International Journal of Engineering & Technology, 14 (1) (2025) 24-29

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
https://doi.org/10.14419/6x85gr51
Research paper

Utilizing machine learning for predictive analysis of emission levels to ensure compliance in refinery operations

Frederick Gyaase [1] *, Olalekan Samuel Okunlola [2], Adeyinka M Olusanya [3], Adeleye Adedokun [3], Olufunmilayo Ifeoluwa Somoye [4], Akpevwe Theophilus Erhieyovwe [5]

[1] Department of Engineering and Physical Sciences, University of Wyoming
[2] Department of Mechanical and Materials Engineering, University of Cincinnati
[3] Department of Geography, Oklahoma State University
[4] Department of Data Science. Northeastern University
[5] Department of Physics Illinois Institute of Tecchnology
**Corresponding author E-mail: fgyaase96@gmail.com*

Received: March 23, 2025, Accepted: April 14, 2025, Published: May 18, 2025

Abstract

This study explores the use of Machine learning (ML) to predict emission levels in refinery operations to support regulatory compliance. Refineries produce a large amount of pollution like CO_2 , SO_x, NO_x, VOCs, and PM that cause environmental degradation and public health concerns. Manual sampling and inspections are siusingmply not real-time and hence at risk for noncompliance. With the advent of ML-based predictive analytics, we can analyze large datasets, predict emission levels, and come up with preventive measures.
When the ML models were applied to the emission prediction, they were found to have certain limitations like the quality of data, computational burden, model interpretability, and data privacy. These models include Linear Regression, Decision Trees, Support Vector Machines (SVMs), Long Short-Term Memory (LSTM) networks, and ensembles of Random Forest and XGBoost. It recommends the integration of ML with traditional monitoring, improvement of data quality, a guarantee of data privacy, and fostering of interdisciplinary collaboration. ML application can be optimized with these strategies, practices can be driven towards sustainability, and compliance can be strengthened in refinery operations.

Keywords: *Machine Learning; Predictive Emission; Refinery Operation; Regulatory Compliance.*

1. Introduction

Refineries are important in meeting global energy demands, but they are major air pollutants that emit CO_2, SOx, NOx, VOCs, and particulate matter [1]. These emissions originate from multiple processes, including refining and burning of fossil fuels, along with routine operations such as flaring and methane leakage during production. As a leading source of global emissions, the industry's activities significantly influence climate change, contributing to global warming, rising sea levels, and severe weather conditions [2]. The emissions caused by these refineries hurt air quality, ecosystem health, and public health, and long-term exposure is associated with respiratory and cardiovascular diseases [3]. With greater awareness of these impacts, regulatory agencies, including the Environmental Protection Agency (EPA) in the United States and the European Environment Agency (EEA), have set very strict emission standards on pollution from refinery operations [4]. Severe penalties, operational shutdowns, and reputational damage are all consequences of non-compliance, which can force refineries to adopt more effective monitoring methods [5].
However, these efforts are still limited in conventional emission monitoring approaches, particularly in terms of delayed data collection, low real-time accuracy and the inability to forecast [6]. To deal with these challenges, there have been advancements in machine learning (ML) and predictive analytics to build the gruesome devices of emission monitoring devices in the line of [7]. Such studies can use ML techniques to process many sensors and historical emission records datasets for the identification of patterns, prediction of emission levels, and recommendation of corrective actions [8]. ML-based predictive analytics is different from traditional methods as it enables real time monitoring, which enables refineries to make timely operational adjustments to avoid regulatory violations [9].
More and more, operations are being optimized and emissions are being minimized by ML models, such as regression analysis, decision trees, random forests, and neural networks [1]. These data-driven techniques allow us to predict accurately, and timely manner refinery processes given the complex interaction between refinery processes and external factors such as weather conditions [7]. When ML is incorporated in emission monitoring, refineries not only become more compliant with environmental regulations but also become more efficient operationally and reduce environmental risks [5]. In an increasing demand for sustainable industrial practices, ML-based predictive analytics provide a viable path for emission reduction and realize the long-term environmental goals.

1.1. Research aim and objectives

The primary aim of this study is to explore how machine learning can be leveraged for predictive analysis of emission levels to ensure compliance in refinery operations. The study will achieve this by examining existing ML techniques, assessing their effectiveness in emission prediction, and evaluating their practical applications in industrial settings.
The objectives are:

1) To critically assess current emission monitoring practices in refinery operations, explore the applicability of machine learning (ML) models for predictive emission analysis, and evaluate the accuracy and reliability of these models compared to traditional methods.
2) To identify the challenges and limitations associated with implementing ML-driven emission monitoring systems in refinery operations and investigate the factors influencing their successful adoption.
3) To develop practical recommendations for integrating machine learning into refinery compliance strategies, aiming to enhance predictive accuracy, regulatory compliance, and overall operational efficiency.

1.2. Scope and limitations of the study

In this research, we are focusing on using machine learning for the prediction of emissions in refinery operations to comply with environmental regulations. The study will investigate the implementation challenges of various ML models and their benefits. Thus, while it provides practically important information regarding the analysis of various data processing methods in managing Geiger modes, it is restricted to secondary data analysis and literature reviews, and experimentation or real-time industrial deployment are not involved. Moreover, these include different ML techniques, while the research does not involve in-depth algorithm development but rather assesses the applicability and effectiveness of these techniques in the industrial setting. This study attempts to do so by addressing these aspects to contribute to the expansion of the field of knowledge on the deployment of AI for environmental monitoring and to support the improvement and sustainability of refinery operations.

2. Literature review

Refining operations are critical to satisfy the world's energy demands, but are important sources of air pollution, providing emissions of CO_2, SO_x, NO_x, VOCs, and PM [1]. They are very damaging to air quality, to public health, and ecosystems, contributing to respiratory and cardiovascular diseases and increased mortality rates [3]. In response to these threats, the Environmental Protection Agency (EPA) and the European Environment Agency (EEA) came forward to set stringent emission limits for the refineries to reduce their environmental impact [4]. Nevertheless, it poses a challenge to compliance due to the limitations of the traditional emission monitoring methods [10].
Current monitoring methods also involve manual sampling, on-site inspections, and periodic reporting [11]. In addition to the use of digital twins for lowering emissions [12]. However, as these methods are often costly, time-consuming, and their data accuracy is not real-time, it is difficult to identify and control emission violations on time [11]. Moreover, these standard methods may have challenges in depicting the fidelity of such dynamic refinery processes without undue underestimation of emissions [6]. Despite strict environmental regulations that lead to innovation, monitoring systems on the ground can break the compliance systems and make operational competitiveness unsuccessful [13]. The limitations that these have created have necessitated the use of the advanced monitoring systems that can deliver accurate, in-time, and cost-effective emission monitoring.
These challenges have led to the promise of machine learning (ML) as an approach to solve the very problem. There are ML techniques, namely regression analysis, decision trees, random forests, and neural networks, that have been able to analyze large datasets, predict emission levels, and suggest corrective measures [7]. ML-based predictive analytics are more proactive than traditional methods that react to situations and that allow for identifying patterns, forecasting potential breaches, and optimizing operational efficiency [14]. [5] researched that applying ML models could substantially increase prediction accuracy in a complex refinery environment, and Olawade et al. (2024) [15] emphasized their application for compliance monitoring. While ML for emission monitoring is being integrated, some challenges, such as data privacy concerns, infrastructure constraints, and scarcity of skilled personnel [1], hinder the process. There is also a lack of literature on how ML can be best incorporated into current monitoring frameworks to provide for comprehensive real-time compliance [16].
Due to the environmental and regulatory pressures, there is an urgent need to seek alternative and more efficient means of monitoring refinery emissions. By examining ML-based predictive analytics to be applied to monitor the emission area, this study attempts to fill the gap by looking into the optimization of compliance and minimizing the environmental impact. This research looks into a hybrid approach that consists of an integration of ML techniques into traditional monitoring systems, allowing ML to boost accuracy sustainably and lead to stable refinery operations. These findings will assist industry practices, environmental protection strategies, and policymakers information on innovative data-driven air emission management solutions.

2.1. Theoretical framework

The Technology Acceptance Model (TAM) and Data-Driven Decision Making (DDDM) provide theoretical perspectives for use in the study of how the use of machine learning (ML) can be used for predictive analysis of emission levels to ensure compliance in refinery operations. The Technology Acceptance Model (TAM), as proposed by [17], is very important in explaining the factors that influence user acceptance of ML-based systems. In TAM, Perceived Usefulness (PU) and Perceived Ease of Use (PEOU) are two core factors that determine users' attitude towards a technology to form their behavioral intention for using the technology. In the context of refinery operation, PU is the perception of how much stakeholders think ML will be able to predict and monitor emission levels to support compliance, and PEOU is the perception of how easy it is to use the technology. The chances of engineers and managers adopting these systems decrease if they view the ML tools as complex or unreliable [18]. However, the ability of ML models to make timely and accurate emission forecasts will demonstrate the capability of these tools to be used in decision-making processes, increasing acceptance of these tools.
As with Data-Driven Decision Making (DDDM) principles, integration of ML for emission prediction also rests on principles of empirical data instead of intuition [19]. DDDM focuses on the collection, analysis, and application of data for the realization of the best outcomes. In refinery operations, ML algorithms process huge datasets to predict emission levels and help with proactive interventions, and DDDM is present. ML models can process real-time and historical records of emission patterns and derive insights that can be used to minimize the regulatory breaches [20]. Nevertheless, DDDM depends on data quality, staff expertise, and organizational readiness. However, data

accuracy or the absence of understanding of how ML analytics work are not challenges that should undermine decision making [21], and such individuals need to have comprehensive training and robust data governance practices. It is important to consider the relationship between TAM and DDDM for the successful implementation of ML in refinery operations. TAM focuses on human factors that affect the acceptance of data-driven technologies, while DDDM provides a rationale for the adoption of data-driven approaches to boost the compliance and operational efficiency. To improve the acceptance of ML systems, stakeholders can be persuaded when they see the practical value of data-driven insights in reducing emissions. At the same time, a data-oriented organizational culture can enhance the perceived worth of ML tools to make it easier to adopt. Thus, ML-based emission prediction systems in refineries should be designed and deployed considering both TAM and DDDM.

3. Methodology

3.1. Search strategy and study selection

This study takes a systematic literature review approach to explore machine learning (ML) applications for predictive analysis of emission levels in refinery operations. The chosen method of systematic review was to ensure a comprehensive and objective synthesis of existing literature. An extensive search of relevant studies was conducted using academic databases such as ScienceDirect, IEEE Xplore, Web of Science, Scopus, Google Scholar, etc. The databases were chosen to accommodate a broad coverage in peer-reviewed journal articles, conference proceedings, and industry reports in ML, emission monitoring, and refinery operations.

To limit the search to articles published from 2015 to 2025, the search was carried out to capture the most recent advances in the field. The combination of keywords such as "machine learning," "predictive analytics," "refinery emissions," "air pollution monitoring," "environmental compliance," and "industrial emission prediction" was applied to refine the search with Boolean operators ("AND," "OR"). English language studies were further restricted to narrow the search, as this would allow for more interested interpretation and comparative analysis.

The study selection was under the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to maintain transparency, reduce selection bias, and increase the validity of the findings. The four stages of the PRISMA process were identification, screening, eligibility, and inclusion. In the identification stage, a wide variety of studies were retrieved and screened by reviewing titles and abstracts to exclude studies that were not relevant. In the stage of eligibility, the full-text articles were read through to determine relevance to the research objectives and the robustness of their methodological framework. Finally, studies that met all criteria were included for a detailed analysis.

3.2. Inclusion and exclusion criteria

To build the credibility and applicability of the systematic literature review, strict inclusion and exclusion criteria were set. We only considered studies published from 2015 to 2025, since this period includes all the latest advancements in the field of machine learning (ML) techniques as well as their use in emission monitoring. Specifically, eligible studies focused on the use of ML to predict or monitor emissions in the refinery operations, with practical applications or empirical evidence rather than pure theoretical discussions. For the data collected, preference was given to peer-reviewed journal articles, conference papers, and authoritative industry reports that have validated and reliable data.

On the other hand, reviews of studies that were based on theoretical frameworks only, or on empirical evidence that was lacking, were excluded to maintain the review's practical relevance. Also excluded from the study were studies not focused on refinery emissions or air pollution monitoring. Review articles, opinion pieces, and editorials were also excluded that are devoid of original data to avoid speculative or anecdotal findings. To ensure consistency and quality of analysis, all duplicate studies and studies without a clear methodological framework were also disregarded. The review was able to synthesize robust, high-quality literature that addresses the research objectives effectively because of these carefully defined criteria.

4. Data extraction and thematic analysis

After the selection process, systematic data extraction was carried out to get the key information from each study, including objectives of research, ML techniques used, data sources, evaluation metrics, and findings. The data were extracted and organized systematically to support the facilitation of comparison and synthesis.

The findings from the included studies were categorized and synthesized by means of a thematic analysis. The approach made it possible to detect recurring patterns, variations, and thematic trends concerning ML-based emission prediction. Specific themes are not disclosed, but insights were grouped into three core themes by means of thematic analysis. From a thorough review of the extracted data, the themes that emerged describe the complexities of applying ML to emission monitoring in refinery operations.

4.1. The role of machine learning and AI in environmental monitoring

With the increasing demand from air pollution to have accurate, timely, and cost-effective ways to monitor them, the integration of machine learning (ML) and artificial intelligence (AI) into environmental monitoring has become relevant. However, the sources of pollution from refinery operations are very important, as they are important sources of CO_2, SOC, NO_x, volatile organic compounds (VOCs), and particulate matter (PM) [1]. They negatively affect air quality, ecosystems, and public health, leading to respiratory and cardiovascular diseases, and an increase in mortality rates [3]. As a result, the Environmental Protection Agency (EPA) and the European Environment Agency (EEA), for example, have to regulate the emission limits to control these impacts [4].

Traditional monitoring methods are mostly based on manual sampling, onsite inspection, and periodic reporting, which are time-consuming, expensive, and lack real-time data [11]. The disadvantages of these limitations may cause underreporting of emissions, and not achieve adequate regulatory compliance or environmental protection [6]. The refinery processes are complex and therefore, more sophisticated, data-driven approaches are needed. Predictive analytics, real-time monitoring, and automation can all be advanced AI and ML systems that will help the emission monitoring to be more accurate and efficient.

Decision trees, random forests, and neural networks, among other machine learning models, have been capable of analyzing large amounts of data, predicting emission levels, and suggesting preventive measures [7]. Unlike traditional methods, which tend to be reactive, ML-based predictive analytics offers the ability of proactive decision making by creating and revealing patterns and falling movements,

forecasting potential breaches as well as optimizing operational performance [11]. A novel approach to environmental management [5] is for instance, used in a complex refinery environment, where ML models contributed to an increase in prediction accuracy as well as in compliance monitoring. However, there are challenges to the advancements made. However, there are challenges when it comes to integrating ML for emission monitoring, including data privacy concerns, lack of infrastructure, and lack of skilled personnel [1]. Furthermore, research on how to effectively combine ML into existing monitoring frameworks is lacking [16].

Using the Technology Acceptance Model (TAM) and Data Driven Decision Making (DDDM), the theoretical framework for the application of ML to the emission monitoring can be explained. Previous research suggests that TAM assumes that the perceived usefulness (PU) and perceived ease of use (PEOU) of ML systems would influence stakeholders' acceptance and utilization of these technologies [17]. In the context of the refinery operations, PU is the ability of ML to accurately predict and monitor emission levels, whereas PEOU relates to the ease of use of the technology by stakeholders [18]. ML can seem too complex or unreliable to engineers and managers, and those systems are not as likely to be adopted. As a result, if verified, the capability of ML models to generate timely and accurate forecasts will increase the acceptance and integration into decision-making processes.

In conjunction with TAM, DDDM emphasizes the use of empirical data for informed decision making rather than intuition [19]. ML algorithms taking real-time data from sensors and historical records can provide emission patterns predictions that allow stakeholders to take interventions in time to avoid regulatory breaches [20]. In this context, the effectiveness of DDDM depends on data and continuous data quality, staff and organizational readiness, and expertise. Erroneous data analysis or the lack of knowledge of ML can adversely affect decision-making [21].

4.2. Machine learning models for emission prediction: effectiveness and comparative analysis

Emission prediction has become a powerful(prediction of emissions) technique that facilitates the use of machine learning (ML) to overcome historically inefficient techniques and slow response time with their monitoring methods. As the Environmental Protection Agency (EPA) and the European Environment Agency (EEA) impose more and stricter measures on reducing the environmental impacts associated with industrial operations [4], these advanced systems are becoming more important. While all of these types of ML models have shown great predictive power in terms of deriving more accurate and timelier emission estimates, some of them, such as traditional methods like linear regression (LR), and others like long short-term memory (LSTM) networks and convolutional neural networks (CNN), have other benefits such as their hardware efficiency since they require less storage and fewer computation. ML allows organizations to optimize compliance and minimize environmental risk as well as informed decision making on real-time data [22].

The selection of different ML models depends highly on both data complexity and computational resources, as well as interpretability requirements. LR is effective for linear relationships but is not appropriate when faced with the complex non-linear dynamics in emission data. The DTs are easy to understand and interpret but tend to overfit, especially when there is little data available. Support Vector Machines (SVM) are highly successful in dealing with nonlinear datasets through kernels but are computationally expensive and rather sensitive to parameter tuning. LSTM is a more advanced neural network, which is good for time series data; they capture long-term dependencies and reduce prediction errors; hence, they are an important tool to be used in tracking the complex emission patterns over time [23]. LSTM networks have been proven to be more efficient than simple models such as SVR and DTs, leading to lower false positives and negatives, which are essential for regulatory compliance.

Reduction of bias and variance has been shown by ensemble methods like Random Forest (RF) and Gradient Boosting Machines (GBM) to outperform in predicting. For instance, RF and GBM have been employed as good predictors of particulate matter ($PM_{2.5}$) and also greenhouse gas emissions compared to simpler models like SVM and LR [24]. It has been proven that the gradient boosting framework XGBoost achieved a high R^2 value of 0.9 and a test accuracy of 99%, which makes it a reliable tool in forecasting emission levels and detecting anomalies [23]. The emission spikes, which can be irregularly wide, can be detected by these advanced models very rapidly, so that timely corrective measures can be taken to prevent environmental harm. Nevertheless, such neural networks as well as ensemble models are usually overly complex, and therefore lack interpretability, which complicates the interpretation of the rationale of the predictions for stakeholders. Because decision-making at regulatory environments must be supported by clear justifications, this lack of transparency can make the acceptance of these models difficult.

Though ML-based emission monitoring promises, model effectiveness relies on the quality of input data. Due to issues such as noise, incomplete datasets, inaccuracies, and more, misleading predictions can arise, which can cause inappropriate interventions or noncompliance with regulatory standards [25]. All of this makes the use of advanced ML models less feasible for smaller refineries or in more resource-constrained areas, especially computational demands and the need for domain-specific expertise [26]. Regarding the performance, complex models such as LSTM and XGBoost perform better, but simpler models like LR and DT can be more practical regarding the contexts where interpretability is more important than prediction accuracy.

A solution to these challenges is to navigate them by a balanced approach, i.e., prioritizing models according to local operational context and regulatory requirements. Domain knowledge's integration into model training can help the predictive accuracy, whereas XAI (Explainable AI) methods could address interpretability issues and increase the stakeholders' trust. In the end, the choice of ML models for emission prediction is limited to considering tradeoffs between accuracy, interpretability, and computational complexity to achieve the best performance under practical real-world applications.

4.3. Challenges and practical considerations in implementing ML-driven emission monitoring systems

Emission monitoring for refinery operations held primarily by machine learning has a lot of technical, ethical, and practical challenges that have to be critically addressed. The quality and availability of data is one of the foremost barriers to the use of ML, as it is essential to the training of robust ML models. Refinery emission data often have incomplete emission data or are even unavailable because of proprietary restrictions [15], [25]. Even more data scarcity is imposed by a lack of monitoring infrastructure in many developing regions, from which it is impossible to develop these accurate models. Such data preprocessing techniques of outlier detection, normalization, and imputation are essential in ensuring a reliable prediction, but come with critical expertise and could be time-consuming [16]. This reliance on high quality data points to a central barrier to success: given no reliable inputs, no matter how sophisticated or complex the ML algorithm (LSTM network; or ensemble method in this case, e.g., XGBoost), it is impossible to achieve stable and accurate emission forecasts [23].

Data issues, however, are just one of them; the computational demand of advanced ML models, especially for deep learning architectures, is another substantial challenge. Implementing and maintaining these systems require high-performance computing infrastructure and technical expertise, which might not be accessible for smaller refineries or those located in resource-constrained regions [26]. These applicational constraints not only limit the scalability of ML solutions but also make it more difficult to integrate such solutions into currently

deployed online operations and invocation of operational systems. Additionally, most complex models — such as the "black box" nature of many of them, like deep neural networks — make it less transparent and understandable. In the regulatory setting, where stakeholders should grasp and justify the decision-making process, this transparency impedes the approval of ML predictions, leading to a lack of trust in such systems [8]. There are ethical and regulatory concerns on top of the technical challenges. As these data sources range from industrial sensors to satellite imagery, the integration of them presents significant privacy issues, especially under consideration of General Data Protection Regulation (GDPR) [1]. Data security is a must, and it is very important to avoid unauthorized access to sensitive operational data, which could result in industrial espionage or other legal repercussions. With such concerns, there are robust data governance frameworks that have to balance analytical accuracy with the requirement to protect sensitive information.

The technology acceptance model (TAM) is one theoretical framework, and data-driven decision making (DDDM) is the other that will help to overcome these challenges. Musa et al. (2024) [18] emphasize that the perceived usefulness and ease of use of ML systems are the main factors that influence stakeholder acceptance. For engineering systems and managers, if it is too complex and not reliable, adoption rates suffer. In contrast, DDM recommends using empirical data to make decisions, and it is only as effective as data quality and organizational readiness [20]. In practice, it turns out that raising a data-driven culture will require continuous learning in a data-driven mindset, protecting data at all costs, and working together strongly in an interdisciplinary team.

The path forward, based on a promising combination of traditional manual monitoring with ML-based predictive analytics, is in hybrid approaches. Within this integration of well-proven methods with new ML approaches, organizations can increase the accuracy of the data, cut down on errors, and minimize cost and compliance management [16]. In the end, ML and AI have the promise to change emission monitoring in refinery operations, but solving problems with data quality, computational requirements, ethical considerations, and stakeholder acceptance are critical to reap the full benefits of these technologies. Solving these problems in an involving, transparent, and ethically charged manner will be important in the progress of sustainable, effective, and proactive emission monitoring systems. One potential path forward is hybrid approaches—combining traditional manual monitoring with ML-based predictive analytics. Organizations can integrate the best of known methods with the latest ML methods to achieve higher data accuracy, lower errors, and lower costs while also meeting compliance [16]. Finally, while ML and AI can transform emission monitoring in refinery operations, hurdles present in data quality, computational strategy, ethical preoccupation, and acceptance by stakeholders should be conquered to derive the best from this. Solutions to such issues will present significant collaborative, transparent, and ethically grounded methods of advancing sustainable, effective, and proactive emission monitoring systems.

4.4. Practical Implications

The practical consequences of deploying ML-based emission monitoring systems are varied, spanning from refinery operations to regulatory compliance to environmental sustainability. Adoption of ML technologies for refineries can have a huge impact on the improvement of operational efficiency, as ML technologies can provide real-time and accurate predictions of emission levels. This capability provides the operators with the ability to proactively choose what to do to address irregular emission spikes, minimize environmental violations, and optimize resource utilization. Research reveals that using simplified advanced ML, e.g., a LSTM network or XGBoost, can increase the accuracy of predictive assumptions to a level sufficient to almost eliminate the false positives and negatives, thus enabling compliance with the stringent environmental regulation [23].

While such systems are shown to be very successful, they rely on substantial investments in more advanced computational infrastructure as well as in skilled personnel and continued training to keep the models accurate. Firms with smaller refineries and limited financial capacity may not be able to afford these investments, and disparities in adoption of technology may result in possible regulatory noncompliance. As a result, this inequality could lead to competitive disadvantages for smaller refineries, which are not as well placed to exploit ML for the sustainable operations of their operations.

From a regulatory perspective, the lack of interpretability of complex ML models can make them less accepted by the stakeholders and hinder their use in formal compliance processes. Because it lacks interpretability, which is typical of deep neural networks, regulators are unable to understand the outcome of the prediction, which could result in disputes or resistance from industry stakeholders [8]. Therefore, it is important to develop explainable AI (XAI) techniques and hybrid approaches that combine ML with traditional monitoring methods, and as a result, build stakeholder trust and accountability.

Second, there should be careful management of the ethical implications of data privacy and security. The attack of unauthorized access to sensitive operational data can cause industrial espionage, legal consequences, and cart notoriety. As a result, the implementation of ML-driven emission monitoring systems requires strict data governance and compliance with privacy regulations like GDPR to protect sensitive information and achieve the full potential of the system.

5. Conclusion

ML-driven emission monitoring systems offer great promise in terms of improving the accuracy, efficiency, and responsiveness of environmental monitoring, especially in industrial settings such as refineries. Proactive compliance management through such models as LSTM networks, XGBoost, and hybrid ML-DL techniques is also possible. Nevertheless, these come with high implementation costs, ethical concerns about data privacy, issues with data quality, and model interpretability. Complex ML models are 'black-box' in nature–stakeholders may resist, and compliance is more difficult. Furthermore, it is hampered by small organizations in resource-scarce regions in need of specialized technical expertise, computational capacity, and sustainable data infrastructure.

This challenge needs an interwoven effort made between them technologically, regrettably, and ethically. To deploy ML more effectively and sustainably, the hybrid monitoring systems, supporting interdisciplinary collaboration, data preprocessing, improving data quality, and investments in capacity building, can be implemented. In addition, Explainable AI techniques and data protection regulations should be developed to improve the model transparency and accountability. ML-driven emission monitoring can make a strategic contribution to environmental sustainability and regulatory adherence in industrial operations with the help of proper planning and collaboration.

5.1. Recommendation

Therefore, several strategic recommendations are proposed to maximize the benefits and minimize the challenges of realizing the ML-driven emission monitoring systems. In the first place, refineries should develop a strong data infrastructure to support the collection, processing, and storage of high-quality data. Technology providers can collaborate with small refineries to develop cost-effective solutions

to provide small refineries with access to advanced ML technologies. In addition, partnering with academic institutions and research organizations can facilitate acquiring domain-specific knowledge, can be used to augment model training, and can address the data shortage problem via initiatives under open data. Additionally, applying a hybrid monitoring technique where traditional techniques are coupled with ML-based prediction will boost the model's reliability and accuracy. Validation of ML predictions is possible using traditional monitoring methods, significantly reducing the chance of false positives or negatives, and ensuring compliance with the regulatory standards. It is also important to develop XAI techniques that assist in making complex ML models more interpretable and attain transparency, thereby enhancing stakeholder trust.

Refineries will need to adopt strict data governance policies that include definitions, privacy by design, and policies that address regulatory and ethical concerns to guarantee the protection of sensitive operational data, for instance, in line with GDPR. More protection against data breaches and unauthorized access can be received if personnel are trained to handle data in a more ethical manner and on cybersecurity, as well as data privacy. Also, collaboration with regulatory agencies to develop clear guidelines around how to use ML-driven systems for compliance monitoring can ease the adoption and reduce resistance to it from stakeholders.

Finally, ML models need constant training and the capacity building of personnel to keep and optimize them. To be equipped with the right skills to operate and maintain ML-driven systems, investment in professional development, workshops, and certifications can be made. Refineries can fully use ML technologies to create a culture of innovation, continuous learning, and ethical responsibility, to improve emission monitoring, support sustainable practices, and remain compliant in an evolving regulatory landscape.

References

[1] F.M. Adebiyi, Air quality and management in petroleum refining industry: A review, Environmental Chemistry and Ecotoxicology 4 (2022) 89–96. https://doi.org/10.1016/j.enceco.2022.02.001.

[2] A. Bello, F. Magi, O. Abaneme, U. Achumba, A. Obalalu, M. Fakeyede, Using Business Analysis to Enhance Sustainability and Environmental Compliance in Oil and Gas: A Strategic Framework for Reducing Carbon Footprint, JETIA 10(50) (2024) 76–85. https://doi.org/10.5935/jetia.v10i50.1303.

[3] Manisalidis, E. Stavropoulou, A. Stavropoulos, E. Bezirtzoglou, Environmental and health impacts of air pollution: A review, Frontiers in Public Health 8 (2020) Article 14. https://doi.org/10.3389/fpubh.2020.00014.

[4] A. Gouldson, A. Carpenter, S. Afionis, Environmental leadership? Comparing regulatory outcomes and industrial performance in the United States and the European Union, Journal of Cleaner Production 100 (2015) 278–285. https://doi.org/10.1016/j.jclepro.2015.03.080.

[5] R.A. Tavella, F.M.R. da Silva Júnior, M.A. Santos, S.G.E.K. Miraglia, R.D. Pereira Filho, A review of air pollution from petroleum refining and petrochemical industrial complexes: Sources, key pollutants, health impacts, and challenges, ChemEngineering 9(1) (2025) Article 13. https://doi.org/10.3390/chemengineering9010013.

[6] A. Ragothaman, W.A. Anderson, Air quality impacts of petroleum refining and petrochemical industries, Environments 4(3) (2017) Article 66. https://doi.org/10.3390/environments4030066.

[7] M.R. Hasan, M.Z. Islam, M.F.I. Sumon, M. Osiujjaman, P. Debnath, L. Pant, Integrating artificial intelligence and predictive analytics in supply chain management to minimize carbon footprint and enhance business growth in the USA, Journal of Business and Management Studies 6(4) (2024) 195–212. https://doi.org/10.32996/jbms.2024.6.4.17.

[8] J.L. Calderon, C. Sorensen, J. Lemery, C.F. Workman, H. Linstadt, M.D. Bazilian, Managing upstream oil and gas emissions: A public health-oriented approach, Journal of Environmental Management 310 (2022) 114766. https://doi.org/10.1016/j.jenvman.2022.114766.

[9] A.H. Al-Moubaraki, I.B. Obot, Corrosion challenges in petroleum refinery operations: Sources, mechanisms, mitigation, and future outlook, Journal of Saudi Chemical Society 25(12) (2021) 101370. https://doi.org/10.1016/j.jscs.2021.101370.

[10] A. Audu, A. Umana, The role of environmental compliance in oil and gas production: A critical assessment of pollution control strategies in the Nigerian petrochemical industry, International Journal of Scientific Research Updates 8(2) (2024) 36–47. https://doi.org/10.53430/ijsru.2024.8.2.0061.

[11] Bell, C., Ilonze, C., Duggan, A., & Zimmerle, D. (2023). Performance of continuous emission monitoring solutions under a single-blind controlled testing protocol. *Environmental Science & Technology*, *57*(14). https://doi.org/10.1021/acs.est.2c09235.

[12] Bello, A. A., Fakeyede, N. M., Gold, O., Eshun, N. V., Akibor, J., & Owusu, N. F. (2025). Optimizing agile collaboration frameworks for carbon-efficient digital twin deployment in oil and gas: Strategies, tools, and challenges in the planning phase. *Global Journal of Engineering and Technology Advances*, *22*(2), 034–045. https://doi.org/10.30574/gjeta.2025.22.2.0025.

[13] Dechezleprêtre, A., & Sato, M. (2017). The impacts of environmental regulations on competitiveness. *Review of Environmental Economics and Policy*, *11*(2), 183–206. https://doi.org/10.1093/reep/rex013.

[14] Samola, M. (2025). ML-based predictive analytics: Enhancing data-driven strategies in various industries. *Cyber and Education Research*. https://www.researchgate.net/publication/389546808_ML-Based_Predictive_Analytics_Enhancing_Data-Driven_Strategies_in_Various_Industries.

[15] Olawade, D. B., Wada, O. Z., Ige, A. O., Egbewole, B. I., Olojo, A., & Oladapo, B. I. (2024). Artificial intelligence in environmental monitoring: Advancements, challenges, and future directions. *Hygiene and Environmental Health Advances*, *12*, 100114. https://doi.org/10.1016/j.heha.2024.100114.

[16] Li, X., Shen, X., Jiang, W., Xi, Y., & Li, S. (2024). Comprehensive review of emerging contaminants: Detection technologies, environmental impact, and management strategies. *Ecotoxicology and Environmental Safety*, *278*, 116420. https://doi.org/10.1016/j.ecoenv.2024.116420.

[17] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008.

[18] Musa, H. G., Fatmawati, I., Nuryakin, N., & Suyanto, M. (2024). Marketing research trends using technology acceptance model (TAM): A comprehensive review of researches (2002–2022). *Cogent Business & Management*, *11*(1), Article 2329375. https://doi.org/10.1080/23311975.2024.2329375.

[19] Mandinach, E., Honey, M., & Light, D. (2006). A theoretical framework for data-driven decision making. Retrieved from https://www.researchgate.net/publication/252996939_A_Theoretical_Framework_for_Data-Driven_Decision_Making.

[20] Elragal, A., & Elgendy, N. (2024). A data-driven decision-making readiness assessment model: The case of a Swedish food manufacturer. *Decision Analytics Journal*, *10*, 100405. https://doi.org/10.1016/j.dajour.2024.100405.

[21] Dodman, S. L., Swalwell, K., DeMulder, E. K., View, J. L., & Stribling, S. M. (2021). Critical data-driven decision making: A conceptual model of data use for equity. *Teaching and Teacher Education*, *99*, 103272 https://doi.org/10.1016/j.tate.2020.103272.

[22] Kumari, S., & Singh, S. K. (2023). Machine learning-based time series models for effective CO_2 emission prediction in India. *Environmental Science and Pollution Research*, *30*(116), 116601–116616. https://doi.org/10.1007/s11356-022-21723-8.

[23] Bharathi, V. P. N., Muthuswamy, K., Natarajan, B., Sheela, M. S., Jothiprakash, G., Shanmugam, K., Loganathan, K., Vasudevan, B., Appavu, S., Marimuthu, R., Rajaram, D., & Ranjan, S. (2025). A comparative analysis and prediction of carbon emission in India using machine learning models. *Global NEST Journal*, *27*(4), 06020.

[24] Zhang, T., He, W., Zheng, H., Cui, Y., Song, H., & Fu, S. (2021) Satellite-based ground $PM_{2.5}$ estimation using a gradient boosting decision tree. Chemosphere, 268, 128801. https://doi.org/10.1016/j.chemosphere.2020.128801.

[25] Ali, M., Mukarram, M. M. T., Chowdhury, M. A., Karin, S., & Faruq, A. N. (2021). Integration & implication of machine learning: Barriers to aid environmental monitoring & management. *Open Access Library Journal*, *8*(6). https://doi.org/10.4236/oalib.1107468.

[26] Aniceto, K. (2025). The role of artificial intelligence (AI) and machine learning (ML) in the oil and gas industry. *Journal of Technology and Systems*, *7*(1), 6–27. https://doi.org/10.47941/jts.2493.