Performance Evaluation of Machine Learning Algorithms for Detecting Gas Leakage System

***Kondireddy Muni Sankar**
Research Scholar, Department of Information Technology, Vels Institute of Science, Technology, and Advanced Studies (VISTAS),Chennai, India.
Dr. B. Booba,
Professor, Research Supervisor, Department of Information Technology, Vels Institute of Science, Technology, and Advanced Studies (VISTAS), Chennai, India.
*Corresponding Author- munisankarkondireddy@gmail.com
*ORCID: 0009-0001-6468-5067

Abstract

Gas leaks in oil and gas plants, particularly within extensive natural gas pipeline networks, pose significant safety and environmental challenges. Traditional leak detection methods, such as acoustic monitoring, infrared imaging, and manual inspections, are often time-consuming, prone to human error, costly, and limited in sensitivity to minor leaks. These drawbacks highlight an urgent need for more advanced solutions. To address these limitations, Artificial Intelligence (AI) and Machine Learning (ML) models have emerged as effective alternatives. These intelligent systems offer higher accuracy, early detection capabilities, cost-effectiveness, and scalability. This research specifically compares and evaluates the efficacy of several ML models, including Linear Regression, Logistic Regression, Random Forest (RF), and K-Nearest Neighbor (KNN). These models are employed to identify minor gas leaks using fundamental operational parameters like pressure and flow data. The models are rigorously compared using established performance metrics across different damage types. Findings indicate that AI-based approaches can detect leaks rapidly and with minimal false alarms. This breakthrough in detection sensitivity is crucial. These intelligent systems significantly enhance safety by enabling prompt leak identification, drastically reduce operational costs through automation, and robustly support regulatory compliance. The numerical outcome showcased that, linear regression has obtained the highest accuracy of 95.02%, followed by Random Forest 92.46%, Logistic regression of 89% and KNN with 87.94%. Ultimately, this technology provides a robust, adaptable solution for gas leak detection in industrial settings, marking a substantial improvement over conventional methods.

Keywords: Gas Leak Detection, Linear Regression, Logistic Regression, Random Forest, K-Nearest Neighbour

I. Introduction

During the course of history, the movement of goods has a crucial requirement for human beings. Since 400 BC, pipelines have functioned as an efficient means for fluid transportation [1]. Specifically for natural gas, pipelines remain the most efficient and cost effective approach for transporting medium to huge volumes over short moderate distances [2]. Thus pipelines play a significant role in the distribution of liquid and gas resources [3-5]. However, leak in a pipeline can lead to severe consequences like wasted resources, economic loss and distribution downtime [6]. Further, in the oil and gas sector, a range of issues and irregularities pose risks to pipelines, potentially leading to human harm and financial setbacks. Common problems encountered in gas plants include corrosion, leaks, and rust, among others [7]. Gas plant leaks not only jeopardize human health and safety but also pose environmental hazards. The release of gases like isobutane and propane into the atmosphere during leaks contributes to ozone depletion and exacerbates global warming, further emphasizing the critical nature of addressing these issues Therefore, in order to avoid these consequences, early detection of gas leaks are very important [8, 9]. Various approaches are undertaken for gas leakage detection using hardware and software techniques such as pressure detection, ultrasonic, bubbling, visual inspection are depicted in Figure 1.

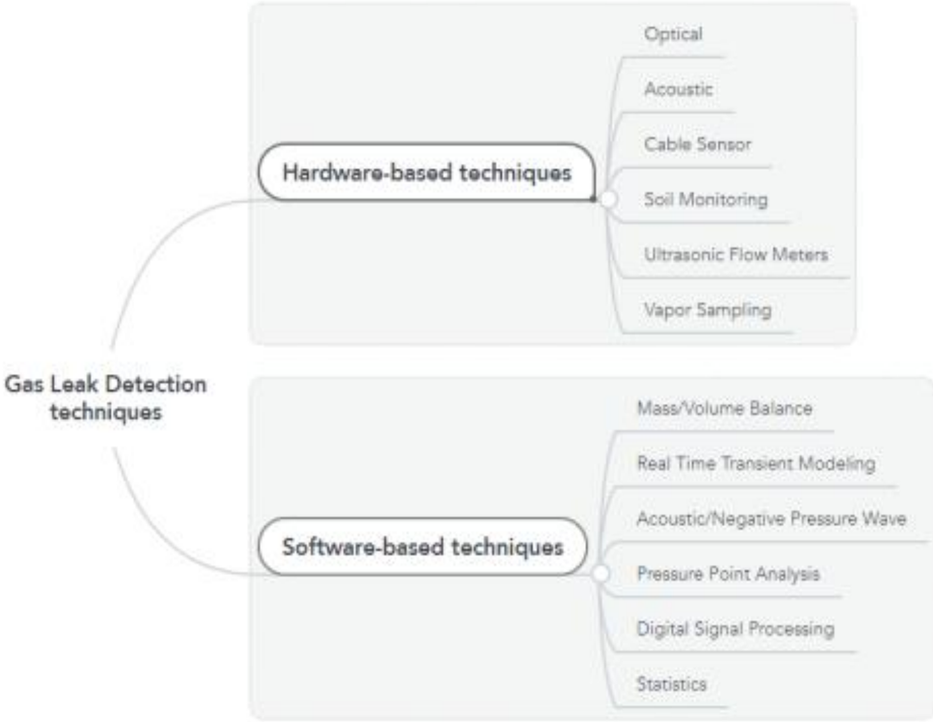

Gas Leak Detection techniques
Hardware-based techniques
Optical
Acoustic
Cable Sensor
Soil Monitoring
Ultrasonic Flow Meters
Vapor Sampling
Software-based techniques
Mass/Volume Balance
Real Time Transient Modeling
Acoustic/Negative Pressure Wave
Pressure Point Analysis
Digital Signal Processing
Statistics

Figure 1. Gas Leak Detection Techniques [10]

Though hardware and software based approaches are opted for gas leakage detection [11, 12], there are certain drawbacks of electing these techniques which includes extremely time consuming, prone to error, subjectivity, limited coverage, heavy dependency on operator skill for effectively detecting gas leaks in gas plants environment. Thus, in order to overcome these drawbacks faced by conventional methods, AI based methods are incorporated. In the oil and gas industry, AI plays a crucial role, with ML- and DL-based techniques being employed to enhance pipeline anomaly detection [13]. AI based models are more employed for gas leakage detection since AI based techniques offers enhanced accuracy, improved reliability for robust detection [14, 15]. AI models can effectively handle detect early gas leak detection allowing for more prompt intervention and mitigation [16]. Further, AI based models [17, 18] can process massive amount of data effectively thereby enabling effective detection and response to gas leaks. Thus, different AI techniques are incorporated for Gas leak detection.

Random forest [19] is used as the optimum combination for gas pipeline leakage detection for reducing economic loss and environmental pollution. Likewise, Faster RCNN model [20] has use for real time automated approach for detecting hydrocarbon gas leak. Along with Faster RCNN model, optical gas imaging technology is also implemented. Although these models are used for gas leakage detection system, there are some of limitations such as overfitting of model, class imbalance. Moreover, gas leakage detection plays a significance role in recent times due to the consequences of gas leaks in current circumstances.

Hence, the originality of this research lies in its comparative analysis of multiple machine learning algorithms for gas leak detection, covering various aspects of damages caused in the environment such as corrosion, excavation damage, incorrect operation, material/weld/equip failure, natural force damage, and other outside force damages. While existing studies have explored AI and ML for pipeline anomaly detection, this work distinguishes itself by rigorously comparing Linear Regression, Logistic Regression, Random Forest, and K-Nearest Neighbor algorithms using real-time data collected from gas plant operators to assess their performance across different damage types. This approach aims to address the limitations of previous models, such as overfitting and interpretability challenges, by providing a comprehensive evaluation of different algorithms for precise and effective gas leakage detection. The focus on real-time data and the comparative analysis across multiple algorithms and damage scenarios contribute to the novelty of this research.

1.1 Research Objectives

The major contributions of the research work are listed as follows,

- To collect and curate comprehensive real-time datasets from gas plant operators, ensuring the inclusion of diverse operational scenarios and potential leakage events. This data collection will form the foundation for robust model development and validation.
- To systematically evaluate and compare the effectiveness of multiple machine learning algorithms including Linear Regression, Logistic Regression, Random Forest (RF), and K-Nearest Neighbors (KNN) in accurately detecting gas leaks in the pipeline by considering various factors such as Corrosion, Excavation damage, Incorrect operation, Material/weld/equip failure, Natural force damage, All other causes and Other outside forces damage. The study will highlight the strengths of each method within the context of gas plant operations.
- To rigorously assess the performance of each predictive model using a suite of evaluation metrics such as precision, recall, F1-score, and accuracy. This multi-metric evaluation will provide a holistic view of each model's reliability, sensitivity, and overall suitability for deployment in real-world gas leakage detection systems.

1.2 Paper Organization

The paper is systematized in the succeeding approach, in which section 2 deals with existing works for detecting gas leaks, section 3 projects comparative analysis implemented for detecting gas leaks, section 4 depicts the outcome obtained by employing research work and section 5 summarizes the research work and future recommendation.

II. Literature Review

Various methods are incorporated for gas leakage detection using various AI are reviewed in the subsequent section.

Gas [21] leakage is crucial to detect. Hence, ML algorithms such as linear regression, logistic regression, RF and SVM [22] has been used in the work for detecting the gas leakage. Though the model has used various algorithms, the limitation of the work deliberates on overfitting of the techniques used and interpretability of the model. Likewise, XGBoost, gradient boosting, Logistic regression with ElasticNet and linear support vector classifier [23] has used in the work for the identifying the gas leakage. Here, the findings obtained by the models are accuracy of 92.3% for XGBoost, 92.5% for gradient boosting, 89.7% for linear support vector classifier and 89% for logistic regression with Elastic Net. However, the accuracy value can be improved further. Besides, different models like decision tree, RF, support vector machine, gradient boosting and artificial neural network [24] has used in the study for leak detection in gas pipelines. All five models have delivered reasonable performance, however, it was noted that, RF technique has deliberated better performance when compared to other models. Moreover, a different ML frameworks [25] have been utilized for gas pipeline leakage mechanism, where the model has used GPLA12_v3 dataset has used and utilized Gaussian Naïve Bayes classifier and RF classifier for classification. Eventually, Gaussian Naïve Bayes classifier has resulted in accuracy of 66% and RF of 76%, which could be improved further in future.

KNN, decision tree, random forest and neural networks [7] were used for reliable gas and water leakage detection. The experimental outcome showcases that, ML techniques delivered considerable outcome for efficient leak detection system in order to avoid consequences such as health risks, economic losses and other aspects as well. Likewise, Study has used linear regression algorithm [26], as ML techniques are very much suitable for leakage detection. Similarly, ML and image processing techniques were used in the study for gas leak detection model in which image processing approach has used for extracting the information from the images and ML based RF [27] algorithm has employed for precise detection of gas leakage. From the experimental outcome, it was identified that employment of RF algorithm demonstrated its ability for automatically detecting and displaying gas leaks in high quality. Study [28] preferred using AE technique for indicating the leakages in pipelines and further, SVM and RVM (Relevance Vector Machine) has used for developing the hyperplane in order to classify the outcomes. Further, the study has also indicated if the leakage is slower or faster.

Correspondingly, different ML algorithms such as linear regression and RF [29] were used for automating gas leak detection system by collecting data from gas plants. However, the experimental outcome has depicted that random forest has delivered better performance than linear regression as accuracy obtained by linear regression was 39%. Moreover, 5 different algorithms [30] were used in the study for gas leak detection in which algorithms like gradient boosting, RF, SVM, ANN and Decision Tree (DT) was used for gas pipeline leaked prediction. However, the outcome has stated that, ANN and SVM model has delivered better outcome than other models. Congruently, SVR based ML model [31] has used for gas leakage system using dataset which was built by PHAST system. Moreover, 5 assorted ML algorithms [32] like RF, SVM, KNN, DT and gradient boosting was used for developing gas pipeline detection method. Though all 5 algorithms delivered reasonable performance for gas leakage detection, much effective outcome was obtained by using SVM model as it could efficiently identify the unusual event of oil and gas pipeline leakage. Besides, the work has adopted using Radial Basis Function Neural Network (RBF-NN) ML model [33] for this process, along with genetic algorithm. The outcome of RBF-NN model has been compared with the Back Propagation Neural Network (BP-NN), in which the result showcased that, RBF-NN obtained decent result when compared to BP-NN.

PAM (Passive Acoustic Monitoring) and ML model [34] was designed in the suggested study for detecting the presence of leakage. The ML models used were RF and gradient boosting. Further, hidden markov model has incorporated for detecting the duration of the leakages. Likewise ML, DL algorithms are used for gas leakage detection system 2D CNN model [35] has used for detecting leakage precisely and timely. This process involved by demonstrating the applicability of the combination of neural networks and detect the changes depending on the vibrations in the pipeline systems. Moreover, thermal IR cameras and UAV [36] has used in the study for detecting oil leakage inside a port environment. The images were fetched from real time dataset and the images were trained using CNN model and CNN model facilitated in frequent inspection of oil leakage on water at low cost. However the model resulted that, better solution has resulted using CNN implementation by generating increased detection rate.

Similarly, YOLO model [37] has used for gas leak detection by employing technology for visualizing the ultrasonic waves generated during gas leaks. The clearness of the ultrasonic images which was collected was decreased due to the increase in measurement of distance. Thus, this model projected as a gas safety management expertise at industrial sites which facilitates precise detection of gas leak status, gas leakage flow and leak position. CNN model [38] is used for recognizing valve internal leakage by using power spectral density images of internal and non-leakage signals under numerous working conditions as input. Similarly, 3 different deep learning models [39] are used for gas leakage detection such as 2D CNN, 3D CNN and ConvLSTM model. Similarly, the model has adopted stand-alone CNN and LSTM approaches [40] for gas leakage detection along with hybrid ML (HML) techniques which consisted of decision tree and XGBoost algorithm. The outcome of the work has stated that, HML resulted in better accuracy (92%) than other algorithms, however, the limitation of the work is the scalability of the model, which needs to be improved by using better algorithms. Further, CNN model has explored in the study with the aim of detecting the leaks and focuses on attaining considerable accuracy by the model [41]. Similarly, CNN [42] has been utilized in the DL assisted gas pipeline leakage detections system (DLGPLDS) for classifying the presence and absence of gas leaks. In which employment of CNN in DLGPLDS framework has resulted in gas pipeline leakage location. Additionally, CNN based process has used in the study for precisely detecting the gas leakage, where combination of layers such as convolution, pooling and fully connected layer were used. Findings of the work has projected in decent performance for gas leakage detection system.

2.1 Gaps identified

Gas leakage system is extremely crucial in recent times due to the consequence faced by the gas leaks in different gas plants, hence, different models were exploited, hoverer, existing ML algorithms encountered challenges related to overfitting, demonstrating strong performance on training datasets but exhibiting poor generalization to novel, unseen data. This necessitated the evaluation of multiple models. Interpretability presented another constraint, particularly with automated processes, as certain machine learning models proved complex and challenging to comprehend, thereby hindering the explanation of their predictive rationale [22]. Furthermore, the accuracy achieved by models such as the Gaussian Naïve Bayes classifier, which yielded 66%, and Random Forest, which achieved 76% [25], alongside the 92% accuracy of HML [40], requires further enhancement and improvement for optimal performance.

Thus, present research work focuses on comparing 4 different ML models like linear regression, logistic regression, KNN and RF for precise and effective prediction of gas leakage detection systems as ML models offer better accuracy, process rapidly, offers adaptability since ML algorithms can adapt to changing conditions and learn from new data, making them suitable for detecting leaks in dynamic environment where gas levels may fluctuate and eventually ML models are cost effective in long run as it reduces the dire need for manual monitoring and maintenance.

III. Proposed Methodology

In the methodology section of this study, a structured approach to building the ML model is outlined. The process begins with data collection and pre-processing as the initial steps. Following this, the present models is trained, and its performance is thoroughly evaluated. The section provides a comprehensive description of each step involved in the methodology. Figure 2 summarizes the methodology of this study.

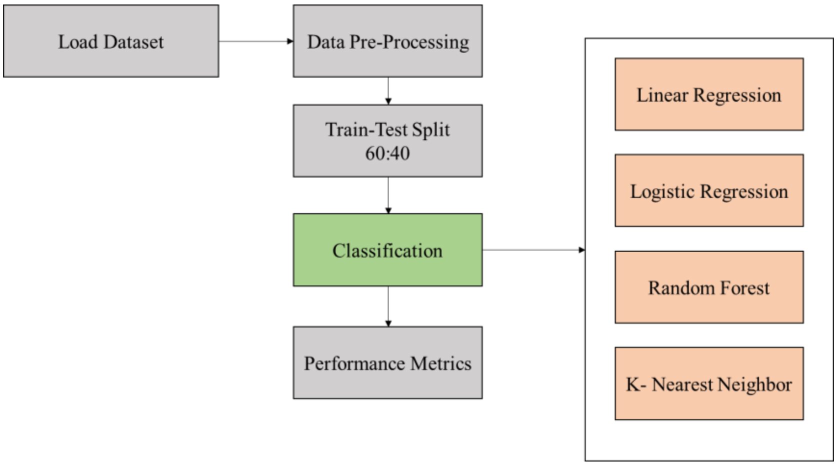


Figure 2. Overall Flow

3.1 Dataset Collection

The data collection process involved collaboration with gas plant operators. Data spanning from 2010 to 2020 were gathered from various operators, resulting in a dataset with 50 features and 3000 instances. This dataset comprises both categorical and numerical attributes, making it suitable for addressing classification challenges. Furthermore, to facilitate model development, the dataset was initially pre-processed and divided into training and testing sets.

3.2 Pre-processing and Data Split

After loading the dataset, the dataset is processed for overcoming the inconsistency issues, handling missing values and outlier challenges by ensuring the data is of high quality and suitable for analysis. Once the data is pre-processed, the model is split as train-test split using 60:40 ratio. In which the train-test split is allocated as 60% training and 40% testing. The training sample is utilized to train the model and improve its understanding of the dataset's complexities. Subsequently, the model's performance is assessed on the testing sample to measure its capability to generalize to unseen data.

3.3 Classification for Gas leakage detection

After train-test split, classification process takes place using 4 different classification algorithm due to its distinct characteristics which helps in the process of gas leakage detection. Hence, subsequent section deals with exhaustive approaches undergone for classification function.

3.3.1 Linear Regression

Linear regression is a statistical approach used for predictive analysis by showcasing the connection between continuous variables. Linear regression demonstrates the linear relationship between the independent as well as dependent variables. Thus, linear regression is the process which comprises fitting a linear model to a set of data points with the aim to establish a relationship between dependent and independent variables, thereby minimizing the sum of squared differences between observed data points and predicted values generated by the model. The process of linear regression is proceeded by defining linear regression model which is depicted in equation,

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \cdots . \beta_n * X_n + \in \qquad (1)$$

Here, Y is defined as the dependent variable, X_1 to X_n is denoted as independent variables, β_0, $\beta_1 .. \beta_n$ is defined as coefficients of the model and error term is stated as $\in$. Thus, Figure 3 represents linear regression process.

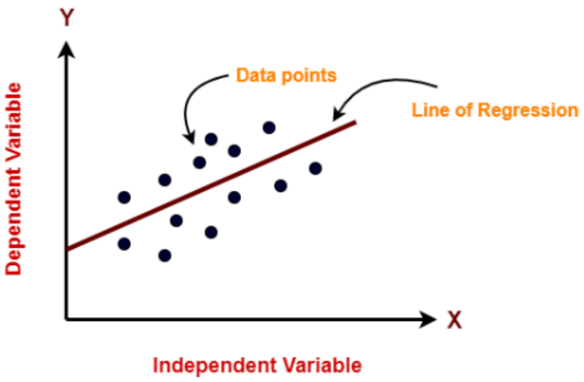


Figure 3. Linear Regression

In present research linear regression is preferred due to its ability to model relationships between variables, making it as a more appropriate technique for examining continuous data and predicting outcomes based on those relationships. Further linear regression is implemented for analyzing the input data and detect the patterns which indicates the presence of gas leak. Thus, by fitting a linear model to the data, it becomes possible to detect the trends and patterns which signify leak and enable in quicker gas leakage detection. Besides, linear regression model can estimate the relationship between the input variables (input data) and output variables (presence of absence of gas leak) which aids as an effective detection algorithm. In addition, linear regression can handle massive dataset easily which makes it easier and effective for real time gas leakage detection as well. Owing to these factors, linear regression is preferred for present work.

Algorithm I: Linear Regression
Step 1: Gather data containing pairs of X values and Y values **Step 2:** Estimate mean for both x values and y values **Step 3:** Gauge the slope of the line **Step 4:** Compute the intercept of the line **Step 5:** Make predictions using calculated slope and intercept in order to make predictions for new x values using straight line. **Step 6:** Assess the performance of the model

3.3.2 Logistic Regression

Logistic regression is considered as one of powerful classification techniques which models the log odds as a linear function of the predictors and uses logistic function to map the probability and the model parameters are calculated using maximum likelihood. Thus, the objective of logistic regression model is to predict the probability that a given input belongs to a specific class depending on one or more predictor variables. In logistic regression, the connection between predictor variables and likelihood of an outcome is modeled using the logistic function, which transform any input value into a probability score between 0 and 1 indicating the probability of the input belonging to the positive class. Thus, equation 2 and 3 shows the mathematical formula used for logistic regression process.

$$\text{Logit}(\pi) = \frac{1}{1+exp(-\pi)} \qquad (2)$$

$$ln(\frac{\pi}{1-\pi)} = \beta_0 + \beta_1 . X_1 + \cdots + \beta_k . X_k \qquad (3)$$

Here, π is denoted as the predicted probability of the outcome interest, exp is defined as the exponential function, Ln is defined as the natural logarithm, β is represented as coefficients of the model and X_1 to X_k is noted as independent variables. Thus, mathematical equation shows that, logit transformation is applied to the model the relationship between the predictor variables and the probability of the event occurring. The coefficients are estimated using methods like maximum likelihood estimation for optimizing the model for best fit of log odds allowing the calculation of predicted probabilities for each observation. Algorithm shows the process involved in logistic regression.

Algorithm II: Logistic Regression
Step 1: Collect the dataset which contains independent variables and binary dependent variables **Step 2**: Initialize bias parameter and weights

```
Step 3: Estimate predicted probabilities by utilizing logistic function depending on weighted sum of features
Step 4: Define loss function like binary cross entropy for measuring error between predicted probabilities and actual target values
Step 5: Update the parameters and weights by minimizing the loss function
Step 6: assess the performance of the model using metrics
```

3.3.3 Random Forest

Like Linear and logistic regression, Random forest is also one of the widely used ML algorithms for classification process due to its robustness and versatility. Generally, RF is a classifier which encompasses of no. of. DT on diverse subsets of the dataset and takes the average of improving the predictive accuracy of the data. Thus, the basic approach of RF is to build a huge no. of. DT during the training process and amass the predictions in order to make final prediction.

The process is initiated by building a decision tree. To build a DT, the algorithm picks a random subset of features at each node and splits the data based on the best feature and split point. This process is repeated iteratively until a termination condition is satisfied, such as reaching the maximum allowable depth or the minimum required number of samples. Then RF uses boot strapping method for creating numerous subsets of the training data to build individual DT. This encompasses of sampling the training data with replacement to create new datasets of the same size as the original dataset. Each DT is trained on different bootstrap sample which aids in introducing diversity among the trees. Then, Random feature selection is opted by the model for reducing the correlation among the trees and enhances the ability of the model to generalize to new data. Therefore, by considering only a subset of features at each node, RF can capture diverse characteristics of the data in each tree. Eventually, after all DT are built, the last step of RF is to aggregate the predictions for making the final predictions. In classification, majority class among the predictions of all the trees is chosen as the final prediction. Algorithm for RF is depicted as follow.

```
Algorithm: Random Forest
Step 1: Select $k$ data points from the training set randomly
Step2 : Construct DT related to selected data points
Step3:  Select the number $N$ for building DT
Step 4: Reiterate step 1 and step 2
Step 5: Detect the prediction of each DT for new data points and assign new data points for categorizing with majority votes
```

Typically, RF is opted for gas leakage detection process as RF can handle both numerical and categorical data effectively, making it more suitable for data collected in gas plants. Moreover, RF model can handle missing values and outliers effectively, thereby making this model as a dependable option for gas leak detection where the quality of the data may vary.

3.3.4 KNN

KNN is a simple ML model which is employed for classification. In KNN, based on the majority class of its K nearest neighbors obtained in feature space, data point for classification is determined. As KNN model stores all the training data points and the label of the data in memory, KNN do not have the need to train the whole dataset, due to this reason, KNN is known as a non-parametric as well as instance based learning algorithm. Further, KNN can also handle non-linear boundaries, thereby making the process effective for classification. The process of KNN involves by choosing k values for making prediction. In the next step, KNN makes prediction depending on the similarity of data points. To determine the nearest neighbor, the algorithm computes the distance between input data points and all other data points in the training set. After calculating the distances, the algorithm detects the KNN to the input data point depending on the chosen distance metric. Eventually, majority voting mechanism used for assigning class labels to the input data point, by doing so effective prediction is made. The process involved in the KNN is showcased in figure.

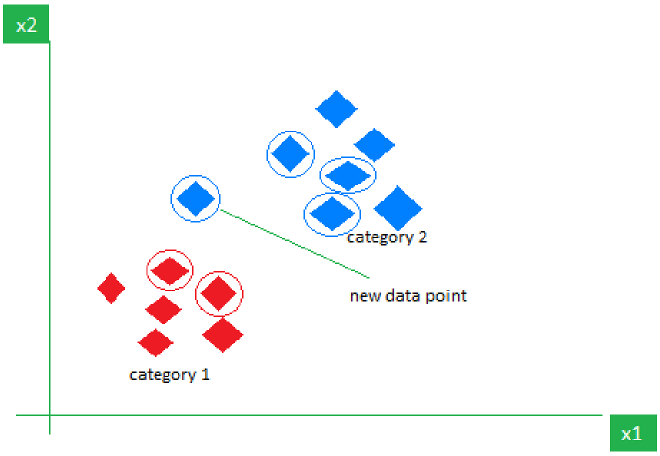


Figure 4. KNN

Figure 4 showcases the process involved in the KNN classification. Where, the data points are categorized as category 1 and category 2 with new data point in center. The Algorithm for KNN is depicted.

```
Algorithm : KNN
Step 1: Choose number k of neighbors
Step 2: Compute ED (Euclidean distance) of $k$ number of neighbors
Step 3: Pick $k$ nearest neighbor as per the estimated ED
Step 4: Determine the no. of. data points belonging to each class among the K nearest neighbor
Step 5: Allocate new data points to that group with the highest no. of. neighbors
Step 6: KNN is equipped for prediction.
```

As these four machine learning models generates various opportunities for gas leakage detection in gas plants, they are opted in the present research work and the results obtained using these models are demonstrated in subsequent section.

IV. Results and Discussion

Results obtained using present model is depicted in subsequent section.

4.1 Performance Metrics

a) Accuracy

Accuracy is characterized as a metric that describes the performance of the models across all classes. Equation 4 depicts the mathematical formula for accuracy,

$$\text{Accuracy} = \frac{\text{TrN+Trp}}{\text{TrN+FaN+TrP+FaP}} \qquad (4)$$

Where TrN , TrP , FaN and FaP is denoted as true negative, true positive, false negative false positive.

b) Recall

Recall is indicated as the solitary of the production metric, which evaluates the total of correct positive groups created out of all the positive classes.

$$\text{Recall} = \frac{\text{TrP}}{\text{FaN+TrP}} \qquad (5)$$

c) Precision (PN)

By calculating the accurate classification count, precision of the model can be calculated. Equation shows the formula for PN.

$$\text{PN} = \frac{\text{TrP}}{\text{FaP+TrP}} \qquad (6)$$

d) F1-Score

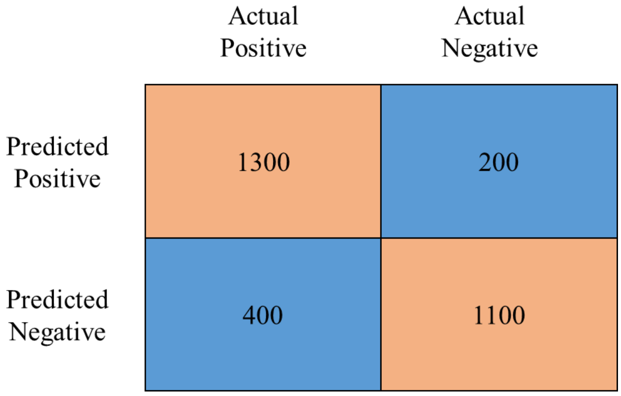
F1-score is represented as measure of harmonic mean of recall and precision value. Mathematical equation for F1 score is depicted in equation,

$$\mathrm{F1-score} = 2 \times \frac{\mathrm{RL} \times \mathrm{PN}}{\mathrm{RL} + \mathrm{PN}} \tag{7}$$

Here, RL is recall and PN is precision

4.2 Performance Analysis

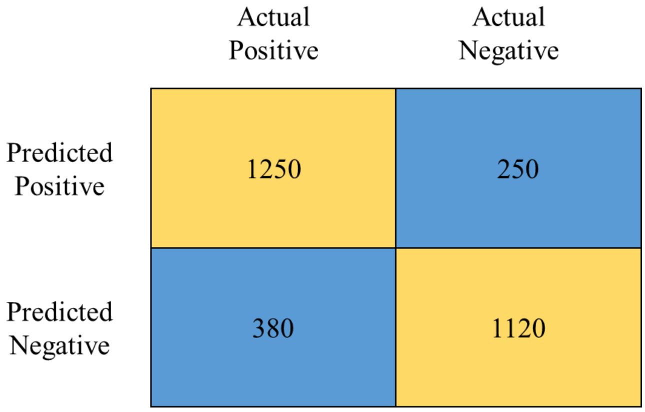
Performance of the different models is assessed in the subsequent section, where metrics are used for estimation. Confusion matrix or error matrix aids in assessing the classification performance in ML process by comparing the predicted values against actual values. Therefore, Figure 5,6, 7 and 8 shows the confusion matrix obtained by linear regression, logistic regression, RF and KNN model for gas leakage detection process.



	Actual Positive	Actual Negative
Predicted Positive	1300	200
Predicted Negative	400	1100

Figure 5. Confusion matrix for Linear Regression

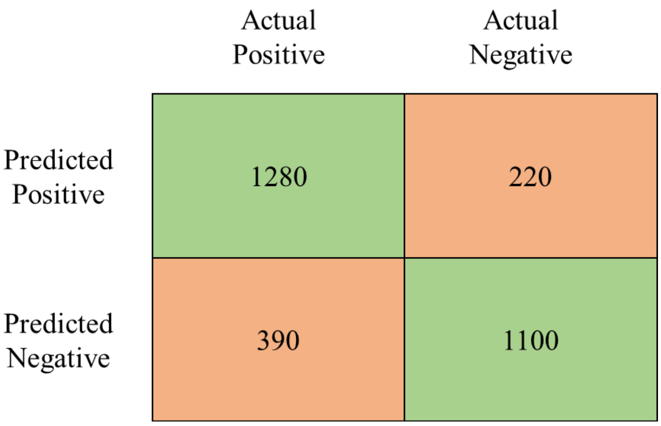
Figure 5 depicts the confusion matrix for linear regression model, where correct and misclassifications are predicted. Here, TP and TN obtained are 1300 and 1100, whereas FN and FP of linear regression is 200 and 400. Likewise, confusion matrix for logistic regression is depicted.



	Actual Positive	Actual Negative
Predicted Positive	1250	250
Predicted Negative	380	1120

Figure 6. Confusion matrix for Logistic Regression

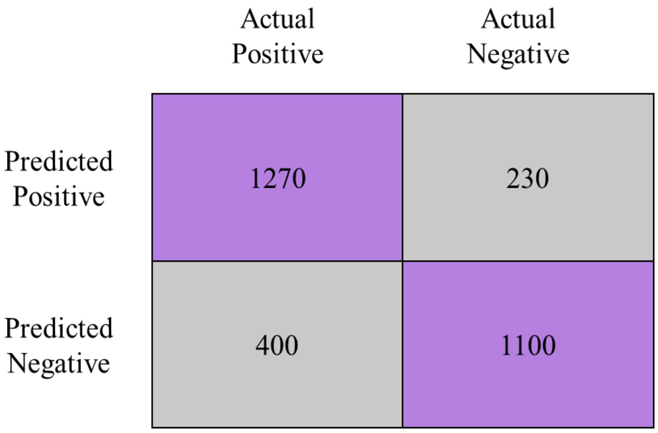
Confusion matrix for Logistic regression is illustrated in Figure 6, where TN and TP of Logistic regression is 1250 and 1120. Likewise, FN and FP obtained is 250 and 380. Here, TP is defined as the number of correct predictions for the positive class and TN is considered as the actual negative class instances accurately predicted as negative. Similarly, FP is represented as negative class instances incorrectly identified as positive cases and FN is projected as actual positive instances erroneously predicted negative.



	Actual Positive	Actual Negative
Predicted Positive	1280	220
Predicted Negative	390	1100

Figure 7. Confusion matrix for RF

Like Linear and Logistic regression, confusion matrix for RF is also projected in Figure 7. Where TN and TP obtained is 1280 and 1100 and FN and FP of RF is 220 and 390. Likewise, confusion matrix for KNN is depicted in Figure 8.



	Actual Positive	Actual Negative
Predicted Positive	1270	230
Predicted Negative	400	1100

Figure 8. Confusion matrix for KNN

Confusion matrix for KNN is illustrated in Figure 8. Here correct classifications obtained are 1270 and 1100, whereas the misclassification obtained by the model is 400 and 230. Thus, from the above confusion matrix it can be identified that, correct classifications are higher than misclassifications. Like confusion matrix, other metrics are used for assessing the efficacy of the model for gas leakage detection using precision, recall, F1 score and ROC-AUC value.

Table 1- Metrics for Linear Regression

Linear Regression				
Types of Damages	**Value of Precision**	**Recall**	**Value of f1-score**	**ROC-AUC**
All other causes	0.9556	0.9184	0.8857	0.7919
Corrosion	0.9423	0.8915	0.9092	0.7953
Excavation damage	0.927	0.8752	0.9119	0.7985
Incorrect operation	0.9322	0.9038	0.9015	0.7876
Material/weld/equip failure	0.9411	0.9287	0.8954	0.792
Natural force damage	0.9085	0.8998	0.9072	0.7934
Other outside forces damage	0.9238	0.9157	0.9122	0.7889

Table 1 depicts different types of damages and the metric values obtained for each damages such as incorrect operation, corrosion, natural force, excavation damage, other outside forces damage, material/weld/equip failure. From the values it can be identified that Excavation damage has generated higher precision rate, recall rate, F1 score and ROC-AUC values for linear regression.

Table 2- Metrics for Logistic Regression

Logistic Regression				
Types of Damages	**Value of Precision**	**Recall**	**Value of f1-score**	**ROC-AUC**
All other causes	0.8783	0.8672	0.8456	0.8413
Corrosion	0.8846	0.8737	0.8422	0.8579
Excavation damage	0.8714	0.8505	0.8089	0.8447
Incorrect operation	0.8692	0.8385	0.817	0.8327
Material/weld/equip failure	0.8632	0.8526	0.8212	0.8168
Natural force damage	0.8606	0.84	0.84	0.8042
Other outside forces damage	0.8571	0.8163	0.8648	0.8405

Similarly, precision value, F1 score value, recall values and ROC-AUC value for different damages are depicted in Table 2, where higher precision rate, F1 score, Recall rate and ROC-AUC is obtained by material/weld/equip failure for logistic regression.

Table 3- Metrics for Random Forest

Random forest				
Types of Damages	**Value of Precision**	**Recall**	**Value of f1-score**	**ROC-AUC**
All other causes	0.9255	0.895	0.9133	0.8791
Corrosion	0.9109	0.8804	0.9089	0.9046
Excavation damage	0.9165	0.8759	0.8944	0.8901
Incorrect operation	0.9019	0.911	0.8794	0.9152
Material/weld/equip failure	0.9192	0.9014	0.8899	0.8856
Natural force damage	0.9084	0.9082	0.8667	0.9024
Other outside forces damage	0.9028	0.912	0.9004	0.8761

Like linear and logistic regression, model of RF and KNN is also assessed in Table 3 and Table 4. In Table 3, performance of Random Forest model is shown, where it is very well noticed different types of damages. It has high precision, recall, f1-score, and ROC-AUC values (mostly above 0.87), meaning it accurately detects and classifies each damage type. Some damage types like "All other causes" and "Material/weld/equip failure" are predicted slightly better than others, but overall, the model works reliably across all categories.

Table 4- Metrics for KNN

KNN				
Types of Damages	**Value of Precision**	**Recall**	**Value of f1-score**	**ROC-AUC**
All other causes	0.8781	0.8487	0.8569	0.8427
Corrosion	0.8737	0.8543	0.8625	0.8384
Excavation damage	0.8866	0.8667	0.8651	0.8208
Incorrect operation	0.8712	0.84	0.853	0.865
Material/weld/equip failure	0.8632	0.8426	0.8712	0.8468
Natural force damage	0.8506	0.82	0.8284	0.8342
Other outside forces damage	0.8671	0.8563	0.8348	0.8405

In Table 4, KNN model performs decently in classifying different types of damages, with precision, recall, f1-score, and ROC-AUC values generally ranging from about 0.82 to 0.88. This means the model is fairly accurate and consistent in identifying most damage categories, though its performance is slightly lower for "Natural force damage" and "Other outside forces damage." Overall, the KNN classifier provides reliable results, but it is a bit less accurate than the Random Forest model, especially for the more challenging damage types.

Table 5- Accuracy Values

Algorithm	**Value of Accuracy (%)**
Linear Regression	95.02%
Logistic Regression	89%
Random forest	92.46 %
KNN	87.94 %

In Table 5, Linear Regression demonstrated the highest accuracy in predicting outcomes, achieving 95.02%. In contrast, Random Forest secured the second-highest accuracy at 92.46%, while Logistic Regression and KNN exhibited comparable, albeit lower, performance, with accuracies of 89% and 87.94% respectively.

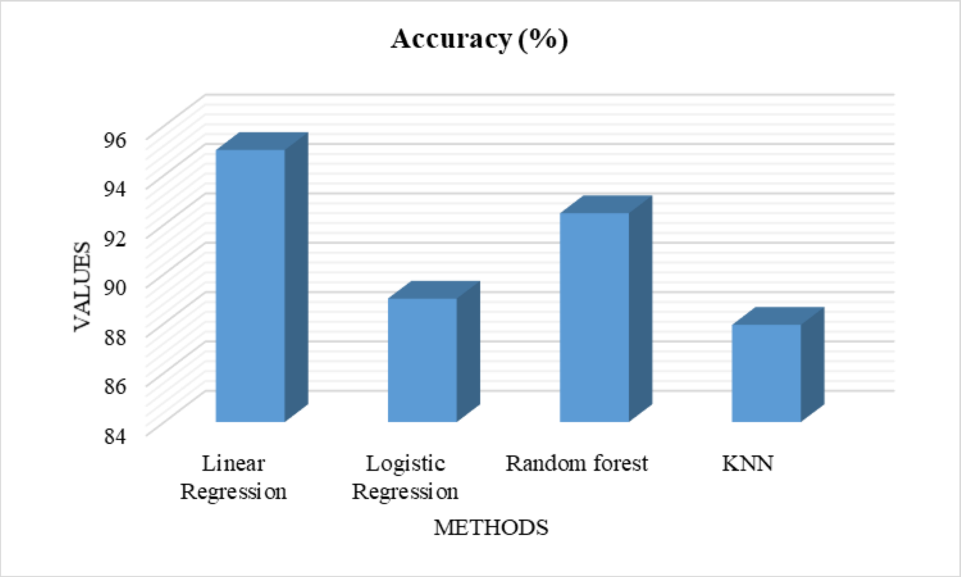


Figure 9. Graphical Representation

From the experimental outcomes Figure 9 it can be observed that, accuracy of linear regression is higher than other models because linear regression model can estimate the relationship between the input variables (input data) and output variables (presence of absence of gas leak) which aids as an effective detection algorithm. In addition, linear regression can handle massive dataset easily which makes it easier and effective for real time gas leakage detection as well.

5. Conclusion

The present work commences with the selection of a relevant dataset, from which multiple predictive models were constructed using Machine Learning algorithms, followed by a comparative analysis to determine the most efficacious model. Four algorithms such as Linear Regression, Logistic Regression, KNN, and Random Forest were employed for this analysis. The study focused on seven types of damages commonly occurring in gas plants like Corrosion, Excavation damage, incorrect operation, Material/weld/equip failure, Natural force damage, other outside forces damage and all other causes evaluating accuracy, precision, recall, and F-score across these different algorithms. The results indicate that Linear Regression demonstrated the highest accuracy at 95.02%. Random Forest followed with an accuracy of 92.46%, while Logistic Regression achieved 89%, and KNN registered 87.94%. These outcomes underscore the efficacy of the present model in industrial data settings, with the primary objective being real-world applicability. Implementing these models could lead to the development of systems capable of promptly identifying gas plant leakage incidents, thereby ensuring operational efficiency and mitigating potential harm to industrial entities and the environment.

Further research will also delve into explainable AI techniques to provide clearer insights into the model's decision-making process, fostering greater trust and facilitating proactive maintenance strategies.

Declaration:

Competing Interests – There is no competing interest for this study.
Funding Information – There is no funding for this study
Author contribution - All the author contribute equally to this paper.
Data Availability Statement – Not applicable
Research Involving Human and /or Animals – Not applicable
Informed Consent – Not application

References

[1] N. Reddy, J. Pothal, R. Barik, P. J. J. o. P. S. E. Senapati, and Practice, "Pipeline slurry transportation system: An overview," *Journal of Pipeline Systems Engineering Practice,* vol. 14, no. 3, p. 03123001, 2023.

[2] N. Yusuf and T. J. E. Al-Ansari, "Current and future role of natural gas supply chains in the transition to a low-carbon hydrogen economy: A comprehensive review on integrated natural gas supply chain optimisation models," *Energies,* vol. 16, no. 22, p. 7672, 2023.

[3] M. Hussain, T. Zhang, M. Chaudhry, I. Jamil, S. Kausar, and I. J. M. Hussain, "Review of prediction of stress corrosion cracking in gas pipelines using machine learning," vol. 12, no. 1, p. 42, 2024.

[4] C.-Y. Luo, S.-Y. Cheng, H. Xu, and P. J. P. c. s. Li, "Human behavior recognition model based on improved EfficientNet," vol. 199, pp. 369-376, 2022.

[5] A. M. Al-Sabaeei, H. Alhussian, S. J. Abdulkadir, and A. J. E. R. Jagadeesh, "Prediction of oil and gas pipeline failures through machine learning approaches: A systematic review," vol. 10, pp. 1313-1338, 2023.

[6] N. Ullah, Z. Ahmed, and J.-M. J. S. Kim, "Pipeline leakage detection using acoustic emission and machine learning algorithms," *Sensors,* vol. 23, no. 6, p. 3226, 2023.

[7] N. Ullah, Z. Ahmed, and J.-M. J. S. Kim, "Pipeline leakage detection using acoustic emission and machine learning algorithms," vol. 23, no. 6, p. 3226, 2023.

[8] S. S. Aljameel *et al.*, "Oil and Gas Pipelines Leakage Detection Approaches: A Systematic Review of Literature," *International Journal of Safety Security Engineering,* vol. 14, no. 3, 2024.

[9] Q. Huang, X. Shi, W. Hu, and Y. J. M. Luo, "Acoustic emission-based leakage detection for gas safety valves: Leveraging a multi-domain encoding learning algorithm," *Measurement,* vol. 242, p. 116011, 2025.

[10] P. Nooralishahi, F. López, and X. J. A. S. Maldague, "A Drone-Enabled Approach for Gas Leak Detection Using Optical Flow Analysis," vol. 11, no. 4, p. 1412, 2021.

[11] B. Lalithadevi and S. J. I. J. o. C. I. S. Krishnaveni, "ExAIRFC-GSDC: An Advanced Machine Learning-Based Interpretable Framework for Accurate Gas Leakage Detection and Classification," *International Journal of Computational Intelligence Systems,* vol. 18, no. 1, pp. 1-33, 2025.

[12] T. Liu, X. Cai, W. Zhou, K. Wang, and J. J. P. Wang, "Enhanced Detection of Pipeline Leaks Based on Generalized Likelihood Ratio with Ensemble Learning," *Processes,* vol. 13, no. 2, p. 558, 2025.

[13] S. S. Aljameel *et al.*, "An anomaly detection model for oil and gas pipelines using machine learning," *Computational Intelligence,* vol. 10, no. 8, p. 138, 2022.

[14] T. Aditiyawarman, A. P. S. Kaban, J. W. J. A.-A. J. o. R. Soedarsono, and P. B. M. E. Uncertainty in Engineering Systems, "A recent review of risk-based inspection development to support service excellence in the oil and gas industry: an artificial intelligence perspective," vol. 9, no. 1, p. 010801, 2023.

[15] J. Yuan, W. Mao, C. Hu, J. Zheng, D. Zheng, and Y. J. E. F. A. Yang, "Leak detection and localization techniques in oil and gas pipeline: A bibliometric and systematic review," vol. 146, p. 107060, 2023.

[16] M. Hussain, A. Alamri, T. Zhang, and I. Jamil, "Application of Artificial Intelligence in the Oil and Gas Industry," in *Engineering Applications of Artificial Intelligence*: Springer, 2024, pp. 341-373.

[17] S. Lee and B. J. S. Kim, "Machine Learning Model for Leak Detection Using Water Pipeline Vibration Sensor," vol. 23, no. 21, p. 8935, 2023.

[18] K. Muthumanickam, P. Vijayalakshmi, S. Kumarganesh, T. Kumaravel, K. M. Sagayam, and L. M. Alkwai, "An efficient gas leakage detection and smart alerting system using IoT," in *Artificial Intelligence and Blockchain in Industry 4.0*: CRC Press, pp. 194-213.

[19] F. Wang *et al.*, "Oil and gas pipeline leakage recognition based on distributed vibration and temperature information fusion," vol. 5, p. 100131, 2021.

[20] J. Shi *et al.*, "Real-time leak detection using an infrared camera and Faster R-CNN technique," vol. 135, p. 106780, 2020.

[21] T. P. Melo, J. Andrade, and K. S. J. C. E. J. Komati, "A Pipeline for Multivariate Time Series Forecasting of Gas Consumption in Pelletization Process," *CLEI Electronic Journal,* vol. 28, no. 3, pp. 2: 1-2: 12, 2025.

[22] I. E. O. N. L. O. Godsday Idanegbe Usiabulu, "Automation of Gas Leak Detection: AI and Machine Learning Approaches for Gas Plant Safety," *Global Journal of Researches in Engineering: J General Engineering,* vol. 24, no. 2, 2024.

[23] J. Enerio, M. Kim, C. Duong, and S. J. A. a. S. Misra, "Machine Learning for Gas Leak Detection and Forecasting," 2024.

[24] O. Akinsete and A. Oshingbesan, "Leak detection in natural gas pipelines using intelligent models," *Society of Petroleum Engineers,* p. D023S009R001, 2019.

[25] V. Yadukrishnan, N. Mohan, K. Soman, and S. S. Kumar, "Machine learning Techniques Based Gas Pipeline Leakage Detection," *IEEE,* pp. 1-5, 2024.

[26] P. P. D. Ferreira, D. P. Kappes, E. M. Oliveira, M. L. da Fonseca, and A. P. S. Medeiros, "LEAK DETECTION SYSTEM USING MACHINE LEARNING TECHNIQUES."

[27] C. Shirley, J. I. J. Raja, S. Evangelin Sonia, I. J. M. T. Titus, and Applications, "Recognition and monitoring of gas leakage using infrared imaging technique with machine learning," pp. 1-14, 2023.

[28] N. K. Banjara, S. Sasmal, S. J. I. J. o. P. V. Voggu, and Piping, "Machine learning supported acoustic emission technique for leakage detection in pipelines," vol. 188, p. 104243, 2020.

[29] G. I. Usiabulu, O. Joel, L. Nosike, V. Aimikhe, E. J. J. o. E. R. Okafor, and Reports, "A New Pressure-Based Modeling Approach for Early Leak Detection in Gas Processing Plants Using Machine Learning," vol. 25, no. 6, pp. 18-27, 2023.

[30] O. Akinsete and A. Oshingbesan, "Leak detection in natural gas pipelines using intelligent models," 2019, p. D023S009R001: SPE.

[31] L. Peng, X. Huang, J. Chen, P. Yang, C. Xing, and C. J. J. o. L. P. i. t. P. I. Zhao, "A method for real-time estimation of gas leakage flow from leakage source based on point detection data," vol. 78, p. 104822, 2022.

[32] S. S. Aljameel *et al.*, "An anomaly detection model for oil and gas pipelines using machine learning," vol. 10, no. 8, p. 138, 2022.

[33] X. Wang, A. Li, Z. Lin, S. Li, Y. J. J. o. P. S. E. Yang, and Practice, "Natural Gas Transmission Pipeline Leak Detection Model Based on Acoustic Emission and Machine Learning," *Journal of Pipeline Systems Engineering Practice,* vol. 15, no. 4, p. 04024047, 2024.

[34] P. Hubert and L. J. a. p. a. Padovese, "A machine learning approach for underwater gas leakage detection," 2019.

[35] C. Spandonidis, P. Theodoropoulos, and F. J. S. Giannopoulos, "A combined semi-supervised deep learning method for oil leak detection in pipelines using iiot at the edge," vol. 22, no. 11, p. 4105, 2022.

[36] T. De Kerf, J. Gladines, S. Sels, and S. J. R. s. Vanlanduit, "Oil spill detection using machine learning and infrared images," vol. 12, no. 24, p. 4090, 2020.

[37] K. P. Yunjeong Gu, Wonhee Lee, Byunghun Song, Jungpyo Hong, Junho Shin and J. J. o. t. K. S. o. H. Mitigation, "Research on Deep learning based Ultrasonic Image Learning to Develop a Gas Leak Detection Model," vol. 23, no. 6, pp. 135-143, 2023.

[38] S.-B. Zhu, Z.-L. Li, X. Li, H.-h. Xu, and X.-m. J. M. Wang, "Convolutional neural networks-based valve internal leakage recognition model," vol. 178, p. 109395, 2021.

[39] J. Wang, J. Ji, A. P. Ravikumar, S. Savarese, and A. R. J. E. Brandt, "VideoGasNet: Deep learning for natural gas methane leak classification using an infrared camera," vol. 238, p. 121516, 2022.

[40] S. C. V. Kutcharlapati, M. Sundararamaiah, P. Manukonda, and K. R. Mudunuru, "IMPLEMENTING MACHINE LEARNING FOR GAS PIPELINE LEAK PREDICTION," *International Research Journal of Modernization in Engineering Technology and Science,* vol. 6, no. 11, 2024.

[41] A. P. Ekong, G. G. James, I. J. J. o. I. S. Ohaeri, and Informatics, "Oil and gas pipeline leakage detection using iot and deep learning algorithm," *Journal of Information Systems Informatics,* vol. 6, no. 1, pp. 421-434, 2024.

[42] S. B. Nuthalapati, M. Arun, C. Prajitha, S. Rinesh, and K. Abubeker, "Computer Vision Assisted Deep Learning Enabled Gas Pipeline Leak Detection Framework," *IEEE,* pp. 950-957, 2024.