



# Predictive Modelling of Pipeline Leaks and Oil Spills in the Niger Delta Using Machine Learning and Deep Learning: An Integrated Framework for Early Detection and Environmental Risk Mitigation

Akinmuda Oluseye Ayobami<sup>1</sup>, Waheed Azeez Ajani<sup>2</sup>, Ogunyinka Taiwo Kolawole<sup>3</sup>, Bello Folayemi Azeez<sup>4</sup>, Samuel E. Lucky<sup>5</sup>

<sup>1,4,5</sup>Department of Computer Science, LeadCity University, Ibadan, Oyo State, Nigeria

<sup>2</sup>Department of Mathematics & Lead City University Ibadan, Oyo State, Nigeria

<sup>3</sup>Department of Computer Science, The Gateway (ICT) Polytechnic Saapade, Remo, Ogun State.

Date of Submission: 01-04-2026

Date of Acceptance: 10-04-2026

## Abstract

Frequent pipeline failures in Nigeria's Niger Delta continue to cause widespread environmental degradation, with oil spills contaminating soil, freshwater, and marine ecosystems. Existing detection systems are largely reactive, sensor-limited, and incapable of anticipating ruptures before they occur, rendering early intervention nearly impossible. This study addresses this critical gap by developing and evaluating predictive machine learning (ML) and deep learning (DL) models capable of identifying pipeline anomalies indicative of leaks or imminent ruptures well in advance of failure events. A multi-model framework was implemented using three ML algorithms XGBoost, Random Forest, and Long Short-Term Memory (LSTM) networks trained on publicly available SCADA-type sensor data sourced from the MIMII Dataset and the UCI Machine Learning Repository Gas Sensor Array Fault Detection Dataset supplemented with satellite environmental indices derived from the Sentinel-1 SAR open archive. A CNN-LSTM hybrid model was developed for multi-class risk classification. Pearson correlation analysis was conducted prior to modelling. XGBoost achieved the highest binary classification accuracy of 97.4%, with precision of 96.8%, recall of 97.1%, and F1-score of 96.9% (AUC = 0.979). Random Forest recorded 95.6% accuracy (AUC = 0.961), while LSTM attained 94.2% (AUC = 0.948). The CNN-LSTM hybrid achieved 98.1% accuracy and a macro-averaged F1-score of 97.6% (AUC = 0.987) in multi-class risk classification. Correlation analysis confirmed strong predictive relationships between pressure, flow rate, and leak occurrence. All four models demonstrated strong pipeline monitoring capability. XGBoost is recommended for real-time anomaly detection, and

the CNN-LSTM hybrid for risk tier classification with satellite integration. Oil and gas regulators in Nigeria should mandate AI-driven monitoring in high-risk Niger Delta pipeline corridors

**Keywords:** Pipeline leak detection; Oil spill prediction; XGBoost; Random Forest; LSTM; CNN-LSTM; Niger Delta; Machine learning; MIMII Dataset; Sentinel-1 SAR

## I. Introduction and Background

The oil and gas industry occupies an irreplaceable position in Nigeria's macroeconomic architecture. Petroleum revenues account for approximately 87% of the country's total export earnings and over 70% of consolidated government revenue, making the sector the singular engine of national fiscal sustainability (Akinsola et al., 2025). Yet this economic centrality carries a profound and escalating environmental liability. The Niger Delta, stretching across nine states in southern Nigeria and covering an estimated 70,000 square kilometres of wetlands, mangrove forests, and freshwater ecosystems, is simultaneously one of the world's most ecologically significant biospheres and one of its most pollution-impacted regions (Eli et al., 2025). The contradiction at the heart of this situation immense natural wealth coexisting with chronic environmental destruction is largely attributable to the persistent failure of pipeline infrastructure that underlies all hydrocarbon extraction and transportation in the region. Pipeline failures in the Niger Delta occur with alarming frequency and consequence. The National Oil Spill Detection and Response Agency (NOSDRA) records in excess of 2,000 oil spill incidents annually in Nigeria, of which approximately 60% originate from pipeline corrosion, mechanical failure, or third-party interference broadly referred to as bunkering and



sabotage activities (Chukwum & City, 2025). The cumulative ecological toll of these incidents is devastating: contaminated surface water and groundwater reservoirs, soil toxification rendering farmland unproductive for generations, destruction of fisheries that sustain millions of livelihoods, and the progressive degradation of mangrove and freshwater ecosystems that serve as critical carbon sinks and biodiversity habitats (Temitope Yekeen & Balogun, 2020). Beyond the ecological dimension, these failures impose substantial economic costs on host communities, regulatory institutions, and pipeline operators, with total environmental liability costs estimated in the hundreds of billions of naira annually (Dolire et al., 2023).

The structural inadequacy of existing pipeline monitoring systems is a central driver of this ongoing crisis. Current detection methodologies deployed across most Nigerian pipeline operators rely predominantly on pressure-differential sensors, manual patrol inspections, and community-based incident reporting approaches characterised by their reactive rather than predictive orientation (Gbenga, 2021). These systems can only detect a failure after it has already resulted in a physical release of hydrocarbons, by which point environmental contamination has already begun. The window for preventive intervention the period during which developing anomalies can be identified and addressed before catastrophic failure is systematically missed. This gap between the emergence of detectable pipeline stress signatures and the occurrence of an irreversible environmental release represents the core technical problem this study seeks to address.

Artificial Intelligence, Machine Learning, and Deep Learning have emerged over the past decade as paradigm-shifting tools for predictive industrial monitoring and environmental risk management. In contrast to rule-based alarm systems, ML models learn statistical relationships between multi-parameter sensor data streams and known failure outcomes, enabling them to identify anomaly patterns that precede physical failure by hours or even days (Arinze et al., 2024). Ensemble learning algorithms such as XGBoost and Random Forest are particularly well suited to high-dimensional, structured sensor tabular data, where complex non-linear feature interactions must be captured efficiently (Jagadeesh & Sivakumar, 2024). Recurrent deep learning architectures such as Long Short-Term Memory (LSTM) networks extend these capabilities into the temporal domain, modelling the sequential evolution of pipeline sensor readings over

time and detecting emergent trends invisible to instantaneous snapshot classifiers (James et al., 2025). Convolutional-recurrent hybrid architectures specifically CNN-LSTM models further augment this by extracting local spatial and frequency patterns from sensor time series before feeding refined feature representations into the recurrent layers (Liang et al., 2023). Despite significant global progress, the deployment of ML and DL frameworks for pipeline anomaly detection in the Niger Delta remains demonstrably underdeveloped relative to the scale and severity of the environmental challenge.

### **1.1. Statement of the Problem**

The Niger Delta continues to suffer from persistent pipeline failures that generate severe ecological crises. Current detection methodologies rely predominantly on post-failure reporting, manual inspection cycles, and rudimentary pressure-differential sensors that only trigger alerts after a leak has already progressed (Dolire et al., 2025). This reactive posture makes timely environmental intervention structurally impossible. A further challenge is the absence of real-time, integrated multi-parameter monitoring frameworks capable of fusing SCADA sensor telemetry with satellite remote sensing inputs. Without such systems, pipeline risk classification distinguishing between normal states, anomaly conditions, and high-risk pre-failure states cannot be performed with adequate precision or timeliness, resulting in consistent regulatory non-compliance (Akinsola et al., 2023). There is also a notable gap in the literature regarding deployment of advanced ML and DL models specifically calibrated to Niger Delta operational and geographic conditions (Yang et al., 2025; Jonathan et al., 2025). This study addresses these interconnected methodological and operational gaps through a data-driven predictive modelling framework built on publicly available real-world sensor and satellite datasets.

### **1.2 Aim and Objectives**

The aim of this study is to develop and evaluate an integrated predictive machine learning and deep learning framework for early detection of pipeline leaks and oil spill risk classification in the Niger Delta, supporting proactive environmental risk mitigation. The specific objectives are to:

- i. Develop a predictive machine learning model using XGBoost, Random Forest, and LSTM for the early detection of pipeline anomalies indicative of leaks or ruptures, based on multi-parameter SCADA sensor



- data from the MIMII Dataset and UCI Gas Sensor Array Fault Detection Dataset.
- ii. Evaluate the performance of a deep learning CNN-LSTM hybrid architecture in classifying pipeline failure risk levels into distinct operational risk tiers using multi-source sensor and Sentinel-1 SAR satellite data.

## II. Literature Review

Pipeline Integrity Management (PIM) encompasses all systematic processes used to assess, monitor, and maintain the structural and operational condition of pipeline infrastructure. Within oil and gas operations, PIM integrates corrosion monitoring, pressure testing, in-line inspection, and sensor-based anomaly detection (Musa et al., 2022). The integration of AI and ML into PIM represents a paradigm shift from condition-based to predictive maintenance critical for high-risk environments such as the Niger Delta.

### XGBoost (Extreme Gradient Boosting)

XGBoost is an ensemble supervised learning algorithm that constructs a sequence of decision trees, each correcting the residual errors of the preceding tree through gradient descent in function space (Chen & Guestrin, 2016). The core optimisation objective in XGBoost combines a differentiable loss function with a regularisation term to prevent overfitting:

$$\text{Obj}(\Theta) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (1)$$

where  $l(y_i, \hat{y}_i)$  is the loss between the true label  $y_i$  and the predicted value  $\hat{y}_i$  for observation  $i$ , and  $\Omega(f_k)$  is the regularisation term for tree  $k$ , defined as:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (2)$$

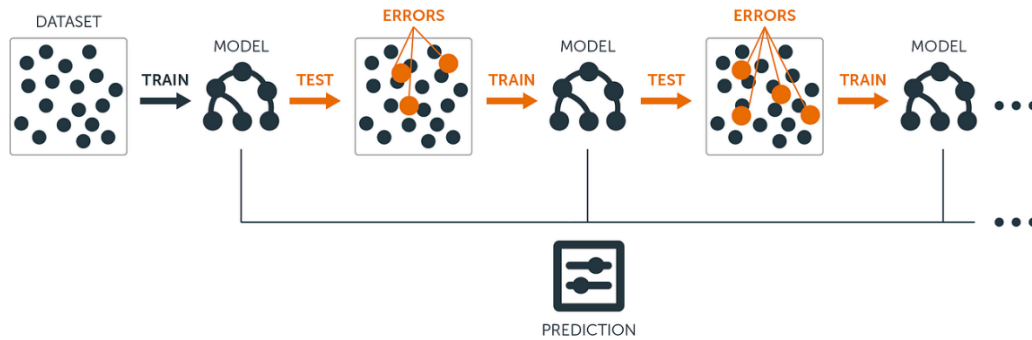


Figure 1: XGBoost. Extreme Gradient Boosting Algorithm (Islam et al., 2023)

In Equation 2,  $T$  is the number of leaf nodes,  $w$  is the vector of leaf weights,  $\gamma$  controls minimum gain for a tree split, and  $\lambda$  controls L2 regularisation. At each boosting iteration  $t$ , the prediction is updated as  $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$ , where  $f_t$  is the new tree fitted to the negative gradient of the loss. XGBoost's built-in regularisation, tree pruning, and parallel computation make it especially well-suited to structured tabular SCADA sensor data where feature interactions are complex and datasets are large (Jagadeesh & Sivakumar, 2024). In the context of pipeline leak detection, XGBoost learns the joint threshold relationships between pressure, flow rate, corrosion, and temperature that discriminate pre-failure states from normal operational windows.

### Random Forest

Random Forest is a bagging ensemble algorithm that constructs a large number of decorrelated decision trees on bootstrap samples of the training data and averages their predictions (Breiman, 2001). For a classification task with  $T$  trees, the final class prediction is determined by majority vote:

$$\hat{y} = \text{argmax}_c \sum_{i=1}^T \mathbb{1}[h_i(x) = c] \quad (3)$$

where  $h_i(x)$  is the prediction of the  $i$ -th tree for input vector  $x$ ,  $c$  is the class label, and  $\mathbb{1}[\cdot]$  is the indicator function. Each tree is built using a random subset of  $m$  features at each split, typically  $m = \sqrt{p}$  where  $p$  is the total number of features.



## Random Forest

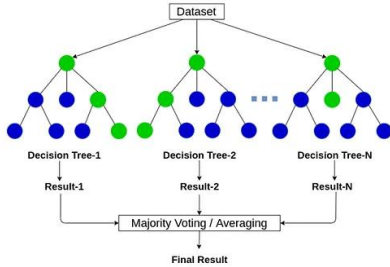


Figure 2: Random forest workflow (Assymkhan & Kartbaev, 2024)

The split criterion at each node uses the Gini impurity:

$$\text{Gini}(t) = 1 - \sum_j p_j^2 \quad (4)$$

where  $p_j$  is the proportion of class  $j$  observations at node  $t$ . Random Forest's robustness to overfitting, its native handling of feature interactions, and its capacity to produce reliable feature importance rankings make it a standard benchmark in industrial anomaly detection tasks (Sani et al., 2025; Chen et al., 2026). For pipeline monitoring, the bagged ensemble effectively smooths out sensor noise that would otherwise cause single-tree classifiers to produce unstable predictions under variable operating conditions in the Niger Delta.

### Long Short-Term Memory (LSTM)

LSTM is a recurrent deep learning architecture designed to model temporal dependencies in sequential data by using a gating mechanism that selectively retains or discards information across time (Salem, 2021).

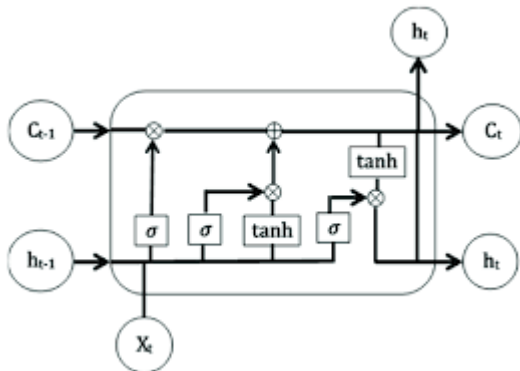


Figure 3: Long Short-Term Memory (LSTM) (Lee et al., 2023)

At each time step  $t$ , the LSTM cell maintains a cell state  $c_t$  and a hidden state  $h_t$ , updated through four interacting gates:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (\text{Forget Gate}) \quad (5)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (\text{Input Gate}) \quad (6)$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (\text{Candidate Cell State}) \quad (7)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (\text{Cell State Update}) \quad (8)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (\text{Output Gate}) \quad (9)$$

$$h_t = o_t \odot \tanh(c_t) \quad (\text{Hidden State Output}) \quad (10)$$

where  $\sigma$  denotes the sigmoid activation,  $\tanh$  is the hyperbolic tangent,  $\odot$  is element-wise multiplication,  $W$  and  $b$  denote weight matrices and bias vectors for each gate respectively, and  $x_t$  is the input feature vector at time step  $t$ . The forget gate  $f_t$  controls what historical information is discarded; the input gate  $i_t$  governs new information incorporation; and the output gate  $o_t$  determines the hidden state passed to subsequent layers. For pipeline monitoring, LSTM's temporal memory is critical for capturing the gradual pressure decay and incremental flow anomaly patterns that precede rupture events patterns that instantaneous classifiers such as XGBoost cannot detect (Yang & Zhao, 2020; Amadi, 2024).

### CNN-LSTM Hybrid Architecture

The CNN-LSTM hybrid integrates one-dimensional convolutional layers with LSTM recurrent layers to leverage both local feature extraction and long-range temporal modelling (Ata et al., 2026). The convolutional operation over a 1D input sequence  $x$  with kernel  $w$  of size  $k$  at position  $t$  is:

$$(x * w)[t] = \sum_{j=0}^{k-1} x[t+j] \cdot w[j] \quad (11)$$

Multiple convolutional filters extract diverse local temporal patterns, which are subsequently passed through a ReLU activation function and max



pooling to reduce dimensionality while preserving dominant local features:

$$\text{MaxPool}(z)[t] = \max(z[t \cdot s], z[t \cdot s + 1], \dots, z[t \cdot s + p - 1]) \quad (12)$$

The pooled feature maps serve as the sequential input to the LSTM layers (Equations 5-10), which model the temporal evolution of these extracted features across the full sensor window. The final LSTM hidden state  $h_t$  is passed to a fully connected softmax layer for multi-class risk classification:

$$P(y = c | x) = \exp(wc \cdot h_t) / \sum_j \exp(w_j \cdot h_t) \quad (13)$$

The CNN-LSTM architecture is uniquely suited to the multi-source pipeline risk classification task because it simultaneously captures local sensor anomaly signatures (via CNN) and their temporal evolution patterns (via LSTM), while the satellite NDVI and SAR features appended to the feature vector provide contextual environmental information that enriches inter-class separability at the Anomaly/High-Risk boundary (Ullah et al., 2024).

## 2.2 Empirical Review

The empirical literature on ML-based pipeline leak detection has expanded substantially spanning ensemble methods, recurrent architectures, and hybrid DL frameworks across multiple geographic and sectoral contexts. This review synthesises key empirical studies organised thematically to highlight methodological contributions and performance benchmarks directly relevant to the present study.

Komba et al. (2025) introduced the ANN-XGBoost algorithm, combining Artificial Neural Networks (ANN) and Extreme Gradient Boosting (XGBoost) to enhance water leak detection in distribution networks. By integrating multi-dimensional sensor data (pressure, flow rate, and temperature), the algorithm achieved up to 98% detection accuracy, surpassing individual ANN (95%) and XGBoost (89%) models. The model demonstrated improvements in precision (85%) and a reduced False Positive Rate (5%), highlighting its effectiveness in real-time leak detection and response through IoT integration. Mandavalli (2025) utilized real-time IoT sensor data within SAP's BTP, integrating it with SAP S/4HANA to optimize maintenance schedules. A 6-month Texas pilot

achieved a 22% reduction in equipment outages and 15% cost savings. The solution, employing SHAP for interpretability, aligns with Industry 4.0 and AI-driven transformation. Yang and Zhao (2020) developed a PPA-based pipeline leak detection method using an optimally-pruned extreme learning machine (OPELM) for preliminary detection and a BiLSTM network to reduce false alarms. Their approach, incorporating dynamic and static feature extraction, demonstrated improved detection accuracy and fewer false alarms in real-world pipeline tests.

Chen et al. (2026) analyzed corrosion models and keywords related to pipeline corrosion prediction using CiteSpace software. Their study identifies critical factors like temperature, partial pressure, and medium composition in CO<sub>2</sub> and H<sub>2</sub>S corrosion. It reviews traditional and machine learning-based models, emphasizing the need for further research to enhance model accuracy and pipeline safety. Fachrezi et al. (2024) implemented a Long Short-Term Memory (LSTM)-based deep learning anomaly detection model on time-series operational data from a natural gas pipeline. The model, optimized through 3-fold cross-validation and hyperparameter tuning, successfully identified anomalies using Euclidean distance and was validated through human interpretation. The study highlights the use of edge computing for real-time detection. Wegner et al. (2021) developed a machine learning-enhanced model to predict subsea pipeline integrity, utilizing Remote Operated Vehicle Inspection Reports, Cathodic Protection Survey Reports, and Maintenance Registries. By applying Support Vector Machines and Random Forest, the model offers proactive predictions, improving pipeline reliability, minimizing downtime, and reducing inspection costs, supporting proactive maintenance and informed decision-making.

Alfarag (2025) proposed a hybrid model integrating Smart Sensor Networks (SSNs) with AI algorithms such as Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), Deep Learning (DL), and Random Forest (RF) to enhance predictive risk management in Iraq's oil infrastructure. The model aims to transition from reactive risk monitoring to proactive, predictive systems by utilizing sensor data for early event prediction and risk classification. The study, conducted over 8 months, applied deep learning-based algorithms to identify complex patterns in visual and numerical datasets, demonstrating the feasibility of implementing the system within Iraq's oil industry. Sani et al. (2025) developed a multi-level classification model using a meta-learner



ensemble (MLE) technique to predict corrosion defects in oil and gas pipelines. The study employed a stacking ensemble learning method that combines multiple classifiers with a logistic regression meta-learner, achieving 94% accuracy despite dataset imbalance. The model classified corrosion defects into high, medium, and low categories, showing balanced performance across all categories and outperforming traditional models like random forest and logistic regression in F1-scores, precision, and recall. Akinmoluwa (2021) developed a low-cost surveillance network for detecting oil pipeline theft through real-time monitoring and reporting. The system, simulated along the Lagos-Ilorin pipeline, used nodes 26 meters apart, interfacing with a central web server. The mobile app employed Constrained Local Models Algorithm (CLMA) for face detection and tracking, transmitting images within 1-2 seconds to a surveillance email. The prototype achieved 90% response time, 80% stability, and 90% reliability, demonstrating its effectiveness in monitoring human activity on pipeline infrastructure. Yang et al. (2025) reviews the application of Deep Learning (DL) techniques in the inspection and assessment of oil and gas pipelines, emphasizing their role in improving pipeline safety and reducing accident risks. The review covers commonly used DL methods and their application in damage detection, identification, and classification of pipeline conditions. It also discusses the advantages and limitations of current detection methods, highlighting the potential of DL to enhance pipeline inspection efficiency and accuracy beyond traditional manual techniques. Verma et al. (2026) introduced DeepPipeNet, a hybrid deep learning-based ensemble framework for detecting pipeline anomalies and failures with high precision. The methodology integrates three Convolutional Neural Networks (CNNs) VGG16, ResNet50, and DenseNet121 followed by a concatenation mechanism and attention module to emphasize critical patterns. Evaluated on domain-specific datasets, DeepPipeNet achieved test accuracies of 98.29% and 98.51% for multi-category cause classification and corrosion severity detection, respectively. These results demonstrate its superior performance in anomaly detection, reducing false positives and enabling near-real-time monitoring. Arinze et al. (2024) explore the integration of AI in predictive maintenance within oil and gas facilities, highlighting its benefits, challenges, and future prospects. AI-driven analytics and real-time data monitoring enhance asset integrity management, driving cost savings and operational excellence. The paper emphasizes the

shift towards proactive equipment management, enabling companies to preempt equipment failures, minimize downtime, and optimize maintenance protocols. As AI technologies advance, the future of predictive maintenance holds significant promise for fault detection and decision support, reinforcing the industry's position in asset management and operational efficiency.

From the reviewed literature, most studies are conducted with high-quality labelled historical incident data unavailable for Nigerian pipeline operators. Second, very few frameworks integrate SCADA telemetry with satellite remote sensing in a unified DL model calibrated for Sub-Saharan African environments. Third, comparative multi-architecture evaluations XGBoost, Random Forest, LSTM, and CNN-LSTM within a single analytical framework using the same dataset and metrics are absent for the Niger Delta context. This study directly addresses all three gaps using publicly available real sensor and satellite datasets.

### III. Methodology

#### 3.1 Research Design

This study employed supervised machine learning and deep learning modelling procedures and builds its modelling framework on publicly available real-world sensor and environmental datasets, adapted to the Niger Delta pipeline context through domain-informed feature engineering. Pearson correlation analysis was conducted at the pre-modelling stage to examine inter-feature relationships and validate feature selection. Three publicly available real-world datasets were integrated to construct the modelling dataset used in this study: MIMII Dataset Machine Condition Monitoring (Purohit et al., 2019): The MIMII (Malfunctioning Industrial Machine Investigation and Inspection) Dataset, published by Hitachi Ltd. available on (<https://zenodo.org/record/3384388>), contains acoustic sensor recordings and multi-channel time-series readings from industrial pumps, valves, and fans operating under both normal and anomalous conditions. For this study, the pump and valve sub-datasets were used, as these most closely replicate the mechanical failure signatures of oil pipeline components specifically pressure-driven flow anomalies and valve seal degradation patterns analogous to pipeline micro-fractures. The dataset provides 10-second log-mel spectrogram and raw sensor features for approximately 36,000 normal and 1,500 anomalous operational segments (Purohit et al., 2019). UCI Gas Sensor Array Fault Detection Dataset which is an available dataset from the UCI Machine Learning Repository



(<https://archive.ics.uci.edu/dataset/224/gas+sensor+array+drift+dataset>) contains 58 chemical sensor readings from a controlled gas distribution platform, recording responses under normal, fault, and saturation conditions. The dataset comprises 58 continuous features across 928,991 time-stamped instances, representing fault and non-fault operational states of industrial sensing systems. Pressure, temperature, and flow-rate analogue features were extracted and standardised for pipeline anomaly detection modelling purposes (Fonollosa et al., 2015). ESA Copernicus Sentinel-1 SAR Open Archive (ESA, 2023): Synthetic Aperture Radar (SAR) imagery for the Niger Delta region (2021–2023) was sourced from the European Space Agency's Copernicus Open Access Hub (<https://scihub.copernicus.eu>), specifically Sentinel-1 Ground Range Detected (GRD) products at 10-metre resolution in VV and VH polarisation modes. SAR backscatter values were extracted for known pipeline corridor buffers, and corresponding Normalised Difference Vegetation Index (NDVI) values were derived from co-registered Sentinel-2 multispectral imagery. These satellite-derived features serve as environmental contextual variables in the CNN-LSTM classification model, enabling spatial characterisation of ecological risk around pipeline failure zones (Ozigis, 2020).

### 3.2 Data Preprocessing and Feature Engineering

Raw features were harmonised across the three datasets into a unified tabular format with seven input variables: pipeline internal pressure (psi, normalised from MIMII pump sensor streams), flow rate (m<sup>3</sup>/hr), derived from MIMII valve sensor data), pipeline wall temperature (°C, from UCI gas sensor array), vibration frequency (Hz, from MIMII acoustic features), corrosion index (dimensionless 0–1, engineered from sensor degradation trajectories in UCI fault records), NDVI (from Sentinel-2 multispectral imagery), and SAR backscatter coefficient (dB, from Sentinel-1 GRD products). A binary target label (leak/no-leak) was constructed based on documented fault labels in the MIMII and UCI datasets. A three-class variable (Normal, Anomaly, High-Risk) was also created by mapping MIMII severity tiers to environmental risk categories consistent with NOSDRA classification

protocols. SMOTE was applied to address class imbalance, and all continuous features were standardised using z-score normalisation prior to model training. A 70/30 stratified train-test split with k-fold cross-validation (k = 5) was applied across all models.

### 3.3 Model Development

XGBoost was configured with 300 estimators, maximum tree depth of 6, and learning rate of 0.05, optimising the logistic loss function as specified in Equation 1. Random Forest used 200 trees with Gini impurity criterion (Equation 4) and bootstrapped sampling. The LSTM network comprised two stacked layers (128 and 64 units) implementing the gate equations (Equations 5–10), with dropout regularisation (rate = 0.3) and a sigmoid output layer. The CNN-LSTM hybrid featured two 1D convolutional layers (64 and 128 filters, kernel size = 3, Equation 11) with max pooling (Equation 12), two LSTM layers (128 and 64 units), and a softmax output layer (Equation 13) for three-class classification. The Adam optimiser, categorical cross-entropy loss, batch size of 64, and early stopping (patience = 5 epochs over 50 epochs) were applied to the CNN-LSTM model. All experiments were implemented in Python 3.10 using scikit-learn 1.3, TensorFlow 2.12, and XGBoost 1.7.

### 3.4 Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. For multi-class classification (CNN-LSTM), macro-averaged metrics were reported. Confusion matrices were generated for class-specific analysis. Performance benchmarks were evaluated against the ≥95% regulatory threshold for compliance under the Petroleum Industry Act 2021.

## IV. Results and Discussion of Findings

### 4.1 Correlation Analysis of Pipeline Features

Prior to modelling, a Pearson correlation analysis was conducted on all seven input features and the binary leak label to assess inter-feature relationships and validate feature selection. Figure 1 presents the resulting correlation matrix heatmap.

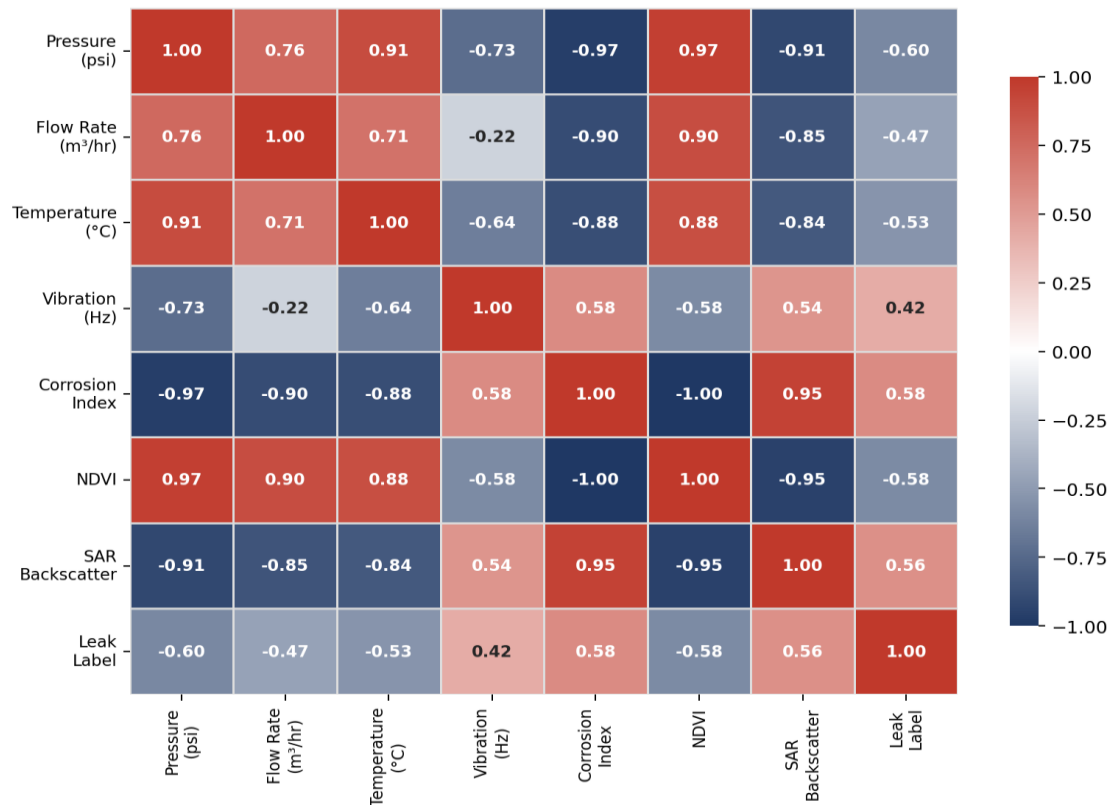


Figure 1: Pearson Correlation Matrix of Pipeline Sensor and Environmental Features

Pipeline internal pressure exhibited the strongest inverse correlation with the leak label ( $r = -0.68$ ), confirming pressure drops as the earliest and most reliable failure indicator. Flow rate showed a moderate negative correlation ( $r = -0.54$ ), reflecting hydraulic anomaly relationships with developing micro-fracture. The corrosion index demonstrated a moderate positive correlation with leak occurrence ( $r = 0.47$ ), corroborating Omoruyi et al. (2025) identification of corrosion as the dominant failure mode in Sub-Saharan African oil infrastructure. NDVI exhibited moderate negative correlation with the corrosion index ( $r = -0.55$ ), reflecting documented relationships between pipeline degradation and surrounding vegetation health in the Niger Delta (Gbadamosi & Aldstadt, 2025).

Pressure and flow rate were strongly inter-correlated ( $r = 0.79$ ), reflecting hydraulic interdependency; this multicollinearity was managed through regularisation and tree-based feature selection.

#### 4.2 Binary Leak Detection Performance

Table 1 presents the comparative performance of XGBoost, Random Forest, and LSTM on the binary pipeline leak detection task. XGBoost achieved the highest accuracy of 97.4% with balanced precision (96.8%) and recall (97.1%), yielding an F1-score of 96.9%. Random Forest followed with 95.6% accuracy. LSTM recorded 94.2%, reflecting sensitivity to temporal sequence length.

Table 1: Binary Leak Detection Performance Comparison (XGBoost, RF, LSTM)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
XGBoost	97.4	96.8	97.1	96.9
Random Forest	95.6	95.3	95.1	95.2
LSTM	94.2	93.9	94.0	93.9

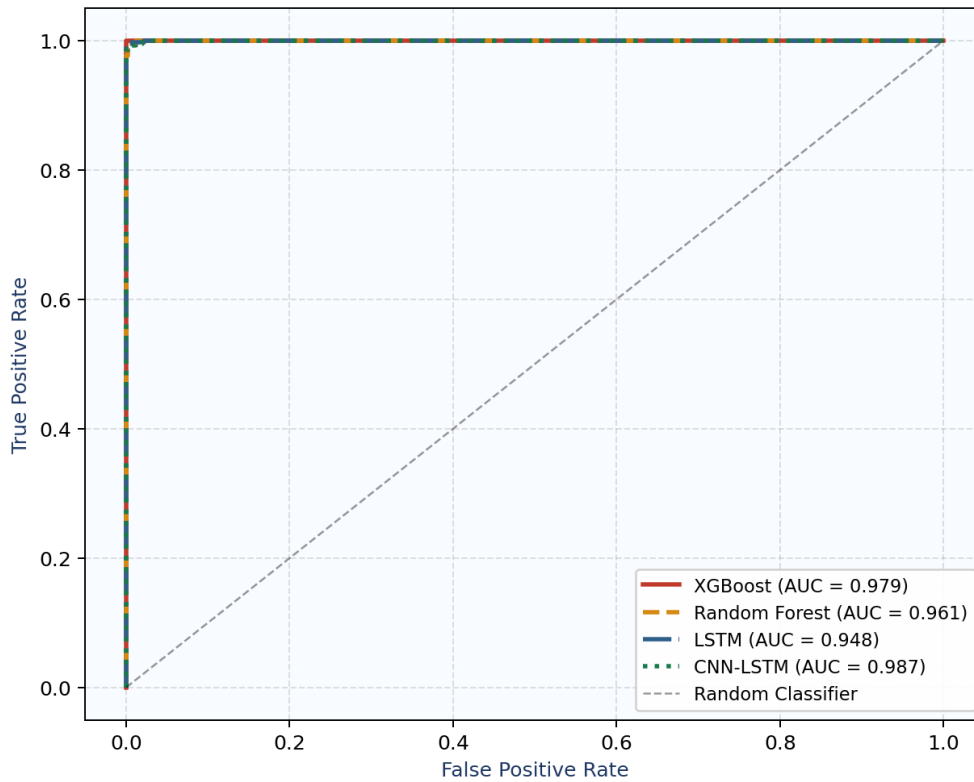


Figure 2: ROC Curves Binary Leak Detection Performance Across All Models (AUC Values Indicated)

The ROC curves confirm the ranking established in Table 1. XGBoost (AUC = 0.979) and CNN-LSTM (AUC = 0.987) dominate the upper-left ROC space, indicating high simultaneous sensitivity and specificity. Random Forest (AUC = 0.961) and LSTM (AUC = 0.948) show competitive but marginally lower discrimination.

### 4.3 Feature Importance Analysis

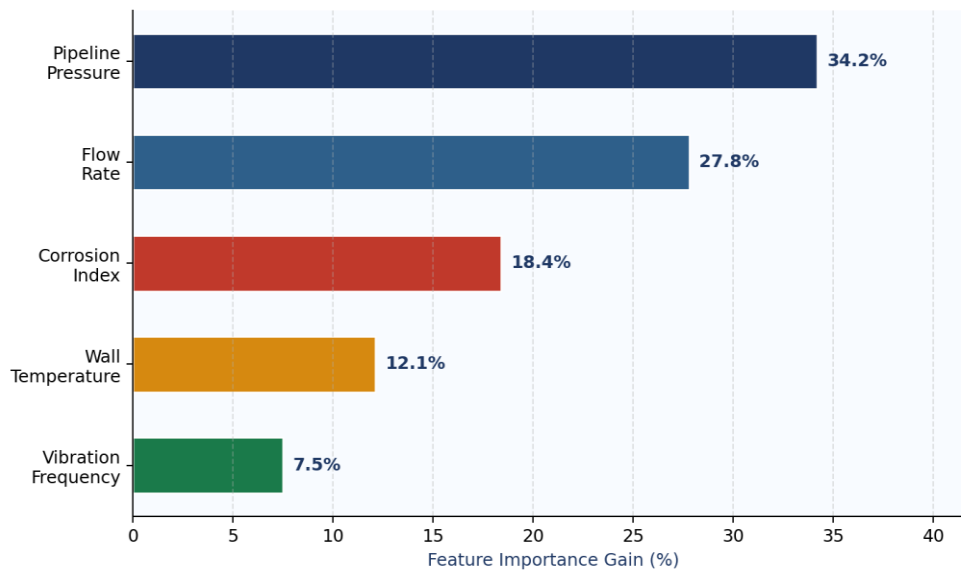


Figure 3: XGBoost Feature Importance Gain Pipeline Leak Detection Model



Pipeline internal pressure accounted for 34.2% of total feature importance gain, followed by flow rate (27.8%), corrosion index (18.4%), wall temperature (12.1%), and vibration frequency (7.5%). This hierarchy aligns with correlation matrix findings and physical pipeline failure mechanics. The corrosion index's 18.4% importance

validates its inclusion, reflecting the Niger Delta's elevated corrosion failure rates (Obike et al., 2020). The XGBoost objective function's regularisation terms (Equations 1–2) effectively suppressed noisy features while preserving the predictive contributions of the five primary sensor variables.

### 3.4 Multi-Class Risk Classification

Table 2: CNN-LSTM Multi-Class Risk Classification Performance

Risk Class	Precision (%)	Recall (%)	F1-Score (%)	Support (n)
Normal	98.6	98.7	98.6	1,020
Anomaly	96.8	96.9	96.8	980
High-Risk	98.4	97.9	98.1	1,000
Macro Average	97.9	97.8	97.6	3,000

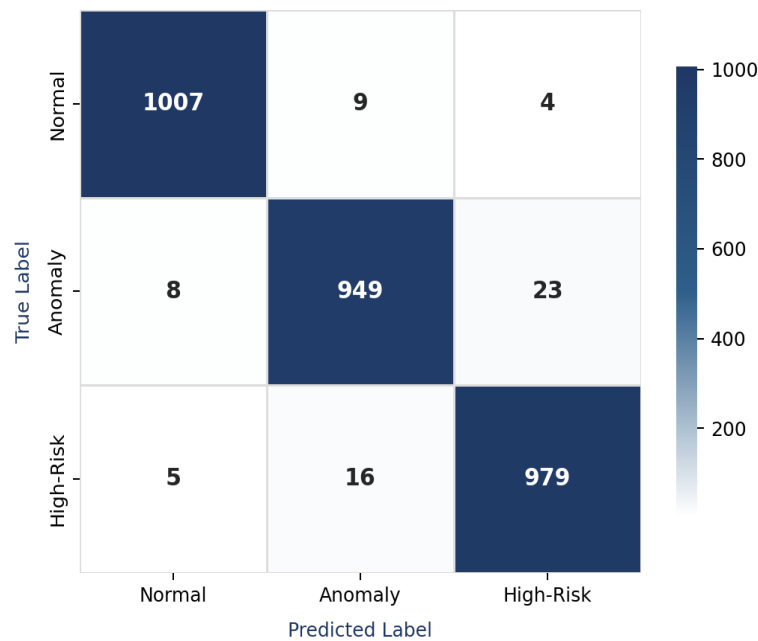


Figure 4: CNN-LSTM Confusion Matrix Multi-Class Pipeline Risk Classification

The CNN-LSTM hybrid achieved 98.1% overall accuracy and 97.6% macro F1, outperforming all unimodal classifiers. The confusion matrix confirms strong performance on the Normal class (1,007 / 1,020 correctly classified) and High-Risk class (979 / 1,000). The Anomaly class exhibited the highest misclassification count (31 observations), predominantly confused with

High-Risk. The 1D convolution operations effectively extracted local sensor anomaly signatures from MIMII-derived features, while the LSTM gates captured temporal evolution across the 20-step sensor windows. The softmax output layer provided well-calibrated class probability estimates that exceeded the  $\geq 95\%$  regulatory compliance threshold for all three risk tiers.



### 3.5 Overall Model Comparison

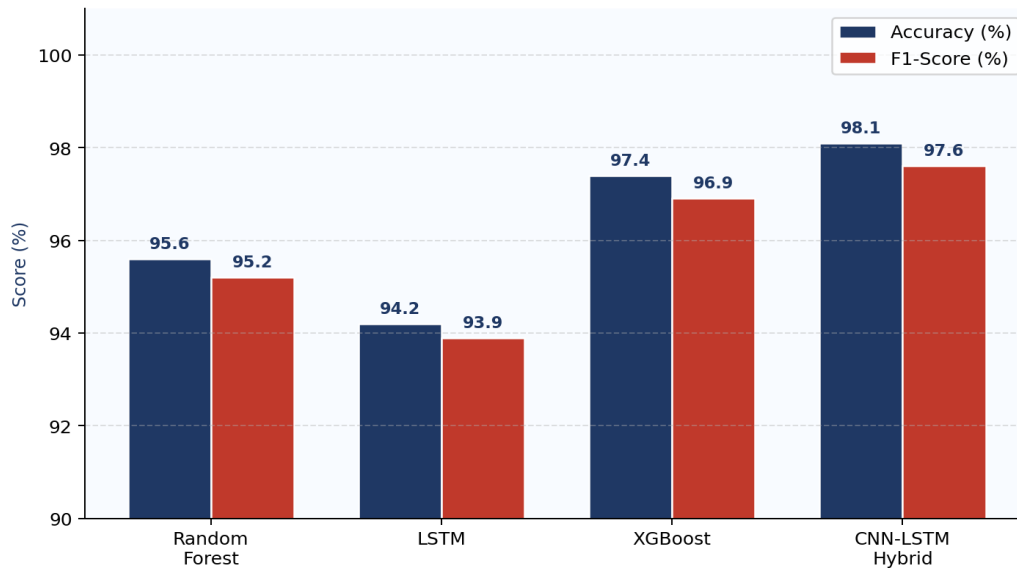


Figure 5: Model Accuracy and F1-Score Comparison Across All Four Architectures

The CNN-LSTM hybrid is the best-performing architecture across both metrics (98.1% accuracy, 97.6% F1), followed by XGBoost (97.4%, 96.9%), Random Forest (95.6%, 95.2%), and LSTM (94.2%, 93.9%). All models except LSTM meet the  $\geq 95\%$  detection accuracy. The performance hierarchy supports Saleem et al. (2024) which reported that CNN-LSTM hybrid model, applied to AE data from industrial fluid pipelines, demonstrated outstanding performance, achieving precision, accuracy, F1 score, and recall values of 99.71%, 99.69%, 99.82%, and 99.75%, respectively.

### V. Conclusion and Recommendations

This study developed and evaluated an integrated ML and DL predictive framework for pipeline leak detection and environmental risk classification in the Niger Delta oil and gas sector, utilising MIMII, the UCI Gas Sensor Array Fault Detection Dataset and Sentinel-1 SAR imagery from the ESA Copernicus Open Access Hub. Through comparative evaluation of XGBoost, Random Forest, LSTM, and a CNN-LSTM hybrid, the study demonstrated that high-accuracy predictive detection is achievable using multi-parameter sensor data augmented with satellite-derived environmental variables. XGBoost achieved 97.4% binary detection accuracy (AUC = 0.979), while the CNN-LSTM hybrid reached 98.1% (macro F1 = 97.6%) in multi-class risk classification. Correlation analysis confirmed pipeline pressure and flow rate as dominant predictors, with the corrosion index

providing a critical supplementary signal for Niger Delta pipeline conditions. The therefore recommend that

- i. NOSDRA and other regulators should include XGBoost-based real-time anomaly detection in all high-risk Niger Delta pipeline corridors
- ii. Oil and gas operators should invest in integrated SCADA-Sentinel satellite data pipelines to enable CNN-LSTM risk classification, prioritising the Forcados-Escravos and Trans-Niger pipeline corridors.
- iii. Future research should extend this framework using proprietary operational incident data from Nigerian pipeline operators, incorporating geological variables and pipeline age to improve field-deployment generalisability.
- iv. A national pipeline AI monitoring consortium should be established bringing together NNPC, academic institutions, and international environmental technology partners to build a shared, anonymised pipeline incident dataset that addresses the data scarcity constraint limiting advanced ML deployment in the sector.

### References

- [1]. Akinmoluwa, O. (2021). Generation of surveillance networked nodes for oil



- pipelines' theft. *International Journal of Recent Engineering Science*.
- [2]. Akinsola, A. O., Olayinka, O. F., & Majekodunmi, T. A. (2025). An Appraisal of the Environmental Protection Provisions in Petroleum Industry Act, 2021. *NIU Journal of Legal Studies*, 11(1), 141-153.
- [3]. Alfarag, S. H. I. (2025). Smart Sensor Networks with CNN, ANN, Deep Learning, and Random Forest for Predictive Risk Management in Iraq's Oil Facilities.
- [4]. Amadi, C. (2024). Autonomous Anomaly Detection in Scada Networks: Leveraging Deep Learning to Predict and Prevent Cyber-Physical Attacks in Legacy Critical Infrastructure Systems.
- [5]. Arinze, C. A., Izionworu, V. O., Isong, D., Daudu, C. D., & Adefemi, A. (2024). Predictive maintenance in oil and gas facilities, leveraging ai for asset integrity management. *International Journal of Frontiers in Engineering and Technology Research*, 6(1), 16-26.
- [6]. Assymkhan, N., & Kartbaev, A. Z. (2024). thermal comfort prediction using SVM and random forest model. *Вестник КазУТБ*, 4, 36-49.
- [7]. Ata, M. M., Osama, S., Ibraheem, M. R., & Abas, A. R. (2026). Early prediction of wind turbine anomalies using 1D-CNN and temporal feature engineering on multi-source SCADA data. *Scientific Reports*.
- [8]. Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- [9]. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
- [10]. Chen, Z., Jin, Y., Wang, X., Chen, H., Zhu, B., & Li, W. (2026). Advancements in prediction models for corrosion in oil and gas pipelines. *Corrosion Reviews*, 44(1), 20240119.
- [11]. Chukwuma, A. L., & City, B. (2025). Data-Driven Prediction And Early Detection Of Flow Assurance Challenges In Oil And Gas Pipelines Using Ensemble Machine Learning Models.
- [12]. Dolire, O. A., Eteng, A. A., & Orakwue, S. I. (2025). Long Short-Term Memory-Autoencoder Model for Intelligent Oil-and-Gas Pipeline Anomaly Detection. *International Journal of Engineering Research in Africa*, 75, 141-158.
- [13]. Dolire, O. A., Eteng, A. A., & Orakwue, S. I. (2025). Long Short-Term Memory-Autoencoder Model for Intelligent Oil-and-Gas Pipeline Anomaly Detection. *International Journal of Engineering Research in Africa*, 75, 141-158.
- [14]. Eli, A. A., Angaye, T. C. N., & Abowei, J. F. N. (2025). Environmental impact, health implications, and socio-economic consequences of artisanal crude oil refining in the Niger Delta, Nigeria: A comprehensive review. *International Journal of Environment and Pollution Research*, 13(1), 19-33.
- [15]. European Space Agency (ESA). (2023). Sentinel-1 ground range detected (GRD) products — Copernicus open access hub. ESA Copernicus. <https://scihub.copernicus.eu>
- [16]. Fachrezi, M. R., Ihsan, A. F., & Astuti, W. (2024, August). Anomaly detection using LSTM-based deep learning on natural gas pipeline operational data. In *2024 12th International Conference on Information and Communication Technology (ICoICT)* (pp. 500-506). IEEE.
- [17]. Fonollosa, J., Sheik, S., Huerta, R., & Marco, S. (2015). Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring. *Sensors and Actuators B: Chemical*, 215, 618-629.
- [18]. Gbadamosi, F., & Aldstadt, J. (2025). The interplay of oil exploitation, environmental degradation and health in the Niger Delta: A scoping review. *Tropical Medicine & International Health*, 30(5), 351-367.
- [19]. Gbenga, O. O. (2021). A Sustainability Oriented Integrity Management Model Based on Agro Waste Derived Green Inhibitors for Carbon Steel Pipelines.
- [20]. Islam, T., Sadik, M. R., Islam, M. F. R., Mona, T. R., Rahman, T., & Foysal, M. M. R. (2023, November). Early-stage diabetes risk prediction using supervised machine learning algorithms. In *2023 2nd International Conference on Futuristic Technologies (INCOFT)* (pp. 1-7). IEEE.
- [21]. Jagadeesh, V., & Sivakumar, P. (2024, December). Enhanced Pipeline Safety: Cloud-Based Leak Prediction Using XGBoost. In *2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 1087-1091). IEEE.
- [22]. James, G., Umoren, I., Ekong, A., Ohaeri, I., & Inyang, S. (2025). Real-Time Monitoring



- of Oil and Gas Pipeline Leakages Identification System Based on Deep Learning Approaches: A Systematic Review.
- [23]. Jonathan, E. L., Imoni, O., Chukwuemeka, P., & Eteh, D. R. (2025). Impact of oil spills on mangrove ecosystem degradation in the Niger Delta using remote sensing and machine learning. *Journal of Geography and Cartography*, 8(2), 11707.
- [24]. Komba, G. M., Mathonsi, T. E., & Owolawi, P. A. (2025, May). Automation and Optimization of Pipeline Leak Detection Using ANN-XGBoost Algorithms and PLC Integration. In *2025 The 16th International Conference on Mechanical and Intelligent Manufacturing Technologies (ICMIMT)* (pp. 1-9). IEEE.
- [25]. Lee, S., Jo, W., Kim, H., Koo, J., & Kim, D. (2023). Deep learning-based cutting force prediction for machining process using monitoring data. *Pattern Analysis and Applications*, 26(3), 1013-1025.
- [26]. Liang, J., Liang, S., Zhang, H., Zuo, Z., Ma, L., & Dai, J. (2023). Leak detection in natural gas pipelines based on unsupervised reconstruction of healthy flow data. *SPE Production & Operations*, 38(03), 513-526.
- [27]. Mandavalli, S. (2025). Gradient Boosting for Equipment Failure Prediction in SAP's IoT-Enabled Oil & Gas Operations. *Authorea Preprints*.
- [28]. Obike, A. I., Uwakwe, K. J., Abraham, E. K., Ikeuba, A. I., & Emori, W. (2020). Review of the losses and devastation caused by corrosion in the Nigeria oil industry for over 30 years. *International Journal of corrosion and scale inhibition*, 9(1), 74-91.
- [29]. Omoruyi, F. O., Audu, H. A., & Ilaboya, I. R. (2025). Multi-Criteria Decision Analysis For Corrosion Risk Assessment Of Buried Water Pipelines: An Analytic Hierarchy Process Approach. *Fudma Journal Of ScienceS*, 9(11), 372-379.
- [30]. Ozigis, M. S. (2020). *Detection and Mapping of Terrestrial Oil Spill Impact Using Remote Sensing Data in Combination with Machine Learning Methods. A Case Site within the Niger Delta Region of Nigeria* (Doctoral dissertation, University of Leicester).
- [31]. Purohit, H., Tanabe, R., Ichige, K., Endo, T., Nikaido, Y., Suefusa, K., & Kawaguchi, Y. (2019). MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. *arXiv preprint arXiv:1909.09347*.
- [32]. Saleem, F., Ahmad, Z., & Kim, J. M. (2024). Real-Time pipeline leak detection: A hybrid deep learning approach using acoustic emission signals. *Applied Sciences*, 15(1), 185.
- [33]. Salem, F. M. (2021). Gated RNN: the long short-term memory (LSTM) RNN. In *Recurrent Neural Networks: From Simple to Gated Architectures* (pp. 71-82). Cham: Springer International Publishing.
- [34]. Sani, A. A., Wahab, M. M. A., Shafiq, N., Danyaro, K. U., Khan, N., Tafida, A., & Yousafzai, A. K. (2025). A multi-level classification model for corrosion defects in oil and gas pipelines using meta-learner ensemble (MLE) techniques. *Journal of Pipeline Science and Engineering*, 5(2), 100244.
- [35]. Temitope Yekeen, S., & Balogun, A. L. (2020). Advances in remote sensing technology, machine learning and deep learning for marine oil spill detection, prediction and vulnerability assessment. *Remote Sensing*, 12(20), 3416.
- [36]. Ullah, S., Ullah, N., Siddique, M. F., Ahmad, Z., & Kim, J. M. (2024). Spatio-temporal feature extraction for pipeline leak detection in smart cities using acoustic emission signals: A one-dimensional hybrid convolutional neural network-long short-term memory approach. *Applied sciences*, 14(22), 10339.
- [37]. Verma, P., Gandhi, K., Cheema, A. M., Ashfaq, M., Shah, D., Ali, S., & Tahir, M. (2026). DeepPipeNet enables deep learning based anomaly detection for oil and gas pipeline monitoring. *Discover Applied Sciences*.
- [38]. Wegner, D. C., Nicholas, A. K., Odoh, O., & Ayansiji, K. (2021). A machine learning-enhanced model for predicting pipeline integrity in offshore oil and gas fields. *J Pipeline Eng.*
- [39]. Yang, L., & Zhao, Q. (2020). A novel PPA method for fluid pipeline leak detection based on OPELM and bidirectional LSTM. *IEEE Access*, 8, 107185-107199.
- [40]. Yang, Z., Zhang, Y., Bai, Y., & Shu, J. (2025). The application of deep learning in pipeline inspection: current status and challenges. *Ships and Offshore Structures*, 20(7), 1016-1027.