

Designing an Integrated Framework Based on Explainable Artificial Intelligence and Data Envelopment Analysis for Evaluating the Performance of Operational Units of the South Zagros Oil and Gas Production Company

Seyed Morteza. Mousavi¹, Javad. Gerami^{2*}, Mohammadreza. Mozaffari², Roya. M. Pour Ahari¹, Mohammadreza. Feylizadeh²

¹ Department of Industrial Engineering, Na.C., Islamic Azad University, Najafabad, Iran

² Department of Industrial Engineering, Shi.C., Islamic Azad University, Shiraz, Iran

* Corresponding author email address: javad_gerami@iau.ac.ir

Article Info

Article type:

Original Research

How to cite this article:

Mousavi, S. M., Gerami, J., Mozaffari, M., M. Pour Ahari, R., & Feylizadeh, M. (2025). Designing an Integrated Framework Based on Explainable Artificial Intelligence and Data Envelopment Analysis for Evaluating the Performance of Operational Units of the South Zagros Oil and Gas Production Company. *Journal of Resource Management and Decision Engineering*, 4(4), 1-21.

<https://doi.org/10.61838/kman.jrmde.441>



© 2025 the authors. Published by KMAN Publication Inc. (KMANPUB). This is an open access article under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

ABSTRACT

This study aimed to design and evaluate an integrated framework combining Data Envelopment Analysis (DEA), machine learning, and Explainable Artificial Intelligence (XAI) for assessing, predicting, and interpreting the operational efficiency of oil and gas production units. The study employed an applied quantitative design using 30 decision-making units (DMUs) evaluated for the 2026–2027 operational year. Four inputs comprising feedstock, energy consumption, water consumption, and labor; two desirable outputs comprising main production and revenue; and five undesirable outputs comprising CO₂, NO_x, SO₂, PM₁₀, and COD were analyzed. DEA was first used to calculate relative efficiency scores and identify efficient and inefficient units. These efficiency scores were then used as the target variable in supervised machine-learning models, including linear regression, support vector regression, random forest, gradient boosting, and XGBoost. Model performance was evaluated using coefficient of determination, mean absolute error, root mean square error, and cross-validation. SHapley Additive exPlanations (SHAP) were subsequently applied to interpret the selected predictive model and determine the contribution of each operational variable to efficiency. The mean DEA efficiency score was 0.879, indicating an average efficiency gap of 12.1% relative to the estimated frontier. Five DMUs achieved an efficiency score of 1.000 and were classified as efficient. XGBoost demonstrated the best predictive performance, with ($R^2=0.944$), MAE=0.017, and RMSE=0.026; its cross-validated (R^2) was 0.889 with RMSE=0.034. SHAP analysis identified energy consumption as the most influential predictor of efficiency (18.6%), followed by main production (16.4%), CO₂ emissions (14.1%), revenue (12.3%), and labor (10.5%). Higher production and revenue generally increased predicted efficiency, whereas excessive energy use, pollutant emissions, and disproportionate labor utilization reduced it. The integrated DEA–machine learning–XAI framework provided a multidimensional, predictive, and interpretable approach to operational performance evaluation.

Keywords: Data Envelopment Analysis; Explainable Artificial Intelligence; Machine Learning; SHAP; Operational Efficiency; Oil and Gas Industry; XGBoost

1. Introduction

The contemporary oil and gas industry operates in an environment characterized by increasing technological complexity, volatility in energy markets, stringent environmental expectations, rising operating costs, and growing pressure to improve productivity while simultaneously reducing resource consumption and ecological impacts. Operational units in upstream and processing systems are required to transform large quantities of feedstock, energy, water, capital, and labor into economically valuable outputs under conditions in which equipment reliability, environmental compliance, process continuity, and production efficiency are closely interdependent. Consequently, conventional performance assessment based solely on production volume, financial indicators, or individual engineering metrics is no longer sufficient for evaluating the overall efficiency of operational units. A comprehensive assessment framework must be capable of simultaneously considering heterogeneous inputs, desirable outputs, undesirable environmental consequences, and nonlinear relationships among operational variables. At the same time, the increasing availability of industrial data has created opportunities for using artificial intelligence and machine learning to complement conventional benchmarking approaches and generate more adaptive, predictive, and interpretable performance-evaluation systems. Recent research on smart manufacturing has emphasized that artificial intelligence can support integrated optimization of production processes by learning complex patterns from operational data and assisting decision-makers in identifying inefficiencies that may remain hidden in conventional monitoring systems (Champa & Segall, 2026). Similar developments in industrial predictive maintenance have demonstrated that machine-learning methods can extract actionable patterns from equipment and process data and predict operational deterioration before it produces serious performance losses (Orazov et al., 2026).

The growing role of data-driven analytics has transformed the conceptual foundations of industrial performance measurement. Instead of evaluating a production system only after inefficiencies occur, modern analytical architectures increasingly seek to anticipate undesirable operational conditions, benchmark performance continuously, and explain the factors responsible for observed differences. Predictive models have already been used effectively to forecast machine failure and support

proactive intervention strategies, demonstrating the potential of machine learning for reducing downtime, maintenance costs, and operational disruptions (Ajayi et al., 2025). Likewise, intelligent optimization methods have been developed for complex scheduling and resource-allocation problems in which multiple conflicting objectives must be addressed simultaneously, showing that learning-based algorithms can improve the coordination of production resources in dynamic environments (Akram et al., 2026). These developments are highly relevant to oil and gas operations, where the interaction among production schedules, equipment constraints, energy consumption, environmental emissions, and maintenance requirements produces a multidimensional optimization problem. However, the use of machine learning in such environments must be accompanied by rigorous benchmarking and validation because predictive performance may vary considerably depending on data characteristics, model architecture, computational requirements, and the operational context in which the model is deployed.

Benchmarking has therefore emerged as a central concern in the development and evaluation of artificial intelligence systems. Recent studies have proposed systematic benchmarks to compare the predictive accuracy, computational efficiency, robustness, and resource requirements of competing models across different domains. For example, benchmarking studies in smart-grid forecasting have emphasized the importance of standardized datasets and comparable performance measures for evaluating artificial intelligence models under realistic operating conditions (Agrawal, 2026). Comparative assessments of supervised machine-learning models have similarly shown that predictive accuracy should not be considered independently from energy efficiency and computational resource consumption, particularly when models are intended for large-scale or continuous industrial deployment (Ali et al., 2026). Benchmarking has also become increasingly important in technically demanding computational environments. RayBench was developed as a performance-analysis framework for GPU rendering and demonstrated the value of structured benchmarking in identifying the hardware and workload configurations associated with optimal performance (Wang & Yu, 2023b), while RenderBench provided a complementary CPU-oriented benchmarking framework based on architecture-independent characteristics (Wang & Yu, 2023a). Although these studies were conducted outside the oil and gas sector, they illustrate a broader methodological principle:

meaningful performance assessment requires a systematic reference framework capable of comparing heterogeneous units or algorithms using multiple performance dimensions.

The need for robust and resource-aware benchmarking becomes even more critical in industrial settings where computational models operate under hardware, communication, or energy constraints. Resource-efficient big-data analytics depends not only on algorithmic accuracy but also on the ability to characterize infrastructure and assign workloads in a manner consistent with computational capacity. The Perona framework, for example, demonstrated that infrastructure fingerprinting can support more resource-efficient big-data processing and reduce unnecessary computational overhead (Scheinert et al., 2022). Similarly, automated deployment and performance benchmarking of machine-learning workloads in distributed computing environments has been proposed as a means of standardizing experimentation and improving reproducibility across heterogeneous computational infrastructures (Rahaman, 2025). Artificial intelligence assessment has also been incorporated into the benchmarking of Internet of Things platforms, reflecting the growing need to evaluate not only technological functionality but also intelligence, interoperability, and computational performance (Crăciun et al., 2024). These developments are particularly relevant for oil and gas companies, where operational information is frequently generated through geographically dispersed sensors, processing facilities, control systems, and industrial Internet of Things infrastructures.

A further consideration is that the usefulness of artificial intelligence in operational environments depends strongly on the quality, structure, and representativeness of the available data. The upstream oil and gas sector generates heterogeneous datasets from production operations, reservoir systems, equipment monitoring, maintenance records, drilling activities, and environmental measurements. A recent machine-learning-centered taxonomy of public datasets for upstream oil and gas applications highlighted substantial differences in data structure, target variables, accessibility, and analytical suitability across available sources, emphasizing the need for careful dataset characterization before model development (Baqer, 2026). More generally, missing observations and incomplete information can substantially alter the validity of machine-learning predictions. Structured approaches to missingness analysis and reconstruction have therefore been proposed to preserve predictive reliability in tabular datasets and prevent inappropriate imputation from distorting model

behavior (Jung & Yoo, 2025). Data quality is equally important in business-process analytics because user-generated and operational feedback often contain information that is not adequately captured in conventional process models. The augmentation of business-process model elements with end-user feedback illustrates how qualitative and quantitative information can be incorporated into organizational process analysis to provide a more complete representation of actual system behavior (Ahmed & Shahzad, 2022). For operational performance assessment, these findings reinforce the importance of rigorous data preprocessing, consistency checking, normalization, and validation before either efficiency measurement or predictive modeling is undertaken.

Beyond data quality, modern industrial analytics must also address transferability and robustness. Oil and gas units frequently operate under different capacities, equipment configurations, environmental conditions, and process constraints; therefore, models trained on one operational context may not generalize effectively to another. Physics-informed few-shot domain adaptation has recently been proposed as an approach to improving cross-domain condition monitoring when labeled data in the target environment are scarce (Wan et al., 2026). Resource-aware meta-learning methods have likewise been developed to enhance anomaly detection under limited-data conditions while controlling computational requirements (Zewail & Mokhtar, 2026). These approaches demonstrate an increasing emphasis on models that remain effective when data are limited, operating conditions change, or the deployment environment differs from the training environment. Such considerations are particularly important in industrial performance evaluation because a model that performs well only on a narrow historical dataset may have limited managerial value if it cannot maintain reliability across facilities, production periods, or operational regimes.

The complexity of modern production systems has also stimulated interest in advanced learning-based optimization. Machine-learning methods are increasingly used for combinatorial optimization problems involving scheduling, routing, assignment, and resource allocation. Nevertheless, reviews of neural and machine-learning approaches to combinatorial optimization have emphasized that model performance must be evaluated critically against conventional methods because learned solvers do not automatically outperform established optimization procedures across all problem classes (Ibrahim et al., 2026). Large language models have also begun to be incorporated

into the evolution and refinement of multi-objective heuristic methods, illustrating the emerging potential of artificial intelligence to participate directly in algorithmic search and optimization processes (Forniés-Tabuenca et al., 2026). These developments are conceptually relevant to operational efficiency analysis because they demonstrate the growing convergence between artificial intelligence and operations research. However, they also suggest that the greatest analytical value may emerge not from replacing established optimization techniques with machine learning, but from integrating the complementary strengths of multiple approaches.

Data Envelopment Analysis represents one of the most suitable operations-research methods for multidimensional performance benchmarking because it evaluates the relative efficiency of comparable decision-making units without requiring the specification of a predefined production function. In an oil and gas setting, operational units consume multiple resources and simultaneously generate economic, productive, and environmental outcomes. Consequently, a relative efficiency measure that incorporates several inputs and outputs is more informative than isolated productivity ratios. Yet DEA is fundamentally a frontier-based benchmarking method rather than a predictive learning algorithm. Although it can identify efficient and inefficient units and establish reference peers, conventional DEA does not directly provide a flexible predictive mechanism for estimating efficiency under new operating conditions, nor does it automatically reveal complex nonlinear patterns among operational variables. Integrating DEA with machine learning can therefore extend the analytical value of efficiency measurement by using DEA scores as empirically derived performance targets and then learning the relationships between those scores and the underlying operational indicators.

Machine learning is particularly suitable for this purpose because nonlinear algorithms can model interactions that are difficult to capture through conventional linear techniques. Research across different fields has repeatedly demonstrated the value of comparing alternative algorithms rather than assuming that a single model is universally optimal. Comparative evaluation of traditional machine-learning and transformer-based models, for instance, has shown that model suitability depends on data characteristics, computational demands, and the structure of the prediction task (Mannan, 2026). Forecasting research has similarly demonstrated that two-stage machine-learning architectures can be useful when data exhibit intermittent or irregular

patterns, highlighting the benefits of model structures that reflect the statistical behavior of the underlying process (Sucuoğlu et al., 2026). In financial decision environments, machine learning has also been used to benchmark climate-risk models and strengthen risk-management processes by identifying patterns that conventional analytical frameworks may overlook (Nimmala, 2023). Expert decision systems have been evaluated in exchange-rate insurance contexts, showing that artificial intelligence can support complex decision processes involving uncertain and multidimensional information (Álvarez-Diez et al., 2025). Collectively, these studies support the use of machine learning as a complementary analytical layer in organizational performance evaluation, provided that model selection is based on transparent comparative criteria such as predictive accuracy, stability, and error magnitude.

Despite their predictive strength, however, complex machine-learning models often create an interpretability problem. In managerial applications, a highly accurate model may have limited practical usefulness when decision-makers cannot determine why it assigns a particular prediction to a specific unit. This problem becomes particularly important in high-stakes industrial environments, where operational recommendations may affect production continuity, capital expenditure, environmental compliance, or workforce allocation. Recent research has increasingly emphasized the need to detect situations in which learned representations are insufficiently explanatory, arguing that predictive success alone does not guarantee that a model has captured decision-relevant structure in a meaningful manner (Raynal et al., 2026). Similarly, deployable artificial intelligence for Internet of Things and Industrial Internet of Things environments faces persistent constraints related to labeling, model transfer, computational resources, and explainability (Rharif et al., 2026). Explainability is therefore not merely an optional visualization feature; it is a core requirement for converting predictive analytics into accountable operational decision support.

The need for transparency is reinforced by developments in other high-risk artificial intelligence domains. Industrial cybersecurity research has emphasized that intelligent systems operating in Industrial Internet of Things environments must be evaluated with respect to architecture, datasets, attack resilience, and implementation constraints because industrial deployment introduces risks that are not apparent in laboratory evaluation (Singhal, 2026). Systematic analyses of adversarial artificial intelligence

have likewise demonstrated that model behavior can be manipulated through adaptive threats, making interpretability and robust evaluation increasingly important in complex computational environments (Patel, 2026). Privacy-preserving approaches such as federated learning further illustrate the tension between analytical utility and responsible data processing, particularly when operational information is distributed across organizations or facilities (Rokade, 2024). Although the present study is focused on efficiency rather than cybersecurity or privacy, these developments demonstrate a broader shift toward artificial intelligence systems that are not only accurate but also transparent, robust, resource-aware, and suitable for real-world governance.

Explainable artificial intelligence provides a practical response to this challenge by quantifying the contribution of individual variables to model predictions. Methods such as SHapley Additive exPlanations can identify which operational indicators increase or decrease predicted efficiency, rank variables according to their overall importance, and provide local explanations for individual operational units. The managerial value of such explanations is substantial because units with similar efficiency scores may suffer from different inefficiency mechanisms. One facility may be constrained primarily by excessive energy consumption, another by disproportionately high labor requirements, and another by environmental emissions or insufficient revenue generation relative to its resource base. An explainable model can therefore transform a predicted efficiency score into an actionable diagnostic profile. This human-centered orientation is consistent with the broader movement toward decision-support systems that provide interpretable and context-sensitive outputs rather than opaque classifications. Fuzzy cognitive maps, for example, have been used to develop human-centered decision-support systems that preserve interpretability while representing complex causal relationships among decision variables (Zakaria, 2026). Similarly, conversational analytical agents have been proposed to allow users to explore complex tabular data through tool-supported language interfaces, reflecting the increasing emphasis on making advanced analytical systems accessible and understandable to domain experts (Yang et al., 2025).

Explainability is also closely related to the broader issue of benchmark transparency. Emerging benchmarking research in artificial intelligence increasingly focuses not only on final predictive scores but also on how models perform across tasks, computational environments, and

evaluation dimensions. FML-bench, for example, has been proposed to benchmark machine-learning agents engaged in scientific research tasks, illustrating the need for structured evaluation frameworks capable of assessing multiple dimensions of intelligent-system performance (Zou, 2025). Benchmarking suites have likewise been developed for quantum learning to provide systematic and structure-based evaluation of learning algorithms under different computational assumptions (Lee & Choi, 2026). CarBench extends the benchmarking principle to neural surrogate models for high-fidelity automotive aerodynamics, where predictive models are evaluated against complex physical simulation outputs (Elrefaie, 2025). Although these domains differ substantially from oil and gas operations, they share a methodological concern directly relevant to the present study: predictive models become more useful when they are embedded in a rigorous benchmarking architecture that clearly defines the target performance measure and permits transparent comparison across units or algorithms.

Accordingly, an integrated DEA-machine learning-XAI framework can address several methodological limitations simultaneously. DEA can establish a multidimensional relative-efficiency frontier using operational resources, desirable outputs, and undesirable environmental outputs. Machine learning can then model the potentially nonlinear relationships between the operational indicators and DEA-derived efficiency scores, enabling prediction and sensitivity analysis beyond the original frontier calculation. Finally, explainable artificial intelligence can reveal the contribution of each operational indicator to predicted efficiency and identify the mechanisms driving high or low performance at both global and unit-specific levels. Such an architecture is particularly appropriate for oil and gas operations because performance is simultaneously influenced by feedstock consumption, energy use, water requirements, labor utilization, production volume, economic returns, and environmental pollutants. A framework that considers only production may reward environmentally costly units, while a framework based exclusively on emissions may overlook economically unsustainable operations. The combined approach therefore enables production, economic, resource, and environmental dimensions to be evaluated within a unified decision-support structure.

This integration is also consistent with the broader evolution of intelligent industrial systems toward data-driven, explainable, and adaptive performance management. The increasing availability of digital operational data makes it technically possible to construct sophisticated predictive

models, but their managerial value depends on whether they produce reliable benchmarks, remain computationally feasible, and generate explanations that can be translated into operational interventions. The emerging literature demonstrates important advances in industrial optimization, machine failure prediction, infrastructure benchmarking, domain adaptation, resource-aware artificial intelligence, decision support, and model explainability; nevertheless, these streams have often developed independently. In particular, there remains a need for frameworks that integrate frontier-based efficiency benchmarking with nonlinear machine-learning prediction and interpretable artificial intelligence in resource-intensive production environments. Addressing this gap is especially important for oil and gas companies, where improving efficiency requires simultaneous attention to productivity, financial performance, energy intensity, water use, labor productivity, and environmental emissions.

Therefore, the aim of this study is to design and evaluate an integrated framework based on Data Envelopment Analysis, machine learning, and explainable artificial intelligence for assessing, predicting, and interpreting the performance of operational units of the South Zagros Oil and Gas Production Company.

2. Methods and Materials

This study employed an applied, quantitative, analytical research design to develop an integrated performance-evaluation framework combining Data Envelopment Analysis (DEA), machine learning, and Explainable Artificial Intelligence (XAI). The principal purpose of the study was to assess the relative operational efficiency of comparable decision-making units (DMUs), predict efficiency performance from operational indicators, and identify the variables exerting the greatest influence on efficiency through an interpretable artificial intelligence approach. The methodological framework was therefore designed to overcome two important limitations associated with conventional performance assessment. First, DEA provides a non-parametric mechanism for benchmarking multiple organizational units in the presence of several inputs and outputs, but its efficiency scores do not by themselves provide a sufficiently detailed explanation of the nonlinear contribution of individual operational factors. Second, machine-learning algorithms can model complex nonlinear relationships and generate accurate predictions, but their outputs may remain difficult for managers to

interpret. The proposed framework integrated these approaches so that DEA generated the relative efficiency benchmarks, machine learning modeled and predicted efficiency, and XAI techniques subsequently explained the contribution and direction of influence of the individual performance indicators. The study was conducted through eight interconnected methodological stages. Initially, the theoretical foundations of organizational efficiency, DEA, machine learning, and explainable artificial intelligence were reviewed and relevant research gaps were identified. Subsequently, the DMUs and the candidate input, desirable-output, and undesirable-output variables were specified on the basis of the operational characteristics of the oil and gas industry and expert judgment. Operational, production, financial, resource-consumption, and environmental data were then collected and subjected to systematic preprocessing. In the next stage, DEA efficiency scores were calculated and efficient and inefficient units were identified. These scores subsequently constituted the target variable for constructing the supervised machine-learning database. A selected machine-learning model was trained to estimate efficiency scores from the operational indicators, its predictive performance was examined using several complementary error and goodness-of-fit criteria, and the resulting model was interpreted using SHapley Additive exPlanations (SHAP). Finally, the DEA benchmarking results and the XAI findings were integrated to identify the operational factors associated with superior or inferior performance and to formulate evidence-based managerial implications regarding resource productivity, production improvement, environmental performance, and operational efficiency.

The units of analysis were operational entities treated as independent DMUs. A total of 30 DMUs were included in the analytical dataset for the 2026–2027 operational year. Each unit was regarded as an organizational entity that transformed broadly comparable resources, including hydrocarbon feedstock, energy, water, and labor, into desirable production and financial outcomes while simultaneously generating environmentally undesirable outputs. Eligibility for inclusion was determined according to several substantive requirements. The DMUs were required to demonstrate sufficient homogeneity in their principal production, processing, exploitation, or related operational functions; employ comparable categories of production resources; generate measurable operational and economic outputs; have measurable environmental indicators; provide complete and verifiable data for all

variables entered into the DEA framework; and possess sufficient managerial or operational independence for performance to be meaningfully evaluated at the unit level. These criteria were intended to satisfy the comparability requirement underlying DEA and to reduce distortions that could arise from evaluating fundamentally dissimilar organizational entities against the same efficiency frontier. The final analytical framework therefore consisted of 30 operational DMUs for which complete information was available across all four inputs, two desirable outputs, and five undesirable outputs.

A census approach rather than probabilistic sampling was adopted. All eligible DMUs for which complete, valid, and comparable information was available were incorporated into the analysis. The use of a census was appropriate because the number of eligible operational units was limited and their administrative and operational records were accessible for performance assessment. Inclusion of all eligible units eliminated conventional sampling error and enabled the DEA frontier to be estimated from the maximum amount of available information. This approach was also particularly relevant to DEA because the identification of efficient frontier units and peer benchmarks depends directly on the set of DMUs included in the production possibility space. Consequently, excluding otherwise eligible units through sampling could modify the estimated frontier and alter the efficiency classification of the remaining DMUs. The number of DMUs was additionally assessed in relation to the dimensionality of the DEA specification. With four input indicators, two desirable outputs, and five undesirable outputs, the inclusion of 30 units was considered sufficient to provide an operationally useful level of discrimination between efficient and inefficient units while maintaining the multivariate structure necessary to reflect economic, resource-consumption, production, and environmental dimensions of performance.

The primary data-collection instrument consisted of a structured researcher-developed data extraction framework designed to retrieve standardized quantitative information from the operational, production, financial, human-resource, energy-consumption, water-consumption, and environmental information systems available for the investigated units. Unlike questionnaire-based performance measurement, the present study relied on objective organizational records and actual operational data. The data-extraction framework was constructed after reviewing the conceptual literature on efficiency measurement in energy-intensive industries and examining the operational

characteristics of oil and gas production and processing activities. Candidate indicators were screened according to three principal criteria: their direct theoretical and operational relationship with oil and gas production processes, the availability and reliability of actual data across all included DMUs, and their comparability across operational units. Expert judgment from individuals familiar with production processes, operational management, efficiency assessment, and environmental performance was also used during the variable-specification stage to ensure that the selected indicators adequately represented the principal resources consumed, the economic and production value created, and the environmental burden generated by the investigated units.

Four variables were specified as DEA inputs. Feedstock, denoted by (X_1), represented the quantity of hydrocarbon material entering or being consumed within the relevant operational process and was measured in tonnes. This variable reflected the principal material resource required for oil and gas production or processing and was therefore considered a major determinant of the scale of operational activity. Energy consumption, denoted by (X_2), represented the electrical and thermal energy used by process equipment, compression facilities, production systems, auxiliary installations, and other operating infrastructure and was expressed in megawatt-hours (MWh). Because energy constitutes an important component of both operating expenditure and environmental burden in oil and gas production, its inclusion allowed the efficiency framework to distinguish between units achieving similar output levels with different energy requirements. Water consumption, denoted by (X_3), represented the amount of water used in processing, cooling, injection, utility services, and related operational activities and was measured in thousand cubic meters. The inclusion of water was particularly relevant because efficient water utilization has both economic and environmental implications in resource-intensive industrial operations. Labor input, denoted by (X_4), represented the human resources employed in production, operation, maintenance, repair, monitoring, and related activities and was measured in person-hours. Measuring labor in person-hours rather than merely the number of employees provided a more direct representation of actual labor resources consumed during the study period.

Two desirable outputs were incorporated into the model. The first was main production volume, measured in tonnes, which represented the principal physical output generated by the operational unit. The second was revenue, measured in

billion Iranian rials, which represented the economic value associated with the unit's productive activities. Simultaneously incorporating physical production and revenue enabled the model to account for both operational productivity and economic performance rather than treating production volume alone as an adequate representation of organizational success. In addition, five environmentally undesirable outputs were incorporated because an operational unit that generates high production or revenue by imposing disproportionately high environmental costs should not necessarily be classified as efficient. These indicators comprised carbon dioxide emissions (CO_2), nitrogen oxides (NO_x), sulfur dioxide (SO_2), particulate matter with an aerodynamic diameter of 10 micrometers or less (PM_{10}), and chemical oxygen demand (COD). The first four indicators represented major atmospheric pollutants associated with fuel combustion, processing operations, industrial equipment, and hydrocarbon production, whereas COD was incorporated as an indicator of the pollution load of industrial wastewater. The simultaneous consideration of desirable and undesirable outputs therefore enabled the efficiency framework to evaluate operational performance from an integrated economic-production-environmental perspective.

Prior to analysis, the extracted dataset underwent systematic quality control and preprocessing. Records were initially examined for completeness, internal consistency, duplicate entries, implausible values, measurement-unit inconsistencies, and coding errors. Missing observations were identified for each variable and each DMU before model estimation. Units lacking the minimum information necessary for valid efficiency assessment were not eligible for inclusion in the final analytical dataset, whereas isolated data-quality issues were cross-checked against the corresponding organizational records wherever possible. Potential outliers were evaluated in conjunction with substantive operational knowledge because extremely high or low observations in industrial datasets may represent either genuine scale differences or erroneous measurements. Consequently, observations were not removed solely because of statistical extremity; exclusion was considered only when the value could be attributed to recording, measurement, or data-entry error rather than a legitimate operational condition. Because the model incorporated variables measured on substantially different scales, including tonnes, megawatt-hours, thousand cubic meters, person-hours, monetary units, and environmental indicators, appropriate scaling and normalization procedures were

applied when required for the machine-learning stage. The original measurement structure was retained for DEA whenever transformation was unnecessary, while the machine-learning dataset was standardized or normalized in a manner compatible with the selected predictive algorithm. All preprocessing procedures were performed consistently across the 30 DMUs to prevent differential treatment of individual units and to ensure that subsequent efficiency comparisons and predictive analyses were based on a coherent analytical dataset.

Data analysis was conducted through an integrated DEA-machine learning-XAI procedure. In the first analytical stage, DEA was employed to estimate the relative efficiency of the 30 DMUs. DEA was selected because it is a non-parametric frontier technique capable of simultaneously incorporating multiple inputs and multiple outputs without requiring the specification of a predefined functional relationship between resources and production outcomes. This characteristic was particularly appropriate for the operational context of the study, in which performance reflected a multidimensional transformation process involving material, energy, water, and labor resources as well as production, financial, and environmental consequences. A DEA specification capable of incorporating undesirable outputs was employed so that reductions in environmental pollutants were treated as improvements in performance rather than as conventional desirable production outputs. The efficiency frontier was constructed from the observed best-performing units, and the relative position of each DMU against this frontier was used to obtain its efficiency score. Units located on the estimated frontier were classified as relatively efficient, whereas units positioned below the frontier were regarded as relatively inefficient. For inefficient DMUs, the DEA results were additionally used to identify efficient reference units and provide information concerning the magnitude and direction of potential improvement in resource utilization, desirable production, and environmental performance.

The specification of the DEA model was aligned with the managerial objective of evaluating the extent to which operational units could improve performance while controlling resource consumption and undesirable environmental consequences. The model simultaneously incorporated the four inputs of feedstock, energy, water, and labor; the two desirable outputs of main production and revenue; and the five undesirable outputs of CO_2 , NO_x , SO_2 , PM_{10} , and COD. This multidimensional configuration allowed a unit to be considered comparatively superior only

when its performance reflected an appropriate balance between resource consumption, productive and financial outcomes, and environmental burden. The resulting DEA efficiency score for each DMU was subsequently stored as the principal performance measure and used as the dependent or target variable in the machine-learning component of the framework. In this manner, DEA served not only as a benchmarking technique but also as the mechanism through which an empirically derived efficiency target was created for subsequent predictive modeling.

Following the DEA stage, a supervised machine-learning dataset was constructed in which the operational, resource, production, financial, and environmental indicators constituted the predictor features and the DEA efficiency score constituted the target variable. Candidate regression algorithms were evaluated according to their ability to represent potentially nonlinear and interactive relationships among the performance indicators. Model development incorporated procedures intended to limit overfitting, particularly because the number of DMUs was relatively limited in comparison with the number of available predictor variables. Cross-validation was therefore used rather than relying exclusively on a single training–testing split. Where hyperparameter optimization was required, tuning was conducted within the model-development process so that model selection was based on out-of-sample predictive performance rather than in-sample fit alone. Data preprocessing operations such as scaling or normalization were performed within the analytical pipeline when required by the selected algorithm so that information from validation observations did not leak into the training stage.

The predictive accuracy of the selected machine-learning model was assessed using the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE), together with cross-validation results. The (R^2) statistic was used to indicate the proportion of variation in DEA-derived efficiency scores explained by the model, whereas MAE quantified the average absolute discrepancy between predicted and observed efficiency scores in their original metric. RMSE provided a complementary error measure that assigned greater weight to larger prediction errors and was therefore useful for identifying models that occasionally produced substantial deviations from the DEA benchmark. Cross-validation performance was examined to assess the stability and generalizability of the model across different subsets of the DMUs. Model selection was based on the joint consideration of these indicators rather than on a single goodness-of-fit

measure, with preference given to a model demonstrating comparatively high explanatory performance, low prediction error, and stable validation results.

After predictive model estimation, explainability analysis was performed using SHapley Additive exPlanations (SHAP). SHAP was employed to decompose each model prediction into the contribution of the individual explanatory variables and thereby transform the machine-learning model from a purely predictive mechanism into an interpretable decision-support instrument. Global SHAP analysis was used to rank the variables according to their overall contribution to predicted operational efficiency, while the distribution and sign of SHAP values were examined to determine whether higher or lower values of each indicator tended to increase or reduce predicted efficiency. This procedure enabled identification of the resource-consumption, production, financial, and environmental factors that exerted the greatest influence on model behavior. Local SHAP explanations were also used to interpret the predictions of individual DMUs, permitting managers to determine why a particular unit received a relatively high or low predicted efficiency score and which variables were primarily responsible for that result. Where nonlinear or threshold relationships emerged, SHAP-based dependence patterns were examined to identify changes in the direction or magnitude of variable effects across different operational levels.

The final analytical stage integrated the DEA and XAI findings. DEA results provided information regarding relative efficiency, frontier membership, peer benchmarking, and the performance gaps of inefficient units, whereas the explainable machine-learning analysis clarified which operational characteristics were most strongly associated with changes in efficiency scores and how these characteristics affected individual DMUs. Convergence between the two analytical components was interpreted as stronger evidence regarding the operational determinants of efficiency. For example, when DEA identified a unit as inefficient and SHAP simultaneously indicated that excessive energy consumption, water consumption, labor utilization, or pollutant generation contributed negatively to its predicted efficiency, the combined framework provided a more diagnostically informative explanation than either technique could produce independently. Conversely, variables showing substantial positive SHAP contributions could be interpreted as characteristics associated with superior performance, subject to the operational context of the individual unit. The integrated results were ultimately

used to formulate managerial recommendations aimed at reducing inefficient resource consumption, improving production and economic performance, controlling environmentally undesirable outputs, strengthening benchmarking practices, and supporting transparent data-driven decision-making across operational units.

3. Findings and Results

The final analytical sample consisted of 30 decision-making units (DMUs) representing operational units included in the performance-evaluation framework. With respect to the organizational distribution of the units, 7 DMUs (23.3%) belonged to the Nar and Kangan operational region, 7 DMUs (23.3%) to the Aghar and Dalan operational region, and 7 DMUs (23.3%) to Bandar Abbas Oil Refinery. In addition, 6 DMUs (20.0%) represented offshore oil and gas platforms, while Parsian Sepehr Refinery, Karun Oil and Gas Production Company, and the National Iranian South Oilfields Company each accounted for one DMU (3.3%). All 30 units had complete data for the four resource inputs, two

desirable outputs, and five undesirable environmental outputs. No DMU was excluded from the final analysis because of missing information after data verification and preprocessing. Examination of the raw operational records showed substantial heterogeneity in the scale of operations across the units, particularly in feedstock utilization, energy consumption, labor input, production volume, and revenue. Such heterogeneity was expected because the DMUs differed in production capacity and technical scale, although all were analyzed within the same resource–output performance framework. Before efficiency estimation, the data were screened for missing values, extreme observations, inconsistencies in measurement units, and implausible records. Extreme observations that reflected genuine differences in operational scale were retained because DEA is designed to compare heterogeneous production scales, whereas records identified as data-entry inconsistencies were checked against their original operational sources. The resulting database was therefore considered suitable for DEA efficiency estimation and subsequent machine-learning analysis.

Table 1

Descriptive Statistics of Inputs, Desirable Outputs, and Undesirable Outputs of the Operational Units

Variable	Role in Model	Unit	Minimum	Maximum	Mean	Standard Deviation
Feedstock	Input	Tonnes	59,599.93	6,817,710.98	1,873,426.54	1,957,318.62
Energy consumption	Input	MWh	41,999.91	5,244,389.48	1,431,856.77	1,503,428.39
Water consumption	Input	Thousand m ³	110.50	11,144.35	3,064.72	3,207.56
Labor	Input	Person-hours	115,699.91	11,668,780.98	3,169,845.63	3,322,716.80
Main production	Desirable output	Tonnes	52,416.38	6,486,574.25	1,762,938.44	1,846,307.52
Revenue	Desirable output	Billion IRR	1,924.16	251,648.73	68,714.52	72,508.31
CO ₂	Undesirable output	Tonnes	11,382.46	642,817.35	188,436.29	173,519.64
NO _x	Undesirable output	Tonnes	38.72	2,846.91	816.35	754.82
SO ₂	Undesirable output	Tonnes	24.31	2,315.67	624.58	607.44
PM ₁₀	Undesirable output	Tonnes	9.18	786.42	214.76	205.91
COD	Undesirable output	Tonnes	16.42	1,247.56	347.82	329.64

As shown in Table 1, substantial variation was observed across the operational units in almost all resource, production, financial, and environmental indicators. Feedstock consumption ranged from 59,599.93 to 6,817,710.98 tonnes, with a mean of 1,873,426.54 tonnes and a standard deviation of 1,957,318.62 tonnes, indicating marked differences in operational scale among the DMUs. Energy consumption showed a similarly wide distribution, ranging from 41,999.91 to 5,244,389.48 MWh, with a mean of 1,431,856.77 MWh. Average water consumption was 3,064.72 thousand cubic meters, whereas the mean labor

requirement was 3,169,845.63 person-hours. Regarding desirable outputs, the units produced an average of 1,762,938.44 tonnes of main product, while the average revenue was 68,714.52 billion IRR. Considerable dispersion was also observed in undesirable environmental outputs. Mean CO₂ emissions were 188,436.29 tonnes, while average NO_x, SO₂, PM₁₀, and COD values were 816.35, 624.58, 214.76, and 347.82 tonnes, respectively. The relatively large standard deviations observed for most variables confirmed that the units operated at different scales and consumed substantially different quantities of resources. However,

because DEA evaluates relative efficiency by comparing the combination of resources consumed with desirable and undesirable outputs produced, these differences did not preclude efficiency estimation. Rather, they enabled the model to distinguish between units achieving superior

production and financial outcomes with comparatively lower resource consumption and environmental burden and those exhibiting inefficient combinations of inputs and outputs.

Table 2

DEA Efficiency Scores and Efficiency Classification of the Decision-Making Units

DMU	Operational Unit	Efficiency Score	Rank	Efficiency Status
DMU1	Nar Gathering Center 1	0.921	9	Inefficient
DMU2	Kangan Gathering Center 2	0.884	14	Inefficient
DMU3	Kangan Gathering Center 3	0.846	20	Inefficient
DMU4	Nar Gathering Center 4	0.903	11	Inefficient
DMU5	Kangan Gathering Center 5	0.812	24	Inefficient
DMU6	Nar Gathering Center 6	1.000	1	Efficient
DMU7	Kangan Gathering Center 7	0.967	6	Inefficient
DMU8	Aghar Gathering Center 1	1.000	1	Efficient
DMU9	Aghar Gathering Center 2	0.932	8	Inefficient
DMU10	Aghar Gathering Center 3	0.875	16	Inefficient
DMU11	Aghar Gathering Center 4	0.793	26	Inefficient
DMU12	Dalan Gathering Center 1	0.958	7	Inefficient
DMU13	Dalan Gathering Center 2	1.000	1	Efficient
DMU14	Dalan Gathering Center 3	0.901	12	Inefficient
DMU15	Atmospheric Distillation Unit	0.857	18	Inefficient
DMU16	Vacuum Distillation Unit	1.000	1	Efficient
DMU17	LPG Unit	0.894	13	Inefficient
DMU18	Sour Water and Merox Treatment Unit	0.768	28	Inefficient
DMU19	Isomax Unit	0.824	22	Inefficient
DMU20	Hydrogen Unit	0.913	10	Inefficient
DMU21	Amine Gas Treatment Unit	0.851	19	Inefficient
DMU22	Parsian Sepehr Process Units	0.979	5	Inefficient
DMU23	Salman Offshore Platform	1.000	1	Efficient
DMU24	SPQ1 Gas Platform	0.941	7	Inefficient
DMU25	Hengam Platform	0.865	17	Inefficient
DMU26	Forouzan Platform	0.821	23	Inefficient
DMU27	Soroush Platform	0.742	30	Inefficient
DMU28	Nasr Platform	0.782	27	Inefficient
DMU29	Karun Oil and Gas Production Unit	0.838	21	Inefficient
DMU30	National Iranian South Oilfields Production Unit	0.754	29	Inefficient

The DEA results presented in Table 2 demonstrated considerable differences in relative efficiency among the 30 investigated operational units. The average efficiency score across all DMUs was 0.879, indicating that, on average, the units were operating approximately 12.1% below the estimated best-practice frontier. Five units, namely DMU6, DMU8, DMU13, DMU16, and DMU23, achieved an efficiency score of 1.000 and were therefore identified as frontier-efficient units. These units jointly constituted the empirical benchmark against which the remaining units were

evaluated. Thus, 16.7% of all DMUs were classified as efficient, while 83.3% demonstrated some degree of relative inefficiency. Among the inefficient units, DMU22 exhibited the highest efficiency score at 0.979 and was therefore positioned very close to the efficiency frontier, suggesting that only relatively modest improvements in its resource allocation or environmental performance would be required for it to attain frontier status. DMU7 and DMU12 also showed comparatively strong performance, with efficiency scores of 0.967 and 0.958, respectively. By contrast, the

lowest relative efficiency was observed for DMU27, with a score of 0.742, followed by DMU30 with 0.754, DMU18 with 0.768, and DMU28 with 0.782. These units represented the greatest potential for performance improvement. Examination of the DEA projections indicated that inefficiency generally reflected a combination of relatively high energy consumption, labor utilization, and environmental emissions relative to the corresponding levels of production and revenue. The efficient units, in contrast, achieved more favorable combinations of desirable outputs and resource requirements while maintaining comparatively

lower levels of undesirable environmental outputs. The results therefore indicated that operational efficiency was not determined simply by the absolute size of production; rather, units were efficient when they generated comparatively high production and economic value from a balanced combination of resources while simultaneously controlling environmental burdens. The frontier units consequently served as meaningful reference benchmarks for the inefficient DMUs and provided the basis for identifying practical resource-saving and emission-reduction targets.

Table 3

Comparison of Machine-Learning Models for Predicting DEA Efficiency Scores

Model	R ²	MAE	RMSE	Cross-Validated R ²	Cross-Validated RMSE
Linear Regression	0.684	0.047	0.061	0.602	0.069
Support Vector Regression	0.812	0.034	0.046	0.748	0.052
Random Forest Regression	0.893	0.025	0.035	0.831	0.041
Gradient Boosting Regression	0.918	0.021	0.031	0.857	0.038
XGBoost Regression	0.944	0.017	0.026	0.889	0.034

Table 3 compares the predictive performance of five candidate machine-learning algorithms used to estimate DEA-derived efficiency scores from the operational, production, financial, and environmental indicators. The results showed that all nonlinear models outperformed conventional linear regression, indicating that the relationships between the performance variables and efficiency scores were not adequately represented by a simple linear functional form. Linear regression produced an (R²) of 0.684, an MAE of 0.047, and an RMSE of 0.061, suggesting moderate explanatory performance but comparatively larger prediction errors. Support Vector Regression improved the coefficient of determination to 0.812 and reduced MAE and RMSE to 0.034 and 0.046, respectively. Random Forest Regression demonstrated stronger predictive performance, explaining 89.3% of the observed variation in DEA efficiency scores, with an MAE of 0.025 and an RMSE of 0.035. Gradient Boosting Regression further improved performance, producing an (R²) of 0.918 and reducing the MAE and RMSE to 0.021 and 0.031, respectively. The highest predictive accuracy was achieved by the XGBoost model, which produced an (R²)

of 0.944, indicating that approximately 94.4% of the variation in DEA-derived efficiency scores was captured by the selected operational indicators. Its MAE of 0.017 indicated that the average absolute difference between actual DEA scores and machine-predicted efficiency scores was only 0.017 efficiency units, while its RMSE of 0.026 confirmed a relatively low frequency of large prediction errors. The superiority of XGBoost remained evident under cross-validation, where the model achieved a mean (R²) of 0.889 and a mean RMSE of 0.034. The relatively small difference between the original and cross-validated performance indices suggested that the predictive results were not solely attributable to overfitting. Accordingly, XGBoost was selected as the final predictive model for the explainable artificial intelligence stage. The improvement of tree-based ensemble algorithms over linear regression further indicated that efficiency was affected by nonlinear relationships and interactions among resource consumption, productive performance, revenue generation, and environmental outputs, thereby justifying the use of an explainable nonlinear machine-learning framework rather than a conventional regression model alone.

Table 4
Global SHAP Importance of Variables in Predicting Operational Efficiency

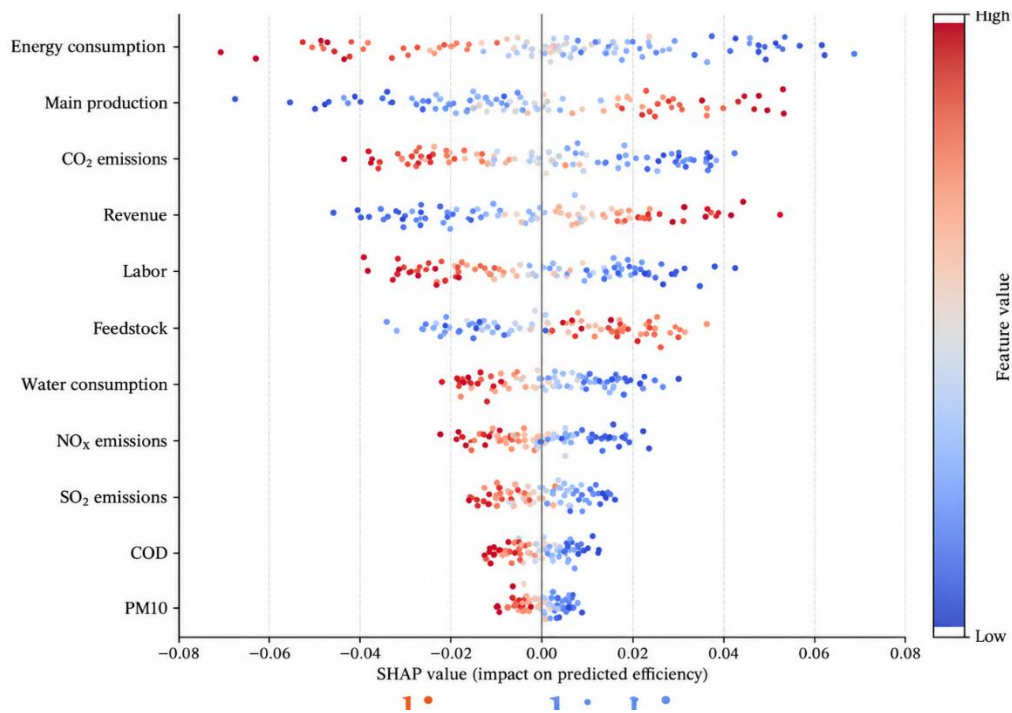
Rank	Variable	Mean Absolute SHAP Value	Relative Importance (%)	Predominant Direction of Effect on Efficiency
1	Energy consumption	0.041	18.6	Negative at high values
2	Main production	0.036	16.4	Positive at high values
3	CO ₂ emissions	0.031	14.1	Negative at high values
4	Revenue	0.027	12.3	Positive at high values
5	Labor	0.023	10.5	Negative when disproportionate to output
6	Feedstock	0.019	8.6	Nonlinear/mixed
7	Water consumption	0.016	7.3	Predominantly negative
8	NO _x emissions	0.011	5.0	Negative
9	SO ₂ emissions	0.008	3.6	Negative
10	COD	0.005	2.3	Negative
11	PM ₁₀ emissions	0.003	1.3	Negative

The SHAP analysis presented in Table 4 identified substantial differences in the contribution of the investigated predictors to the efficiency predictions generated by the XGBoost model. Energy consumption emerged as the most influential variable, with a mean absolute SHAP value of 0.041 and a relative importance of 18.6%. Higher energy consumption generally produced negative SHAP values when it was not accompanied by a proportionate increase in production or revenue, indicating that excessive energy intensity reduced predicted operational efficiency. Main production was the second most influential variable, accounting for 16.4% of total explanatory importance. Higher production values generally shifted predictions toward greater efficiency, particularly when increased production was achieved without a commensurate escalation in inputs and environmental outputs. CO₂ emissions represented the third most important variable, contributing 14.1% of total importance. Elevated CO₂ emissions were predominantly associated with negative SHAP contributions, demonstrating that the inclusion of environmental performance materially affected the efficiency assessment. Revenue ranked fourth, accounting for 12.3% of predictive importance, and generally exerted a positive influence on predicted efficiency. Labor input

represented 10.5% of total importance; its influence was particularly negative in DMUs where relatively high person-hours were associated with moderate production and revenue, suggesting the existence of labor-productivity inefficiencies. Feedstock accounted for 8.6% of total importance and demonstrated a nonlinear effect. Moderate and operationally proportionate feedstock levels were associated with positive contributions, whereas exceptionally high feedstock consumption without corresponding output gains reduced predicted efficiency. Water consumption contributed 7.3% of total explanatory importance and generally exerted a negative effect at higher levels. Environmental pollutants other than CO₂ had smaller but consistently negative contributions, with NO_x, SO₂, COD, and PM₁₀ accounting for 5.0%, 3.6%, 2.3%, and 1.3% of total SHAP importance, respectively. Collectively, these results demonstrated that the machine-learning model did not interpret economic or production performance independently of environmental and resource-consumption indicators. Instead, the highest efficiency predictions were associated with a combined pattern of higher production and revenue, lower energy and water intensity, greater labor productivity, and reduced pollutant emissions.

Figure 1

SHAP Summary Plot of the Effects of Operational, Economic, Resource-Consumption, and Environmental Variables on Predicted Efficiency Scores



The SHAP summary plot illustrated both the relative importance of each predictor and the direction of its contribution to the predicted DEA efficiency score. The variables were ordered according to their mean absolute SHAP values, with energy consumption, main production, CO₂ emissions, revenue, and labor occupying the highest positions. The distribution of observations across positive and negative SHAP values showed that high levels of energy consumption were predominantly concentrated on the negative side of the efficiency prediction, whereas high levels of main production and revenue were generally concentrated on the positive side. High CO₂ emissions consistently shifted predictions toward lower efficiency, confirming that environmental burdens constituted an important source of inefficiency rather than a peripheral performance indicator. The plot also revealed nonlinear patterns for feedstock and labor. Increased feedstock was not inherently associated with lower efficiency; rather, its contribution depended on whether the corresponding production and revenue gains justified the additional input requirement. A similar pattern was observed for labor, where relatively high person-hour values had limited adverse influence in high-output units but exerted stronger negative effects when coupled with lower production productivity.

Water consumption and the remaining environmental pollutants showed more uniformly negative effects at elevated levels. Therefore, the SHAP visualization confirmed that predicted performance was determined through a multidimensional interaction among production scale, economic returns, input intensity, and environmental externalities rather than by any single operational characteristic.

The integration of DEA and explainable machine-learning results provided a more detailed interpretation of the performance differences among the operational units. DEA identified the units positioned on the best-practice frontier and quantified the relative efficiency gap for each inefficient unit, whereas XGBoost and SHAP clarified the operational mechanisms underlying these differences. The five frontier-efficient units shared a general pattern of comparatively favorable production-to-input ratios, higher revenue generation relative to resource utilization, and lower environmental burdens relative to their production scale. By contrast, the lowest-performing DMUs were characterized predominantly by excessive energy requirements, relatively unfavorable labor productivity, higher carbon intensity, and weaker conversion of feedstock into economically valuable output. The integration of both analytical components

therefore demonstrated that inefficient performance could not be attributed solely to insufficient production. In several units, production was relatively high in absolute terms, but resource consumption and undesirable outputs increased at a greater rate, thereby reducing relative efficiency. This finding was particularly important because it showed that managerial policies focusing exclusively on production expansion would not necessarily improve organizational performance unless accompanied by improvements in energy efficiency, labor productivity, water management, and emission control.

A closer examination of the inefficient units indicated different patterns of performance deficiencies and therefore suggested that improvement strategies should be unit-specific rather than uniform. Units with efficiency scores close to one, such as DMU22, DMU7, and DMU12, showed comparatively small efficiency gaps and required relatively modest adjustments to move toward the frontier. For these units, incremental energy conservation, process optimization, preventive maintenance, and targeted emission reduction could potentially generate sufficient improvement to achieve frontier-level performance. In contrast, the units with efficiency scores below approximately 0.80, particularly DMU27, DMU30, DMU18, and DMU28, exhibited broader inefficiency patterns involving more than one resource or environmental dimension. SHAP-based local interpretation showed that their lower efficiency predictions were generally driven by combinations of high energy consumption, comparatively high labor requirements, elevated environmental emissions, and insufficient desirable output relative to total resource utilization. These units consequently require more fundamental operational interventions, including process redesign, technological modernization, optimization of equipment loading, reduction of avoidable downtime, improved energy-management systems, and stronger environmental control measures.

The efficiency frontier also provided an empirical basis for benchmarking. DMU6, DMU8, DMU13, DMU16, and DMU23 represented efficient reference units because they achieved the most favorable combinations of resource use, production, revenue, and environmental performance observed within the sample. However, efficiency scores of 1.000 should be interpreted as relative rather than absolute efficiency. These units were classified as efficient because no other observed unit or feasible combination of observed units dominated their performance under the specified model. Their status therefore does not imply that additional

improvement is impossible. Nevertheless, they provide internally relevant benchmarks because their resource-conversion patterns were superior relative to the remaining investigated units. Benchmarking inefficient units against technically comparable efficient units can therefore support the identification of realistic improvement targets rather than relying exclusively on externally imposed standards.

The predictive analysis further demonstrated that DEA efficiency could be approximated with high accuracy using operational indicators. The selected XGBoost model explained 94.4% of the variation in DEA-derived efficiency scores and maintained a comparatively strong performance under cross-validation. This result indicated that the operational characteristics contained sufficient information to predict efficiency patterns with relatively low error. More importantly, the SHAP analysis converted the predictive model into an interpretable decision-support mechanism. Rather than merely assigning a predicted efficiency score to each unit, the integrated framework identified the variables responsible for moving an individual unit's prediction upward or downward. This characteristic is particularly relevant for operational management because two units with similar overall efficiency scores may have completely different underlying inefficiency structures. One unit may be primarily constrained by excessive energy consumption, while another may suffer from low labor productivity or high environmental emissions. Consequently, identical managerial interventions would not necessarily be appropriate for both units.

Overall, the combined results supported the usefulness of integrating DEA, machine learning, and explainable artificial intelligence for multidimensional operational performance evaluation. DEA established the relative efficiency frontier and identified efficient and inefficient DMUs; XGBoost captured nonlinear relationships between the operational indicators and DEA-derived efficiency; and SHAP revealed the relative importance, direction, and unit-specific contribution of each predictor. The findings showed that energy consumption was the strongest factor differentiating efficiency levels, followed by production volume, CO₂ emissions, revenue, and labor utilization. The results also demonstrated that economically strong performance alone was insufficient to guarantee efficiency when it was accompanied by disproportionate resource consumption or environmental pollution. Accordingly, the integrated framework provided a more comprehensive representation of performance than conventional financial or production indicators and enabled simultaneous

consideration of productivity, economic value creation, resource conservation, and environmental sustainability.

4. Discussion and Conclusion

The present study developed and evaluated an integrated framework combining Data Envelopment Analysis (DEA), machine learning, and Explainable Artificial Intelligence (XAI) to assess the performance of operational units in the oil and gas sector. The findings demonstrated substantial heterogeneity in operational efficiency across the 30 decision-making units. The mean DEA efficiency score was 0.879, indicating that, on average, the investigated units operated approximately 12.1% below the estimated best-practice frontier. Five units, namely DMU6, DMU8, DMU13, DMU16, and DMU23, achieved an efficiency score of 1.000 and were therefore classified as relatively efficient frontier units, whereas the remaining units exhibited different levels of inefficiency. The lowest efficiency scores were observed for DMU27, DMU30, DMU18, and DMU28. These results show that substantial opportunities exist for reducing resource consumption, improving desirable outputs, and controlling environmental burdens across the less efficient operational units. The dispersion of efficiency scores also confirms the usefulness of a relative benchmarking framework for identifying performance gaps that may remain concealed when production volume or financial output is examined independently.

The DEA results further indicated that operational efficiency was not solely associated with the absolute level of production. Some units with relatively large production scales remained inefficient because they consumed disproportionately high levels of energy, labor, water, or feedstock or generated comparatively high levels of undesirable environmental outputs. This finding supports the multidimensional conceptualization of industrial performance in which efficiency depends on the relationship between resources consumed and the totality of desirable and undesirable consequences generated. Contemporary artificial intelligence and industrial optimization research similarly emphasizes that production-system performance should be evaluated through multiple objectives rather than a single output indicator. In smart manufacturing, integrated AI-based optimization frameworks have been developed specifically because production quality, resource consumption, process reliability, and system responsiveness are simultaneously relevant to operational performance

(Champa & Segall, 2026). Likewise, learning-based multi-objective scheduling approaches have demonstrated that improvements in one operational dimension may be accompanied by deterioration in another unless the production system is evaluated through a balanced set of objectives (Akram et al., 2026). Thus, the present DEA findings reinforce the importance of simultaneously considering production, revenue, resource utilization, and environmental externalities.

The five frontier-efficient units represent internally observed examples of comparatively favorable resource conversion and environmental performance. Their efficiency status does not indicate absolute technical perfection; rather, it means that no other observed DMU or feasible combination of units dominated their performance under the specified DEA structure. This distinction is important because benchmarking is inherently relative. Recent benchmarking research across artificial intelligence and computational systems similarly emphasizes that benchmark scores derive meaning from a defined comparison environment rather than representing universally optimal performance. Structured benchmark suites developed for GPU and CPU rendering illustrate how system performance must be understood relative to alternative configurations and workloads (Wang & Yu, 2023a, 2023b). Similar reasoning underlies AI benchmark development in scientific research and smart-grid forecasting, where models are compared under common datasets and evaluation criteria rather than judged through isolated performance measures (Agrawal, 2026; Zou, 2025). In the present study, the DEA frontier therefore provides a context-specific performance benchmark that can be used to identify realistic peer units and operational improvement targets.

The machine-learning stage revealed that nonlinear predictive models were considerably more effective than conventional linear regression in reproducing the DEA efficiency scores. Linear regression explained 68.4% of the variation in efficiency, whereas Support Vector Regression, Random Forest, Gradient Boosting, and XGBoost achieved progressively stronger performance. XGBoost produced the best results, with an (R^2) of 0.944, an MAE of 0.017, and an RMSE of 0.026. The model also retained comparatively strong performance under cross-validation, with a mean cross-validated (R^2) of 0.889 and an RMSE of 0.034. These findings suggest that operational efficiency in oil and gas units is governed by nonlinear and interactive

relationships that cannot be adequately represented through simple additive linear effects.

This result is consistent with a broad body of recent machine-learning research showing that advanced models can capture complex performance relationships across heterogeneous datasets. Comparative evaluation of traditional machine-learning and transformer-based methods has demonstrated that model effectiveness depends strongly on the underlying structure of the prediction problem and that nonlinear methods may substantially outperform simpler approaches when relationships among variables are complex (Mannan, 2026). Similarly, two-stage machine-learning architectures have shown advantages in forecasting settings characterized by irregular and intermittent patterns (Sucuoğlu et al., 2026). Studies in industrial predictive maintenance have likewise demonstrated that machine-learning models can identify nonlinear patterns in operational variables that are predictive of equipment deterioration and production disruption (Ajayi et al., 2025; Orazov et al., 2026). The present findings extend this reasoning from failure prediction to efficiency prediction, indicating that the interaction of resource use, output generation, and environmental burden can be learned effectively through nonlinear ensemble models.

The strong predictive performance of XGBoost also demonstrates that DEA-derived efficiency scores contained a sufficiently systematic relationship with the underlying operational indicators to support supervised learning. This result is methodologically important because it suggests that DEA and machine learning can be used as complementary rather than competing techniques. DEA establishes an empirical efficiency frontier without requiring a predetermined functional relationship, whereas machine learning can approximate the resulting efficiency surface and provide predictive flexibility. Reviews of machine learning for optimization have emphasized that learned models should not automatically replace established analytical approaches; rather, hybridization can be more advantageous when different methods contribute distinct capabilities (Ibrahim et al., 2026). Emerging approaches that combine artificial intelligence with heuristic and multi-objective optimization similarly demonstrate the potential of hybrid analytical architectures for solving complex operational problems (Forníes-Tabuenca et al., 2026). Accordingly, the present framework capitalizes on the benchmarking strengths of DEA and the nonlinear predictive capability of machine learning.

The XAI results added a further interpretive layer to the model by identifying energy consumption as the most influential predictor of efficiency, with a relative SHAP importance of 18.6%. Higher energy consumption generally reduced predicted efficiency when it was not accompanied by proportional improvements in production or revenue. This finding is operationally plausible because oil and gas production and processing are highly energy-intensive activities, and excessive energy use directly affects operating costs and indirectly contributes to environmental emissions. The importance of computational and physical energy efficiency has also been emphasized in recent benchmarking research. Comparative analyses of supervised machine-learning models have shown that energy consumption should itself be treated as a performance criterion alongside predictive accuracy (Ali et al., 2026). Resource-aware artificial intelligence research similarly demonstrates that computational effectiveness depends on balancing predictive capability with resource consumption (Zewail & Mokhtar, 2026). Although these studies concern computational rather than industrial energy use, they support the more general principle that efficiency requires achieving output gains without disproportionate resource expenditure.

Main production ranked as the second most influential SHAP variable, accounting for 16.4% of total explanatory importance, and higher production generally contributed positively to efficiency predictions. Revenue ranked fourth and also exerted a predominantly positive effect. These findings confirm that productive and economic outputs remain central to efficiency, but their influence must be interpreted relative to the resources consumed to generate them. This balance between output maximization and input restraint is consistent with the logic of intelligent optimization and resource allocation in modern production systems. AI-based smart manufacturing research emphasizes the simultaneous optimization of productivity and resource efficiency rather than the maximization of throughput alone (Champa & Segall, 2026). Similarly, intelligent decision systems have been used in complex financial and operational environments to integrate multiple criteria and improve the quality of decision-making under uncertainty (Álvarez-Díez et al., 2025). Therefore, the positive effects of production and revenue observed in the present study should not be interpreted independently from the negative effects associated with excessive resource use and environmental burden.

CO₂ emissions emerged as the third most important predictor, with a relative contribution of 14.1%, and higher

values were generally associated with lower efficiency. NO_x, SO₂, PM₁₀, and COD also showed predominantly negative relationships with predicted efficiency. This finding demonstrates that incorporating undesirable outputs materially changed the interpretation of operational performance. A unit could not achieve high efficiency solely through production expansion if the associated environmental costs increased disproportionately. This is an important methodological advantage because conventional performance indicators may classify high-output units as successful even when their resource intensity and emissions are excessive. The present framework therefore embeds environmental responsibility within operational efficiency rather than treating it as an independent compliance issue.

The relevance of environmental and contextual variables to machine-learning performance is increasingly recognized in other domains as well. Climate-risk benchmarking has shown that predictive models can improve decision quality when environmental risk variables are explicitly incorporated into financial risk evaluation (Nimmala, 2023). The broader movement toward robust and deployable AI likewise stresses that real-world systems must account for contextual constraints rather than optimize narrowly defined predictive objectives (Rharif et al., 2026). Within oil and gas operations, the environmental indicators examined in the present study act as performance penalties when elevated emission levels are not justified by proportional desirable outputs. This feature makes the integrated framework more compatible with contemporary sustainability-oriented performance management.

Labor represented the fifth most important SHAP predictor and showed a predominantly negative effect when labor utilization was disproportionately high relative to production and revenue. This result should not be interpreted as suggesting that a smaller workforce is inherently preferable. Rather, the model indicates that labor efficiency depends on the relationship between person-hours and generated outputs. Excessive labor requirements may reflect inefficient work organization, equipment problems, high maintenance intensity, process complexity, or avoidable operational downtime. Machine-learning-based process analysis increasingly recognizes the value of integrating operational and user-level information to understand process inefficiencies more accurately (Ahmed & Shahzad, 2022). The present result similarly suggests that human-resource indicators should be interpreted within the broader production process rather than considered in isolation.

Feedstock demonstrated a nonlinear and mixed relationship with efficiency. Higher feedstock use was not necessarily detrimental because increased material input can support higher production and revenue. However, excessive feedstock consumption without corresponding output gains reduced predicted efficiency. This nonlinear pattern helps explain why XGBoost substantially outperformed linear regression. The same general issue appears in domain-adaptation and scientific benchmarking studies, where the relationship between model inputs and performance is context-dependent rather than uniformly linear (Lee & Choi, 2026; Wan et al., 2026). High-fidelity surrogate modeling research likewise demonstrates that complex physical systems often require nonlinear models capable of capturing interactions among multiple input conditions (Elrefaie, 2025). The mixed SHAP effect of feedstock therefore reflects the physical reality that input scale contributes to output only when conversion efficiency remains favorable.

One of the most important contributions of the present study is the interpretability achieved through SHAP. The findings showed not only which variables were globally important but also how they influenced the predictions of individual DMUs. This is particularly valuable because two operational units with similar DEA scores can exhibit completely different inefficiency structures. A unit may be inefficient because of excessive energy use, while another may experience low efficiency primarily because of environmental emissions, labor intensity, or inadequate desirable output. Explainable AI therefore transforms a predictive model from a black-box scoring mechanism into a diagnostic decision-support tool. Recent work on representational vigilance has emphasized that predictive success alone may conceal explanatory insufficiency and that learned representations should be examined to determine whether they support meaningful interpretation (Raynal et al., 2026). Similar concerns have been raised in deployable AI for industrial environments, where model explainability is identified as a prerequisite for operational trust and implementation (Rharif et al., 2026). Human-centered intelligent decision-support systems also prioritize interpretability because recommendations are more useful when decision-makers can understand the causal or contributory structure underlying model outputs (Zakaria, 2026).

The explainability component is especially relevant for oil and gas operations because managerial interventions often entail substantial financial, technical, and environmental consequences. An opaque model that merely

labels a unit as inefficient offers limited guidance regarding which resource or process should be targeted. By contrast, the integrated DEA–XGBoost–SHAP framework allows the efficiency deficit to be linked with specific operational variables. The broader development of tool-enabled analytical agents and interactive tabular-data systems similarly reflects a shift toward artificial intelligence that supports human inquiry and interpretation rather than delivering unexplained predictions (Yang et al., 2025). In industrial environments, this distinction is critical because decision support must remain understandable to engineers and managers responsible for implementing corrective actions.

The results also highlight the importance of robust data architecture. The predictive validity of the integrated framework depends directly on the completeness, consistency, and comparability of the operational data used to estimate efficiency. Oil and gas machine-learning research has demonstrated substantial heterogeneity in public datasets, including differences in data quality, structure, and suitability for specific modeling tasks (Baquer, 2026). Missing-data research has similarly shown that structured analysis of missingness is essential for avoiding biased predictions and maintaining model reliability (Jung & Yoo, 2025). The present study addressed these concerns through data checking, preprocessing, normalization where required, and cross-validation. Nevertheless, the findings reinforce the need for organizations to establish reliable and standardized operational data systems if integrated AI-based performance management is to become sustainable.

Finally, the findings support the broader transition toward resource-aware, benchmark-driven, and accountable artificial intelligence. Automated machine-learning deployment and benchmarking research has shown that reproducibility and performance comparison require standardized evaluation procedures (Rahaman, 2025), while IoT platform benchmarking has demonstrated the value of systematic AI assessment in distributed technological environments (Crăciun et al., 2024). Industrial cybersecurity research similarly emphasizes that deployable AI should be evaluated in relation to real operational constraints rather than predictive accuracy alone (Singhal, 2026). The increasing importance of privacy-preserving and adversarially robust machine learning further shows that trustworthy AI must satisfy multiple dimensions of performance simultaneously (Patel, 2026; Rokade, 2024). Within this broader methodological context, the present study demonstrates that a hybrid DEA–machine learning–

XAI framework can provide a transparent and multidimensional foundation for operational performance evaluation in the oil and gas sector.

Several limitations should be considered when interpreting the findings. First, the analysis was based on 30 decision-making units observed for a single operational period, which limits the ability to examine temporal changes in efficiency or distinguish persistent inefficiency from short-term fluctuations caused by maintenance shutdowns, unexpected production interruptions, market conditions, or temporary operational constraints. Second, although the selected indicators represented important production, financial, resource-consumption, and environmental dimensions, operational performance in the oil and gas industry may also be influenced by variables such as asset age, reservoir characteristics, equipment reliability, maintenance expenditure, technological maturity, process complexity, safety incidents, unplanned downtime, and geographic conditions. Third, DEA efficiency scores are relative to the units and variables included in the model, meaning that a change in the composition of the DMUs or performance indicators could alter the estimated frontier. Fourth, the relatively limited number of observations increases the importance of cautious machine-learning model validation even though cross-validation was used to reduce overfitting risk. Finally, the study relied on organizational records, and any measurement errors or inconsistencies in underlying operational and environmental databases may affect both the DEA scores and the predictive model.

Future studies should extend the framework by using panel or longitudinal data covering multiple operational years so that changes in productivity and efficiency can be investigated dynamically. Incorporating Malmquist productivity indices or other dynamic DEA approaches could help distinguish technological change from changes in technical efficiency. Further research could also include additional operational indicators such as equipment availability, unplanned downtime, maintenance costs, process safety performance, reservoir pressure, asset age, and production losses. Larger datasets would enable more extensive comparison of machine-learning methods, including deep learning, stacking, and hybrid ensemble architectures. Future research could also compare alternative treatments of undesirable outputs and investigate robust, stochastic, network, or dynamic DEA specifications. External validation in other oil and gas companies or across different national operating environments would provide

important evidence regarding generalizability. In addition, future studies should examine causal relationships rather than relying exclusively on predictive associations, particularly when the objective is to determine whether changing a specific operational factor will produce a measurable improvement in efficiency.

Managers should use the proposed framework as a diagnostic and benchmarking system rather than as a simple ranking instrument. Efficient frontier units can be used as operational references for identifying realistic resource-consumption and production targets, while SHAP explanations can help managers determine which interventions are most relevant for each inefficient unit. Priority should be given to improving energy efficiency because energy consumption emerged as the most influential determinant of efficiency, particularly in units where high energy use was not matched by proportional production gains. Units with high carbon and pollutant intensity should integrate emission-reduction programs with process optimization rather than treating environmental performance as a separate compliance activity. Labor, feedstock, and water indicators should also be monitored in relation to output rather than through absolute consumption targets. Finally, companies should establish standardized digital data infrastructures capable of continuously integrating operational, financial, production, maintenance, and environmental information so that DEA efficiency scores, predictive models, and explainability analyses can be updated periodically and incorporated into routine managerial decision-making.

Authors' Contributions

Authors contributed equally to this article.

Declaration

In order to correct and improve the academic writing of our paper, we have used the language model ChatGPT.

Transparency Statement

Data are available for research purposes upon reasonable request to the corresponding author.

Acknowledgments

We would like to express our gratitude to all individuals helped us to do the project.

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethics Considerations

In this research, ethical standards including obtaining informed consent, ensuring privacy and confidentiality were considered.

References

- Agrawal, R. (2026). BQEB ForecastBench: Benchmarking AI Models for Smart Grid Forecasting Using BQEB-Data V1. <https://doi.org/10.21203/rs.3.rs-10484554/v1>
- Ahmed, S., & Shahzad, K. (2022). Augmenting Business Process Model Elements With End-User Feedback. *IEEE Access*, 10, 115635-115651. <https://doi.org/10.1109/access.2022.3216418>
- Ajayi, O. O., Kurien, A., Djouani, K., & Dieng, L. (2025). A Proactive Predictive Model for Machine Failure Forecasting. *Machines*, 13(8), 663. <https://doi.org/10.3390/machines13080663>
- Akram, K., Bhutta, M. U., Butt, S. I., Rizwan, M., Khan, D. M. B., Khan, M., & Khan, A. (2026). A Novel Multi-Objective Dynamic Flexible Job Shop Scheduling Algorithm Using Reinforced Learning Based Black Widow Spider Algorithm. *PLoS One*, 21(4), e0347108. <https://doi.org/10.1371/journal.pone.0347108>
- Ali, A., Qamar, R., Asif, R., & Hina, S. (2026). Benchmarking Energy Efficiency of Supervised Machine Learning Models on Multi-Domain Classification Datasets. *Information*, 17(7), 652. <https://doi.org/10.3390/info17070652>
- Álvarez-Díez, S., Baixauli-Soler, J. S., & Kondratenko, A. (2025). Evaluating Expert Decision Systems for Exchange Rate Insurance. *International Journal of Business Analytics*, 12(1), 1-25. <https://doi.org/10.4018/ijban.378390>
- Baqer, M. (2026). A Machine Learning-Centric Taxonomy and Structured Characterization of Public Datasets for Upstream Oil and Gas. *Big Data and Cognitive Computing*, 10(6), 188. <https://doi.org/10.3390/bdcc10060188>
- Champa, S. S., & Segall, R. S. (2026). Artificial Intelligence and Machine Learning Framework for Smart Manufacturing Optimization. *International Journal of Artificial Intelligence*, 2(1), 1-46. <https://doi.org/10.4018/ijaibm.410302>
- Crăciun, R.-A., Pietraru, R. N., & Moisescu, M. A. (2024). Internet of Things Platform Benchmark: An Artificial Intelligence Assessment. *Revue Roumaine Des Sciences Techniques — Série Électrotechnique Et Énergétique*, 69(1), 97-102. <https://doi.org/10.59277/rst-ee.2024.1.17>
- Elrefaie, M. (2025). CarBench: A Comprehensive Benchmark for Neural Surrogates on High-Fidelity 3D Car Aerodynamics. <https://doi.org/10.48550/arxiv.2512.07847>
- Forniés-Tabuenca, D., Uribe, A., Otamendi, U., Artetxe, A., Rivera, J. C., & Lacalle, O. L. d. (2026). REMoH: A Reflective Evolution of Multi-Objective Heuristics Approach

- via Large Language Models. <https://doi.org/10.21203/rs.3.rs-8837965/v1>
- Ibrahim, M. E. A., Ahmed, A. E. S., & Daadaa, Y. (2026). Beyond Neural Solvers: A Critical Review of Machine Learning for Combinatorial Optimization. *Mathematics*, 14(12), 2208. <https://doi.org/10.3390/math14122208>
- Jung, H., & Yoo, P. D. (2025). SMART: Structured Missingness Analysis and Reconstruction Technique for Credit Scoring. *Scientific reports*, 15(1). <https://doi.org/10.1038/s41598-025-99997-4>
- Lee, D., & Choi, B.-S. (2026). A Structure-Based Benchmarking Suite for Quantum Learning. *Concurrency and Computation Practice and Experience*, 38(10). <https://doi.org/10.1002/cpe.70758>
- Mannan, K. A. (2026). Comparative Evaluation of Traditional Machine Learning and Transformer Models for Fake News Detection. <https://doi.org/10.21203/rs.3.rs-10953606/v1>
- Nimmala, R. (2023). Enhancing Financial Risk Management: Utilizing Machine Learning in Climate Risk Model Benchmarking. *Journal of Mathematical & Computer Applications*, 2(1), 1. [https://doi.org/10.47363/jmca/2023\(2\)146](https://doi.org/10.47363/jmca/2023(2)146)
- Orazov, B., Nobatov, A., & Mamiyev, A. (2026). Application of Machine Learning Models for Predictive Maintenance in Industrial Production Systems. 10. <https://doi.org/10.1117/12.3109286>
- Patel, D. (2026). Adversarial Intelligence: A Systematic Review of AI-Driven Cyber Threat Detection, Adaptive Malware Evasion, and Arms Race Dynamics in Zero Trust Security Environments. <https://doi.org/10.21203/rs.3.rs-10137205/v1>
- Rahaman, R. (2025). Automated Deployment and Performance Benchmarking of Machine Learning Workloads on Hadoop Clusters Using Ansible. <https://doi.org/10.21203/rs.3.rs-7003490/v1>
- Raynal, J., Slangen, P., Raynal, E., & Margerit, J. (2026). Detecting Explanatory Insufficiency in Learned Representations: A Framework for Representational Vigilance. <https://doi.org/10.48550/arxiv.2606.13172>
- Rharif, A., Charafi, Z., Zaydi, M., Belkhala, S., & Maleh, Y. (2026). Deployable AI for IoT/IIoT Security: A Systematic Review of Labeling, Transfer, Resource, and Explainability Constraints. <https://doi.org/10.21203/rs.3.rs-9929399/v1>
- Rokade, M. D. (2024). Advancements in Privacy-Preserving Techniques for Federated Learning: A Machine Learning Perspective. *Journal of Electrical Systems*, 20(2s), 1075-1088. <https://doi.org/10.52783/jes.1754>
- Scheinert, D., Becker, S., Bader, J., Thamsen, L., Will, J., & Kao, O. (2022). Perona: Robust Infrastructure Fingerprinting for Resource-Efficient Big Data Analytics. 209-216. <https://doi.org/10.1109/bigdata55660.2022.10020860>
- Singhal, S. (2026). AI-driven Cybersecurity for Industrial Internet of Things: Architectures, Challenges, Datasets, and Future Research Directions. *Frontiers in Big Data*, 9. <https://doi.org/10.3389/fdata.2026.1938279>
- Sucuoğlu, B. Y., Beyca, Ö. F., & Kosanoglu, F. (2026). Forecasting Intermittent Sales in Fashion Retail: A Two-Stage Machine Learning Approach. *Forecasting*, 8(4), 56. <https://doi.org/10.3390/forecast8040056>
- Wan, J., Yar, K. P., Low, M. Y. H., Xu, C., Doan, N. C. N., Ng, H. Y., & Wang, W. (2026). PI-FSL: Physics-Informed Few-Shot Domain Adaptation for Robust Cross-Domain Condition Monitoring. *Technologies*, 14(3), 167. <https://doi.org/10.3390/technologies14030167>
- Wang, P., & Yu, Z. (2023a). RayBench: An Advanced NVIDIA-Centric GPU Rendering Benchmark Suite for Optimal Performance Analysis. *Electronics*, 12(19), 4124. <https://doi.org/10.3390/electronics12194124>
- Wang, P., & Yu, Z. (2023b). RenderBench: The CPU Rendering Benchmark Suite Based on Microarchitecture-Independent Characteristics. *Electronics*, 12(19), 4153. <https://doi.org/10.3390/electronics12194153>
- Yang, A. I., Woo, J., Zhang, R., Mach, A., Ramkumar, P. N., & Ma, Y. (2025). Tool-Wielding Language-Model-Based Agent Offers Conversational Exploration of Clinical Tabular Data. <https://doi.org/10.64898/2025.12.01.25341392>
- Zakaria, A. (2026). A Human-Centric Fuzzy Decision Support System for Medical Diagnosis Using Fuzzy Cognitive Maps. *Scientific reports*, 16(1). <https://doi.org/10.1038/s41598-026-51590-z>
- Zewail, R., & Mokhtar, B. (2026). Resource-Aware Contrastive Scattering Meta-Learning for Efficient Few-Shot Acoustic Anomaly Detection. *Advanced Intelligent Systems*, 8(9). <https://doi.org/10.1002/aisy.202501454>
- Zou, Q. (2025). FML-bench: Benchmarking Machine Learning Agents for Scientific Research. <https://doi.org/10.48550/arxiv.2510.10472>