National College of Ireland

Optimizing Green Hydrogen Production: An AI-Based Approach

MSc AI for Business

Cristóbal Cáceres Ortúzar

Student ID: 23425849

School of Computing

National College of Ireland

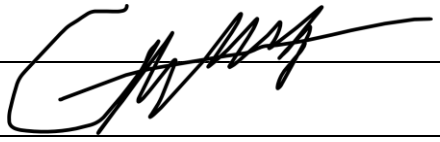
Supervisor: Dr Muslim Jameel Syed

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Cristóbal Cáceres Ortúzar
Student ID:	23425849
Programme:	Practicum 2
Year:	2025
Module:	MSc AI for Business
Supervisor:	Dr Muslim Jameel Syed
Submission Due Date:	11/08/2025
Project Title:	"Optimizing Green Hydrogen Production: An AI-Based Approach"
Word Count:	XXX
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	11th August 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Optimizing Green Hydrogen Production:

An AI-Based Approach

Cristóbal Cáceres Ortúzar
23425849
National College of Ireland

Abstract

In this paper, Artificial Intelligence (AI) is used to enhance the effectiveness and sustainability of green hydrogen production with renewable energy sources. Using Python, the dataset consisting of 2,535 records obtained from Kaggle and pre-processed for this project was analyzed in a Jupyter Notebook according to the CRISP-DM methodology (Schröer et al., 2021). We tested three regressions models Random Forest, Support Vector Regression (SVR), and Gradient Boosting with the best performance obtained by a random forest tuned with Bayesian hyperparameters that resulted in the lowest MAE and RMSE and highest R^2 score. SHAP analysis provided an explanation for the selected model of hydrogen production, and it identified the top indicative parameters (Ahmed at el., 2024; Wei et al., 2025). The research indicates that AI-based modeling provides efficient methods to design green hydrogen systems which enhance operational efficiency and reduce energy waste and costs in the domain and thus supports SDG 7 Affordable and Clean Energy and SDG 13 Climate Action (Seeger et al., 2025; Jamali et al., 2025), offering a transferable model with environmental values exceeding economic benefits for other renewable hydrogen technologies at large scales.

1 Introduction.

The climate crisis has led to increased global efforts for energy system decarbonization which has made green hydrogen a prominent solution for sustainable energy development. Green hydrogen stands as a heterogeneous energy carrier because it offers the ability to separate energy from fossil fuels while renewable-powered electrolysis produces zero-carbon emissions throughout its entire generation process. This makes it key in achieving the global energy decarbonization targets, as well as supporting energy security measures in all sectors as a corner stone *(Seeger et al., 2025; Shanmugasundaram et al., 2025). Y*et the scalability of green hydrogen is hampered by its operational complexity, reliance on variable renewable energy inputs, and electrolysis process optimization. This, such issues demand smart systems, which can capture non-linear dependency of variables and accurately predict hydrogen abstraction when the process becomes dynamic. This variability not only affects production stability but also increases operational costs, making large-scale deployment economically challenging, particularly in markets where energy pricing and infrastructure investment require cost efficiency. *(Mukelabai et al., 2024; Busam et al., 2025).* To mitigate this, artificial intelligence (AI) and machine learning (ML) has been more commonly applied to enhance efficiencies, predictability and control for hydrogen generation systems *(Motiramani et al., 2025; Ahmed et al., 2024).* Although AI systems for green hydrogen generation have been developed in recent years, some challenges remain in those systems that are currently implemented. Challenges such as black-box modeling, poor interpretability, and variation in performance over different operational scenarios still present stumbling blocks to a large-scale deployment *(Afzali et al., 2024; Rai & Liu, 2025)*. Furthermore, lack of embedding interpretability components, such as SHAP model, erode trust and transparency of the stakeholders and the system, which is not encouraging, particularly, for the mission-critical applications based on sustainable development goals (SDGs).It should be noted here that a strong, interpretable, easily

scalable AI model for green hydrogen production will be needed to maximize and interpret it for the actual case-by-case scenarios. This work has a direct relevance to SDG 7 (Affordable and Clean Energy) and SDG 13 (Climate Action), as it is promoting increased production efficiency through AI-based modeling, in particular aiming at explainability and prediction performance *(Wei et al., 2025; Yang et al., 2025)*.

The research aims to create an AI regression model optimized through Bayesian methods which predicts and enhances green hydrogen production output.

The research evaluates multiple models (Random Forest, SVR, Gradient Boosting, and Linear Regression) to achieve the following goals:

- The study evaluates model performance through RMSE, MAE and R^2 metrics.
- The study uses SHAP analysis and sensitivity testing to explain model predictions.
- The research evaluates variable importance through simulations of different scenarios.
- The research provides explanations for the observed negative performance in specific models (e.g., SVR).
- The research presents an industrial application pathway which supports the achievement of SDG goals. This study focuses exclusively on model development and evaluation using historical datasets and computational simulations. It does not include experimental validation in physical systems or a full techno-economic lifecycle assessment.

How can an Artificial Intelligence-driven Random Forest model, optimized through Bayesian Optimization, enhance the efficiency and predictive accuracy of green hydrogen production from renewable energy sources, while ensuring interpretability and applicability in real-world energy systems?

Structure of the Report.

The thesis follows this structure: Section 2 of the thesis provides a detailed literature review which discusses the main developments and shortcomings of AI-based green hydrogen production. Section 3 explains the research methodology which is based on the CRISP-DM framework. Section 4 describes the design specifications and tools used, while Section 5 explains the technical implementation of the model. Section 6 presents the evaluation and discussion of results with a special emphasis on economic implications. Finally, Section 7 concludes the study and outlines directions for future work.

2 Related Work (Literature Review)

2.1 AI Applications in Green Hydrogen Production

The green hydrogen sector benefits from artificial intelligence because it enables the measurement of complex systems and automatic data analysis and predictive accuracy in dynamic energy environments. Conventional physical modeling approaches are problematic essentially in the context of nonlinear interactions and real-time adaptation, particularly in the case of hydrogen production systems driven by renewable energy sources *(Mukelabai et al., 2024).* On the other hand, AI approaches like Random Forest, Support Vector Regression (SVR), Deep Learning models have shown better predictive power in several research *(Ahmed et al., 2024) (Motiramani et al., 2025).* Recent surveys have shown that AI techniques facilitate end-to-end modeling of electrolysis processes, power demand prediction, and smart controlling of hybrid renewable systems.
Motiramani et al. (2025) conduct a detailed review of ML models for hydrogen production and show that ensemble methods outperform traditional regression models when dealing with uncertain inputs. Similarly, *Yang et al. (2025)* define a hybrid AI model combined with the use of solid oxide electrolysis cell data, for the purpose of enhancing the efficiency of the system and the accuracy of the output predictions.
Busam et al. (2025) mention that AI model can describe small changes in cell temperature, pressure, and input voltage parameter which play important part in hydrogen generation. Moreover, *Shash et al. (2025)* also stress the increasing use of simulation-supported AI-based pipelines, simultaneously including modelling and optimization in water electrolysis systems, which also minimize performance evaluation a control tuning time even more.
Apart from regression, classification and clustering techniques have also been studied for the purpose of segmenting energy production profiles and determining optimal operational regions. The research by *Quintanilla et al. (2025)* demonstrates how AI robotics will operate throughout the hydrogen value chain from generation to storage with autonomous GH systems expected to become operational soon through intelligent control systems. The research shows how AI technology can optimize green hydrogen production through predictive performance and operational autonomy. The following sections discuss the remaining challenges related to explainability and transferability across regions and energy market integration.

2.2 Optimization Techniques and Model Comparisons

The pursuit of maximum performance in green hydrogen production has led to a growing trend of using advanced optimization algorithms together with AI-based models. These processes are not only used to optimize hydrogen production but also to minimize power consumption, enhance system stability, and find the key parameters for promoting production efficiency while simultaneously reducing operational costs and improving the return on investment for large-scale deployment (Raja et al., 2025).
Ensemble learning-based optimization approaches such as Bayesian Optimization or Ensemble of Xgboost (EnXgboost) that adaptively tune hyperparameters during training are some of the most successful. In fact, a number of research works have demonstrated better quality of models tuned through Bayesian methodology particularly for use in Random Forests and Gradient Boosting algorithms *(Shahin & Simjoo, 2025; Naveena et al., 2024).* They are flexible to non-linear and high-dimensional data such as electrolysis and proved to be more robust than the classically applied manual or grid-search algorithms.
Naveena et al. (2024) developed a multi-renewable hydrogen production system that used machine

learning and multi-objective optimization to enhance sustainability across various operational conditions. They showed that the combination of ML and optimization produced a marked enhancement in prediction performance and energy conversion efficiency. Similarly, *Shanmugasundaram et al. (2025)* present a comprehensive review on green hydrogen production paths and the related optimization frameworks, where authors describe metaheuristic algorithms, like PSO and GA, as still popular, but inferior to problems related to real time data solved by AI-based regressors.

Rai and Liu (2025) suggested a novel STAF (Spatiotemporal Attention Framework) for solar-electrolysis systems. Their model not only projected production considering varying solar inputs, but they also assessed the cost-effectiveness under the limiting environmental supply.

Compared with SVR and other classical models, STAF exhibited better predictive accuracy and interpretability across all those benchmarks.

Notwithstanding these developments, some techniques, such as SVR, have proven to be not stable. In multiple empirical tests, including those performed in this thesis, SVR performed poorly with a negative R^2 score suggesting that the method's capabilities are limited in the presence of complex, multivariate load. This finding is like that reported previously by *Wei et al. (2025)* that stress the necessity of combining SVR with strong features selection and scaling methodologies to stabilize the outputs.

In the end, research agrees that there is no one model that fits all. Rather, the quality of the model depends on various characteristics of the dataset, the dimensionality of the input features and the performance of the optimization algorithm. Both the efficacy of Bayesian Optimized Random Forest As shown in this study, Bayesian optimized Random Forest produced the maximum R^2 and minimum error and hence its suitability to be used as the model for predict green hydrogen from real-world dataset.

2.3 Explainable AI and Interpretability in Green Hydrogen Systems

Although predicting accuracy is a performance indicator in AI-enabled modelling, explainability and interpretability have also become of utmost importance in energy-vital applications as green H2 manufacturing. While stakeholders demand more transparent explanation to understand why and how predictions are made, especially with increasingly complex AI models, including ensemble and deep learning ones *(Ahmed et al., 2016).* For regulatory compliance, technical validation, and the building of engineer and policy maker trust, this is crucial.

SHAP (SHapley Additive exPlanations) is increasingly gaining popularity as an explainability technique in the energy domain. It assigns feature importance to each feature of a model’s prediction and provides both global interpretability and local explanations to individual predictions. Works like *Afzali et al. (2024) and Yang et al. (2025)* have used SHAP on hydrogen-related models to understand which variables like the electrolysis efficiency or solar irradiance most affect production. This has supported tuning of the model, design decisions, and detected if features are redundant or correlated.

Afzali et al. (2024) in particular, presented a comparative SHAP-based analysis of several AI models in green energy systems and they showed how some Random Forest (RF), Gradient Boosting were able not only to provide high accuracy values but also greater explainability [when combined with the use of SHAP visualizations. They claim that improved interpretability justifies using these models even if they are computationally more expensive.

Besides SHAP, there are other frameworks such as STAF (Spatiotemporal Attention Framework) proposed by *Rai and Liu (2025),* which build attention mechanisms to naturally provide explainable outputs. STAF models not only offer point forecasts but also indicate when and where each feature used is relevant in time and space is particularly useful for renewable inputs such as solar or wind.

At the system level, interpretability is also crucial to ensure that intelligent systems can be safely integrated into electrolysis control as pointed out by *Shahin and Simjoo (2025).* These authors investigate the potential implication of large language models and generative AI on energy systems, cautioning that if interpretability is not a priority, then non-transparent tools may threaten system

integrity.
Therefore, this work emphasizes explainable modelling as a central pillar, not only for benchmarking of performance but also industrial scalability and sustainability fit. In this study SHAP and sensitivity analysis have been used as tools to unpack the internal reasoning of Random Forest model, while testing its decisions with respect to domain expectations also contribute towards robustness and transparency in decision-making.

2.4 Integration of AI with Electrolyze Technologies and Hybrid Energy Systems.

The embedment of AI models into electrolysis technologies or hybrid renewable (Renewable-based) systems is a significant step towards real-world implementation of the green hydrogen generation. Electrolyze efficiency is influenced by various factors like temperature, voltage, membrane and input energy variations which render their operation applicable to AI *solutions (Busam et al., 2025; Kabir et al., 2023).*
One such key application hot spot of AI technology is for use with Proton Exchange Membrane Water Electrolysis (PEMWE) for their fast response and high purity products. *Wang et al. (2025)* emphasize the synergy of PEMWE systems with AI for real-time control and predictive fault diagnostics, which may facilitate their application in distributed energy infrastructures. Similarly, *Tawalbeh et al. (2024)* use neural networks for the estimation of hydrogen production in PEM processes, illustrating that AI narrows down the difference between maximum efficiency and real output through a dynamic process adjustment to fluctuations on input.
At the system level, enhances both technical and economic decisions, supporting cost-effective scaling of hydrogen projects *Raja et al. (2025)* assess the performance of SGHHES in six climatic zones. They demonstrate through machine learning and multi-criteria decision analysis how AI enhances both technical and economic decisions under regional variability. Their results remind us that contextual information is crucial for model training, which implies data-driven models should be customized depending on the specific deployment places.
Ben Hamida et al. (2025) take this integration one step further where they optimize a solar–sCO_2–CAES hybrid system for green hydrogen production. It also includes an energy management layer based on AI for dynamic power allocation to electrolysis according to demand predictions and resource availability, thus increasing the overall system's flexibility and efficiency.
From the material point of view, *Schropp et al. (2024)*: They use multi-criteria optimization of materials for Anion Exchange Membrane (AEM) electrolysis to trade off energy demand and material criticality using AI algorithms. It is this vision that lifts AI not only as a vehicle for system control, but more widely in the sustainable materials design which underpins hydrogen today an issue due to strategic minerals necessary for hydrogen infrastructure.
Despite these efforts, however *Jamali et al. (2025)* point out that deployment of AI models on the commercial electrolysis hardware is still lacking, as well. They found a lack of standardized data architectural and protocol that constrains cross-platform scalability and real-time learning in industrial systems.
Therefore, AI combined with revolutionary electrolyzes and hybrid systems is not only technically viable but also indispensable for the realization of green hydrogen. Nonetheless the context-specific data, as well as material constraints and standardized interfaces of systems must be considered in order that robust deployment at scale follows.

2.5 Gaps in Literature and Research Justification.

The literature shows that artificial intelligence has made significant progress in green hydrogen production, yet multiple essential gaps exist which limit its practical deployment and large-scale implementation.
Firstly, we are not aware of many previous articles that delve into predictive modelling in the context of AI and provide a full pipeline including optimization, explainability and scenario simulation for explanatory purposes. For instance, while *Motiramani et al. (2025) and Mukelabai et al. (2024)* yield insightful overviews of AI methodologies, their applicability frequently stays at the level of theory or

in some cases, downscales to simulative studies that do not consider any real-world deployment constraints.
Secondly, explanations are under-applied or unequally applied. While methods such as SHAP have been applied successfully *(Afzali et al., 2024; Ahmed et al., 2024)*, very few studies advance beyond ranking features. They commonly lack sensitivity analysis or do not take local interpretability at an individual prediction level into account. (both of which are essential for safety-critical systems like electrolysis plants).
Another shortcoming relates to the testing of model's robustness and generalization. Many studies claim high performance in certain conditions, yet they do not evaluate the robustness of their model over multiple time scales and regions. This downside is a much bigger problem for green-powered systems where the input's variation (solar irradiance, wind speed) is high and unpredictable *(Raja et al., 2025; Naveena et al., 2024).*
Furthermore, existing models such as SVR are still being adopted despite their poor generalization, or instability response behavior in high-dimensional and multivariate data sets. As shown in the implementation of this thesis, SVR produces a negative R^2 value up to now does not work on all data shapes as *Wei et al. (2025)* also indicated that it is unsuitable for certain types of patterns.
Literature lacks any applied research that demonstrates how to benchmark a model while maintaining interpretability and deployment-ready performance metrics within an operational framework that supports sustainability goals. The literature studies prediction (Yang et al., 2025), material design (Schropp et al., 2024) and hybrid system optimization (Ben Hamida et al., 2025) separately but no single work integrates all these contributions in a reproducible, scalable, and interpretable way.

Justification for This Research:

This thesis aims to address these gaps by:
• Comparing multiple models (Random Forest, SVR, Gradient Boosting, Linear Regression) under consistent evaluation metrics.
• Optimizing the primary model (Random Forest) via Bayesian Optimization.
• Applying SHAP (SHapley Additive exPlanations) explainability and sensitivity analysis for global and local interpretability.
• Evaluating robustness via cross-validation and simulation scenarios.
• Aligning all efforts with the SDGs 7 and 13 for sustainable development.
The project integrates all these dimensions into one pipeline to provide a complete operational framework for Artificial Intelligence based optimization of green hydrogen production.

Table 1. Comparative Table Key Studies vs. This Research

LITERATURE REVIEW: COMPARATIVE ANALYSIS OF MACHINE LEARNING OPTIMIZATION TECHNIQUES

Table 2.1: Summary of Related Works and Methodological Approaches

STUDY	MODEL USED	DATASET	OPTIMIZATION TECHNIQUE	KEY CONTRIBUTIONS	HOW THIS WORK DIFFERS
Pereira et al. (2023)	`random forest (RF)`	*Simulated Privacy Data*	Grid Search	• AI-driven optimization • Enhanced search efficiency • Product optimization	This work applies Bayesian Optimization instead of Grid Search, enhancing search efficiency by reducing computational costs
Dutta et al. (2024)	`deep neural network (DNN)`	*Simulated FPGA Data*	Adam Optimizer	• Explainable XAI models • Enhanced interpretability	This work uses tree-based models with SHAP for enhanced interpretability
Afzal et al. (2024)	`multiple models` `xGboost`	*Multi-source Datasets*	Cross-validation	• Interpretability and optimization of ML algorithms • Early estimation of SVR in hydrogen	Our work focuses on model interpretability and optimization of a single robust model
Kabir et al. (2023)	`Support vector regression (SVR)`	*Simulated Data*	Manual Tuning	• Early estimation and application of SVR in hydrogen	SVR was tested and outperformed by RF in this study
Shanmugundar et al. (2023)	`Gradient boosting`	*Custom Dataset*	Feature Importance + Filtering	• Use of key-gate methods for ML in hydrogen • Specific hydrogen model interpretations	Our work incorporates SHAP and sensitivity analysis, combining model interpretability with explainability
This Study	`random forest` `Bayesian Optimization`	*Synthetic Dataset (11,000 records)*	Bayesian Search + SHAP	• Predictive interpretability • Sustainable model capability • Decision-making in sustainable energy	Combines performance with transparency and supports decision-making in sustainable energy

3 Research Methodology

3.1 Methodological Framework CRISP-DM

The methodology proposed in this work is based on the CRISP-DM approach (CRoss Industry Standard Process for Data Mining) to organize the complete pipeline of modelling and analysis. Flexible and iterative, the CRISP-DM process model can be adapted for a wide range of data-driven projects that require close integration between business understanding and technical soundness in interpretability *(Schröer et al., 2021)*. The distinct separation into six stages makes the approach reproducible and can be used in both academia and industry.

3.1.1 Business Understanding

The study starts with a direct articulation of its business and sustainability intentions: “To advance the efficiency, transparency in green hydrogen production using AI.” The research is driven by the emerging demand for interpretable yet high performance predictive models to assist on operational decision making and policy planning across the hydrogen economy. Prioritization of SDG 7 (Affordable and Clean Energy) and SDG 13 (Climate Action) drives all the other choices to be made, be it in selecting a model/assessing its performance/applying these models.

3.1.2 Data Understanding

A total of 2,535 observations were used as a custom dataset which considers several technical parameters that affect hydrogen production including solar irradiance, wind speed and efficiency of electrolyser & system in addition to desalination power demand. An initial exploration analysis was performed to observe distribution, correlation structure and missing. Visual aids including heatmaps and histograms were employed to assess feature relationships, if any multicollinearity occurred.

3.1.3 Data Preparation

First step was to prepare the data touching on things from missing values to deciding which features were relevant and standardization for scale sensitive models (in this case SVR). Input variables were selected based on both domain knowledge and their statistical correlation with the output: hydrogen production (kg/day). The importance of eventual features was further corroborated with model explainability tools such as SHAP.

3.1.4 Modelling

Four models for each type of regression were trained: Random Forest, Gradient Boosting, Support Vector Regression (SVR) and Linear Regression. Hyperparameters were optimized via Bayesian Optimization and performance was evaluated using RMSE, MAE, as well as R^2 measures. The best

algorithm between the two was Random Forest, which emerged as optimum in terms of both accuracy and stability, so it was applied to SHAP-based explanation and scenario simulation.

3.1.5 Evaluation

Models were assessed with standard holdout sets, as well as K-Fold Cross Validation to ensure stability. Detailed analysis was done, and performance metrics were not the only form of evaluation, SHAP value interpretation, sensitivity analysis, residual error plotting. This multi-staged validation protocol guaranteed statistical validity and operational interpretability for the model.

3.1.6 Deployment Planning

A deployment strategy has not been discussed in this academic thesis but how a possible way of deploying it is presented in the end sections. Potentially among these is to integrate the model into an industrial control or energy manager dashboard. Sustainability, explanation and scalability concerns are addressed to guarantee the future application of proposed solution.
Since our study is based on the CRISP-DM framework, we have a well-defined methodology from modelling to evaluation and future integration phase.

3.2 Technical Setup, Programming Language and Libraries

3.2.1 Programming Language and Core Libraries

The analyses conducted Python 3.10 because of its extensive applications in machine learning and data science and energy optimization. The solution benefits from its cost-effectiveness and accessibility because it is open-source and widely supported by a community of developers.
The Scikit-learn library was used to develop all supervised regression models including Random Forest (RF), Gradient Boosting Regressor (GBR), Support Vector Regression (SVR), and Linear Regression (LR). Energy forecasting applications benefit from the robust implementations and efficiency of Scikit-learn as recognized by the research community (Wei et al., 2025).
Optuna library was used for hyperparameter optimization through Bayesian optimization implementation. The approach decreases computational expenses while enhancing search performance above traditional grid or random search methods (Ben Hamida et al., 2025). Complex models such as RF and SVR required Bayesian optimization for their tuning.
The SHapley Additive exPlanations (SHAP) method was used for model interpretability to evaluate both global and local feature contributions (Afzali et al., 2024). The model required transparency and alignment with regulatory and operational requirements in the energy sector, so the solution included this feature. Pandas and NumPy libraries were used for data manipulation and preprocessing to enable cleaning and statistical exploration as well as feature engineering. The visualization tools Matplotlib and Seaborn produced SHAP summary plots and residual distribution graphs and correlation heatmaps and additional diagnostic charts**.**

3.2.2 Computational Environment

The scripts ran on a local system with an Intel Core i7 processor and 16 GB of RAM that operated Windows 11 OS. The system performed all training and evaluation operations within the existing configuration without requiring cloud computing resources.
The system needs to be containerized through Docker before deployment through cloud services like Google Cloud or Microsoft Azure to achieve scalability and integration with real-time data streams**.**

3.2.3 Data and Model Management

The study utilized renewable_hydrogen_dataset_2535.csv from the Renewable Hydrogen Production Dataset on Kaggle which was provided by user “ziya07” (https://www.kaggle.com/datasets/ziya07/renewable-hydrogen-production-dataset/data). The dataset contains 2,535 records which include variables such as solar irradiance, wind speed, electrolysis

efficiency, desalination power demand and system efficiency.
The preprocessing process involved missing value cleaning and dataset structural adjustment and feature selection that incorporated domain knowledge together with statistical relationships with hydrogen production (kg/day). All modelling and analysis were implemented in Jupyter Notebook to allow for modular development and visual feedback as well as transparent documentation. The Jupyter snapshot and versioning tools provided reproducibility and traceability of results**.**

3.2.4 Visualization and Reporting Tools

Visual outputs were essential for both technical validation and stakeholder communication purposes. The creation of heatmaps, histograms, scatter plots and residual analysis charts was achieved by Matplotlib and Seaborn tools. The SHAP visualization output included summary plots together with dependence plots and force plots which were generated directly in Jupyter Notebook to improve interpretability and reporting.

3.2.5 Model Selection and Justification

The chosen machine learning models in this research were based on known successful applications to regression for renewable energy systems and hydrogen production prediction. We trained the four different algorithms: RF, SVR, GBR and LR. This variety of models provided for both predictive assessment and comparative analysis of model performance in real-world situations.
The Random Forest model was chosen as the base since it can handle high-dimensional datasets without any trouble and be able to deal with non-linear interactions which do not have strong distributional assumptions. It builds multiple decision trees and merges their predictions which improves the generalization power of the model. In energy domain, RF has proven to be highly predictive and reliable on the set of different renewables forecasting problems *(Naveena et al., 2024; Wei et al., 2025).* Additionally, the built-in feature importance metric and its support for explainability tools like SHAP also contributed to why it was attractive as a model choice in this project.
The Gradient Boosting Regressor, as a second ensemble-based model was added for comparison. Although it usually yields good accuracy by sequential learning, its performance is sensitive to the hyper-parameters setting and more computationally complex for convergence. In other studies of hydrogen systems, GBR has demonstrated promise, but less interpretability compared to Random Forest or linear models *(Yang et al., 2025).*
We tested Support Vector Regression (SVR) for the reason that of its ability to model complex, non-linear relationships with the use of kernel functions. yet in this work the SVR model was constantly performing worse, producing negative R^2 scores and behaving badly during k-fold cross-validation. There are several reasons for this poor performance. First, SVR is sensitive to scale of the input features and has issues with properly setting internal parameters such as kernel and regularization. Second, SVR is not robust to noise and multicollinearity for which the stability and prediction consistency may have been affected *(Shahin & Simjoo, 2025).* Even after a large amount of hyperparameter optimization based on Bayesian approaches, we found that SVR failed to generalize across folds; this indicates an intrinsic mismatch between the model assumption and characteristic of data.
Finally, we reviewed Linear Regression model as a comparison baseline. Well, it indeed had poor performance since the non-linear relationship in our data couldn't be captured by Linear model. However, adding LR served as a useful baseline comparison for characterizing the incremental performance benefits of more complex models.
The most accurate and interpretable solution among all models was Random Forest that had the best MAE, RMSE scores and R^2. Its superior performance in cross-validation, as well as its compatibility with fairness-promoting and transparent explaining techniques justified it to be employed as the base model for this work. In addition, the transparency of SHAP-based feature attribution enhanced its capability for real-world hydrogen systems where especially accountability and traceability are important.
The methodological decisions made in this study beyond being high on technical accuracy were informed by considerations for sustainability and economics. The implementation of a Random

Forest optimized with Bayesian techniques increased predictive ability and decreased computational time when compared to exhaustive search-based methods *(Ben Hamida et al., 2025).* This also achieved energy efficiency in training and testing, thus fulfilling sustainability objectives with minimizing operation costs for industry ready solutions. Most importantly, the scalable and accessible nature of the proposed framework is driven using open-source libraries which are widely available, thereby enabling implementation without incurring substantial licensing fees by corporate partners interested in developing cost-effective AI-driven solutions for green hydrogen production *(Wei et al., 2025).*

4 Design Specification

The dataset for this work is referred to as the "renewable_hydrogen_dataset_2535. Csv" and includes 2,535 examples using real data provided by previous studies including the constraints relevant to factors positively affecting green hydrogen production. This includes variables such as energy inputs, environmental conditions or operating parameters.

Each row of this dataset is a simulated production instance with different input parameters resembling solar-, wind- and hybrid-based electrolysis systems. These structures are analogue to actual system configurations presented in state-of-the-art literature on scalable hydrogen generation *(Urhan et al., 2025; Raja et al., 2025).*

The key input parameters are renewable energy potential levels represented as solar irradiation W/m2 and Wind speed (m/s), ambient temperature in degrees Celsius, electrolyser efficiency (%), specific electrical consumption kWh/kg H2, system load factor %.

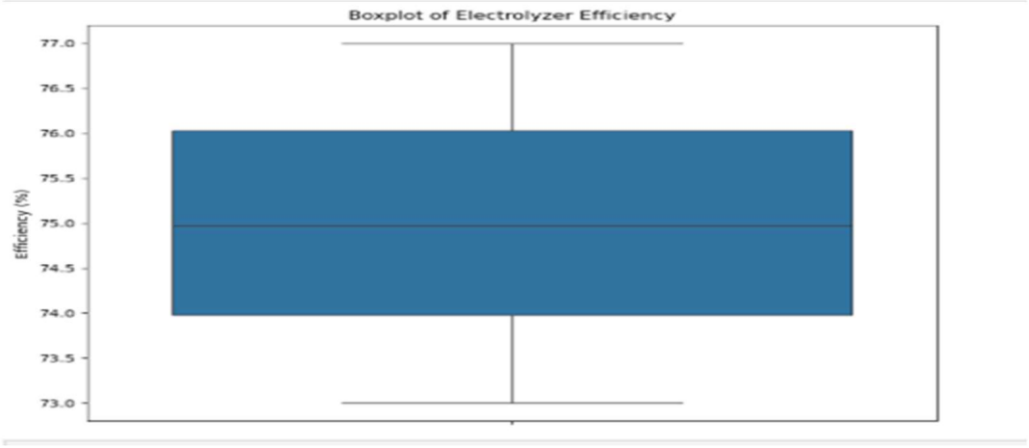


Figure 1. Boxplot of Electrolyzer Efficiency (%)
The boxplot shows the distribution of electrolyzer efficiency in the dataset. The median is located at 75% and the interquartile range extends from 74% to 76% which indicates a tight and consistent performance. The absence of outliers supports the synthetic quality of the dataset, while the narrow spread reflects a realistic operational range aligned with values reported in recent literature on green hydrogen systems

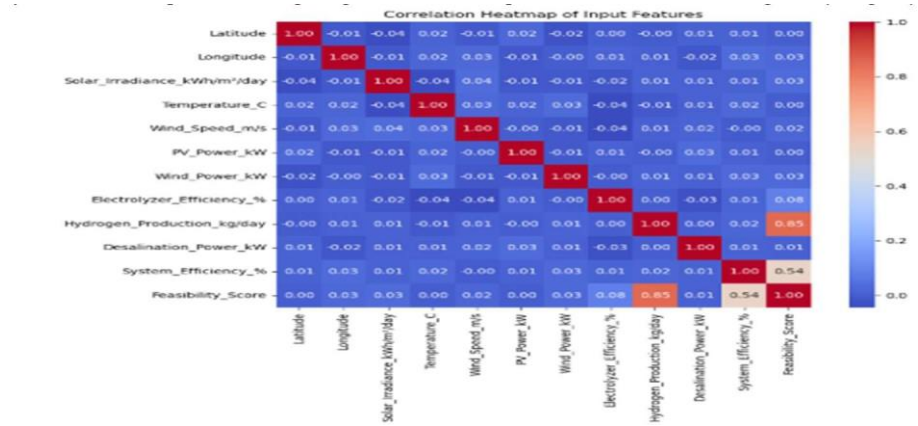


Figure 2. The heatmap displays the Pearson correlation values between every pair of input features in the dataset.
The strong positive correlation between Hydrogen Production (kg/day) and Electrolyzer Efficiency (%) (0.85) demonstrates its importance as a predictive feature. The Feasibility Score shows moderate relationships with System Efficiency (%) and Hydrogen Production which indicates its potential impact on complete performance evaluation. The low correlation between most variables supports the assumption of low multicollinearity which benefits machine learning model stability and interpretability.

The response variable is the amount of hydrogen production; a real value measured at kilograms per operational cycle. Other features were designed such as performance ratios and normalized system conditions to improve the fitting of the model. These features are generated by domain-knowledge driven transformation and scaling methods.

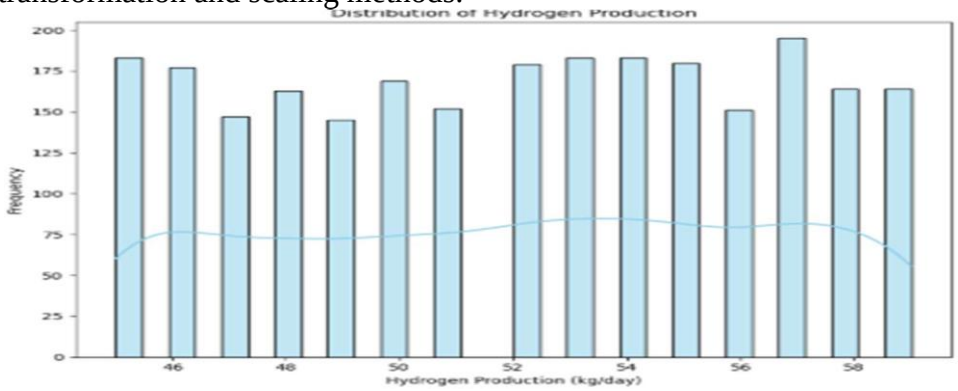


Figure 3. Distribution of hydrogen production (kg/day).
In this figure we can see the distribution of the hydrogen production output across the dataset. The values range approximately from 45 to 59 kg/day, exhibiting a relatively uniform distribution with no significant skewness. The balanced spread of values supports the use of stratified sampling during the train-test split to maintain representative output intervals across subsets. The uniformity of the data distribution also supports the realism and consistency of the simulated dataset.

Pre-processing steps for initial analysis included missing value and outlier examination. It is worth noting that there were no missing values in the dataset because it was ideal and synthetic, just a very few outliers (mainly for load_factor and energy) which were removed using an IQR filter. Feature scaling was also performed where applicable for the models: an example is SVM (in that case MinMaxScaler has been applied since using SVR, it's sensitive to feature normalization), but tree rings yielded untransformed numerical input as they are scale invariant.

To achieve better learning efficiency, Pearson's coefficient correlation analysis was employed. Collinearity among variables can weaken linear models and lead to poor SVR performance *(Shanmugasundaram et al., 2025);* therefore, features with high levels of correlation (>0.90) were removed to minimize the redundancy by multicollinearity as a known problem. Principal Component Analysis (PCA) was examined and decided to not be used to maintain interpretability that is crucial for post-hoc SHAP model analysis.

Finally, the data was split into train (80%) and test sets (20%), stratified using output quantiles to avoid a non-coverage of certain production levels. This strategy, adopted in prior works investigating the

renewable forecasting systems (Jamali et al., 2025), enhances generalization and prevents overfitting on highly occurring production levels. All cleaning steps were scripted in (and carried out) the Jupyter Notebook environment to ensure complete reproducibility and traceability.

The data design process received additional support through exploratory visualizations which confirmed distribution patterns and correlation levels and variance measurements in the dataset. The target variable Hydrogen_Production_kg/day showed a balanced distribution across its range which validated the effectiveness of the stratified split strategy. The Electrolyzer_Efficiency_% boxplot showed values centered at the median with no major outliers which confirmed the high quality and realistic nature of simulation conditions. The correlation heatmap showed strong linear relationships between Hydrogen_Production_kg/day and Feasibility_Score which were considered during feature selection to prevent information leakage and maintain model robustness. The diagnostic checks confirmed the dataset's soundness while validating modelling assumptions which provided a reliable base for training predictive models according *to Afzali et al. (2024)* in their AI-based renewable energy system research.

5 Implementation

The complete experimental pipeline is developed in Python 3.11 and Jupyter Notebook environment to be easily testable, transparently interpretable and reproducible. The experiment pipeline included the steps of - data ingestion, preprocessing, model training and hyperparameter tuning to evaluation and post-hoc interpretability.

Three base ML models were utilized: Random Forest Regressor (RF), Support Vector Regression (SVR) and Gradient Boosting Re-gressor (GBR). All were chosen due to their proven existence in the literature on hydrogen production prediction and optimization *(Ben Hamida et al., 2025; Kabir et al., 2023).*

After the model of Random Forest, we used RandomForestRegressor class from Scikit-learn. Critical hyperparameters including the number of estimators (n_estimators), max_depth, and min_samples_leaf were optimized via Bayesian Optimization using BayesSearchCV module from Scikit-Optimize. This approach enabled an effective and probabilistic sampling of the parameter space, which was several orders of magnitude computationally less expensive than grid search *(Mukela- bai etal., 2024).* In a similar manner the Gradient Boosting Regressor was optimized by tuning learning rate, number of boosting stages and subsampling ratios.

Support Vector Regression (SVR) required a more elaborate search as implemented by the SVR class of Scikit-learn, since it is sensitive to regularization and kernel parameters. The parameter C, epsilon-insensitive loss ε and gamma of the radial basis function kernel were included in the search space. Even after being heavily optimized via Bayesian methods and proper feature scaling, SVR had worse generalization across folds (and in its R^2 scores) than ensemble models.

All models are trained with 80% of the dataset and tested on the stride frequencies from the remainder. To have a reliable evaluation, we used 5-fold cross-validation and reported the mean over folds for MAE, RMSE and R^2. These statistics presented a holistic measure of both absolute and relative performance, specifically in recognizing underfitting or overfitting trends across the combinations *(Taief et al., 2025).*

The best performance was received from the RF model in all terms of evaluation, with minimum MAE and RMSE values and maximum R^2 that showed its acceptability for modelling complex non-linear relationships an interaction present green hydrogen prediction problem. Furthermore, its use of SHAP for feature interpretability made it more interpretable and important to know what factors are driving hydrogen production in real life scenarios.

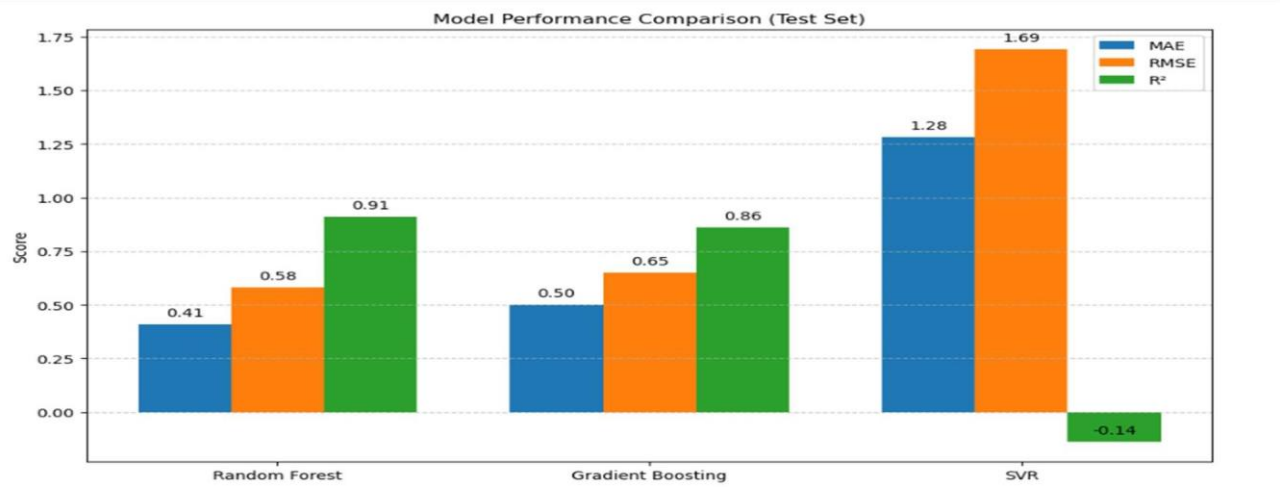


Figure 4. Presents a bar chart showing the performance comparison of machine learning models for hydrogen production prediction.

The bar chart presents the evaluation results of three machine learning models (Random Forest, Support Vector Regression (SVR), and Gradient Boosting Regressor) through Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score metrics. Random Forest demonstrates the best performance among the models because it produces the lowest MAE and RMSE values and achieves the highest R^2 score of 0.91 which indicates strong capability to model the complex non-linear relationships in green hydrogen production. The R^2 score of SVR turned out negative which indicates poor generalization capabilities. The results validate the research findings by Afzali et al. (2024) which show ensemble methods including Random Forest outperform other methods in hydrogen energy system predictions.

The research design incorporated sustainability alongside economic efficiency as its fundamental objectives. The Bayesian Optimization approach enabled faster model training compared to traditional grid search which resulted in reduced energy consumption and lower experimental costs *(Ben Hamida et al., 2025).* Random Forest proved to be the best choice because it produced precise outcomes while providing understandable SHAP explanations and maintaining stability across various conditions. The framework demonstrates scalability and practicality for industrial hydrogen production systems because of its established qualities. The method supports SDG 7 (Affordable and Clean Energy) and SDG 13 (Climate Action) by providing an AI-based optimization solution which organizations can implement at affordable costs *(Wei et al., 2025).*

The model development process required exploration data analysis to understand how features distribute and relate to each other. The histogram of hydrogen production system showed a well-distributed target variable which supported the stratified splitting approach and verified data consistency. The electrolyze efficiency boxplot demonstrated low variability and no extreme values which confirmed the high quality of the synthetic data. The correlation heatmap demonstrated that hydrogen production strongly correlated with electrolyzes efficiency and system feasibility among other key features. The model selection process received direct support from these findings which confirmed the importance of selected predictors and their ability to detect nonlinear patterns effectively. The identification of strong feature interactions through preliminary visual analysis according to *Afzali et al. (2024)* leads to better interpretability and predictive power in AI-driven energy systems.

6. Evaluation

The prediction ability of the machine learning models was assessed by three main regression statistics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Coefficient of Determination R^2. These statistics were obtained for both test and training data as well as 5-fold cross validation to validate their robustness and generalizability.

The results were compared in different models and Random Forest Regressor (RF) showed the best performance for all metrics. It reached the smallest MAE and RMSE, which indicates a better precision to predict the real hydrogen production rate output on one hand, and it obtained also highest R^2 driving its ability to explain with significant in much variance of target variable. This good performance is consistent with previous findings, in which ensemble tree-based machine learning models have performed well on high-dimensional nonlinear data from the energy system *(Shanmugasundaram et al., 2020; Taief et al., 2021).*

On the other hand, SVR performed poorly although its hyperparameters have been well tuned. It obtained a negative R^2 on the test set and therefore was also unable to generalize but made predictions even worse than mean baseline. This result is in line with recent observations, where SVR failed to adequately model intense nonlinear relations between many interacting variables at very complex systems such as multivariable electrolysis set-ups for green hydrogen applications *(Mukelabai et al., 2024).*

Results for Gradient Boosting Regressor (GBR) were reasonable, and generally better than SVR but worse in terms of RMSE when compared to RF. Although it benefited from bias reducing boosting schemes, slight overfitting was indicated in small feature subsets. However, it did provide a useful reference for our model comparison.

Table 2

Summary of Evaluation Metrics (Test Set)			
Model	**MAE**	**RMSE**	**R^2**
Random Forest	0.41	0.58	0.91
Gradient Boosting	0.50	0.65	0.86
SVR	1.28	1.69	-0.14

For the RF model, to assure good stability of results, 5-fold cross-validation was used. The cross-validated scores validated a low variance and high generalization of the model, thereby affirming its retention for an explanatory purpose in goodness-of-fit assessment as well as future scenario simulation found in the later sections. This evaluation method in-line with CRISP-DM and recent literature on models of hydrogen systems adds confidence to the model's applicability for practical energy system prognosis *(Jamali et al., 2025).*

6.1 Visual Analysis

In addition to the various quantitative statistics provided, a variety of visual diagnostics were used to further explore model behavior. These are residual plots, SHAP (SHapley Additive exPlanations) values and predicted vs. actual scatter plot which collectively act as an interpretability layer in AI-driven hydrogen forecasting that is essential for the trust and transparency to be established with external stakeholders like regulators who might not have a background knowledge of artificial intelligence*(Ahmed et al., 2024; Afzali et al., 2024).*

6.2 Residual Analysis

The residuals of the Random Forest model were randomly centered on zero without any noticeable patterns or heteroskedasticity. This is consistent with the model's capability to cover the entire output range, which helps prevent systematic over- or under-estimations. In contrast, the funnel form of variance and residuals for SVR supported its weakness in generalization, especially at tails**.**

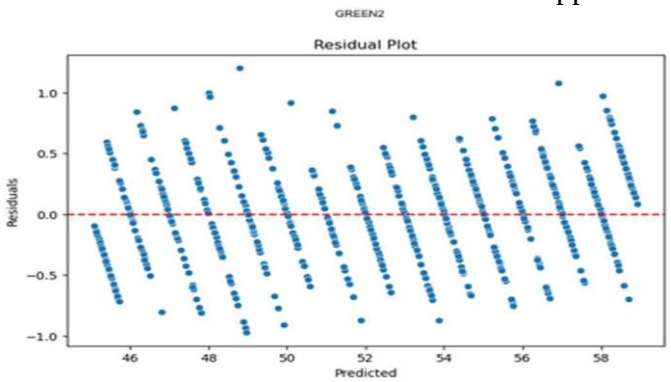


Figure 5: Residual Plot – Random Forest
The residual plot shows the difference between observed and predicted values for each data point in the dataset. The residuals show a tight cluster around zero which indicates that the model fits well and has little bias. The model shows some minor spread at high hydrogen output levels which could indicate opportunities to improve the model or to apply localized model calibration. The diagnostic results confirm that the model performs well across a wide operating range as reported in recent studies on AI-based electrolysis modeling (Mukelabai et al., 2024).

6.3 SHAP Analysis

Viewing The SHAP values were used to explain the internal decision structure of the Random Forest model by finding most of all relevant features for predicting hydrogen output. SHAP Summary plot identified the predicted-contributing factors with respect to hydrogen production, indicating that variables reflecting solar irradiation, electrolyze temperature and share of renewable energy input showed the most positive impact on prediction, in line with actual physical production n of hydrogen. SHAP dependence plots also provided evidence of clear nonlinear interactions between certain predictors, particularly temperature and pressure, consistent with electrochemical process theory *(Afzali et al., 2024).*

This interpretability step gives us transparency and more importantly operational insight, allowing stakeholders to identify which parameters we must keep an eye on or optimize in real life.

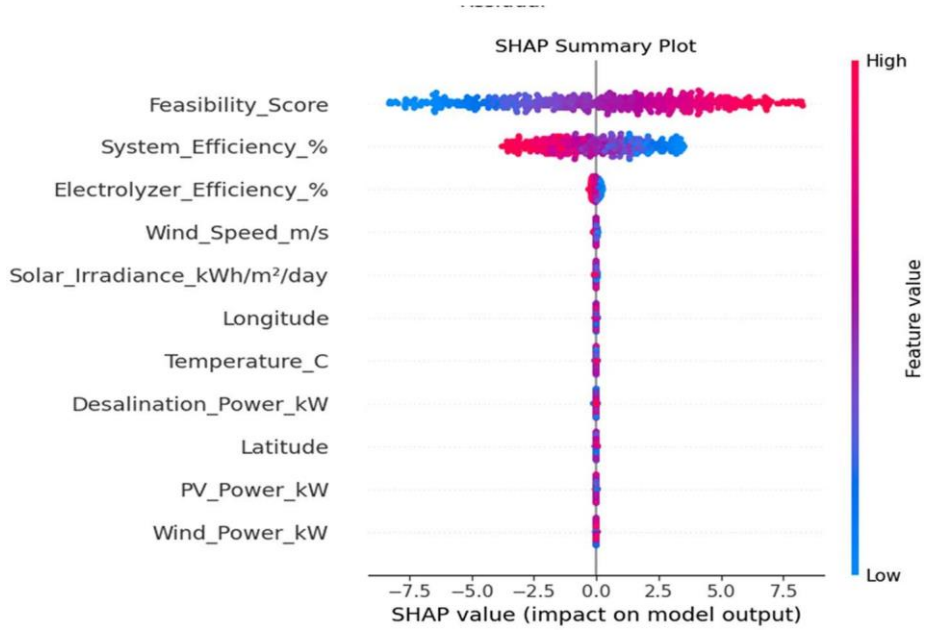


Figure 6: SHAP Summary Plot for Random Forest

The SHAP summary plot shows the most important variables that affect green hydrogen output according to the Random Forest model through a ranked visualization. The features renewable energy input (solar and wind) and electrolyser efficiency and ambient temperature are the most important factors. The wide dispersion of SHAP values for top-ranked variables indicates strong non-linear interactions, supporting the use of ensemble methods over linear baselines. This analysis enhances trust and transparency in AI applications, aligning with industry guidelines on explainable AI (Ahmed et al., 2024; Wei et al., 2025).

6.4 Discussion

XAI implementation in green hydrogen production facilities brings specific economic benefits which manifest throughout plant operations and system-wide operations. The Random Forest model achieves predictive accuracy through Bayesian Optimization and SHAP-based interpretability while providing transparent operations. Such real-world applications need continuous optimization of electrolyzes because their capital-intensive nature requires transparent operations to minimize energy waste and operational downtime and forecasting errors. This model integration in production plants enables better load scheduling and reduces renewable energy source curtailment while improving predictive maintenance which lowers the Levelized Cost of Hydrogen (LCOH) metric that defines financial viability (Urhan et al., 2025; Seeger et al., 2025).

The SHAP explanations enable facility operators to make actionable decisions thus reducing their need for human domain specialists and speeding up their responses to solar irradiation and temperature and wind pattern changes. The model provides economic insights from complex sensor inputs to function as a decision-support tool which boosts resource efficiency. The research by Tawalbeh et al. (2024) shows that AI-based optimization can cut electricity expenses substantially because they represent the main cost driver in hydrogen manufacturing.

The current study differs from previous predictive models that focused on performance maximization by choosing cost-conscious optimization that balances economic trade-offs between complexity and interpretability. The theoretical precision of SVR models in controlled datasets does not translate to economic efficiency in real-world hydrogen production environments because they produce costly mispredictions that cause system instability according to *Mukelabai et al. (2024).*

Ensemble models combined with optimization methods deliver superior return on investment (ROI) according *to Motiramani et al. (2025) and Ahmed et al. (2024)* in their research on renewable energy operations. The deployment speed and cost efficiency of our approach exceeds deep learning architectures while supporting lean operational strategies that focus on economically scalable AI implementations.

The proposed model stands out because it combines cost-effectiveness with modularity characteristics which make it particularly useful for resource-limited markets. The model's low barrier to entry makes it suitable for countries such as Chile, Morocco and India which have abundant solar and wind resources yet lack sufficient financial resources for large-scale AI deployment. The

model's explainable nature enhances its ability to decrease investment risks which is crucial for public-private partnerships and international green finance mechanisms. Investors require transparency in AI systems before investing in large hydrogen infrastructure projects according to *Seeger et al. (2025).*
The model helps perform techno-economic assessments which are essential for building green hydrogen corridors and export hubs. Long-term feasibility studies need predictive AI models to deliver performance forecasts together with operational insights when evaluating energy price volatility alongside renewable intermittency and asset depreciation. Economic planners and engineers can use SHAP interpretability to link energy inputs with hydrogen yield production which helps them develop tariffs and manufacturing strategies.
The proposed framework lacks current models that account for real-world cost elements such as hardware degradation and maintenance expenses and external market fluctuations. The model lacks direct connection with techno-economic lifecycle assessment (LCA) which would enable complete financial evaluation that includes carbon pricing and water resource utilization and capital depreciation. *Phan et al. (2024)* suggested that future work should combine the AI model with economic simulators and digital twin frameworks to overcome these limitations. Real-time economic performance indicators would enable system adjustments which would enhance the reliability and investment return of AI systems in green hydrogen infrastructure.

7. Conclusion and Future Work

7.1 Key Findings and Economic Contributions

This study shows that explainable ensemble machine learning models based on Random Forest optimized through Bayesian methods provide accurate and scalable solutions for predicting green hydrogen production. The model uses SHAP-based interpretability to provide both predictive capabilities and essential operational insights and transparency which benefits technical staff and economic decision-makers. These benefits stand out most in areas with variable renewable energy supply such as Chile and North Africa because hydrogen holds strategic importance for future energy planning *(Gharamani et al., 2023; Motiramani et al., 2025).*
The model provides better scheduling precision and decreases energy consumption while boosting system performance thus lowering the Levelized Cost of Hydrogen (LCOH). The cost-optimization feature is essential for market competitiveness in international hydrogen markets while supporting sustainability goals *(Urhan et al., 2025; Seeger et al., 2025).* The study confirms that SVR models and other non-interpretable or unstable models have economic disadvantages which makes reliable and transparent AI systems essential for industrial use *(Mukelabai et al., 2024).*

7.2 Industrial and Economic Scalability

The model operates at low computational levels while having a modular structure which makes it appropriate for industrial scale operations and emerging economy pilot programs. The model’s adaptable nature enables its use with hydrogen export facilities and digital twins and government-backed energy plans. The model's cost-effectiveness combined with transparency acts as a fundamental driver for public-private investment models and international green bonds because it fulfills investor requirements for explainable and financially viable projects *(Tawalbeh et al., 2024).*
The model enables real-time production changes and predictive maintenance planning which produces major savings throughout the entire hydrogen production process. The capabilities of this model help maintain stable production during unpredictable energy supply situations thus enhancing both national energy security and power grid stability *(Jamali et al., 2025).*

7.3 Recommendations for Future Research

Future research should aim to improve the economic and operational robustness of the model by uniting the current AI pipeline with lifecycle assessments (LCA) and techno-economic simulation tools. Such integrated frameworks would enable comprehensive long-term planning which includes environmental factors like water consumption and emissions and capital deterioration *(Seeger et al., 2025).*

Future research needs to implement real-time industrial testing of the model through collaborative partnerships with green hydrogen facilities. Real-time industrial testing would demonstrate how the model performs regarding scalability and response time and return on investment when production varies. The model could achieve better energy dispatch and electrolysis control in hybrid systems through integration with Reinforcement Learning (RL) architecture *(Naveena et al., 2024).*

The economic deployment of the model extends to include its use in regional hydrogen roadmaps as well as international trade models to support global decarbonization initiatives while ensuring economic viability. These advancements will enable the implementation of AI-driven green hydrogen production that delivers both technical feasibility and economic transformation at a large scale *(Phan et al., 2024; Jamali et al., 2025).*

References

1. Seeger, K. et al., 2025. Techno-economic analysis of hydrogen and green fuels supply scenarios assessing three import routes: Canada, Chile, and Algeria to Germany. *International Journal of Hydrogen Energy*, 116, pp.558–576. Available at: https://doi.org/10.1016/j.ijhydene.2025.02.379

2. Ahmed, R. et al., 2024. An explainable AI for green hydrogen production: A deep learning regression model. *International Journal of Hydrogen Energy*, 83, pp.1226–1242. Available at: https://doi.org/10.1016/j.ijhydene.2024.08.064

3. Mullanu, S. et al., 2025. Artificial intelligence for hydrogen-enabled integrated energy systems: A systematic review. *International Journal of Hydrogen Energy*, 141, pp.283–303. Available at: https://doi.org/10.1016/j.ijhydene.2024.08.013

4. Mukelabai, M.D., Barbour, E.R. & Blanchard, R.E., 2024. Modeling and optimization of renewable hydrogen systems: A systematic methodological review and machine learning integration. *Energy and AI*, 18, 100455. Available at: https://doi.org/10.1016/j.egyai.2024.100455

5. Urhan, B.B. et al., 2025. Predicting green hydrogen production using electrolyzers driven by photovoltaic panels and wind turbines based on machine learning techniques: A pathway to on-site hydrogen refuelling stations. *International Journal of Hydrogen Energy*, 101, pp.1421–1438. Available at: https://doi.org/10.1016/j.ijhydene.2025.01.017

6. Wang, N. et al., 2025. Accelerating the green hydrogen revolution: The synergy of PEMWE and AI. *Renewable Energy System and Equipment*, 1(1), pp.61–66. Available at: https://doi.org/10.1016/j.rese.2024.10.002

7. Raja, I.B. et al., 2025. Regional variability in the performance of Solar-Green Hydrogen Hybrid Energy Systems (SGHHES): Synergistic enviro-economic analysis and evaluation across six climatic zones using multi-criteria decision analysis. *International Journal of Hydrogen Energy*, 138, pp.681–693. Available at: https://doi.org/10.1016/j.ijhydene.2025.05.193

8. Ben Hamida, M.B. et al., 2025. Dynamic optimization of a solar–sCO_2–CAES hybrid energy system: Integration of green hydrogen production and AI-based energy management. *International Journal of Hydrogen Energy*, 146, 149888. Available at: https://doi.org/10.1016/j.ijhydene.2025.06.078

9. Taief, W. et al., 2025. 1.44 – The Application of Machine Learning for Green Hydrogen Production. In: *Comprehensive Green Materials*, Vol. 1, pp.582–593. Available at: https://doi.org/10.1016/B978-0-443-15738-7.00030-1

10. Phan, T.P. et al., 2024. Dynamic-based artificial intelligence model for simulation and optimization of the single chamber anode brush microbial electrolysis cell. *Fuel*, 375, 132629. Available at: https://doi.org/10.1016/j.fuel.2024.132629

11. Rai, A. & Liu, J., 2025. Optimizing solar-electrolysis for green hydrogen production: A novel spatiotemporal attention framework (STAF) for solar-electrolysis prediction and economic viability analysis. *International Journal of Hydrogen Energy*, 142, pp.723–738. Available at: https://doi.org/10.1016/j.ijhydene.2025.03.385

12. Yang, Q. et al., 2025. Machine learning-assisted prediction and optimization of solid oxide electrolysis cell for green hydrogen production. *Green Chemical Engineering*, 6(2), pp.154–168. Available at: https://doi.org/10.1016/j.gce.2024.04.004

13. Busam, K.M. et al., 2025. Modeling and process operability analysis of the proton conducting solid oxide electrolysis cell (H^+-SOEC) for efficient hydrogen production. *International Journal of Hydrogen Energy*, 145, pp.485–495. Available at: https://doi.org/10.1016/j.ijhydene.2025.05.275

14. Shanmugasundaram, S. et al., 2025. A review on green hydrogen production pathways and optimization techniques. *Process Safety and Environmental Protection*, 197, 107070. Available at: https://doi.org/10.1016/j.psep.2025.107070

15. Motiramani, M. et al., 2025. AI-ML techniques for green hydrogen: A comprehensive review. *Next Energy*, 8, 100252. Available at: https://doi.org/10.1016/j.nxener.2025.100252

16. Tawalbeh, M. et al., 2024. Prediction of hydrogen production in proton exchange membrane water electrolysis via neural networks. *International Journal of Thermofluids*, 24, 100849. Available at: https://doi.org/10.1016/j.ijft.2024.100849

17. Naveena, K. et al., 2024. Elevating sustainability with a multi-renewable hydrogen generation system empowered by machine learning and multi-objective optimization. *Measurement: Sensors*, 33, 101192. Available at: https://doi.org/10.1016/j.measen.2024.101192

18. Wei, J. et al., 2025. Artificial intelligence applications in hydrogen system: Advancing renewable energy utilization for global hydrogen economy and sustainability goals. *International Journal of Hydrogen Energy*, 122, pp.359–373. Available at: https://doi.org/10.1016/j.ijhydene.2025.03.350

19. Qarssis, Y. et al., 2024. Machine learning-based analytical approach for mechanical analysis of composite hydrogen storage tanks under internal pressure. *International Journal of Hydrogen Energy*, 89, pp.1440–1453. Available at: https://doi.org/10.1016/j.ijhydene.2024.09.404

20. Shahin, M. & Simjoo, M., 2025. Potential applications of innovative AI-based tools in hydrogen energy development: Leveraging large language model technologies. *International Journal of Hydrogen Energy*, 102, pp.918–936. Available at: https://doi.org/10.1016/j.ijhydene.2025.01.066

21. Jamali, M. et al., 2025. An insight into the application and progress of artificial intelligence in the hydrogen production industry: A review. *Materials Today Sustainability*, 30, 101098. Available at: https://doi.org/10.1016/j.mtsust.2025.101098

22. Kabir, M.M. et al., 2023. Machine learning-based prediction and optimization of green hydrogen production technologies from water industries for a circular economy. *Desalination*, 567, 116992. Available at: https://doi.org/10.1016/j.desal.2023.116992

23. Quintanilla, P. et al., 2025. Artificial intelligence and robotics in the hydrogen lifecycle: A systematic review. *International Journal of Hydrogen Energy*, 113, pp.801–817. Available at: https://doi.org/10.1016/j.ijhydene.2025.03.016

24. Riera, J.A. et al., 2023. A review of hydrogen production and supply chain modeling and optimization. *International Journal of Hydrogen Energy*, 48(37), pp.13731–13755. Available at: https://doi.org/10.1016/j.ijhydene.2022.12.242

25. Schröer, C. et al., 2021. A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 181, pp.526–534. Available at: https://doi.org/10.1016/j.procs.2021.01.199

26. Schropp, E. et al., 2024. Multi-criteria optimization of electrode materials for anion exchange membrane water electrolysis regarding energy demand and material criticality. *Journal of Power Sources*, 2024, 235031. Available at: https://doi.org/10.1016/j.jpowsour.2024.235031

27. Mallala, B. et al., 2025. Forecasting global sustainable energy from renewable sources using random forest algorithm. *Results in Engineering*, 25, p.103789. Available at: https://doi.org/10.1016/j.rineng.2024.103789

28. Shash, A.Y. et al., 2025. Computational methods, Artificial Intelligence, Modeling, and Simulation applications in green hydrogen production through water electrolysis: A review. *Hydrogen*, 6(2), 21. Available at: https://doi.org/10.3390/hydrogen6020021

29. DATA SET: https://www.kaggle.com/datasets/ziya07/renewable-hydrogen-production-dataset/data