12º CONGRESSO BRASILEIRO DE PESQUISA E DESENVOLVIMENTO EM PETRÓLEO E GÁS

TÍTULO DO TRABALHO:

EXPLORING DATA-DRIVEN SOLUTIONS FOR HYDROGEN SAFETY: PREDICTIVE MODELING OF EXPLOSION THRESHOLDS

AUTORES:

Sandrely Pereira da Silva[1,2,*], Caio Souto Maior[1,3], Isis Didier Lins[1,2], Márcio das Chagas Moura[1,2]

INSTITUIÇÃO:

[1]Centro de Estudos e Ensaios em Risco e Modelagem Ambiental (CEERMA), [2]Departamento de Engenharia de Produção, Universidade Federal de Pernambuco (UFPE), [3]Núcleo de Tecnologia, UFPE

EXPLORING DATA-DRIVEN SOLUTIONS FOR HYDROGEN SAFETY: PREDICTIVE MODELING OF EXPLOSION THRESHOLDS

Resumo

O hidrogênio é uma forma limpa de energia que oferece uma alternativa sustentável com zero emissões de carbono. Possui uma ampla gama de aplicações, incluindo transporte, processos industriais e soluções de armazenamento de energia. Embora apresente vantagens significativas, o hidrogênio introduz desafios de segurança devido à sua elevada sobrepressão, combustão rápida e ampla faixa de inflamabilidade, exigindo medidas de segurança em todas as fases do seu gerenciamento. Este estudo explora aplicações de Machine Learning (ML) para prever e analisar os limites de explosão de misturas de hidrogênio-oxigênio (H2-O2), com o objetivo de garantir a integridade das operações de hidrogênio. Para a análise foi utilizado um conjunto de dados disponível na literatura, composto por dados tabulares de pressão, temperatura e concentrações de N_2, He, Ar e H_2O. Técnicas de pré-processamento, como seleção de características, data augmentation e padronização, foram aplicadas para melhorar a qualidade dos dados e o desempenho do modelo. Posteriormente, vários algoritmos preditivos de ML foram analisados para avaliar sua eficácia na classificação de misturas de H2-O2 como explosivas ou não explosivas. O estudo alcançou acurácia máxima de 96.39% utilizando Support Vector Machine, com alta precisão e recall, indicando sua robusta capacidade em identificar corretamente casos positivos e negativos. Estas descobertas contribuem para o desenvolvimento de sistemas de hidrogênio confiáveis, prevendo e mitigando os riscos de explosão associados às misturas de hidrogênio, aumentando assim a segurança e a eficiência do manuseamento de hidrogênio.

Palavras-chave: hidrogênio, misturas H2-O2, aprendizado de máquina, segurança.

Abstract

Hydrogen is a clean form of energy that offers a sustainable alternative with zero carbon emissions. It has a wide range of applications, including transportation, industrial processes, and energy storage solutions. Although it presents significant advantages, hydrogen introduces safety challenges due to its high overpressure, rapid combustion, and wide flammability range, demanding safety measures at every stage of its management. This study explores Machine Learning (ML) applications to predict and analyze the explosion thresholds of hydrogen-oxygen (H2-O2) mixtures to ensure the integrity of hydrogen operations. A dataset available in the literature was utilized for the analysis, comprising tabular data on pressure, temperature, and concentrations of N_2, He, Air, and H_2O. Preprocessing techniques such as feature selection, data augmentation, and standardization were applied to improve data quality and model performance. Subsequently, various ML predictive algorithms were evaluated to assess their effectiveness in classifying H2-O2 mixtures as explosive or non-explosive. The study achieved a maximum accuracy of 96.39% using Support Vector Machine, with high precision and recall, indicating its robust capability in correctly identifying both positive and negative cases. These findings contribute to the development of reliable hydrogen systems by predicting and mitigating explosion risks associated with hydrogen mixtures, thereby enhancing the safety and efficiency of hydrogen handling.

Keywords: hydrogen, H2-O2 mixtures, machine learning, safety.

Introduction

Hydrogen offers significant potential for reducing greenhouse gas emissions and combating climate change, becoming a sustainable alternative to fossil fuels. It can generate energy without producing pollutant gases since its only byproduct is water vapor (Acar and Dincer, 2020). Besides, hydrogen can be stored and transported in various forms, including compressed gas, liquid hydrogen, and chemical carriers (i.e., ammonia or metal hydrides). Due to hydrogen's physical and chemical properties, safety challenges need to be considered (Aziz, 2021).

One of the most concerning aspects of hydrogen is its high overpressure and the extreme shock waves generated upon detonation. The lightweight molecules of hydrogen facilitate rapid combustion, resulting in significantly higher overpressures (Shirvill et al., 2012). Moreover, hydrogen's wide flammability range and its susceptibility to ignition across various concentrations, when combined with air, facilitate the formation of potentially explosive mixtures (Tashie-Lewis and Nnabuife, 2021). Thus, the risks associated with hydrogen handling and storage reinforce the need for safety measures throughout its supply chain (Moradi and Groth, 2019). In this sense, it is important to understand and manage explosion thresholds of hydrogen-oxygen (H2-O2) mixtures to delineate safe operating conditions from potentially hazardous scenarios, as seen in Mével et al. (2016).

Nowadays, data-driven strategies, such as statistical analysis and machine learning algorithms, are being used to analyze large volumes of data to extract patterns and findings and make informed decisions (Sarker, 2021). Thus, Machine Learning (ML) can be applied to predict and analyze explosion thresholds of H2-O2 mixtures. The use of ML allows for the identification of critical parameters that influence hydrogen's explosive behavior, providing the development of mitigation strategies (Benson and Obasi, 2023). Additionally, data-driven techniques contribute to enhancing the design of hydrogen storage and handling systems, optimizing emergency response planning, and improving the reliability and security of hydrogen-related operations (Silva et al., 2024).

This study focuses on predictive modeling to understand the variables related to explosions of H2-O2 mixtures by exploring ML applications capable of distinguishing between explosive and non-explosive hydrogen mixtures in a classification process. Moreover, it is expected to enhance the accuracy of model predictions and, consequently, ensure more reliable and secure systems. The study is structured as follows: section 2 involves the methodologies applied, including data analysis, preprocessing techniques, and predictive algorithms used. Section 3 summarizes the main findings of the study and section 4 provides some concluding remarks.

Methodology

The proposed methodology consists of analyzing the dataset, preprocessing data, tuning of hyperparameters, training predictive models, and results analysis, as presented in Figure 1.

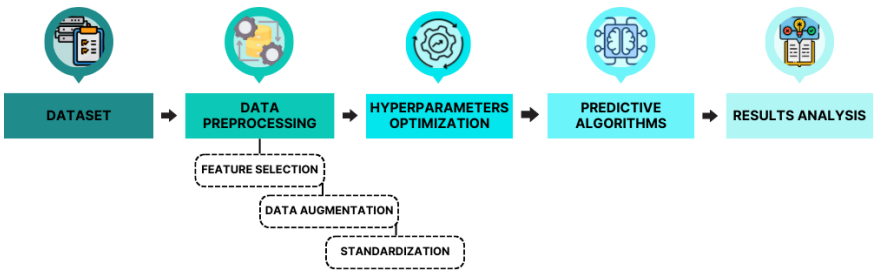

Figure 1: Methodology steps utilized.

Dataset

Data is essential for extracting important information, especially when analyzing the behavior of gas mixtures (Lv et al., 2023). The dataset used in this study presents 3000 rows and 9 columns, including a binary label, sourced from Li et al. (2024). It includes measurements of pressure, ranging from 101 to 108 Pa, and temperature, from 600 to 1000 K. A temperature increase of 50 K within 0.1s was used by the authors as the criterion to classify the dataset into explosive and non-explosive labels. It also encompasses concentrations of N2, He, Air, and H2O, ranging from 0 to 8.0 units.

Preprocessing

Data preprocessing involves transforming raw data into a clean and usable format to ensure the ML models perform effectively (Zhou et al., 2017). This study uses feature selection, data augmentation, and standardization preprocessing techniques.

Feature Selection

Feature selection involves the identification of the most relevant variables in a dataset. It improves model performance, reduces overfitting, and minimizes computational costs (El-Hasnony et al., 2020). Extra Trees classifier is a learning method used for feature selection. It works by constructing multiple decision trees from a random subset of features and samples during training. The algorithm evaluates the importance of each feature by measuring how much it reduces the impurity of the split at each node (Mienye and Jere, 2024). Figure 2 shows the features ordered by its importance values.

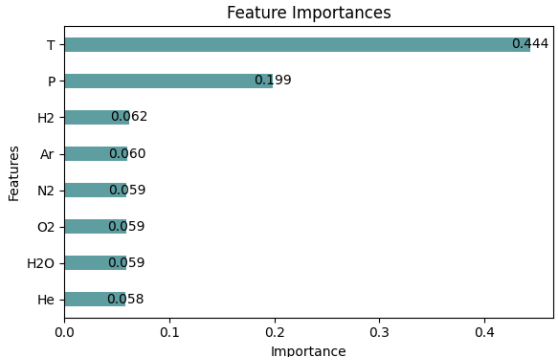

Figure 2: Features ordered by importance values

For the analysis in this study, we defined 0.062 as the threshold for selecting features. Therefore, the three most important features (i.e., ‘T’, ‘P’, and ‘H2’) were selected.

Data augmentation

The imbalance presented in datasets influences model training towards the majority class (Lango and Stefanowski, 2022). In the dataset utilized in this study, class 0, representing non-explosive cases, was the majority class with 75.7% of the data, while class 1 comprised 24.3% of the data (i.e., explosive cases). In this context, we performed a data augmentation process with random oversampling. Random oversampling is a data augmentation technique to handle class imbalance in datasets by duplicating samples from the minority class. This technique ensures a balanced 50% distribution for each class, preventing model bias towards the majority class (Wibowo and Fatichah, 2021).

Standardization

Standardization is a preprocessing technique used to transform data into a standardized scale, typically with a mean of 0 and a standard deviation of 1. This process ensures that all features contribute equally to model training and avoids bias towards variables with larger scales (Habib and Okayli, 2024). In this study, we applied the StandardScaler algorithm, a specific implementation of standardization in scikit-learn Python library, simplifying the transformation of numerical data for ML techniques (Mankodi et al., 2023).

Optimization of hyperparameters

Hyperparameter optimization methods focus on optimizing the architecture of an ML model by detecting the optimal hyperparameter configurations. Thus, GridSearchCV is an algorithm to systematically evaluate a specified set of hyperparameters to find the optimal group for a given model (Singh et al., 2024). It works by exhaustively searching through a predefined grid of hyperparameter values and performing cross-validation to assess the performance of each combination, identifying the best hyperparameters that maximize the generalization of the models (Yang and Shami, 2020).

Predictive algorithms

ML predictive algorithms identify patterns and relationships within data to predict unknown or future events. Their ability to analyze large datasets and generate significant findings makes them great tools for data-driven solutions (Sarker, 2021). In the current study, we implemented different algorithms: boostrap aggregating (Bagging), multilayer perceptron (MLP), support vector machine (SVM), k-nearest neighbors (KNN), gradient boosting (GB), and extra gradient boosting (XGB). Table 1 presents a brief description of each model.

Bagging	Combines predictions from multiple models to improve accuracy and reduce overfitting by averaging results from different subsets of data.
MLP	A neural network with multiple layers of nodes, used for complex classification and regression tasks.
SVM	Finds the optimal hyperplane to separate data into different classes, effective for both linear and non-linear problems.
KNN	Classifies data points based on the majority class among their k-nearest neighbors.
GB	Builds models sequentially, where each new model improves on the errors of the previous ones, using weak learners like decision trees.
XGB	An optimized gradient boosting algorithm that builds decision trees sequentially to creates a strong predictive model through iterative refinement by correcting errors.

Table 1: Brief description of the predictive algorithms.

Results

The study explored different preprocessing techniques, such as feature selection (FS), data augmentation (DA), and hyperparameter tuning, to assess their impact on the accuracy of classification ML models. Our evaluation included the use of balanced accuracy, calculated through (1).

$$Balanced\ accuracy = \frac{1}{2} \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad (1)$$

Our analysis demonstrates that all models achieved relatively high accuracy, indicating their strong ability to classify most examples correctly. SVM is the best model with an accuracy of 0.9639, demonstrating high precision (0.9635) and recall (0.9611), indicating that SVM is capable of identifying true positives while minimizing false positives, making it highly reliable for classification tasks. The MLP also performed well with an accuracy of 0.9583, indicating a balanced approach with slightly lower overall accuracy but higher precision (0.9640) and recall (0.9633). The Bagging model, while not achieving the highest accuracy, demonstrated strong precision (0.9600) and recall (0.9600). Conversely, the KNN model presented reasonable precision (0.9477) but underperformed in recall (0.9467) compared to the other models. Table 2 underscores these results.

	Accuracy	Precision	Recall
SVM	0.9639	0.9635	0.9611
MLP	0.9583	0.9640	0.9633
KNN	0.9477	0.9477	0.9467
BAG	0.9476	0.9600	0.9600

Table 2: Results from models with best performance.

When analyzing the confusion matrices of the three models with higher accuracy, we observe that both SVM and MLP exhibit high true positive rates, indicating strong performance in correctly identifying positive cases. These two models maintain very low rates of false positives and false negatives, confirming their high precision and recall. In contrast, the KNN model demonstrates higher rates of both false positives and false negatives, suggesting it misclassifies more negative cases as positive and, conversely, positive cases as negative. The matrices are shown in Figure 3.

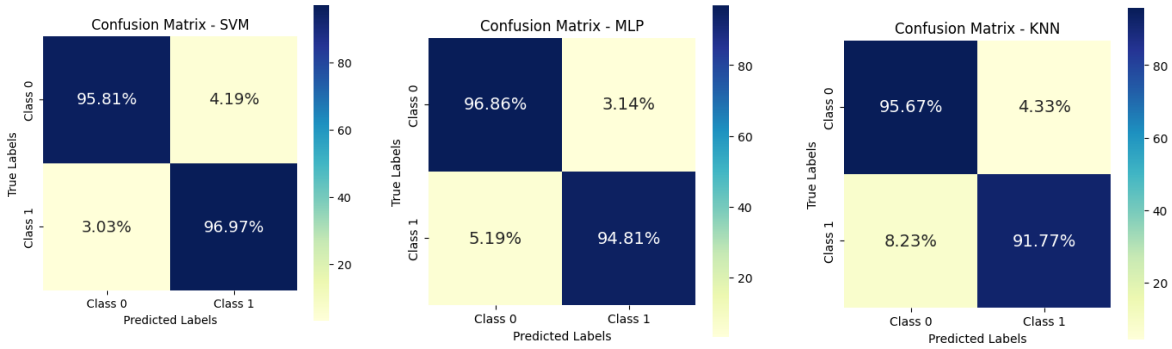

Figure 3: Confusion matrices.

Table 3 shows all accuracy results under different preprocessing configurations.

		without tuning		with tuning	
		without DA	with DA	without DA	with DA
Bagging	without FS	0.9411	0.9475	0.9411	0.9482
	with FS	0.9448	0.9476	0.9448	0.9476
MLP	without FS	0.9108	0.9602	0.9333	0.9462
	with FS	0.9253	0.9574	0.9325	0.9583
SVM	without FS	0.8405	0.8871	0.8679	0.8402
	with FS	0.8863	0.9510	0.9513	**0.9639**
KNN	without FS	0.7912	0.8223	0.7965	0.8206
	with FS	0.9230	0.9371	0.9152	0.9372
XGB	without FS	0.9513	0.9468	0.9499	0.9477
	with FS	0.9391	0.9468	0.9391	0.9419
GB	without FS	0.9492	0.9538	0.9375	0.9317
	with FS	0.9383	0.9386	0.9383	0.9166

Table 3: All accuracy results.

Conclusion

This study evaluated various Machine Learning (ML) models to assess their performance in classifying H2-O2 mixtures as explosive or non-explosive, considering variables related to chemical elements and environment (i.e., temperature, pressure). We explored preprocessing techniques (e.g., feature selection, data augmentation) and hyperparameter tuning. The SVM model exhibited the highest accuracy of 96.39%, high precision and recall, indicating its robust capability in correctly identifying both positive and negative cases. The MLP model also performed well, though with slightly lower accuracy compared to the SVM. However, some models did not significantly improve with data augmentation, indicating the need for more investigations. It is important to recognize the practical challenges of applying these models in real-world scenarios, where operational variability may impact model performance. For instance, adapting the models to different scenarios by training further models for each specific condition can be a significant challenge. For future analysis, we recommend using an alternative data augmentation technique, considering the generation of synthetic samples from the minority class that are close to the decision boundary, to improve models' results. Furthermore, exploring different sets of features or hyperparameter grids may contribute to finding better performance of the models. The advancement of this methodology can lead to reliable hydrogen systems by predicting and mitigating risks associated with hydrogen mixtures, including potential explosion hazards.

Acknowledgments

The authors thank the Brazilian research funding agencies Human Resources Program (PRH 38.1) entitled "Risk Analysis and Environmental Modeling in the Exploration, Development and Production of Oil and Gas", financed by 'Agência Nacional de Petróleo (ANP)' and managed by FINEP, the 'Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq): 402761/2023-5', and 'Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)' - Finance Code 001 for the financial support through research grants.

References

Acar, C., and Dincer, I. (2020). The potential role of hydrogen as a sustainable transportation fuel to combat global warming. *International Journal of Hydrogen Energy*, *45*(5), 3396–3406. https://doi.org/10.1016/j.ijhydene.2018.10.149

Aziz, M. (2021). Liquid Hydrogen: A Review on Liquefaction, Storage, Transportation, and Safety. *Energies*, *14*(18), 5917. https://doi.org/10.3390/en14185917

Benson, C., and Obasi, I. (2023). *The Efficacy of Digital Tools in Detecting and Minimizing Risks Associated with Hydrogen Safety*. https://ssrn.com/abstract=4617484

El-Hasnony, I. M., Barakat, S. I., Elhoseny, M., and Mostafa, R. R. (2020). Improved Feature Selection Model for Big Data Analytics. *IEEE Access*, *8*, 66989–67004. https://doi.org/10.1109/ACCESS.2020.2986232

Habib, M., and Okayli, M. (2024). Evaluating the Sensitivity of Machine Learning Models to Data Preprocessing Technique in Concrete Compressive Strength Estimation. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-024-08776-2

Lango, M., and Stefanowski, J. (2022). What makes multi-class imbalanced problems difficult? An experimental study. *Expert Systems with Applications*, *199*, 116962. https://doi.org/10.1016/j.eswa.2022.116962

Li, J., Liang, W., and Han, W. (2024). Predicting the explosion limits of hydrogen-oxygen-diluent mixtures using machine learning approach. *International Journal of Hydrogen Energy*, *50*, 1306–1313. https://doi.org/10.1016/j.ijhydene.2023.10.204

Lv, Q., Zhou, T., Zheng, R., Nakhaei-Kohani, R., Riazi, M., Hemmati-Sarapardeh, A., Li, J., and Wang, W. (2023). Application of group method of data handling and gene expression programming for predicting solubility of CO2-N2 gas mixture in brine. *Fuel*, *332*, 126025. https://doi.org/10.1016/j.fuel.2022.126025

Mankodi, A., Bhatt, A., and Chaudhury, B. (2023). *An AutoML Based Algorithm for Performance Prediction in HPC Systems* (pp. 108–119). https://doi.org/10.1007/978-3-031-29927-8_9

Mével, R., Sabard, J., Lei, J., and Chaumeix, N. (2016). Fundamental combustion properties of oxygen enriched hydrogen/air mixtures relevant to safety analysis: Experimental and simulation study. *International Journal of Hydrogen Energy*, *41*(16), 6905–6916. https://doi.org/10.1016/j.ijhydene.2016.03.026

Mienye, I. D., and Jere, N. (2024). A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access*, *12*, 86716–86727. https://doi.org/10.1109/ACCESS.2024.3416838

Moradi, R., and Groth, K. M. (2019). Hydrogen storage and delivery: Review of the state of the art technologies and risk and reliability analysis. *International Journal of Hydrogen Energy*, *44*(23), 12254–12269. https://doi.org/10.1016/j.ijhydene.2019.03.041

Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. *SN Computer Science*, *2*(5), 377. https://doi.org/10.1007/s42979-021-00765-8

Shirvill, L. C., Roberts, T. A., Royle, M., Willoughby, D. B., and Gautier, T. (2012). Safety studies on high-pressure hydrogen vehicle refuelling stations: Releases into a simulated high-pressure dispensing area. *International Journal of Hydrogen Energy*, *37*(8), 6949–6964. https://doi.org/10.1016/j.ijhydene.2012.01.030

Silva, S. P., Maior, C. S., Lins, I. D., and Moura, M. C. (2024). Analysing Hydrogen Embrittlement in Steels Using Machine Learning on Environmental and Material Aspects from Open Datasets. *8th International Symposium on Natural Hazard-Triggered Technological Accidents.* Trondheim, Norway.

Singh, J., Sandhu, J. K., and Kumar, Y. (2024). An Analysis of Detection and Diagnosis of Different Classes of Skin Diseases Using Artificial Intelligence-Based Learning Approaches with Hyper Parameters. *Archives of Computational Methods in Engineering*, *31*(2), 1051–1078. https://doi.org/10.1007/s11831-023-10005-2

Tashie-Lewis, B. C., and Nnabuife, S. G. (2021). Hydrogen Production, Distribution, Storage and Power Conversion in a Hydrogen Economy - A Technology Review. *Chemical Engineering Journal Advances*, *8*, 100172. https://doi.org/10.1016/j.ceja.2021.100172

Wibowo, P., and Fatichah, C. (2021). An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, *7*(1), 63. https://doi.org/10.26594/register.v7i1.2206

Yang, L., and Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Zhou, L., Pan, S., Wang, J., and Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*, 350–361. https://doi.org/10.1016/j.neucom.2017.01.026