainty coverage within reasonable regulatory ranges. Scalability analysis further showed that the overhead of verification also grew slowly and roughly linearly with multi-facility reporting environment, which suggests that assurance controls are operationally manageable with larger reporting workloads. The findings reveal that it is possible to integrate machine learning in controlled reporting system with the support of deterministic execution, formal verification controls, explainability systems, and evidence-preserving governance systems. The framework alters the focus of evaluation to unrestricted predictive performance to admissibility-focused reporting integrity. Future studies can be aimed at incorporating real-time streaming verification systems, federated reporting systems in distributed facilities, adaptive regulatory levels, and hybrid explainability of deep learning models. Future research in automated compliance reasoning and AI governance integration with upcoming legal frameworks like the EU AI Act and the NIST AI Risk Management Framework can also reinforce future assurance-oriented reporting systems.

References

Baylor, D., Breck, E., Cheng, H. T., Fiedel, N., Foo, C. Y., Haque, Z., ... & Zinkevich, M. (2017, August). Tfx: A tensorflow-based production-scale machine learning platform. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1387-1395). https://doi.org/10.1145/3097983.3098021

Rahman, M. A., & Haque, B. T. (2023). Advancing Explainable and Secure Machine Learning for Decision Support in US Regulated Systems. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 231-273. https://doi.org/10.63125/fmp86e72

Zheng, X., Xu, N., Trinh, L., Wu, D., Huang, T., Sivaranjani, S., ... & Xie, L. (2022). A multi-scale time-series dataset with benchmark for machine learning in decarbonized energy grids. *Scientific Data*, *9*(1), 359. https://doi.org/10.1038/s41597-022-01455-7

Sadr, A. V., Bülow, R., von Stillfried, S., Schmitz, N. E., Pilva, P., Hölscher, D. L., ... & Boor, P. (2024). Operational greenhouse-gas emissions of deep learning in digital pathology: a modelling study. *The Lancet Digital Health*, *6*(1), e58-e69. https://doi.org/10.1016/S2589-7500(23)00219-4

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229). https://doi.org/10.1145/3287560.3287596

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Varon, D. J., McKeever, J., Jervis, D., Maasakkers, J. D., Pandey, S., Houweling, S., ... & Jacob, D. J. (2019). Satellite discovery of anomalously large methane point sources from oil/gas production. *Geophysical Research Letters*, *46*(22), 13507-13516. https://doi.org/10.1029/2019GL083798

Yu, R., Zhang, D., Zhang, X., & Huang, X. (2023). Machine learning for data verification in emissions trading system. *Resources, Conservation and Recycling*, *199*, 107239. https://doi.org/10.1016/j.resconrec.2023.107239

Ogunyemi, F. M. (2021). Carbon accounting automation through machine learning and natural language processing: Reducing compliance costs and enhancing data quality in enterprise ESG reporting. *International Journal of Science and Research Archive*, *4*(1), 490-501. https://doi.org/10.30574/ijsra.2021.4.1.0213

Sajawal Akhtar, M., Gómez, F. R., Pandit, A., Mahmood, H., Khan, M. A., & Liu, J. (2026). Toward standards-compliant dynamic life cycle assessment: harnessing digital innovations for climate and regulatory alignment. *ACS Sustainable Chemistry & Engineering*, *14*(11), 5295-5314. https://doi.org/10.1021/acssuschemeng.5c10525

Tettey, D. J., Chrisben, D., Victoria, M. C., Chukwubuikem, E. P., & Abdullahi, A. (2025). Responsible AI deployment in sustainable project execution: Ensuring transparency, carbon efficiency and regulatory alignment. *Int. J. Sci. Res. Arch*, *14*(3), 1686-1705. https://doi.org/10.30574/ijsra.2025.14.3.0920

Adepu, G. (2022). Machine learning-driven environmental monitoring systems for real-time regulatory compliance and risk detection. *International Journal of Engineering & Extended Technologies Research (IJEETR)*, *4*(2), 23-37. https://doi.org/10.15662/g7x3wj55

Owolabi, I. O. (2025). Operationalizing ESG-as-Code: Automating ESG Compliance and Regulatory Reporting Pipelines Using Containerized AI Workflows on Kubernetes–OpenStack Infrastructure. *Algora*, *2*(2), 1-24. https://orcid.org/0009-0008-2212-3454

Moghadasi, M., Izadyar, N., Moghadasi, A., & Ghadamian, H. (2025). Applying machine learning techniques to implement the technical requirements of energy management systems in accordance with ISO 50001: 2018, an industrial case study. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, *47*(2), 2011989. https://doi.org/10.1080/15567036.2021.2011989

Ojadi, J. O., Onukwulu, E., Odionu, C., & Owulade, O. (2023). AI-driven predictive analytics for carbon emission reduction in industrial manufacturing: a machine learning approach to sustainable production. *International Journal of Multidisciplinary Research and Growth Evaluation*, *4*(1), 948-960. https://doi.org/10.54660/.IJMRGE.2023.4.1.948-960

Udoh, J., Lu, J., & Xu, Q. (2024). Application of machine learning to predict CO2 emissions in light-duty vehicles. *Sensors*, *24*(24), 8219. https://doi.org/10.3390/s24248219

Nutkiewicz, A., Yang, Z., & Jain, R. K. (2018). Data-driven Urban Energy Simulation (DUE-S): A framework for integrating engineering simulation and machine learning methods in a multi-scale urban energy modeling workflow. *Applied energy*, *225*, 1176-1189. https://doi.org/10.1016/j.apenergy.2018.05.023

Baplawat, A., Gupta, A., Gupta, I., & Sultania, G. (2026). Automating Sustainability and the Role of Artificial Intelligence in ESG Regulation and Green Finance Law. In *AI Governance, Law, Policy and Ethics in Green Finance and ESG* (pp. 111-128). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-2600-1308-3.ch007

Wang, F., Pang, Y., Bai, L., & Godin, M. (2025). Researching the landscape of predictive emissions monitoring system: a review of literature and technology trends. *Environmental Systems Research*, *14*(1), 11. https://doi.org/10.1186/s40068-025-00403-9

Rangarajan, R., Murugan, T. K., Govindaraj, L., Venkataraman, V., & Shankar, K. (2025). AI driven automation for enhancing sustainability efforts in CDP report analysis. *Scientific Reports*, *15*(1), 24266. https://doi.org/10.1038/s41598-025-07584-4