IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 January 2025, accepted 8 March 2025, date of publication 13 March 2025, date of current version 21 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3551259

RESEARCH ARTICLE

Predicting Atmospheric Dispersion of Industrial Chemicals Using Machine Learning Approaches

MARÍA VALLE[1], JAIRO A. CARDONA[1], (Member, IEEE), CÉSAR VILORIA-NUÑEZ[2], (Senior Member, IEEE), AND CHRISTIAN G. QUINTERO M.[1], (Senior Member, IEEE)

[1]Department of Electrical and Electronics Engineering, Universidad del Norte, Barranquilla 081007, Colombia
[2]School of Digital Transformation, Universidad Tecnológica de Bolívar, Cartagena 130001, Colombia

Corresponding author: Christian G. Quintero M. (christianq@uninorte.edu.co)

This work was supported in part by the Project Strengthening Mutual Help for Industrial Sector Companies Through a Information and Communication Technology (ICT) System for Risk Reduction and Technological Emergency Management With Impact on Communities of Atlantico, Colombia—Royalty Project under Grant 2021-53723.

ABSTRACT This study presents an intelligent framework for assessing atmospheric dispersion in industrial accident scenarios involving chemical substances. The research focuses on modeling the dispersion of key chemicals, such as chlorine, methanol, and propane, under various accident conditions, including leaks, fires, and explosions. Atmospheric and contextual variables, such as wind speed, air temperature, tank specifications, and chemical release parameters, were thoroughly characterized to construct a robust database using experimental data and software simulations. Machine learning techniques were rigorously trained and tested to predict atmospheric dispersion, emphasizing hyperparameter optimization to enhance model performance. Dimensionality reduction methods, such as principal component analysis and correlation-based dimensionality reduction, were implemented to improve computational efficiency, reduce data noise, and maintain essential information. Results demonstrate the effectiveness of the proposed approach, with satisfactory predictions across all evaluated risk areas. Key contributions include the development of a replicable framework adaptable to diverse industrial scenarios, applying hyperparameter tuning to optimize model accuracy, and integrating dimensionality reduction techniques to streamline data processing. These advancements establish a foundation for future studies to incorporate additional chemicals and accident scenarios, improving the flexibility and reliability of atmospheric dispersion modeling. Future work will explore hybrid machine learning models and advanced dimensionality reduction methods to enhance the system's applicability to complex industrial environments.

INDEX TERMS Atmospheric dispersion, intelligent prediction, industrial emergencies, technological risks, machine learning models.

I. INTRODUCTION

The rapid industrial expansion observed in recent decades has increased the frequency of sudden accidents, often involving releasing hazardous chemical substances. These incidents pose significant risks to the environment and nearby populations [1]. As a result, accurately predicting the atmospheric dispersion of these substances has become critical for assessing their potential impact and implementing preventive measures.

Emergencies involving hazardous materials are a reality and can occur in any context [1]. For this reason, it is necessary to strengthen the disaster risk management (DRM) stages carried out before, during, and after these events. Such efforts aim to minimize their consequences, protect public health and safety, and preserve environmental balance.

In the industrial sector, tools like the ALOHA software are widely used to simulate the atmospheric dispersion of chemical accidents and forecast dispersion patterns. While

The associate editor coordinating the review of this manuscript and approving it for publication was Frederico Guimarães.

ALOHA offers valuable insights, it is not without limitations. For example, the software does not support integration with external platforms, preventing automatic data entry and the extraction of simulation results. Furthermore, its database of optional chemicals for simulation is static and cannot be expanded, restricting its adaptability to new industrial scenarios.

Given the widespread use of ALOHA across industries nationwide, there is a clear need for innovative approaches to enhance the atmospheric dispersion modeling process. Adopting next-generation tools, such as machine learning techniques, is a promising solution. These methods, characterized by their efficiency and robust data analysis capabilities, can optimize dispersion predictions without relying on the predefined equations required by traditional software like ALOHA. Machine learning improves adaptability and ensures more accurate and efficient processing of large datasets.

This study proposes developing a machine-learning-based framework to predict the dispersion of hazardous chemical substances in various industrial accident scenarios. By leveraging advanced algorithms, the goal is to generate accurate and efficient predictions of atmospheric dispersion patterns under variable conditions.

The contributions of this work are threefold:

1) A comprehensive characterization of atmospheric dispersion processes and the inherent challenges in their modeling.
2) A proposed solution applying machine learning techniques to address these challenges, emphasizing accuracy and computational efficiency.
3) Validation of the proposed framework's effectiveness using experimental data and software simulations, demonstrating its potential for industrial applications.

The remainder of this paper is organized as follows: Section II presents a theoretical background about atmospheric dispersion, ALOHA, and Machine Learning concepts. Section III presents background information on atmospheric dispersion models and their limitations, emphasizing the ALOHA software. Section IV describes the machine learning techniques employed in the proposed framework. Section V presents the proposed approach implementation. Section VI presents the results and validation process, while Section VII concludes with a discussion of key findings and potential future research directions.

II. THEORETICAL BACKGROUND

A. ATMOSPHERIC DISPERSION

Atmospheric dispersion refers to the spread of pollutants in the atmosphere. Understanding this process is crucial for air quality, public health, and regulatory decisions. Various models, such as Gaussian, Euler, CFD (Computational Fluid Dynamics), and Lagrangian, predict the concentration and trajectory of the pollutant. These models help forecast pollutant levels, trace sources of pollution, and support regulatory and epidemiological applications related to air quality [2].

The limitations of atmospheric dispersion modeling include the following:

- **Simplified assumptions**: Dispersion models use simplified assumptions about air pollutant behavior and atmospheric conditions, which may not capture the complexities of the real world, leading to inaccuracies [3].
- **Uncertainties in meteorological estimations**: Meteorological estimations in dispersion models have limitations and uncertainties. The precision of weather forecasts, the representativeness of meteorological data, and the variability of atmospheric conditions introduce uncertainties into the results [3].
- **Sensitivity to input parameters**: Dispersion models are sensitive to input parameters like emission rates, stack heights, and meteorological data. Small parameter changes can significantly affect results, underscoring the need for accurate and reliable data [4].
- **Lack of real-time data**: Dispersion models often use historical or averaged meteorological data, missing short-term atmospheric variations. Real-time data offers more accurate, current information for modeling [4].

Despite these limitations, atmospheric dispersion modeling is crucial in air quality management. It offers insights into pollutant dispersion and helps assess emission impacts on air quality. Recognizing these limitations allows modelers to make informed decisions and interpretations.

Accurate atmospheric dispersion modeling needs various data to predict air pollutant concentrations. Key data requirements include:

- **Meteorological data**: Atmospheric dispersion models need meteorological data like wind speed, direction, air temperature, stability, and mixed layer heights [4].
- **Emissions data**: Dispersion models need precise emissions data from sources like industrial plants, vehicular traffic, or accidental chemical releases [5].
- **Topography**: Dispersion models may need local topography data, such as the height and slope of nearby hills or mountains [6].

B. ALOHA® SOFTWARE

ALOHA® (Areal Locations of Hazardous Atmospheres) is the hazard modeling program for the CAMEO® software suite (Computer-Aided Management of Emergency Operations), which is used widely to plan for and respond to chemical emergencies, it receives information about an actual or potential chemical release and then generates estimations about the threat zones (see Figure 1). ALOHA can model toxic gas clouds, flammable gas clouds, BLEVE (Boiling Liquid Expanding Vapor Explosions), jet fires, swimming pool fires, and vapor cloud explosions. Hazard area estimates are displayed on a grid in ALOHA and can also be plotted on maps in MARPLOT, Esri's ArcMap, Google Earth, and Google Maps [7].

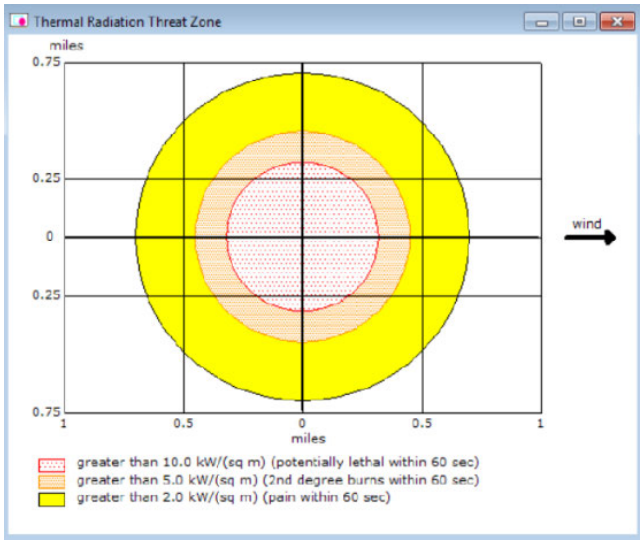

FIGURE 1. ALOHA threat zone chart [7].

ALOHA has several limitations, including its inability to integrate with other platforms for automated data input and result extraction and its incapacity to incorporate chemicals not included in the software's predefined list.

C. MACHINE LEARNING CONCEPTS

Machine learning, a subfield of AI, uses data and algorithms to enable computers to learn from experience without explicit programming [8].

It employs statistical techniques to identify patterns and make predictions or decisions based on data. Algorithms are trained on datasets to make accurate predictions or decisions. Training data help adjust the algorithm's settings to reduce errors and enhance accuracy. Models are evaluated on their predictions for new, unseen data [9]. The evaluation identifies model strengths and weaknesses to enhance performance. Optimization adjusts parameters to minimize errors and boost accuracy [9]. Machine learning's key concepts include supervised learning, unsupervised learning, reinforcement learning, evaluation, and optimization.

This study employs Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Multi-Layer Perceptron (MLP) as selected machine learning techniques. The Multi-Layer Perceptron (MLP) model has three layers: input, hidden, and output. The input layer processes signals, the output layer handles tasks like prediction and classification, and the hidden layers perform the main computations. Data flow forward from input to output. Neurons are trained using backpropagation. MLPs approximate continuous functions and solve nonlinearly separable problems, with applications in pattern classification, recognition, prediction, and approximation [10].

Support Vector Regression (SVR), derived from Support Vector Machines (SVM), extends SVM to regression and function estimation. SVR aims to minimize the deviation between the predicted and actual responses for the training and unseen data. Its foundation in structural risk minimization (SRM) reduces both empirical risk (training error) and the Vapnik-Chervonenkis (VC) dimension for better generalization. SVR uses a high-dimensional feature space and a training approach that minimizes a generalized error bound, including training errors and a regularization term [11]. This makes SVR robust and effective for various applications.

A Random Forest is a collection of independent tree-structured classifiers, each constructed using identically distributed random vectors. These vectors guide the tree growth and each tree votes for the most popular class [12].

The K-Nearest Neighbors (KNN) algorithm is a simple and effective nonparametric classification technique used in supervised learning. It relies on a labeled dataset divided into classes to predict the class of unlabeled data. In regression, KNN predicts continuous values by averaging the values of its K nearest neighbors. KNN is useful in data sets with distinct groups, especially when prior knowledge of the data is lacking [13].

III. RELATED WORK

The research about machine learning for atmospheric dispersion prediction is crucial. In toxic dispersion, PHAST™ - based approaches (a paid software for risk analysis) have created a database of 30,022 toxic release scenarios with 19 chemicals. This database supports a prediction model for downwind dispersal distances, which showcases the effectiveness of deep learning in toxic dispersion analysis [14].

Models also predict the behavior of chemical substances using machine learning. Reference [15] presents an ecological risk assessment framework for river ecosystems, emphasizing the potential of machine learning to predict the reproductive toxicity of chemical endocrine disruptors. The results show that the SVM model outperforms the correlation model in predictive precision.

Reference [16] used SVR models to estimate the gas dispersion. However, a considerable challenge lies in determining suitable input variables for these models. Consequently, two specific inputs were meticulously selected: the original monitoring parameters and the integrated Gaussian parameters. In particular, the findings of this study revealed that the adoption of integrated Gaussian parameters resulted in enhanced performance within the machine learning models employed.

Predicting atmospheric dispersion limits is essential. A study develops a neural network model for predicting dispersion, focusing on variables, time aspects, and facility design. Analyzing these elements creates a robust framework for forecasting dispersion patterns. Advanced techniques capture complex variable relationships, improving prediction accuracy and understanding of dispersion, enabling proactive risk mitigation [17]. Model combinations can refine results. For instance, Reference [18] presents a prediction algorithm

combining SVM with a Gaussian model, which adequately predicts atmospheric dispersion. Conversely, an investigation applied the five most prevalent machine learning methodologies for air pollution prognosis, wherein the MLP demonstrated superior performance [19].

Reference [20] presents results that predict air pollution using three years of air quality and meteorological data. The nonlinear autoregressive neural network outperformed other models. The authors highlight the need to select relevant meteorological variables and continuously update the model. Alternative approaches to predictive modeling, such as those explored by the authors in [19], involve the application of regression techniques to determine the extent of pollution under various meteorological conditions. Similarly, the work of [21] adapts this model within the agricultural sector to forecast potential dispersion patterns and provide preemptive alerts to farmers about potential plant diseases. Reference [22] proposes a superior multipoint deep learning model using convolutional long short-term memory for predicting air pollution. Effective air quality management requires ambient air monitoring to assess current parameters, enforce regulations, identify non-compliance, and devise corrective strategies [23]. It aids in understanding natural processes like pollution dilution, dispersion, and wind-driven movement. The Air Quality Index (AQI) quantifies air pollution and informs the public. A higher AQI signifies greater health risks. Accurate prediction of AQI is based on air monitoring and pollutant levels over specific intervals. This study uses data preprocessing, the ANOVA F-test for feature selection, and ensemble-based classifiers to forecast AQI from historical air quality data. The model significantly improves the accuracy of the AQI prediction, improving atmospheric forecasting.

In [24], the challenge of predicting atmospheric dispersion in accidental hazardous chemical releases is addressed. It reviews past incidents and key concepts, summarizes experiments, and classifies prediction models into three categories. Special attention is given to atmospheric detection research. Relevant professional software is introduced, and its efficiency and accuracy are discussed. Future trends highlight image processing, database creation, algorithm improvements, and a comprehensive emergency response system.

Managing pure hydrogen leaks is crucial due to fire or explosion risks when hydrogen contacts air. In this context, [25], [26] characterized turbulent hydrogen flows from leaks using machine learning. An empirical-analytical-numerical model described the mass, momentum, and concentration flows through integral formulas, transformed into ordinary differential equations and solved numerically. This facilitated acquiring critical parameters like the molar fraction of hydrogen. ML algorithms, including linear regression (LR), ANN, SVR, KNN, RF, RT, and REP tree (REPT), were evaluated to predict hydrogen concentration dispersal in the air. The RF algorithm proved to be the best for forecasting hydrogen leak distribution.

TABLE 1. Related work compilation.

Reference	Prediction	Artificial intelligence technic	Correlation study	Dataset	Evaluation metrics
[14]	Toxic dispersion	RF, NN	Yes	30022	RMSE, MSE, R2
[15]	Chemical reproductive toxicity	SVM	No	-	MAE, MSE, EVS, R2
[16]	Gas dispersion	BP, SVR	Yes	23900	R2, NMSE
[17]	Leak gas dispersion	NN	No	150	Accuracy
[18]	Contaminant dispersion	BP, SVM, RBF	Yes	-	MSE, COR
[19]	Air pollution	MLP, SVR, RF, KNN	Yes	-	RMSE
[20]	Air pollution	SVM, ANN	Yes	-	R2
[19]	Gaussian plume	Regression model	No	-	MSE
[21]	Air pollution	Regression model	Yes	-	ESDD
[22]	Air quallity	NN	Yes	~ 500000	RMSE, MAE
[23]	Air quallity	Ensemble learning	No	-	Accuracy,
[24]	Atmospheric dispersion	ANN, KNN, RF	Yes	-	RMSE
[25]	Hydrogen Leak	DT, RF, GBRr y ANN	Yes	-	RMSE, MSE, R2, MAE
[26]	Hydrogen Leak	LR, ANN, SVR, KNN, RF, RT	Yes	-	R, MAE, RAE, RMSE, RRSE

Table 1 summarizes the related work, highlighting machine learning techniques, evaluation metrics, and dataset size, among other factors.

Common models include MLP, RF, SVM, and SVR. Many cited studies use correlation analyses to reduce dimensionality and enhance model performance. In particular, data set sizes vary in the reviewed literature. Studies such as [14], [16], and [22] use extensive data sets with thousands to hundreds of thousands of data points, while [17] shows effective training with smaller data sets. This highlights the versatility and effectiveness of machine learning methods in different scenarios, even with limited data.

IV. PROPOSED APPROACH

The industrial boom in recent decades has led to more frequent accidents, releasing chemicals harmful to the environment and nearby populations [1]. Understanding their atmospheric dispersion is crucial for assessing impacts and implementing preventive measures.

Worldwide, many companies use ALOHA software to model atmospheric dispersion during chemical accidents and predict dispersal patterns. Using modern tools to enhance this process is essential. Machine learning tools offer an efficient and data-driven alternative, eliminating the need for

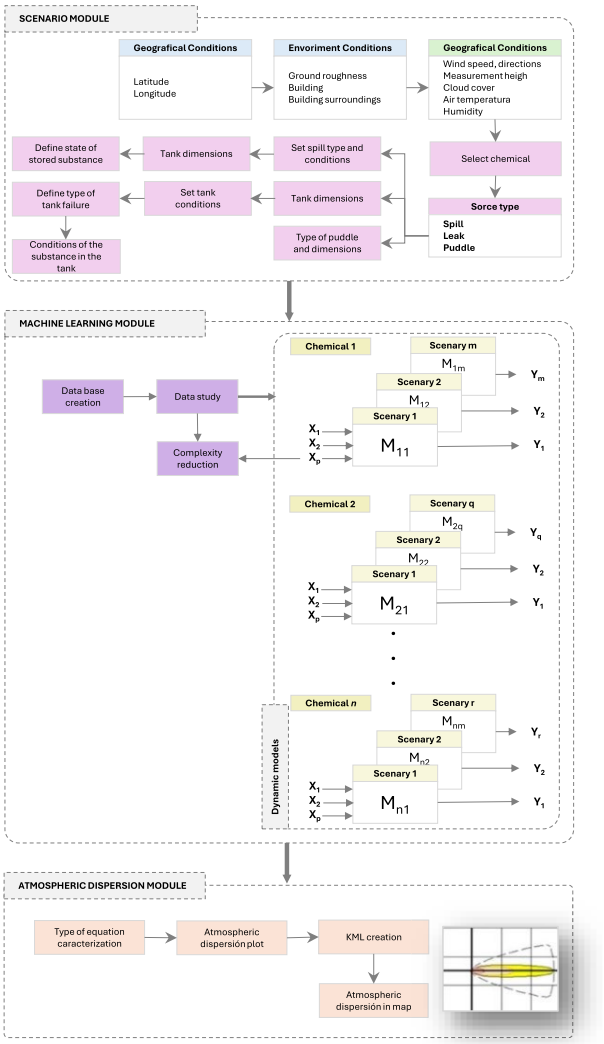

FIGURE 2. Block diagram of the proposed approach.

ALOHA's formulas to simulate scenarios. This study investigates the development of a machine-learning solution for simulating chemical substances in various technological risk scenarios. The objective is to produce accurate atmospheric dispersion graphs according to the scenario conditions. In this sense, this study aims to determine potential impact distances for implementing proactive measures such as community preparedness, alert systems, and environmental protection. Figure 2 shows the three key modules of the system. The "*Scenario module*" defines accident scenarios with various chemicals. The "*Machine Learning module*" then uses the scenario to predict dispersion distances, displayed in the *"Atmospheric Dispersion Plot module"*.

A. SCENARIO MODULE

The selection of substances for this study is based on specific criteria to ensure both the relevance and compatibility of the modeled scenarios with the ALOHA tool. First, chemical substances reported by companies in the APELL Barranquilla Group are evaluated according to their hazard classification under the United Nations Globally Harmonized System (GHS) [27]. Only substances classified as hazardous in liquid or gaseous states are considered, as ALOHA does not support simulations involving solid substances [28].

A second key criterion is the availability of these substances in ALOHA's library, which contains over 800 chemical compounds. This filtering process verified the compatibility of reported substances with the tool, resulting in the exclusion of more than 80% of the initially listed substances, as they are not included in the simulator's database.

Finally, the selection prioritized substances that represented relevant scenarios in terms of industrial usage frequency and potential impact in the event of accidents. Preliminary simulations also helped identify substances with broader dispersion footprints, ensuring their representativeness for analyzing high-risk chemical scenarios.

Based on these criteria, three main substances are selected. Propane is chosen due to its extensive use as a fuel in industrial processes and its potential for frequent flammable leaks [29]. Methanol, a highly flammable substance commonly used in chemical applications, represents critical scenarios involving liquid spills [30]. On the other hand, chlorine is selected for its gaseous dispersion capacity, as evidenced by simulations showing more extensive dispersion footprints than those of the other substances, making it a representative case for high-impact situations [31].

To simplify the computational modeling process, this study focused on approximately circular dispersion footprints [32]. While this approach reduces computational complexity, it provides a practical and reliable basis for assessing the behavior of hazardous substances under typical industrial accident scenarios. However, more complex dispersion shapes, such as irregular or elongated patterns, could be analyzed using advanced computational resources.

The modeled scenarios reflect common industrial accidents in Barranquilla, such as propane leaks, chlorine spills, chlorine leaks, methanol spills, and methanol leaks. Additionally, global events underscore the relevance of these substances: for instance, a train derailment in Tampa carrying 115,000 liters of propane gas (Figure 3-A) [33], a chlorine gas leak at a port in Jordan causing 13 fatalities (Figure 3-B) [34], and the sinking of a barge carrying methanol in the Ohio River (Figure 3-C) [35]. In addition, Figure 3-D captures a chlorine truck crash in Medellin's Western Tunnel [36]; Figure 3-E depicts a truck fire involving 800 liters of methanol in Mexico City, successfully contained by firefighters [37], and Figure 3-F shows a derailed train carrying propane near the Florida airport, with no injuries or leaks reported [38].

The methodology used is sufficiently flexible to accommodate other substances and scenarios, provided they meet the compatibility requirements of ALOHA or similar tools. This includes expanding the simulator's database or employing alternative modeling software to incorporate substances not originally considered.

FIGURE 3. Recent accidents involving selected chemicals.

This flexibility ensures that the methodology can be applied to a broader range of scenarios, considering diverse industrial processes, storage conditions, and environmental settings. Future research could explore these extensions to enhance the utility and generalization of the simulator for chemical risk management.

B. MACHINE LEARNING MODULE

Scenarios and their parameters are defined, then Python code generates the cases from ALOHA simulation and records the outputs in CSV files to build the models. Studying the correlation between inputs and outputs is crucial to reduce data set complexity and identify influential inputs. The system uses synthetic data from ALOHA scenarios. Machine learning techniques like RF, KNN, SVR, and MLP are implemented to find the most suitable model for the experiments, selected from a state-of-the-art review due to their satisfactory results in similar problems. Given the problem's complexity, determining the best technique beforehand is challenging. Steps in the machine learning module are the following:

1) Obtain scenario variables.
2) Pre-process available data.
3) Identify significant variables.
4) Implement machine learning models.
5) Select the best model.
6) Make predictions.

Equations (1) and (2) represent the mathematical representation of the overall machine learning module. Here, Q is the chemical, S the scenario, M the machine learning model, X the input variables, and Y the output. In addition, p is the number of inputs (X), n the number of chemicals (Q), and m the number of scenarios (S).

$$X = [x_1, x_2, \ldots x_p] \tag{1}$$

$$\forall Q \exists S : M_{nm}(X) \approx Y_m \tag{2}$$

C. ATMOSPHERIC DISPERSION MODULE

Python is used to generate a KML (Keyhole Markup Language) file that graphically describes danger zones for each scenario. The KML document is designed for the representation of geographic data in applications like Google Earth and GIS (Geographical Information Systems). It defines threat zones with different chemical concentrations, labeled "*Yellow Threat Zone*," "*Orange Threat Zone*," and "*Red Threat Zone*." Each zone is a polygon with geographical coordinates, style information, and metadata, including an identifier ("*MarplotID*"). In addition, there is an ALOHA source point on the map.

This KML document visually depicts hazardous areas and source points relevant to chemical dispersion modeling in GIS or mapping tools.

TABLE 2. Related nomenclature for chemicals, scenarios, and models.

Chemical		Scenario		Model
Name	Nomenclature	Name	Nomenclature	Nomenclature
Chlorine	Q_1	Spill	S_1	M_{11}
		Leak	S_2	M_{12}
Methanol	Q_2	Puddle	S_3	M_{21}
		Leak	S_4	M_{22}
Propane	Q_3	Toxic Area	S_5	M_{31}
		Blast	S_6	M_{32}

TABLE 3. Characterization of inputs meteorological variables.

Variables	Range or Options
Building	Enclosed office building, Single storied building, Double storied building
Building surroundings	Unsheltered surroundings, Sheltered surroundings.
Location	Vía 40
Wind speed (m/s)	3, 4, 5, 6
Wind direction	NE, NNE, ENE, ESE
Measurement height (m)	3
Ground Roughness	Urban or forest, Open country
Cloud cover (Scale)	0, 3, 5, 7, 10
Air temperature (°C)	Between 18 and 35
Humidity (%)	Between 77 and 87

V. IMPLEMENTATION

A. DATA COLLECTION

Table 2 aligns the accidents with the nomenclature in Figure 2. The following sections will use this nomenclature for chemical accidents.

Table 3 lists the meteorological variables needed to reproduce chemical accidents. It also includes other variables based on the accident's origin and their variation ranges, determined by software options, geographic area, and expert judgment.

Tables 4 to 6 list the variables needed for accident simulations based on the scenarios and chemicals evaluated.

The characterized variables form the input vector (X) for the proposed models.

TABLE 4. Q_1 scenarios.

Scenario	Variables	Range or Options
S_1	Source	Instantaneous or continue
	Amount of pollutant	Instantaneous: 1ton. Continue: 1ton/min, 68kg/min
	Source height (m)	Length of the tank
S_2	Tank direction	Horizontal or vertical
	Length (m)	Between 0.665 and 1.8
	Diameter (inches)	Between 0.25 and 0.35
	Chemical state	Liquid, gas or unknown
	Temperature within the tank (°C)	Ambient temperature
	Amount of gas (Kg)	1 ton or 68 Kg
	Liquid volume (%)	Between 0 and 100
	Hole or short	Pipe or valve
	Bottom of the leak (%)	Between 0 and 100

TABLE 5. Q_2 scenarios.

Scenario	Variables	Range or Options
S_3	Type of puddle	Burning puddle (poolfire)
	Puddle diameter (m)	Between 0 and 100
	Puddle volume (m3)	Between 0 and 10000
	Initial puddle temperature (°C)	Between 0 and 40
S_4	Tank direction	Horizontal or vertical
	Length (m)	Between 0.665 and 1.8

TABLE 6. Q_3 scenarios.

Scenario	Variables	Range or Options
S_5	Tank direction	Horizontal
	Length (m)	1
	Diameter (inches)	1,25
	Chemical state	Liquid
	Temperature within the tank (°C)	Ambient temperature
	Amount of gas (Kg)	1 ton or 68 Kg
	Liquid volume (%)	Between 0 and 95
	Type of tank failure	Leaking tank, not burning
	Opening shape	Circular
	Opening diameter (inches)	Between 6 and 24
	Release time (minute)	1
	Max average sustained release rate (g)	Between 700 and 880
	Hole or short	Pipe or valve
	Bottom of the leak (%)	Between 0 and 100
	Hazard to analyze	Toxic area of vapor cloud
S_6	Tank direction	Horizontal
	Length (m)	1
	Diameter (inches)	1,25
	Chemical state	Liquid
	Temperature within the tank (°C)	Ambient temperature
	Amount of gas (Kg)	1 ton or 68 Kg
	Liquid volume (%)	Between 0 and 95
	Type of tank failure	Leaking tank, not burning
	Opening shape	Circular
	Opening diameter (inches)	Between 6 and 24
	Release time (minute)	1
	Max average sustained release rate (g)	Between 700 and 880
	Hole or short	Pipe or valve
	Hazard to analyze	Blast

B. DATA PREPROCESSING

First, the CSV files are cleaned to ensure that no columns have empty data. Variables with constant values are removed. To handle categorical features, one-hot encoding transforms each category into a separate binary column (0 or 1), indicating its presence or absence. This enables the regression model to effectively use categorical features. One-hot encoding prevents artificial order in categorical data and helps the model capture relevant relationships between categories and the target variable in the regression.

Next, two methods are evaluated to reduce the dimensionality of the input data: studying variable correlations and analyzing the principal components.

1) CORRELATION STUDY

A correlation study is performed on the variables of each scenario to evaluate the reduction in input dimensionality and identify the key variables that influence the outputs.

Figures 4 – 9 show the correlation matrix for each scenario. The variables in these matrices are described in Table 7.

- **Chlorine spill**: For this scenario (see Figure 4), there is a high correlation (0.67 to 0.81) between "*Amount pollutant entering the atmosphere*" and output distances. A 0.78 correlation is also seen between "*Duration*" and "*Release duration*", highlighting its importance. Additionally, there is a 0.80 correlation between "*Duration*" and spill type (*continuous* or *instantaneous*). Since "*Duration*" is zero for instantaneous spills and greater than zero for continuous spills, the variables "*Continuous*" and "*Instantaneous*" can be discarded.

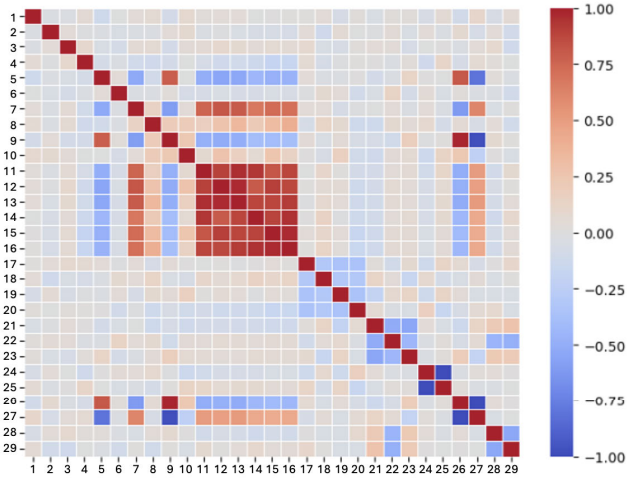

FIGURE 4. Chlorine spill correlation.

TABLE 7. Table summarizing variables for different scenarios.

Variable Number	Chlorine spill	Chlorine leak	Methanol leak	Methanol puddle	Propane toxic area	Propane BLAST
1	Wind speed	Wind speed	Wind speed	Wind speed	Wind speed	Wind speed
2	Cloud cover	Cloud cover	Cloud cover	Cloud cover	Cloud cover	Cloud cover
3	Air temperature	Air temperature	Air temperature	Air temperature	Air temperature	Air temperature
4	Humidity	Humidity	Humidity	Humidity	Humidity	Humidity
5	Duration	Diameter	Diameter	Puddle Diameter	Liquid volume	Liquid volume
6	Length	Length	Length	Puddle volume	Opening diameter	Opening diameter
7	Amount pollutant entering the atmosphere	Liquid Volume	Liquid Volume	Initial Puddle Temperature	Bottom of the leak	Bottom of the leak
8	Rate	Bottom of the leak	Shape diameter	Red Vertical distance	Max avg release time	Max avg release time
9	Release duration	Shape diameter	Opening length	Yellow Vertical distance	Temperature within tank	Temperature within tank
10	Total amount released	Opening length	Opening width	Orange Vertical distance	Release Duration	Release Duration
11	Red Vertical distance	Opening width	Bottom of the leak	Red Horizontal distance	Max Average Sustained Release Rate	Max Average Sustained Release Rate
12	Yellow Vertical distance	Red Vertical distance	Temperature within tank	Yellow Horizontal distance	Total Amount Released	Total Amount Released
13	Orange Vertical distance	Yellow Vertical distance	Red Vertical distance	Orange Horizontal distance	Yellow Vertical distance	Yellow Vertical distance
14	Red Horizontal distance	Orange Vertical distance	Yellow Vertical distance	ENE	Orange Vertical distance	Orange Vertical distance
15	Yellow Horizontal distance	Red Horizontal distance	Orange Vertical distance	ESE	Yellow Horizontal distance	Yellow Horizontal distance
16	Orange Horizontal distance	Yellow Horizontal distance	Red Horizontal distance	NE	Orange Horizontal distance	Orange Horizontal distance
17	ENE	Orange Horizontal distance	Yellow Horizontal distance	NNE	ENE	ENE
18	ESE	Double storied	Orange Horizontal distance	Urban or forest	ESE	ESE
19	NE	Enclosed office	ENE	Open country	NE	NE
20	NNE	Single office	ESE	Double storied	NNE	NNE
21	Double storied	Sheltered	NE	Enclosed office	Double Storied	Double Storied
22	Enclosed office	Unsheltered	NNE	Single office	Enclosed office	Enclosed office
23	Single office	Urban or forest	Urban or forest	Sheltered	Single office	Single office
24	Urban or forest	Open country	Open country	Unsheltered	Urban or forest	Urban or forest
25	Open country	Hole	Circular	–	Open country	Open country
26	Continue	Valve	Rectangular	–	Horizontal	Horizontal
27	Instantaneous	Circular	Double storied	–	Vertical	Vertical
28	Sheltered	Rectangular	Enclosed office	–	Sheltered	Sheltered
29	Unsheltered	–	Single storied	–	Unsheltered	Unsheltered
30	–	–	Sheltered	–	Hole	Hole
31	–	–	Unsheltered	–	Valve	Valve

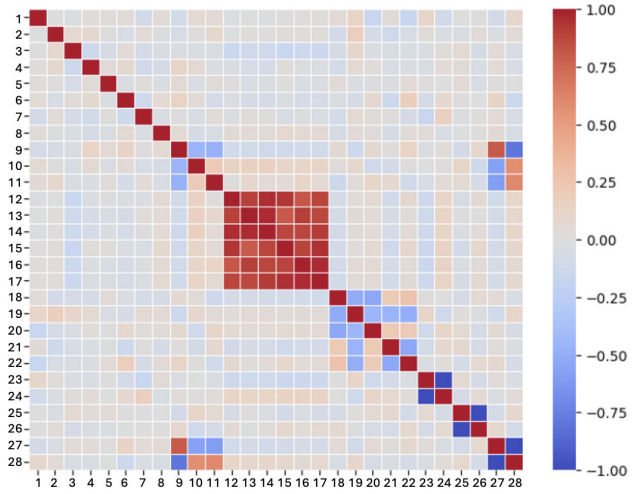

FIGURE 5. Chlorine leak correlation.

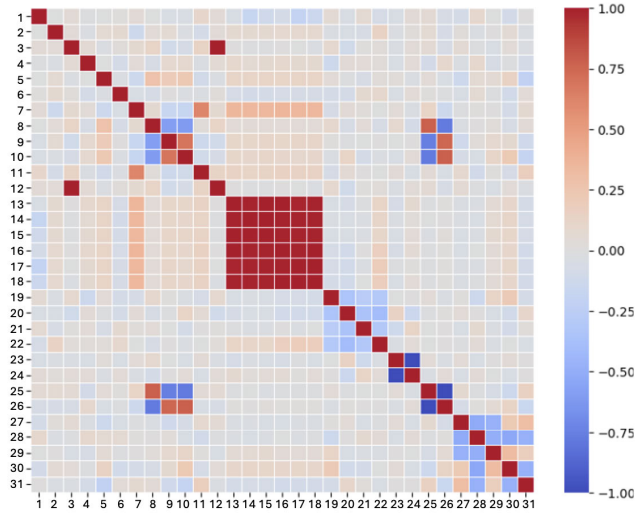

FIGURE 6. Methanol leak correlation.

- **Chlorine leak**: No high correlations exceeding 0.8 are observed (see Figure 5).
- **Methanol leak**: Variables "*Air temperature*" and "*Temperature within tank*" are totally correlated; hence, one variable is redundant (see Figure 6).
- **Methanol puddle**: Figure 7 shows a high correlation between "*Puddle Diameter*" and all predicted distances, indicating its significant impact on atmospheric dispersion. There is also a total correlation between "*Initial Puddle Temperature*" and "*Air Temperature*," making one of them redundant.
- **Propane toxic area**: For a propane leak (see Figure 8), a 0.97 correlation between "*Air temperature*" and "*Temperature within tank*" suggests redundancy. "*Liquid Volume*" significantly affects the "*Total amount released*" output.
- **Propane BLAST**: In a propane BLAST leak (Figure 9), a 0.95 correlation between "*Air temperature*" and "*Temperature within the tank*" suggests one variable could be eliminated. "*Liquid Volume*" is also highly relevant to the "*Total amount released*".

Table 8 summarizes the correlation study in our research. It describes various models with initial input variables and their reduction in correlation. In particular, the correlation reduction is modest across all models. This suggests that correlation-based dimensionality reduction is insufficient for our objectives. Thus, we should explore alternative methods like Principal Component Analysis (PCA) to refine our models.

2) PRINCIPAL COMPONENTS ANALYSIS

PCA is a technique to reduce dimensionality in data analysis and machine learning [39]. It simplifies the representation of

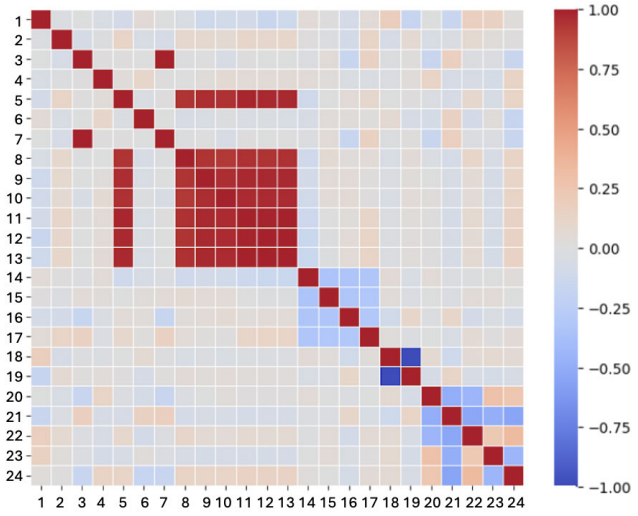

FIGURE 7. Methanol puddle correlation.

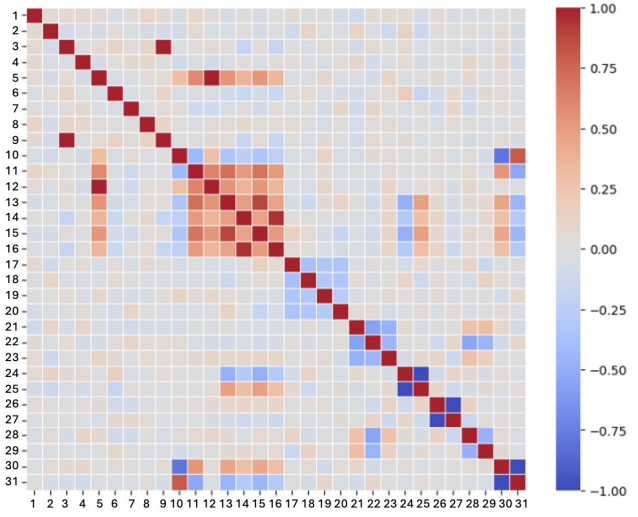

FIGURE 8. Propane toxic area correlation.

TABLE 8. Resume of the correlation study.

Model	Input variables	Correlation reduction
M_{11}	20	18
M_{12}	25	25
M_{21}	18	17
M_{22}	25	24
M_{31}	22	21
M_{32}	26	25

the data set while retaining most of the original variability by transforming the data into principal components, which are linear combinations of the original variables. PCA uses the covariance matrix of the original features to understand their relationships. Features with high covariance change together, while those with low covariance vary independently.

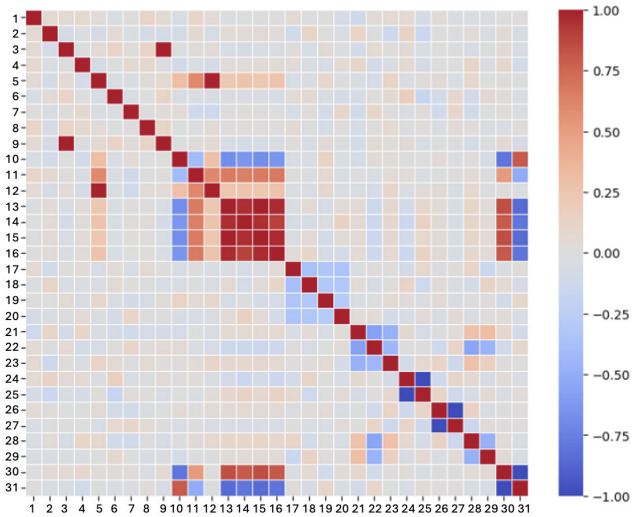

FIGURE 9. Propane BLAST correlation.

Comprehensive experiments are conducted with specific configurations of the machine learning model. A MLP with a single hidden layer of 10 neurons uses the ReLU activation function, a regularization term (alpha) of the maximum iterations of 0.001, 100 and a random seed (`random_state`) of 1 for reproducibility.

A RF model with 10 trees is used, with a maximum tree depth of 5 and a minimum of 2 samples to split a node. A random seed ensures the consistency of the experiment. The KNN algorithm uses $K = 5$, considering the five nearest neighbors for predictions. Uniform weights are used, and the optimal neighbor search method is auto-selected. Leaf size is set to 20 for `kd_tree` and `ball_tree` algorithms and Euclidean distance ($p = 2$) is used for similarity calculations. A SVR with a RBF kernel is configured with $C = 1.0$ and *epsilon* $= 0.1$ to optimize predictions.

Each model is evaluated in PCA experiments to determine their effectiveness in dimensionality reduction and output prediction. These configurations provided a robust framework for comparing and selecting suitable models for our study.

Figures 10 to 15 show PCA application in the research scenarios, illustrating its effectiveness in reducing data dimensionality while preserving relevant information. The graph ranges equal to the original data size for each scenario, showing how PCA condenses information more compactly and simply, aiding analysis and interpretation.

MAE (Mean Absolute Error) evaluates model performance in the Figures by quantifying the average absolute differences between predictions and actual values. MAE is chosen for its ability to measure average prediction errors without regard to direction, making it suitable for assessing PCA model accuracy. It offers a clear measure of model effectiveness.

Tables 9 and 12 summarize the optimal number of principal components for each scenario, based on the MAE metric, showing how the choice of components impacts the quality of the prediction.

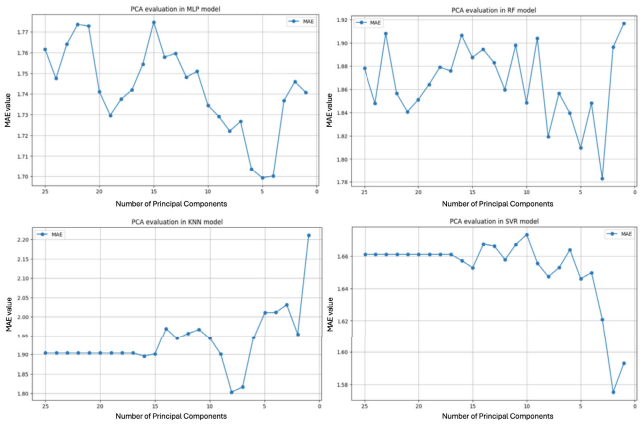

FIGURE 10. Chlorine spill PCA application.

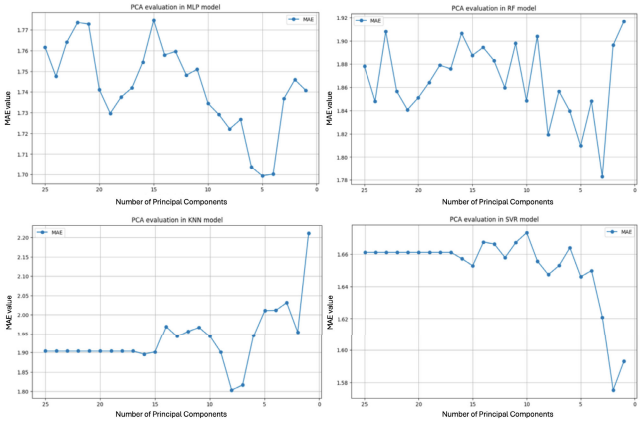

FIGURE 11. Chlorine leak PCA application.

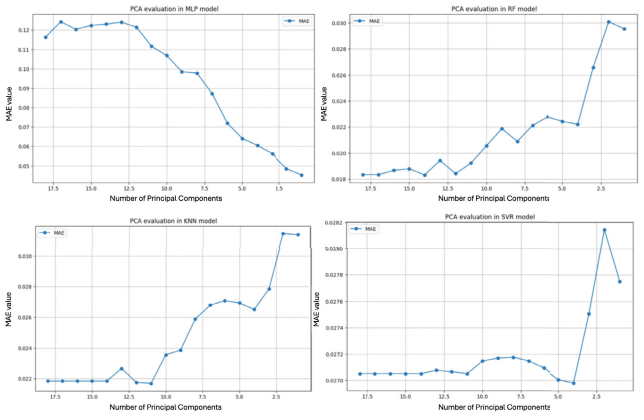

FIGURE 12. Methanol Puddle PCA application.

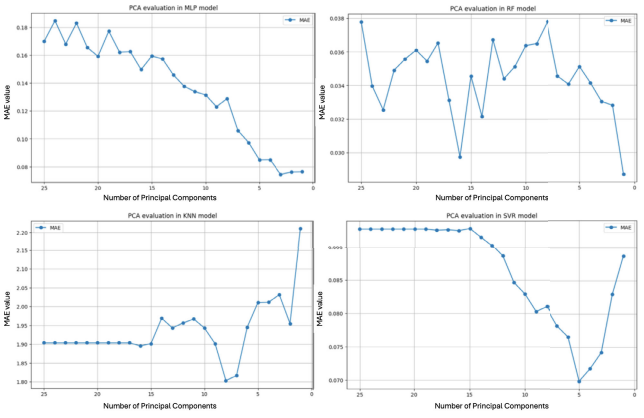

FIGURE 13. Methanol leak PCA application.

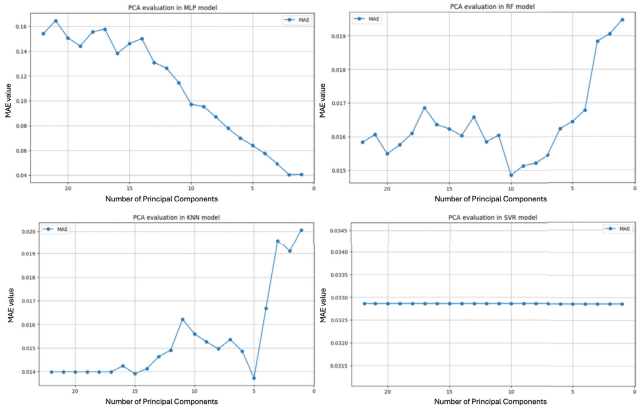

FIGURE 14. Propane toxic PCA application.

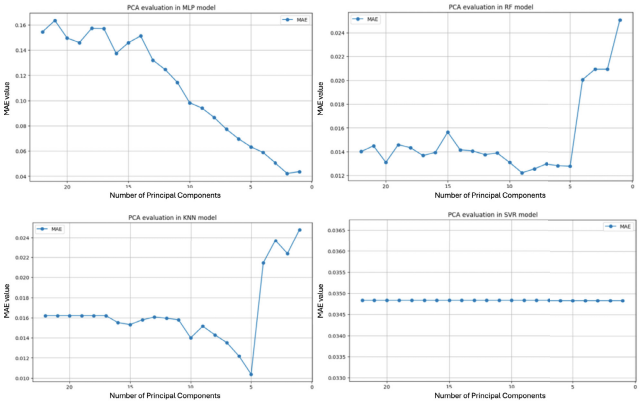

FIGURE 15. Propane BLAST PCA application.

Figures 16 to 21 show correlation matrices for the model with the fewest components per scenario. The correlation analysis between PCA components and target output assesses the relationship between the condensed information in the principal components and variables of interest. This correlation helps to determine how well the PCA components capture relevant variability for predictions. A significant correlation between a principal component and a target variable indicates that the component is crucial to predicting that variable. Conversely, a low or non-existent correlation suggests that the component contributes minimally to the prediction. This analysis aids in the identification of the most pertinent components for modeling and can guide feature selection to enhance the effectiveness of prediction models. The correlation analysis enriches the understanding of how

TABLE 9. PCA results for MLP.

Chemical	Scenario	N components	Best MAE
Q_1	S_1	16	1,01
Q_1	S_2	5	1,69
Q_2	S_3	1	0,04
Q_2	S_4	3	0,07
Q_3	S_5	2	0,04
Q_3	S_6	2	0,04

TABLE 10. PCA results for RF.

Chemical	Scenario	N components	Best MAE
Q_1	S_1	10	0,94
Q_1	S_2	3	1.78
Q_2	S_3	14	0,02
Q_2	S_4	1	0,02
Q_3	S_5	10	0,01
Q_3	S_6	9	0,01

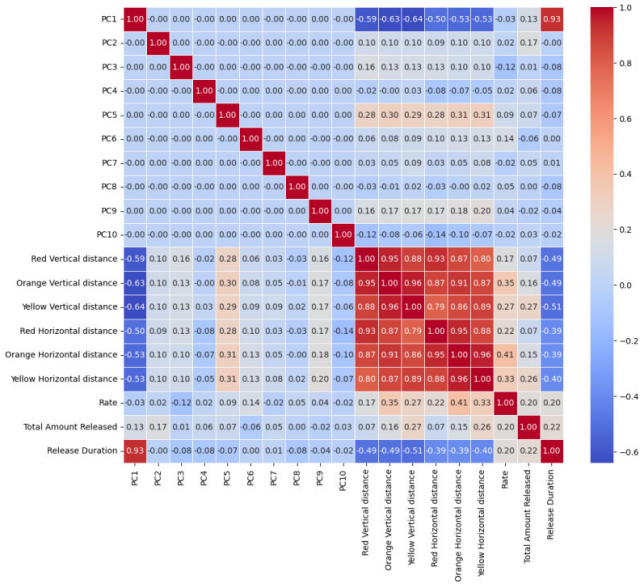

FIGURE 16. Chlorine spill PCA correlation.

TABLE 11. PCA results for KNN.

Chemical	Scenario	N components	Best MAE
Q_1	S_1	16	1,18
Q_1	S_2	8	1.80
Q_2	S_3	11	0,02
Q_2	S_4	14	0,02
Q_3	S_5	5	0,01
Q_3	S_6	5	0,01

PCA simplifies the data and its ability to precisely inform about the variables of interest.

TABLE 12. PCA results for SVR.

Chemical	Scenario	N components	Best MAE
Q_1	S_1	16	0,94
Q_1	S_2	2	1.57
Q_2	S_3	4	0,02
Q_2	S_4	5	0,06
Q_3	S_5	4	0,03
Q_3	S_6	4	0,03

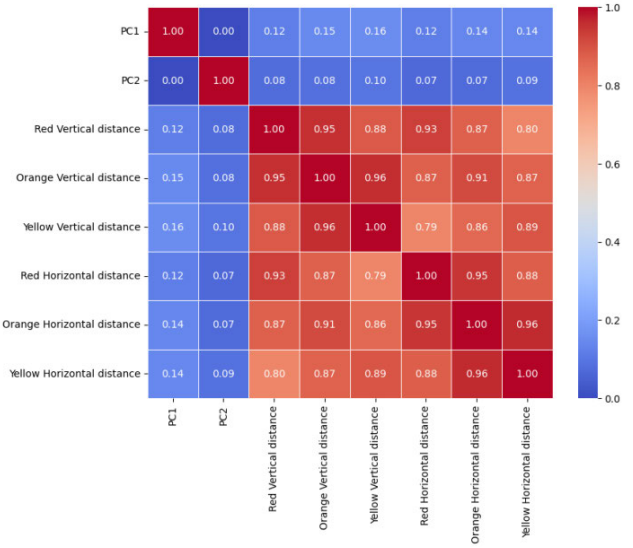

FIGURE 17. Chlorine leak PCA correlation.

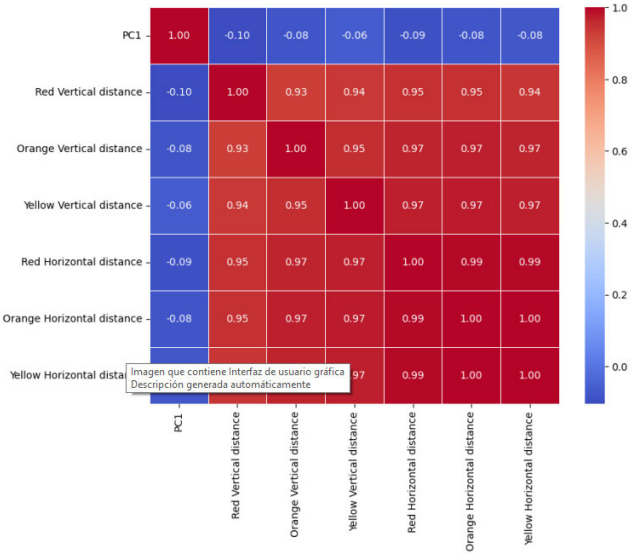

FIGURE 18. Methanol puddle PCA correlation.

C. HYPERPARAMETER TUNING

The performance of machine learning models largely depends on proper hyperparameter tuning. Methods like Grid Search, Random Search, and Bayesian Optimization are available

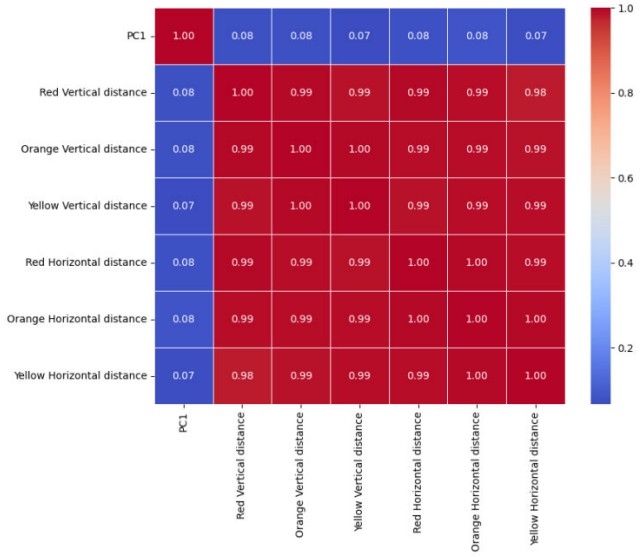

FIGURE 19. Methanol leak PCA correlation.

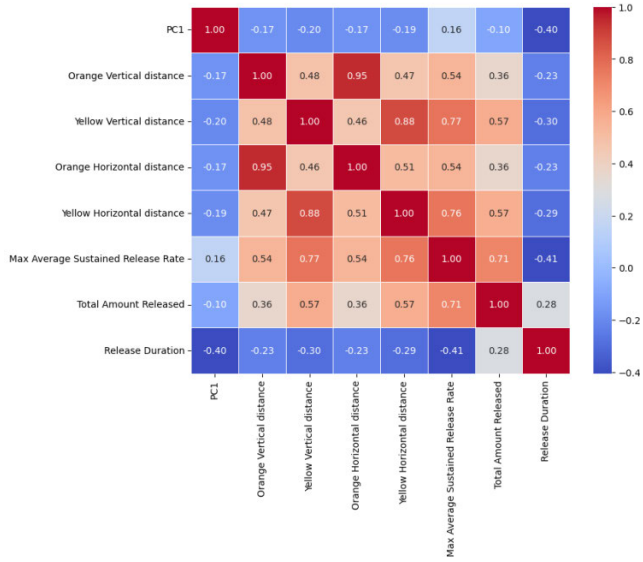

FIGURE 20. Propane toxic area correlation.

for this task. Optuna stands out as a powerful and versatile tool for hyperparameter optimization [40]. Optuna is a hyperparameter optimization framework designed for machine learning, featuring an imperative define-by-run-style user API for high modularity and dynamic search space construction, and advanced techniques like adaptive sampling and tree-based search to efficiently discover the best hyperparameter configurations, greatly reducing tuning time and improving model performance.

The following summarizes experimental results using various machine learning models for predicting chemical dispersion incidents. Each model is defined by hyperparameters, which influence its performance and generalization. These hyperparameters can be adjusted to fine-tune the model to

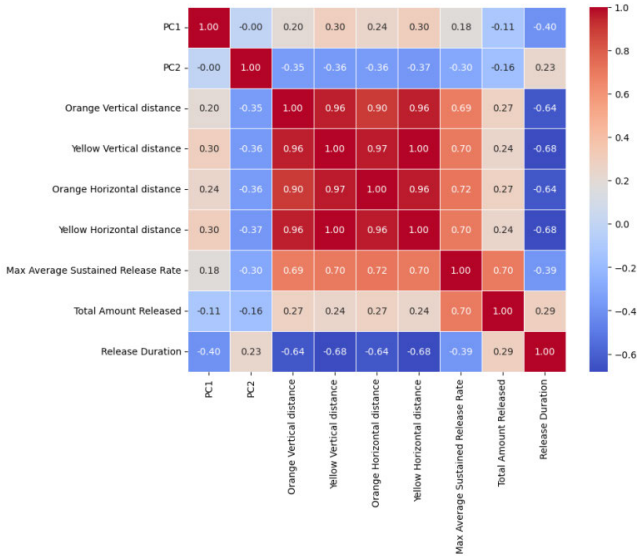

FIGURE 21. Propane blast PCA correlation.

the data and task requirements. Below is an explanation of the key hyperparameters used in each model:

1) **MLP (Multilayer Perceptron)**
   - **Activation**: The activation function in the hidden layers controls signal propagation. The `activation` parameter is tuned among `identity`, `logistic`, `tanh`, and `relu`.
   - **Alpha**: Alpha is a regularization term that penalizes network weights to prevent overfitting, explored on a logarithmic scale from 1^{-5} to 1^{-1}.
   - **Hidden layer size:** The number of neurons in hidden layers is controlled by the `n_layers` parameter, which varies from 1 to 5. Each layer can have 1 to 200 neurons.
2) **RF (Random Forest)**
   - **N estimator**: The number of trees in the random forest. The n estimator is varied from 1 to 200.
   - **Max depth**: The maximum depth of each tree, tuned within 1 to 32.
   - **Min samples split**: The minimum samples needed to split a node, varied between 0.1 and 1.0 during tuning.
3) **KNN (K-Nearest Neighbors)**
   - **N neighbors**: Number of neighbors for classification or regression, ranging from 1 to 20.
   - **Weights**: Weights assignment strategy: `Distance` or `Uniform`.
   - **P**: Controls the distance metric, varied between 1 and 2 during hyperparameter tuning.
4) **SVR (Support Vector Regression)**
   - **Kernel**: Kernel to map data to a higher-dimensional space: `linear`, `poly`, `rbf`, `sigmoid`.

- **C**: The penalty parameter C controls the trade-off between margin maximization and error minimization. Values of C vary logarithmically from 1^{-5} to 1^{2}.
- **Epsilon**: Defines the margin of error in regression, with hyperparameter tuning exploring values from 1^{-5} to 1^{1} logarithmically.

To improve the quality and interpretability of machine learning models, Tables 13 to 16 summarize the configuration of each model. They specify hyperparameters like activation functions for MLP, the number of estimators and maximum depth for RF, neighbors for KNN, and kernel selection for SVR. Proper hyperparameter selection enhances model performance, generalization, and prediction accuracy.

TABLE 13. Hyperparameter tuning results for MLP.

Chemical	Scenario	Activation	Alpha	Hidden layer size	Solver
Q_1	S_1	Tanh	0,00021	3	Adam
Q_1	S_2	Logistic	0,08558	3	Sgd
Q_2	S_3	Logistic	0,09825	3	Adam
Q_2	S_4	Logistic	220.153	2	Adam
Q_3	S_5	Logistic	0,07706	3	Sgd
Q_3	S_6	Tanh	0,06841	5	Sgd

TABLE 14. Hyperparameter tuning results for RF.

Chemical	Scenario	N estimator	Max depth	Min samples split
Q_1	S_1	145	2	0,25489
Q_1	S_2	70	2	0,54784
Q_2	S_3	80	13	0,14753
Q_2	S_4	82	14	0,78963
Q_3	S_5	102	10	0,45218
Q_2	S_6	111	9	0,93540

TABLE 15. Hyperparameter tuning results for KNN.

Chemical	Scenario	N neighbors	Weights	P
Q_1	S_1	3	Distance	1
Q_1	S_2	16	Uniform	1
Q_2	S_3	2	Distance	2
Q_2	S_4	19	Uniform	1
Q_3	S_5	7	Distance	1
Q_2	S_6	11	Distance	1

Tables 17 to 20 present the PCA experiment results on chemical dispersion scenarios, detailing hyperparameters for MLP, RF, KNN, and SVR. Tables show the PCA's effect on hyperparameters and model performance, highlighting the critical role of selecting the right number of principal components for effective dimensionality reduction and model prediction accuracy.

TABLE 16. Hyperparameter tuning results for SVR.

Chemical	Scenario	Kernel	C	Epsilon
Q_1	S_1	Linear	10	0,1
Q_1	S_2	Linear	5	0,3
Q_2	S_3	Poly	0,1	0,04
Q_2	S_4	Rbf	2	0,001
Q_3	S_5	Poly	3	0,1
Q_2	S_6	Poly	0,5	1

TABLE 17. Hyperparameter tuning results for MLP after PCA implementation.

Chemical	Scenario	Activation	Alpha	Hidden layer size	Solver
Q_1	S_1	Tanh	763.417	3	Adam
Q_1	S_2	Identity	0,00472	3	Sgd
Q_2	S_3	Identity	0,00871	3	Adam
Q_2	S_4	Logistic	0,01763	5	Adam
Q_3	S_5	Logistic	0,09973	2	Sgd
Q_2	S_6	Tanh	0,00053	3	Adam

TABLE 18. Hyperparameter tuning results for RF after PCA implementation.

Chemical	Scenario	N estimator	Max depth	Min samples split
Q_1	S_1	150	6	0,27854
Q_1	S_2	76	2	0,81107
Q_2	S_3	84	15	0,63752
Q_2	S_4	85	16	0,94337
Q_3	S_5	108	8	0,10037
Q_2	S_6	112	12	0,19025

TABLE 19. Hyperparameter tuning results for KNN after PCA implementation.

Chemical	Scenario	N neighbors	Weights	P
Q_1	S_1	13	Distance	1
Q_1	S_2	13	Uniform	1
Q_2	S_3	15	Uniform	1
Q_2	S_4	12	Uniform	1
Q_3	S_5	8	Distance	1
Q_2	S_6	8	Distance	1

D. OPTUNA ANALYSIS

The Optuna dashboard provides charts and information to analyze the hyperparameter optimization performance for

TABLE 20. Hyperparameter tuning results for SVR after PCA implementation.

Chemical	Scenario	Kernel	C	Epsilon
Q_1	S_1	Linear	1	0,01
Q_1	S_2	Rfb	50	0,01
Q_2	S_3	Poly	0,01	0,4
Q_2	S_4	Rbf	0,5	2
Q_3	S_5	Linear	10	0,1
Q_2	S_6	Linear	0,1	1

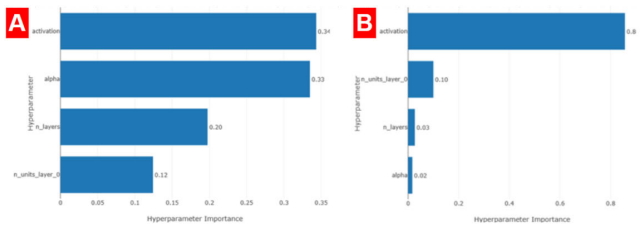

FIGURE 22. Hyperparameter importance graph for chlorine spill with MLP.

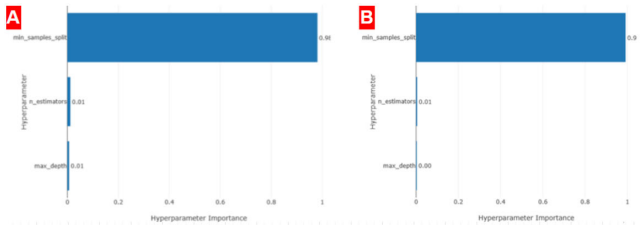

FIGURE 23. Hyperparameter importance graph for chlorine spill with RF.

various models. Here, we used the Chlorine Spill model with Optuna. Figures 22 to 25 show the importance of each hyperparameter for the chlorine spill scenario (S_1), highlighting their impact on the model performance.

In Figure 22 A, for an MLP model with correlation-based dimensionality reduction, the key hyperparameters are `activation` and `alpha`, both at 0.34. In Figure 22 B, with PCA applied, `activation` is the most important in 0.8.

In Figure 23 A, for an RF model using correlation-based dimensionality reduction, the "`min simple split`" hyperparameter is most important (importance: 0.98). Similarly, in Figure 23 B, where PCA is applied, "`min simple split`" also predominates (importance: 0.99).

Figure 24 A shows that for a KNN model with correlation-based dimensionality reduction, "`n neighbors`" is the most crucial hyperparameter (importance = 0.67). In contrast, Figure 24 B indicates that with PCA, `P` is the key hyperparameter (importance = 0.63).

Figure 25 A shows that for a SVR model with correlation-based dimensionality reduction, the most important hyperparameter is `kernel` (importance: 0.50). In Figure 25 B, with PCA applied, `C` is the most important hyperparameter (importance: 0.84).

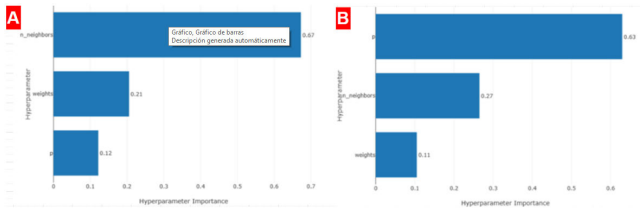

FIGURE 24. Hyperparameter importance graph for chlorine spill with KNN.

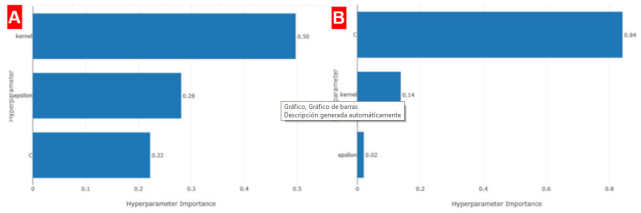

FIGURE 25. Hyperparameter importance graph for chlorine spill with SVR.

TABLE 21. Summary of results for chlorine scenarios.

		PCA reduction			Correlation reduction		
Scenario	Model	RMSE	Accuracy	MAE	RMSE	Accuracy	MAE
S_1	RF	0,36	95	0,94	0,31	96	0,78
	SVR	0,36	91	0,94	**0,29**	**98**	**0,37**
	KNN	0,70	89	1,18	1,12	90	1,35
	MLP	0,52	92	1,01	0,48	96	0,74
S_2	RF	0,93	85	1,78	**0,09**	**99**	**0,01**
	SVR	0,83	86	1,57	0,98	88	1,16
	KNN	1,04	87	1,80	0,58	85	5,41
	MLP	0,88	93	1,69	0,09	93	0,01

VI. EXPERIMENTS RESULTS AND ANALYSIS

Tables 21 to 23 summarize the results for the two scenarios analyzed for each chemical compound. Each Table presents RMSE, Accuracy, and MAE values for four models: RF, SVR, KNN, and MLP. Results are shown for both PCA and correlation-based dimensionality reduction. The best-performing model in each scenario is highlighted for clarity.

A. ANALYSIS

For chlorine in scenario 1 (S_1), the SVR model with correlation-based dimensionality reduction performs best with an RMSE of 0.29, Accuracy of 98, and MAE of 0.37. In scenario 2 (S_2), the RF model with correlation-based dimensionality reduction excels with RMSE of 0.09, Accuracy of 99, and MAE of 0.01. In both scenarios, superior results are achieved without PCA.

For methanol in scenario 3 (S_3), the RF model excels with an RMSE of 0.13, an accuracy of 97, and an MAE of 0.02. In scenario 4 (S_4), the KNN model excels with an RMSE of 0.16, an accuracy of 95, and an MAE of 0.02.

TABLE 22. Summary of results for methanol scenarios.

		PCA reduction			Correlation reduction		
Scenario	Model	RMSE	Accuracy	MAE	RMSE	Accuracy	MAE
S_3	RF	**0,13**	**97**	**0,02**	11,45	89	0,26
	SVR	0,14	96	0,02	1,65	75	1,45
	KNN	0,16	92	0,02	0,64	84	0,04
	MLP	0,20	95	0,04	1,21	69	0,85
S_4	RF	0,18	85	0,02	1,27	76	1,23
	SVR	0,27	90	0,06	1,81	88	1,67
	KNN	**0,16**	**95**	**0,02**	1,72	91	0,03
	MLP	0,30	91	0,07	0,47	65	0,33

TABLE 23. Summary of results for propane scenarios.

		PCA reduction			Correlation reduction		
Scenario	Model	RMSE	Accuracy	MAE	RMSE	Accuracy	MAE
S_5	RF	0,07	98	0,01	15,38	74	0,64
	SVR	0,15	96	0,03	1,87	92	0,05
	KNN	**0,06**	**99**	0,01	75,59	65	0,86
	MLP	0,19	94	0,04	27,40	85	0,23
S_6	RF	0,07	91	0,04	6,50	93	0,02
	SVR	0,15	90	0,03	8,33	81	0,24
	KNN	0,19	92	0,05	5,68	90	0,09
	MLP	**0,07**	**95**	**0,01**	16,34	79	0,41

For propane in scenario 5 (S_5), the KNN model is the top performer with an RMSE of 0.06, Accuracy of 99, and MAE of 0.01. In scenario 6 (S_6), the MLP model excels with an RMSE of 0.07, Accuracy of 95, and MAE of 0.01.

The best results for methanol and propane are achieved using PCA for dimensionality reduction, highlighting PCA's importance in improving model performance for these compounds.

B. FUNCTIONALITY

This section examines the system's capabilities by comparing ALOHA and our machine learning models for atmospheric dispersion estimation, and discussing the development of a Flutter-based web interface. The interface allows users to simulate chemical accidents and estimate atmospheric dispersion, rate, release duration, and total amount released.

1) **Methanol puddle**: In this simulated methanol puddle incident using ALOHA, a significant threat zone is identified. With wind at $4m/s$ from *NE*, the zones are: red (lethality) up to $146m$, orange (second-degree burns) up to $165m$, and yellow (pain) up to $219m$, all within $10sec$. Figure 26 compares the ALOHA chart (A) with the RF model chart (B) using *KML* files in Google Earth according to results in Table 22. Table 24 compares vertical and horizontal diameters for each zone with the model predictions.
2) **Methanol leak**: In this simulated incident using ALOHA software, a significant threat zone is identified. With methanol release and wind at $7m/s$ from *NNE*, the threat zones have three levels: The red zone, up to $150m$, poses lethality within $60sec$. The orange zone extends to $170m$ with a risk of second-degree burns. Beyond that, the yellow zone reaches $230m$, where pain from methanol vapor exposure is possible within $60sec$. Figure 27 compares the ALOHA chart A and the KNN model chart B, both derived from *KML* files in Google Earth.
Table 25 compares the vertical and horizontal diameters of each zone with the model's predictions.

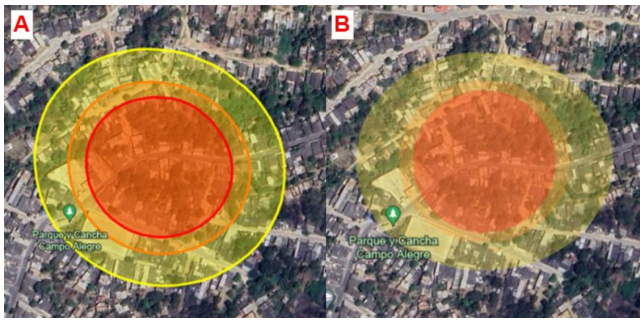

FIGURE 26. Methanol puddle accident in ALOHA vs machine learning model.

TABLE 24. Comparison of results for methanol puddle accident.

Distances	ALOHA results (Km)	Predicted results (Km)
Horizontal red	0,146	0,144
Vertical red	0,145	0,141
Horizontal orange	0,165	0,164
Vertical orange	0,165	0,170
Horizontal yellow	0,219	0,221
Vertical yellow	0,261	0,255

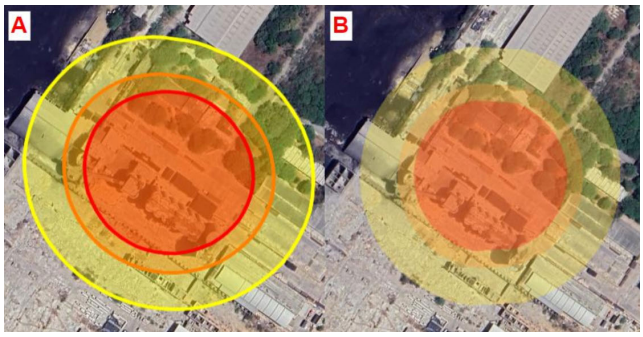

FIGURE 27. Methanol leak accident in ALOHA vs machine learning model.

A web interface developed in Flutter demonstrates the system's functionality by simulating a chemical accident and estimating atmospheric dispersion. Figure 28 shows the interface that simulates a chlorine spill.

The interface left frame defines all necessary variables based on the chemical and the scenario. The red section selects the chemical, the green section edits the geographical

TABLE 25. Comparison of results for methanol leak accident.

Distances	ALOHA results (Km)	Predicted results (Km)
Horizontal red	0,150	0,155
Vertical red	0,147	0,147
Horizontal orange	0,170	0,171
Vertical orange	0,162	0,160
Horizontal yellow	0,230	0,233
Vertical yellow	0,220	0,222

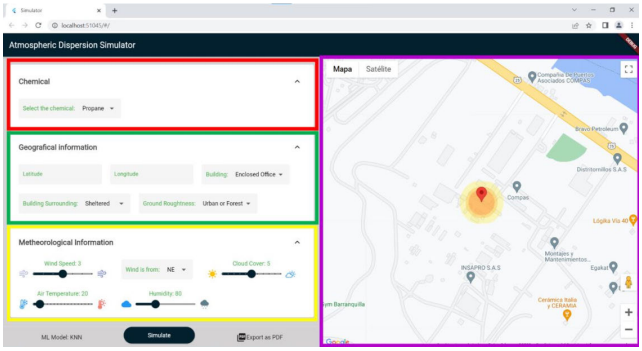

FIGURE 28. Web page for the implementation of the proposed system.

info, and the yellow section modifies the meteorological variables. Pressing "`Simulate`" lets the algorithm choose the best machine learning model based on evaluation metrics and returns the predicted results. The right side displays a map with atmospheric dispersion from a *KML* file.

VII. DISCUSSION

This study addresses the critical issue of atmospheric dispersion of chemical substances in technological risk scenarios, proposing an innovative solution based on machine learning techniques. The importance of understanding and modeling atmospheric dispersion to mitigate the impacts of industrial accidents on human health and the environment is emphasized throughout the work. Traditionally tackled with tools such as ALOHA software, this challenge requires new approaches to overcome its limitations.

One of the primary contributions of this study lies in integrating machine learning models to enhance the accuracy and efficiency of chemical dispersion pattern predictions. The results demonstrate that algorithms such as Random Forest, K-Nearest Neighbors, Support Vector Regression, and Multi-Layer Perceptron can provide more robust representations of simulated scenarios, surpassing the constraints of traditional ALOHA-based methods, such as the inability to integrate real-time data or expand its database of simulable chemical substances.

The proposed approach aligns with the current need to strengthen the stages of the disaster risk management (DRM) cycle, particularly in preparation, response, and recovery from hazardous material emergencies. This underscores the relevance of incorporating next-generation tools, such as machine learning, to comprehensively address the challenges of technological risk management.

However, the study also highlights significant challenges. The dependence on high-quality data to train machine learning models underscores the need to ensure the availability and accuracy of inputs, such as meteorological data and terrain characteristics. Furthermore, adopting these technologies in industrial settings will require considerable effort in terms of training and adapting existing processes.

In terms of applicability, the results obtained could have significant practical implications, not only for the industry but also for governmental and non-governmental organizations responsible for emergency management. The potential to integrate predictive models into broader decision-making systems suggests a promising path toward more effective and preventive management of technological risks.

Finally, this work lays the groundwork for future research exploring the implementation of these technologies in real time and their validation in real-world scenarios. It also emphasizes the importance of developing collaborative solutions involving multiple stakeholders to maximize the impact of these tools in risk reduction and the protection of the environment and communities.

VIII. CONCLUSION

This study has achieved a comprehensive characterization of atmospheric dispersion processes, identifying key challenges inherent in their modeling and emphasizing the limitations of traditional approaches like ALOHA software. While ALOHA has contributed significantly to the field, its lack of integration with other platforms and inability to update its chemical database has highlighted the need for more adaptable and efficient solutions. Addressing these gaps, this research introduces a machine learning-based framework that enhances both the accuracy and computational efficiency of dispersion modeling.

The implementation of machine learning techniques, including Random Forest and Support Vector Regression, has demonstrated significant improvements in prediction accuracy. Furthermore, the integration of dimensionality reduction techniques such as Principal Component Analysis (PCA) has enhanced the computational efficiency of the proposed approach, enabling robust analysis even in complex scenarios. The validation through experimental data and software simulations confirms the practicality and effectiveness of this framework, showcasing its potential for real-time application in industrial decision-making processes.

In addition to achieving suitable modeling, this work emphasizes the scalability and adaptability of machine-learning techniques in industrial contexts. The results indicate that the proposed approach not only optimizes atmospheric dispersion predictions but also provides a pathway for integrating intelligent tools into risk management systems, paving the way for safer and more efficient industrial practices.

Despite its contributions, the study recognizes the necessity of future work to refine and extend the proposed framework. Key areas for further research include:

- Expanding the scope of training datasets to encompass a wider range of chemical substances and environmental scenarios.
- Developing real-time integration capabilities to enhance decision-making during emergencies.
- Validating the framework in diverse industrial contexts to ensure its robustness and adaptability across different sectors.
- Establish collaborations with regulatory agencies and industry partners to obtain anonymized and real-world accident datasets, enabling direct validation of model predictions and refining their accuracy.

In conclusion, this study makes significant contributions to advancing atmospheric dispersion modeling through the application of machine learning techniques. It establishes a foundation for improving technological risk management, offering solutions that are not only precise but also scalable and adaptable to evolving industrial and environmental safety demands.

ACKNOWLEDGMENT

The authors would like to thank the Universidad del Norte and the ANDI Group, sponsors of the royalties project.

REFERENCES

[1] Y. Xie, J. Kuang, and Z. Wang, "Atmospheric dispersion model based on GIS and Gauss algorithm," in *Proc. 29th Chin. Control Conf.*, Jul. 2010, pp. 5022–5027.

[2] J. B. Johnson, "An introduction to atmospheric pollutant dispersion modelling," *Environ. Sci. Proc.*, vol. 19, no. 1, p. 18, Jul. 2022. [Online]. Available: https://www.mdpi.com/2673-4931/19/1/18

[3] EE Agency. (Oct. 14, 2023). *Annex 5.3 Limitations and Uncertainties in Meteorological Estimates Using Dispersion Models. European Environment Agency*. [Online]. Available: https://www.eea.europa.eu/publications/TEC11a/page015.html

[4] Meteorological Data. (Oct. 14, 2023). *The South Coast Air Quality Management District*. [Online]. Available: https://www.aqmd.gov/home/air-quality/meteorological-data

[5] USEP Agency. (Oct. 14, 2023). *Air Quality Dispersion Modeling. Environmental Protection Agency*. [Online]. Available: https://www.epa.gov/scram/air-quality-dispersion-modeling

[6] R. Bhattacharya. (Oct. 14, 2023). *Atmospheric Dispersion. Asian Nuclear Safety Network*. [Online]. Available: https://ansn.iaea.org/Common/Topics/OpenTopic.aspx?ID=13012

[7] USEP Agency. (Jul. 2023). *Computer-Aided Management of Emergency Operations. United States Environmental Protection Agency*. [Online]. Available: https://www.epa.gov/cameo

[8] S. Brown. (Oct. 14, 2023). *Machine Learning, Explained. MIT Sloan*. [Online]. Available: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

[9] DOMO. (Oct. 14, 2023). *Machine Learning Basics*. [Online]. Available: https://www.domo.com/glossary/what-are-machine-learning-basics

[10] S. Abirami and P. Chitra, "Energy-efficient edge based real-time healthcare support system," in *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases* (Advances in Computers), vol. 117, P. Raj and P. Evangeline, Eds., Amsterdam, The Netherlands: Elsevier, 2020, pp. 339–368. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0065245819300506

[11] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," *Neural Inf. Processing-Letters Rev.*, vol. 11, no. 10, pp. 203–224, 2007.

[12] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[13] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICCS)*, May 2019, pp. 1255–1260.

[14] D. Song, K. Lee, C. Phark, and S. Jung, "Spatiotemporal and layout-adaptive prediction of leak gas dispersion by encoding-prediction neural network," *Process Saf. Environ. Protection*, vol. 151, pp. 365–372, Jul. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957582021002615

[15] D. Ma and Z. Zhang, "Contaminant dispersion prediction and source estimation with integrated Gaussian-machine learning network model for point source emission in atmosphere," *J. Hazardous Mater.*, vol. 311, pp. 237–245, Jul. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304389416302370

[16] M. R. Delavar, A. Gholami, G. R. Shiran, Y. Rashidi, G. R. Nakhaeizadeh, K. Fedra, and S. Hatefi Afshar, "A novel method for improving air pollution prediction based on machine learning approaches: A case study applied to the capital city of Tehran," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 2, p. 99, 2019. [Online]. Available: https://www.mdpi.com/2220-9964/8/2/99

[17] L. W. Young, C. N. Ho, S. K. Il, H. S. Ok, and K. J. Suk, "The reliability of pollution prediction with regression analysis and the possibility of dispersion and receptor models," in *Proc. 7th Int. Conf. Properties Appl. Dielectric Mater.*, vol. 3, 2003, pp. 1035–1038.

[18] R. Miriyagalla, Y. Samarawickrama, D. Rathnaweera, L. Liyanage, D. Kasthurirathna, D. Nawinna, and J. L. Wijekoon, "On the effectiveness of using machine learning and Gaussian plume model for plant disease dispersion prediction and simulation," in *Proc. Int. Conf. Advancements Comput. (ICAC)*, Dec. 2019, pp. 317–322.

[19] Ö. F. Ertuğrul and M. E. Tağluk, "A novel version of K nearest neighbor: Dependent nearest neighbor," *Appl. Soft Comput.*, vol. 55, pp. 480–490, Jun. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494617300984

[20] I. Mokhtari, W. Bechkit, H. Rivano, and M. R. Yaici, "Uncertainty-aware deep learning architectures for highly dynamic air quality prediction," *IEEE Access*, vol. 9, pp. 14765–14778, 2021.

[21] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, and classification," *IEEE Trans. Neural Netw.*, vol. 3, no. 5, pp. 683–697, Sep. 1992.

[22] T. Abbott et al., "Dark energy survey year 1 results: Constraints on extended cosmological models from galaxy clustering and weak lensing," *Phys. Rev. D, Part. Fields*, vol. 99, no. 12, Jun. 2019, Art. no. 123505, doi: 10.1103/PhysRevD.99.123505.

[23] E. M. R. Devi, R. Shanthakumari, R. Rajadevi, V. Hari, and S. Lakshmanan, "Forecasting air quality pollutants using ensemble learning models," in *Proc. 2nd Int. Conf. Vis. Towards Emerg. Trends Commun. Netw. Technol. (ViTECoN)*, May 2023, pp. 1–6.

[24] Z. Dou, Z. Liu, L. Li, H. Zhou, Q. Wang, J. Zhang, and L. Chen, "Atmospheric dispersion prediction of accidental release: A review," *Emergency Manage. Sci. Technol.*, vol. 2, no. 1, pp. 1–20, 2022, doi: 10.48130/EMST-2022-0009.

[25] M. F. El-Amin, "Detection of hydrogen leakage using different machine learning techniques," in *Proc. 20th Learn. Technol. Conf.*, Jan. 2023, pp. 74–79.

[26] M. F. El-Amin and A. Subasi, "Forecasting a small-scale hydrogen leakage in air using machine learning techniques," in *Proc. 2nd Int. Conf. Comput. Inf. Sci. (ICCIS)*, Oct. 2020, pp. 1–5.

[27] T. Pinthong, M. Ketcham, T. Ganokratanaa, P. Pramkeaw, and N. Chumuang, "Globally harmonized system label detection using color segmentation," in *Proc. IEEE Int. Conf. Cybern. Innov. (ICCI)*, Mar. 2023, pp. 1–6.

[28] R. Jones, W. Lehr, D. Simecek-Beatty, and M. Reynolds, *ALOHA (Areal Locations of Hazardous Atmospheres) 5.4.4: Technical Documentation*. Washington, DC, USA: NOAA Office of Response and Restoration, 2013.

[29] I. D. Rodionov, M. A. Gomorev, I. P. Rodionova, A. I. Rodionov, V. L. Shapovalov, D. V. Shestakov, and M. G. Golubkov, "Remote detection of emergency emissions and gas leaks," *Russian J. Phys. Chem. B*, vol. 18, no. 5, pp. 1389–1395, Oct. 2024.

[30] R. H. Vaivads, M. F. Bardon, and V. Battista, "A computational study of the flammability of methanol and gasoline fuel spills on hot engine manifolds," *Fire Saf. J.*, vol. 28, no. 4, pp. 307–322, Jun. 1997.

[31] I. Mohammadfam, O. Kalatpour, and K. Gholamizadeh, "Quantitative assessment of safety and health risks in HAZMAT road transport using a hybrid approach: A case study in Tehran," *ACS Chem. Health Saf.*, vol. 27, no. 4, pp. 240–250, Jul. 2020.
[32] M. I. V. Rada, B. G. Granados, C. G. Quintero M, C. Viloria-Nũñez, J. Cardona-Peña, and M. Á. J. Paba, "Atmospheric dispersion prediction for toxic gas clouds by using machine learning approaches," in *Proc. Int. Conf. Smart Technol., Syst. Appl.* Cham, Switzerland: Springer, Jan. 2023, pp. 185–198.
[33] Semana. (Feb. 2023). *Tren Cargado De Gas Propano Se Descarril En Florida: Autoridades Buscan Evitar Una Explosin*. [Online]. Available: https://www.semana.com/mundo/noticias-estados-unidos/articulo/tren-cargado-de-gas-propano-se-descarrilo-en-florida-autoridades-buscan-evitar-una-explosion/202340/
[34] (Jun. 2022). *Chlorine Gas Leak Kills 13 in Jordan Port*. [Online]. Available: https://www.france24.com/en/live-news/20220628-chlorine-gas-leak-kills-13-in-jordan-port
[35] A. Phillips. (Mar. 2023). *Ohio River Disaster As Barge With Tons of Toxic Methanol Sinks*. [Online]. Available: https://www.newsweek.com/ohio-river-disaster-barge-toxic-methanol-louisville-kentucky-1791064
[36] (2023). *Video Del Momento Exacto En El Que Se Accident Un Camin Lleno De Cloro En El Tnel De Occidente2023*. [Online]. Available: https://telemedellin.tv/video-accidente-camion-tunel-occidente/628708/
[37] L. R. Online. (Mar. 2021). *(Videos) Se Incendia Camin Con 800 Litros De Metanol En CDMX*. [Online]. Available: https://www.razon.com.mx/ciudad/videos-incendia-camion-800-litros-metanol-cdmx-426522
[38] T. Nguyen and Perez-CarrilloMelissa. (Mar. 2023). *Train Carrying Propane Derailed Near Florida Airport; No Leaks Detected, Officials Say*. [Online]. Available: https://www.usatoday.com/story/news/nation/2023/02/28/sarasota-florida-derailment-propane-officials/11368082002/
[39] S. Wold, K. H. Esbensen, and P. Geladi, "Principal component analysis," *Chemometric Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, Aug. 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0169743987800849
[40] Optuna. (Oct. 26, 2023). *Optuna: A Hyperparameter Optimization Framework*. [Online]. Available: https://optuna.readthedocs.io/en/stable/

MARÍA VALLE received the B.S. degree (cum laude) in electronics engineering and the M.Sc. degree (cum laude) in electronic engineering from the Universidad del Norte, Colombia, in 2021 and 2024, respectively. She is currently a Computer Programmer. Additionally, she actively participates in research projects exploring applications of machine learning within the artificial intelligence (AI) field. Her experience working on Internet of Things (IoT) projects demonstrates her commitment to staying at the forefront of technological advancements.

JAIRO A. CARDONA (Member, IEEE) received the B.S. degree in telecommunications and electronics engineering and the M.Sc. degree in telematics from the Universidad del Cauca, Colombia, in 1995 and 2005, respectively. He is currently pursuing the Ph.D. degree with the Doctoral Program in Computer Science. He is currently a member of the Telecommunications and Signal Processing Group and an Assistant Professor with the Electrical and Electronics Department, Universidad del Norte (UN), Barranquilla, Colombia. He is also a Counselor and an Evaluator of the Minciencias National Program in Electronics, Telecommunications, and Computer Science. His research and teaching interests include the development of computational intelligence approaches applied to several application domains and technology. His research expertise is related to computer networks, software defined networks, and quality of experience.

CÉSAR VILORIA-NUÑEZ (Senior Member, IEEE) received the B.S. degree (cum laude) in electronics engineering and the M.Sc. degree in computer science from the Universidad del Norte, Colombia, in 2008 and 2010, respectively. He is currently pursuing the Ph.D. degree with the Doctoral Program in Economics and Innovation Management. He is currently the Dean of the Digital Transformation School, Universidad Tecnológica de Bolívar (UTB), Cartagena, Colombia. He is also a Counselor and an Evaluator of the Minciencias National Program in Electronics, Telecommunications, and Computer Science. His research and teaching interests include the development of computational intelligence approaches applied to several application domains and technology and engineering management. His research expertise is related to digital transformation, enterprise architecture, and the implementation of digital solutions.

CHRISTIAN G. QUINTERO M. (Senior Member, IEEE) received the B.S. degree (cum laude) in electronics engineering from the Industrial University of Santander, Colombia, in 2001, the M.Sc. degree in information technology from the University of Girona (UdG), Spain, in 2005, and the Ph.D. degree (cum laude) from the Doctoral Program in Information Technology, Department of Electronics, Computer Science and Automatic Control, UdG, in 2007. He is currently a member of the Robotics and Intelligent Systems Group and a Professor in automatic control and intelligent systems design with the Universidad del Norte (UN), Barranquilla, Colombia, where he is also an Associate Professor with the Department of Electrical and Electronics Engineering. He is a Counselor and an Evaluator of the Minciencias National Program in Electronics, Telecommunications, and Computer Science. His research and teaching interests include the development of computational intelligence approaches applied to several application domains. His research expertise is related to the definition, design, negotiation, implementation, evaluation of technological projects, and project management of technology teaching.