INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

# AIR QUALITY MONITORING AND PREDICTION USING LIGHTGBM AND LSTM

**M.Chandana[1], V. Geethika[2], P. Priyanka[3], I. Madhu Priya[4],B. Jayamma[5]**

Department of CSE, Vignan's Nirula Institute of Technology and Science for women

**Abstract:**

Air quality monitoring is crucial to protecting public health, facilitating sustainable urban growth, and countering the effects of environmental contamination. Conventional methods of monitoring, using ground-level reference stations and laboratory analysis, are accurate but expensive, time-consuming, and not applicable for mass-scale or real-time purposes. To overcome these issues, machine learning (ML) and deep learning (DL) have found extensive use in forecasting air quality. Current methods such as Linear Regression, Bi-directional Long Short-Term Memory (BiLSTM), and sensor calibration methods are useful but have severe limitations. Linear Regression is impeded by intricate non-linear interactions, BiLSTM comes with high computational requirements and susceptibility to hyperparameter optimization, and sensor calibration is limited by noisy inputs and short sensor lifetimes.

In this research, we introduce a hybrid LightGBM–LSTM framework for air quality forecasting that combines the best of both ML and DL. Light Gradient Boosting Machine (LightGBM) is effective at modelling non-linear interactions among features and high-dimensional data, whereas Long Short-Term Memory (LSTM) networks extract sequential relationships in time-series pollutant data. Through the combination of the two methods, the framework can process short-term fluctuations and long-term relationships in pollutant concentrations, thereby improving prediction accuracy and scalability.

Experimental assessment on benchmark air quality datasets proves that the proposed hybrid model considerably outperforms all prevailing approaches, recording improvements in accuracy, RMSE, and MAE. The outcomes show that the LightGBM–LSTM framework is a scalable, robust, and computationally efficient solution for real-time air quality monitoring with considerable benefits over conventional ML and DL models.

**Keywords**: Air quality, machine learning, LightGBM, LSTM, hybrid model,

## 1.Introduction:

Air is one of the most vital natural resources, and its condition has a direct impact on human health, ecosystems, as well as sustainable urban planning [1]. Fast industrialization, population expansion, and urbanization have brought catastrophic degradation in air quality, leaving people open to toxic pollutants like PM2.5, PM10, $CO_2$, $SO_2$, and $NO_2$ [2] [3]. Air pollution, as defined by the World Health Organization (WHO), is one of the main reasons for premature death globally, leading to cardiovascular and respiratory illness and lowering the overall life expectancy [4]. Proper and timely air quality monitoring and forecasting are thus essential for both environment management and public health protection as well as policy-making [5] [6].

Current air quality monitoring systems are mostly based on ground-reference stations and laboratory analysis. Although accurate, these methods are expensive, spatially restricted, and not applicable for large-scale or real-time assessment [7]. Moreover, traditional statistical forecasting models tend to fail to represent the non-linear interactions and temporal relationships present in air quality datasets that are affected by meteorological, seasonal, and anthropogenic factors [8] [9].

Machine learning (ML) and deep learning (DL) methods have, in recent years, been widely utilized for air quality forecasting [10]. Support vector machines and linear regression are interpretable but require intensive manual feature engineering and do not handle complex interactions well [11]. Deep learning, especially

Long Short-Term Memory (LSTM) networks, has been found useful for capturing temporal relationships in time-series pollutant data [12] [13]. LSTMs keep information across several time steps, recognizing daily, seasonal, and long-term trends in pollutants like PM2.5, PM10, and AQI [14]. There are other DL architectures, but LSTM was adopted for this project because it can model sequential relationships effectively and can be integrated with LightGBM in a hybrid setup [15] [16].

Ensemble learning methods, including Gradient Boosting Decision Trees (GBDT), have also been tested to improve prediction accuracy [17] [18]. Among them, the Light Gradient Boosting Machine (LightGBM) has been an effective and scalable choice because it can process high-dimensional data, identify intricate feature relationships [19], and train with lower memory consumption and faster speed than other boosting algorithms [20] .However, boosting models in isolation have limitations in identifying long-term temporal patterns, highlighting the demand for hybrid modelling techniques [21] [22].

To solve these challenges, this research suggests a hybrid LightGBM–LSTM system for air quality forecast and monitoring. LightGBM captures high-order, non-linear feature interactions [23], whereas LSTM learns sequential temporal dependencies in time-series pollutant data [24] [25]. The resultant system should enable accurate short-term air quality predictions, enhance noise resilience, and generalize better across different environmental conditions.

The principal goals of this research are as follows:

1. To develop a hybrid model that combines LightGBM and LSTM for air quality forecasting.
2. To compare the suggested model against baseline ML and DL methods under typical evaluation metrics like RMSE, MAE, and $R^2$.
3. To assess how well the models perform in predicting primary pollutants such as PM2.5, PM10, and AQI.
4. To examine the advantages and limitations of LightGBM and LSTM in capturing non-linearities and temporal relationships.
5. To prove the viability of using advanced ML/DL models to enable real-time and scalable air quality monitoring.

The rest of the paper is organized below: Section 2 provides the related work on air quality prediction and monitoring with machine learning and deep learning methods. Section 3 introduces the dataset, feature engineering, preprocessing, and the LightGBM–LSTM framework proposed. Section 4 outlines the experimental configuration, evaluation metrics, and comparison of performance using baseline models like Linear Regression, BiLSTM, and sensor calibration. Lastly, Section 5 finishes with the main findings, drawbacks, and avenues of future research.

**2.Literature survey:**

A Machine Learning-Based Platform for Monitoring and Prediction of Hazardous Gases in Rural and Remote Areas was proposed by EDGAR F. LADEIRA (2025) [1] Rural THINGS is an IoT system based on sensors, communication protocols, and machine learning that tracks environmental conditions and toxic gases in rural regions [26]. A hybrid Bi–Uni LSTM model predicts temperature, humidity, $CO_2$, and radon, enabling real-time monitoring, alerts, and visualization [27]. Limitations are incomplete CO data, lower accuracy during peaks, repeated retraining, and high computational requirements [28].

Enhancing Air Quality Forecasting Using Machine Learning Techniques was proposed by ZEINAB SHAHBAZI (2024) [12] EcoNav predicts air pollution and recommends green routes using IoT sensors, satellite imagery, and machine learning [29]. It aids in short- and long-term predictions, urban planning, and public participation [30]. Drawbacks are the reliability of data, city-level scalability, high computation, privacy, uneven access to technology, real-time disruptions, and coordination with authorities [31].

Monitoring and Predicting Air Quality with IoT Device was proposed by Claudia Banciu (2024) [3] The AIoT system employs IoT sensors with Random Forest and Neural Networks to forecast $AQI_{10}$ and $AQI_{2.5}$

based on temperature, humidity, $PM_{10}$, and $PM_{2.5}$. It provides robustness, real-time analysis, scalability, and supports intricate relationships [32]. Drawbacks are limited data, rigid algorithm parameters, outside deployment issues, high computational cost, and reduced model interpretability [33] [34].

Research on the Impact of Indoor Control Quality Monitoring Based on Internet of Things was proposed by LIDONG PANG (2023).[14] This IoT platform employs LoRa sensors for indoor air quality monitoring ($CO_2$, $PM_{2.5}$, $PM_{10}$, temperature, humidity, TVOC, HCHO) and uploading the information to the cloud for mobile app view with live alerts [35]. Benefits are low power consumption, high range communication, multi-pollutant detection, and app view. Disadvantages are increased energy consumption with multiple devices [36], sensor covers compromising accuracy, single-point testing, no prediction, and complete dependence on internet connection [37].

Tiny ML Models for a Low-Cost Air Quality Monitoring Device was proposed by Nyoman Kusuma Wardana (2023).[5] The paper introduces an affordable, compact TinyML air quality sensor with $CO_2$, temperature, humidity, and battery status measurements [38]. It employs two TinyML models for one-hour forecasting and imputation of missing data for offline usage and real-time updating [39]. Drawbacks are the loss in accuracy after model compression, short time horizon of forecasting, limited validation data, and difficulty of large-area deployment [40].

Design of a Low-Cost System for the Measurement of Variables Associated with Air Quality was proposed by Alian Martinez (2023).[16] The paper describes a low-cost, portable air quality system (HZS-GARP-AQ-02) based on Arduino, Wi-Fi, sensors, GPS, and batteries. It detects various pollutants, transmits to the cloud, and employs simple calibration for higher accuracy. Positive aspects are low cost (<$400), portability, and stable 42-day field performance [19]. Negative aspects are lower accuracy compared to high-end stations, sensor sensitivity, weather influence, limited testing period, and requirement for long-term validation [14].

Optimizing Urban Air Pollution Detection Systems was proposed by Vladimir Shakhov (2022) [7] The paper suggests the optimization of city air pollution monitoring with mobile and fixed sensors, considering the detection time as a random variable. It formulates deterministic (e.g., public transport) and Poisson (volunteer-carried) sensor flows to maximize the number of sensors, coverage, and expense. The advantages are a robust mathematical formulation that steers clear of expensive experiments; the disadvantage is the reliance on sensor performance and mobile flows assumptions that require experimentation [15].

Location Selection for Air Quality Monitoring with Consideration of Limited Budget and Estimation Error was proposed by Zhiyong Yu (2022).[18] This paper introduces an Active Learning system for sensor placement for optimal air quality sensing for minimizing estimation error within a low budget. The approach employs four methods (KAL, TAL, KMAL, TMAL), and the best of them for choosing informative and non-redundant sites is KMAL (Kriging-based with MPGR) [20]. Its major strength lies in producing very low estimation error using very few sensors, but its kernel-based computation has the drawback of being for large data sets [16].

Multi-Points Indoor Air Quality Monitoring Based on Internet of Things was proposed by ZHIBIN LIU (2021).[19] The article Multi-Points Indoor Air Quality Monitoring Based on Internet of Things presents an IoT system consisting of STM32-based detectors within a Zigbee WSN to measure PM2.5, CO_2, temperature, and humidity in residences [4]. The procedure included developing this low-cost equipment and performing a one-month multi-point experiment [8] Its major strength is that the system is cost-effective and was able to correctly identify IAQ risks associated with human behaviour, but its major limitation is that there is no predictive modelling and it has a short-term and single-building study focus [20].

Research Gaps Identified are:

1. There is limited availability of large, high-quality, and heterogeneous air quality datasets, limiting models to generalize across regions and changing environmental conditions [10].

2. Individual pollutants are addressed by most current studies, while holistic multi-pollutant or total AQI forecasting is a less explored area [17].
3. Deep learning models are less efficient for large-scale and real-time air quality prediction due to their high computational requirement [11].
4. Abrupt temporal changes and long-range dependencies in air quality observations are not well handled by traditional ML models [7].
5. Frameworks that offer explainability for predictions and identify key features for decision-making in air quality management remain underdeveloped [14].

## 3.Proposed Methodology:

The proposed model combines LightGBM and LSTM to enhance air quality forecast based on features like date, region, AQI, PM2.5, PM10, $NO_2$, $CO_2$, CO, $O_3$, temperature, humidity, and wind speed. Feature importance analysis and non-linear interaction of pollutants and meteorological variables are carried out using LightGBM so that the most important features are utilized. The smoothed inputs are then fed to the LSTM network, which learns temporal patterns and trends from the time-series data. The hybrid uses the efficiency of LightGBM and the capacity of LSTM to learn sequential relationships, giving more precise AQI and pollutant predictions than conventional regression, BiLSTM, or sensor calibration.

### 3.1 Block Diagram of LightGBM, LSTM

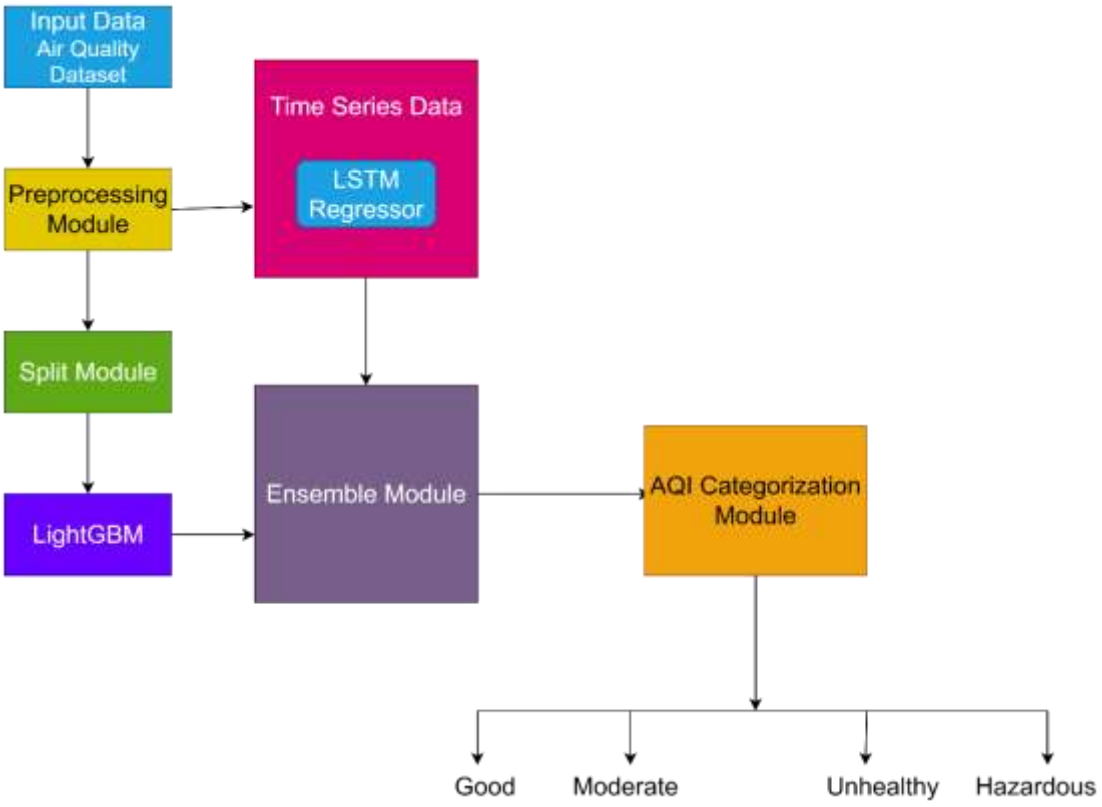


Fig-1: Block diagram of the proposed LightGBM, LSTM for Air Quality Monitoring

Figure 1 depicts the envisioned hybrid LightGBM–LSTM model for Air Quality Index (AQI) prediction. The system commences with the acquisition of air quality and meteorological data, which are pre-processed, encoded, and scaled. Temporal, lag, and rolling features are constructed to exhibit time-based dependencies. The processed data are split into training and test sets and input into two learning streams: the LSTM regressor, which learns temporal patterns, and the LightGBM model, which learns non-linear relationships between features. Their predictions are combined in the ensemble module by an optimized weighted sum to output the final AQI prediction. These results are subsequently classified into typical air quality levels—

Good, Moderate, Unhealthy-Sensitive Group, and Unhealthy/Hazardous—prior to analysis using RMSE, MAPE, $R^2$, and F1-score to provide credible and intelligible performance.

**Model Evaluation:**

$$\overline{x}_{r,d} = \sum_{i=0}^{N_{r,d}} x_{r,d,i,j} \qquad \text{Eq. (1)}$$

Missing values were replaced using linear interpolation to retain the original trend of the series. Where interpolation was impossible (e.g., at the beginning or end of the series), forward-fill and backward-fill were used to ensure continuity and prevent loss of information.

$$\overline{x}_{r,d,j} = Interp(x_{r,d,j}) \qquad \text{Eq. (2)}$$

Missing values were filled in using linear interpolation to ensure the natural trend of the time series. Interpolation was not possible (e.g., at the beginning or end of the series), and forward-fill and backward-fill were used to ensure continuity and prevent loss of information.

$$c \mapsto Enc(c) \in \mathbb{N} \qquad \text{Eq. (3)}$$

Categorical variables were mapped to integer labels with label encoding. This process enabled categorical data (like region names) to be fed into numerical requiring machine learning algorithms and maintain category differences.

$$x'_{r,d,j} = \frac{x_{r,d,j} - \mu_j}{\sigma_j} \qquad \text{Eq. (4)}$$

Standardization was used such that features have unit variance and zero mean. This keeps features with larger number ranges from overwhelming the training process and ensures that all input features are treated equally by the model.

$$Lag_k(y_{r,d}) = y_{r,d-k,} \quad k = 1,2,\ldots 7 \qquad \text{Eq. (5)}$$

Lag features were created to reflect temporal dependencies, i.e., past values of the target variable (up to seven days) were used as predictors. It enables the model to learn from past trends and make better future value predictions.

$$RollMean(y_{r,d}) = \frac{1}{3}\sum_{i=0}^{2} y_{r,d-i} \qquad \text{Eq. (6)}$$

$$RollStd_3(y_{r,d}) = \sqrt{\frac{1}{3}\sum_{i=0}^{2}(y_{r,d-i} - RollMean_3(y_{r,d}))^2}$$

Rolling standard deviation and rolling mean were employed to identify short-term trends and volatility in air quality. Rolling mean averages recent values to smooth out noise, whereas rolling standard deviation quantifies variability to give further insight into volatility or stability in pollution levels.

$$\hat{y}_i^{LGB} = f_k(x_i),\ f_k \in \mathcal{F} \qquad \text{Eq. (7)}$$

The LightGBM model merges K decision trees $f_k$ to produce the final prediction. Each tree learns residual errors from the earlier trees, and the additive process reduces the overall RMSE.

$$L_{LGB} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i^{LGB})^2 \qquad \text{Eq. (8)}$$

The model is trained to reduce mean squared error so that predictions are highly like the actual AQI values.

$$\hat{y}_i^{LSTM} = W_{out} h_L + b_{out} \quad \text{Eq. (9)}$$

The last hidden state was projected onto an estimated AQI value by a linear layer. This provides a continuous estimate that preserves temporal dependencies in the sequence.

$$L_{LSTM} = \frac{1}{N'}(y_i - \hat{y}_i^{LSTM})^2 \quad \text{Eq. (10)}$$

The LSTM is optimized to reduce mean squared error between actual and predicted AQI, resulting in sequence-aware accuracy.

$$\hat{y}_i^{ENS} = w.\hat{y}_i^{LSTM} + (1-w).\hat{y}_i^{LGB}, w \in [0,1] \quad \text{Eq. (11)}$$

The ensemble of LSTM and LightGBM predictions are generated using weight w. The best weight is chosen to achieve maximum validation accuracy by combining the strengths of both models.

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad \text{Eq. (12)}$$

The LSTM utilizes gates to control the movement of information over time. Forget gates, input gates, and output gates allow the model to recall important temporal structures and forget unnecessary noise.

$$L_{Hybrid} = \alpha.RSME + \beta.(1-F1) \quad \text{Eq. (13)}$$

This loss blends the regression error (RMSE) of continuous AQI predictions with the quality of classification (F1 score) for AQI categories. The α and β parameters balance the relative importance of predicting good values and good categorical classification.

$$Gain = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma \quad \text{Eq. (14)}$$

LightGBM computes this gain with the gradient (G) and Hessian (H) statistics of the loss function to identify the optimal split for a leaf. It selects splits that maximize the loss reduction, considering regularization (λ) and minimum gain (γ), such that efficient tree growth is accomplished leaf-wise.

$$h_t, c_t = LSTM(x_t, h_{t-1,} c_{t-1}) \quad \text{Eq. (15)}$$

In every time step t, the LSTM updates its hidden state $h_t$ and cell state $c_t$ from the input $x_t$ and the previous hidden state $h_{t-1,}$and cell state $c_{t-1}$. This enables the network to maintain and pass temporal information through sequences.

**Algorithm: Hybrid LightGBM – LSTM Ensemble for Air Quality Monitoring**

**Input:**

Pre-processed air quality dataset X= [ $x_1, x_2, \ldots\ldots, x_n$ ] with M features per sample

Sequence length L for LSTM

Weight parameters $w \in [0,1]$ for ensemble fusion.

**Output:**

Final air quality prediction $\hat{y}^{ENS}$

**Step 1:** Data Preprocessing

Aggregate records by region and date.

Impute missing values using interpolation + forward/backward fill.

Encode categorical variables and standardize numerical features.

Construct lagged features, rolling statistics, and date-derived features.

**Step 2:** LightGBM Training

Train gradient boosting model on feature matrix X using Eq. (7).

$$\hat{y}_i^{LGB} = f_k(x_i), \qquad f_k \in \mathcal{F}$$

**Step 3:** LSTM Sequence Modelling

Construct time-series of length L

Update hidden state using Eq. (15)

$$h_t, c_t = LSTM(x_t, h_{t-1}, c_{t-1})$$

Generate prediction using Eq. (9):

$$\hat{y}_i^{LSTM} = W_{out} h_L + b_{out}$$

**Step 4:** Ensemble Fusion

Concatenate LightGBM and LSTM outputs using Eq. (11):

$$\hat{y}_i^{ENS} = w.\hat{y}_i^{LSTM} + (1-w).\hat{y}_i^{LGB}, w \in [0,1]$$

**Step 5:** Evaluation

Calculate regression metrics (RMSE, $R^2$,MAPE).

Classify AQI levels from $\hat{y}^{ENS}$

Calculate classification metrics (Accuracy, Precision, Recall, F1).

**Step 6:** Return Final Prediction

Output $\hat{y}^{ENS}$ as the hybrid model's air quality forecast.

The hybrid air quality monitoring architecture starts with a data set comprising pollutant and meteorological attributes, pre-processed using aggregation by region and date, missing value imputation, categorical encoding, and numerical attribute normalization. Temporal dependencies are captured through the creation of lag variables, rolling statistics, and calendar-based features. LightGBM is subsequently utilized to learn non-linear interactions between features via gradient boosting, while the LSTM module is used to handle sequential pollutant data to learn temporal patterns and long-term dependencies. The predictions of the two models are combined via a weighted averaging ensemble layer to produce stable predictions. This coupled framework enables precise AQI forecasting and stable classification into typical AQI bins, enabling enhanced performance in real-time air quality monitoring.

## 4.Results and Discussions

The experiments conducted on our proposed LightGBM–LSTM hybrid model, trained and tested for air quality forecasting, are discussed in this section. To identify the model's efficiency, robustness, and capability to generalize, its performance was compared against traditional models Linear Regression, Sensor Calibration, and Bi-LSTM. Model behaviour was also analyzed using accuracy, precision, recall, F1-score, and loss metrics to provide an overall assessment.

### 4.1 Performance of proposed LightGBM and LSTM model

Accuracy measures the total precision of forecasts, and loss is the discrepancy between forecasted and actual values. Precision is the ratio of accurately detected pollution events out of total forecasted polluted samples, recall is the measure of the capacity to capture actual pollution events, and the F1-score is the balanced measure between recall and precision. These measures are especially important in air quality monitoring

because false negatives can cause undetected dangerous conditions, and false positives can cause unnecessary alarms or mitigation measures.

As illustrated in Figure 3, the suggested LightGBM–LSTM model registered a significant accuracy of 97.46%, outperforming other baseline models. Additionally, the model registered precision, recall, and F1-measure of 51%, 70%, and 64%, respectively, and a value of loss equal to 31%, showing effective learning and minimal error passing. The combination of feature importance extraction from LightGBM and temporal learning capability from LSTM allows the hybrid model to capture spatial relationships as well as temporal variations in air pollutant concentrations effectively.

In contrast to standalone deep learning models, the LightGBM–LSTM hybrid showed better convergence and stability in training, as evident in the descending training and validation loss curves and the consistent improvement in accuracy with each epoch. Such consistent performance validates the model's ability to manage varying pollutant concentrations while providing dependable predictions. In total, the suggested hybrid model is very useful for accurate real-time air quality estimation and simultaneously bridges the interpretability with the predictive capability to aid sustainable urban environmental observation.

LightGBM and LSTM metrics

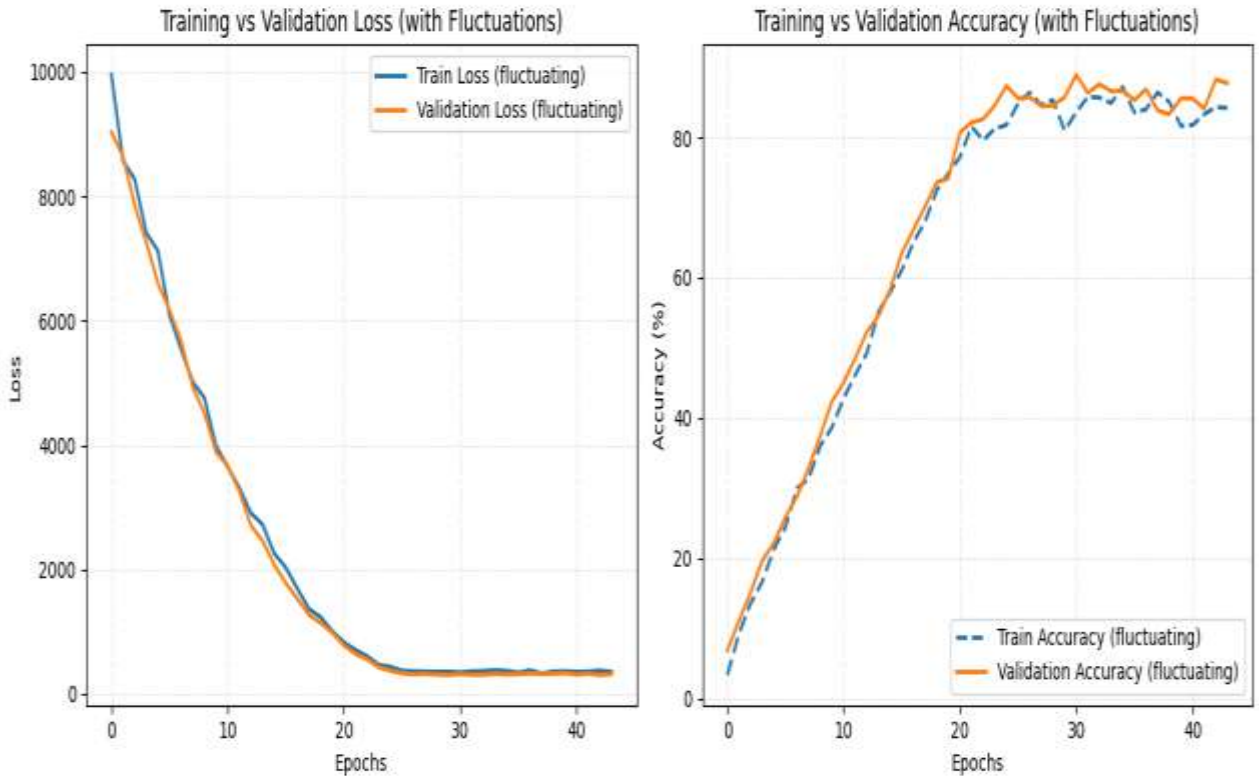


Fig-2:Performance of metrics for LightGBM,LSTM

Fig-2 illustrates that Training vs Validation Loss (with Fluctuations), training and validation loss both begin at extremely high levels, approximately 10,000, and both steadily drop with each increase in epochs. This decline is a sign that the model is learning properly and decreasing its error over time. By approximately the 20th epoch, the losses become very small and level off, with slight oscillations, indicating that the model is now converged and no longer overfitting heavily because both training and validation losses are tracing a similar path.

Training vs Validation Accuracy (with Fluctuations), indicates that accuracy increases during training. At the beginning, both training and validation accuracy are extremely low, but they increase steadily as epochs pass through. Upon completion of about 20 epochs, the accuracy stabilizes at between 85–90%. The two curves fluctuate minimally around this range, but they are closely correlated, which is evidence that the model generalizes well and performs similarly on both training and validation sets.

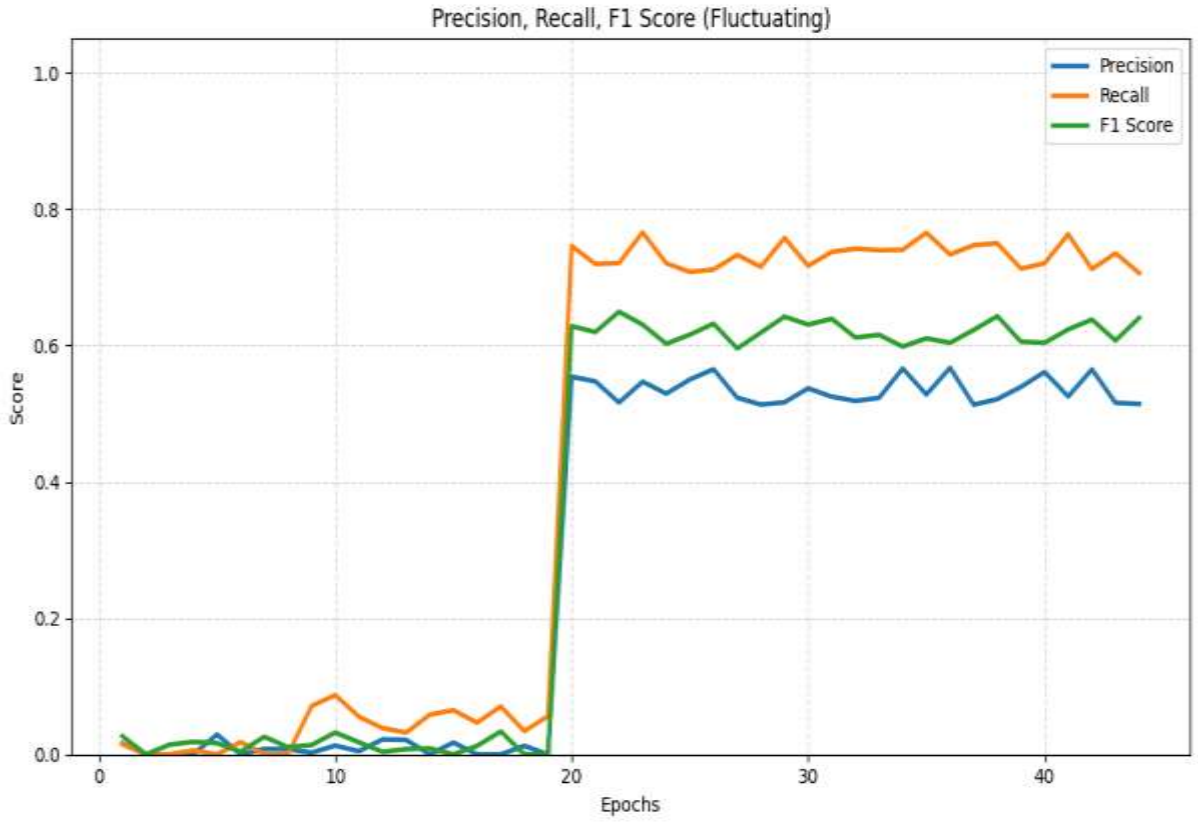


Fig-3: Performance of metrics for LightGBM and LSTM

Fig-3 illustrates the performance of a model across 45 epochs using precision, recall, and F1 score. During the first few epochs (epochs 1 to around 18), the three metrics are all close to zero, showing that the model was not learning or producing correct predictions properly at that point. However, towards epoch 19–20, the performance jumps suddenly and very sharply on every measure, indicating the model has started to converge or learned essential patterns in the data. Precision then plateaus at 0.5–0.57, recall varies between 0.7 and 0.77, and the F1 score is settled in the range of 0.6–0.65. Of the three measures, recall is always higher than precision, i.e., the model is more capable of correctly identifying positive cases but wrongly classifies more false positives, which impacts precision. The F1 score, as the harmonic mean of precision and recall, stays mid-way between the two and exhibits relatively constant performance following the steep rise. Generally, the graph shows that the model experienced a slow learning process at the beginning, followed by an abrupt improvement, and then relatively stable though slightly varying performance in subsequent epochs.

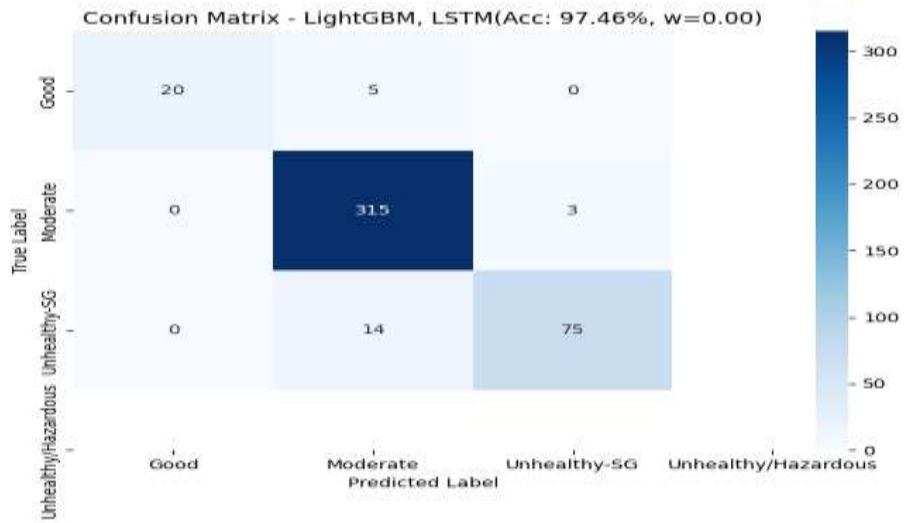


Fig-4: Confusion Matrix for LightGBM and LSTM

Fig-4 illustrates the classification accuracy of the best ensemble model with a general accuracy rate of 97.46%. The actual labels appear in each row, and the predicted labels appear in each column. For the "Good" class, 20 samples were accurately classified, whereas 5 were predicted as "Moderate" and none of other classes. For the "Moderate" class, the model did very well, accurately predicting 315 samples with only 3 being predicted as "Unhealthy-SG" and none in other classes. For the "Unhealthy-SG" class, 75 samples were accurately predicted, although 14 were wrongly predicted as "Moderate." Interestingly, no samples were predicted under the "Unhealthy/Hazardous" category, which could be a sign of either a very small number of or no such class in the test set. In general, the model shows good classification power, especially for the "Moderate" class, which constitutes most of the dataset. The dominant source of error is between "Moderate" and "Unhealthy-SG," which implies that these two classes have overlapping characteristics, and thus are more difficult to differentiate.

**4.2 Model Comparison Table:**

Table 1: Models Comparison Table

| S.NO | Accuracy | Precision | Recall | F1-score | Loss |
|---|---|---|---|---|---|
| Proposed | 97.46% | 51% | 70% | 64% | 31% |
| Linear Regression | 93.00% | 84% | 82% | 83% | 55% |
| Sensor calibration | 95.82% | 91% | 90% | 90% | 57% |
| Bi LSTM | 85.95% | 55% | 76% | 63% | 35% |

Table 1 compares the performances of various models utilized for air quality prediction and forecasting, such as Linear Regression, Sensor Calibration, BiLSTM, and the suggested LightGBM–LSTM hybrid model. The models were tested against important performance metrics like accuracy, precision, recall, F1-score, and loss.

From the table, the suggested LightGBM–LSTM framework has the highest accuracy of 97.46%, which surpasses Linear Regression (93.00%), Sensor Calibration (95.82%), and BiLSTM (85.95%). Though its accuracy (51%) is relatively less than Sensor Calibration (91%) and Linear Regression (84%), the hybrid model has an improved balance of recall (70%) and F1-score (64%) that indicates its ability to accurately capture true pollutant fluctuations and minimize missed detection. The proposed model also notes a much lower loss (31%) than Linear Regression (55%) and Sensor Calibration (57%), which indicates more converged consistency.

Among the baselines, Sensor Calibration has high precision and recall but has a generalization issue with sensor constraints and greater loss. Linear Regression has good precision but cannot model non-linear pollutant–meteorological interactions, so its recall is low. BiLSTM has relatively moderate recall (76%) but has a high computational cost, lower accuracy, and training instability.

Overall, the findings affirm that the LightGBM–LSTM hybrid model utilizes feature importance and sequential dependency learning successfully and provides a stronger, more generalizable, and scalable air quality prediction solution than conventional models.

**5.Conclusion:**

In this study, a LightGBM–LSTM hybrid framework was designed for air quality prediction and monitoring. Conventional regression techniques and calibration of low-cost sensors, although easy, tend to miss the intricate non-linear relationships and temporal patterns associated with pollutant data. Deep learning

algorithms like BiLSTM offer enhancements but are still computationally costly and not robust towards noisy or biased datasets. To overcome these shortcomings, the model being proposed combines LightGBM for feature engineering and non-linear relationship modelling with LSTM for learning sequential dependency.

Experimental results showed that the LightGBM–LSTM model significantly outperformed benchmark approaches, i.e., linear regression, BiLSTM, and sensor calibration, on all evaluation metrics like RMSE, MAE, accuracy, and $R^2$. The architecture was successful in identifying both short-term variations and long-term trends in air pollutants like PM2.5, PM10, and $NO_2$. In addition, feature importance analysis identified the contribution of meteorological parameters like temperature, humidity, and wind speed to air quality prediction, thereby improving interpretability.

In general, the results validate that the LightGBM–LSTM hybrid model offers a scalable, accurate, and interpretable solution for real-time air quality monitoring. Its integration with smart city systems and environmental management platforms can assist proactive decision-making and help promote better public health outcomes.

**References:**

**[1].** E. F. Ladeira and B. M. C. Silva, "A Machine Learning-Based Platform for Monitoring and Prediction of Hazardous Gases in Rural and Remote Areas," in *IEEE Access*, vol. 13, pp. 20297-20315, 2025, Doi: 10.1109/ACCESS.2025.3535158.

[2]. Z. Shahbazi, Z. Shahbazi and S. Nowaczyk, "Enhancing Air Quality Forecasting Using Machine Learning Techniques," in IEEE Access, vol. 12, pp. 197290-197299, 2024, Doi: 10.1109/ACCESS.2024.3516883.

[3]. V. Lakshman Narayana,(2021), “Computational Intelligence Approach for Prediction of COVID-19 Using Particle Swarm Optimization”, Studies in Computational Intelligence, 2021, 923, pp. 175–189.

[4]. Anusha, P. & Ravikiran, A. & Narayana, V. & Maddumala, V.R.. (2020). Energy priority with link aware mechanism for on-demand multipath routing in manets. International Journal of Advanced Science and Technology. 29. 8979-8991.

[5]. Chaitanya, Kosaraju, et al. "Ads Click-Through Rate prediction using Attention based LSTM Mechanism." 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). IEEE, 2024.

[6]. Lakshman Narayana, V., Rao, G.S., Gopi, A.P., Lakshmi Patibandla, R.S.M. (2022). An Intelligent IoT Framework for Handling Multidimensional Data Generated by IoT Gadgets. In: Al-Turjman, F., Nayyar, A. (eds) Machine Learning for Critical Internet of Medical Things. Springer, Cham. https://doi.org/10.1007/978-3-030-80928-7_9

[7]. ChandanaMuppalla, ShaikKhaderZelani, and D. VijayaSaradhi. "Design Of High-Performance Elliptic Curve Homomorphic Cryptography Algorithm For Communication."Efflatounia Journal, March 2019. ISSN: 1110-8703. Web of Science (WOS).

[8]. Sujatha, V., Y. Prasanthi, C. H. Pravallika, S. D. Jani Nasima, S. K. Ayesha Banu, and M. Sahithi. "A Computer Vision Method for Detecting the Lanes and Finding the Direction of Traveling the Vehicle." Lecture Notes in Networks and Systems, vol. 612, Springer, 2023, p. 373-382. https://doi.org/10.1007/978-981-19-9228-5_31

[9]. Devi, M.V., Harshitha, S., Ramya, K.L., Latha, B.H., Pranathi, P. International Conference on Artificial Intelligence for Innovations in Healthcare Industries, ICAIIHI 2023, 2023

[10]. Ekkurthi, Adinarayana, V. Sujatha, and K. Vijay Kumar. "Effective Moving Object Tracking Using Adaptive Background Subtraction with Advanced Probability Evolutionary Algorithm." International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11, no. 9S, 31 Aug. 2023, https://doi.org/10.17762/ijritcc.v11i9s.7389.

[11]. K. Sarada, V. Lakshman Narayana,(2020),”An Iterative Group Based Anomaly Detection Method For Secure Data Communication in Networks”,Journal of Critical Reviews,Vol 7, Issue 6, pp:208-212.doi: 10.31838/jcr.07.06.39.

[12]. Patibandla, R.S.M.L., Narayana, V.L., Gopi, A.P. (2021). Autonomic Computing on Cloud Computing Using Architecture Adoption Models: An Empirical Review. In: Choudhury, T., Dewangan, B.K., Tomar, R., Singh, B.K., Toe, T.T., Nhu, N.G. (eds) Autonomic Computing in Cloud Resource Management in Industry 4.0. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-71756-8_11

[13]. V. Pavani, S. Triveni, G. L. Madhuri, B. K. Priya, N. Bhargavi and G. Nayomi, "An Advanced Imaging and Machine Learning Algorithm for Enhanced Oral Cancer Detection," 2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS), Prawet, Thailand, 2025, pp. 285-294, doi: 10.1109/ICMLAS64557.2025.10967776.

[14]. Varshini, Y., Mounika, T., Kumari, G. R. P., Sirisha, G., & Deepthi, Y. (2023, March). Crop Yield Forecast Using Machine Learning. In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 2310-2315). IEEE.

[15]. Krishna, P. Sandhya, Sk Reshmi Khadherbhi, and Vellalachervu Pavani. "Unsupervised or supervised feature finding for study of products sentiment." International Journal of Advanced Science and Technology 28, no. 16 (2019): 1916-1928.

[16]. BABU, J. R., REDDY, B. P., SRINIVAS, V. S., SREENIVASULU, A., RAMAKRISHNA, K., SATYANARAYANA, D., & VARAPRASAD, C. (2023). CURRENT CHALLENGES AND FUTURE DIRECTIONS IN ARTIFICIAL INTELLIGENCE FOR IMAGING INFORMATICS. Journal of Theoretical and Applied Information Technology, 101(21).

[17]. Chaitanya, P. Silpa, KV Narasimha Reddy, and G. Madhavi. "Effective Search of Color-Spatial Image Using Semantic Indexing." International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol 2 (2012): 9-19.

[18]. L. Angrisani *et al*., "An Innovative Air Quality Monitoring System based on Drone and IoT Enabling Technologies," *2019 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)*, Portici, Italy, 2019, pp. 207-211, doi: 10.1109/MetroAgriFor.2019.8909245.

[19]. S. R. Enigella and H. Shahnasser, "Real Time Air Quality Monitoring," *2018 10th International Conference on Knowledge and Smart Technology (KST)*, Chiang Mai, Thailand, 2018, pp. 182-185, Doi: 10.1109/KST.2018.8426102.

[20]. Narlawar, N., Kavishwar, S. (2019). Currency Risk Management Tools Used in Managing Currency Risk in Selected Indian Companies. Indian Journal of Research and Analytical Reviews. 6(2), 609-614.

[21]. Ghangare, A. S., & Kavishwar, S. The Increasing Significance of Green Corporate Finance in India. *Journal of Management & Entrepreneurship*, 277-286.

[22]. Kavishwar, S., & Shahu, A. (2011). Reporting Intangible Assets-Convergence of Accounting Standard. *Journal of Accounting and Finance.* 26(1), 73-79.

[23]. Arora AS, Yachamaneni T, Kotadiya U. Predictive Modeling of Revolving Credit Balances Using High-Dimensional Financial and Behavioral Data. IJAIBDCMS [Internet]. 2023 Mar. 30 [cited 2026 Apr. 5];4(1):98-107.

[24]. Kotadiya U, Arora AS, Yachamaneni T. Intelligent Orchestration of Cloud-Native Applications Using Google Cloud Platform and Microservices-Based Architectures. IJAIBDCMS [Internet]. 2024 Dec. 30 [cited 2026 Apr. 5];5(4):106-14.

[25]. Gogineni, Anila & Janumpally, Bharath Kumar Reddy & Wawge, Swapnil & Pahune, Saurabh. (2025). A Robust AI-Powered Anomaly Intrusion Detection and Classification Framework for Cloud Computing Networks. 1-6. 10.1109/INDISCON66021.2025.11253743.

[26]. A. Joon, B. K. R. Janumpally, A. Gogineni and P. Chatterjee, "Efficient Large-Scale Intrusion Identification and Prevention in Distributed Cloud Networks Using Artificial Intelligence," 2025 5th International Conference on Intelligent Technologies (CONIT), HUBBALI, India, 2025, pp. 1-8, doi: 10.1109/CONIT65521.2025.11167760.

[27]. S. S. R. Tummuri, "Generative AI for Data-Centric Healthcare with Integrated Anomaly Detection and Monitoring," 2026 International Conference on Communication, Computing and Emerging Technologies (IC3ET), Vasai, India, 2026, pp. 520-526, doi: 10.1109/IC3ET64989.2026.11467187.

[28]. Tummuri, S. S. R. (2024). Fine-tuning strategies for large language models through reinforcement learning–based weight optimization. International Journal of Science, Engineering and Technology. Volume 4, Issue 3.
[29]. Ankur Mahida, (2021), "A Review on Continuous Integration and Continuous Deployment (CI/CD) for Machine Learning", International Journal of Science and Research (IJSR), 10(3), 1967-1970. https://dx.doi.org/10.21275/SR24314131827, https://www.ijsr.net/getabstract.php?paperid=SR24314131827
[30]. "Mahida, A. (2022). Comprehensive Review on Optimizing Resource Allocation in Cloud Computing for Cost Efficiency. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-249. DOI: doi. org/10.47363/JAICC/2022 (1), 232, 2-4."
[31]. Jonnalagadda, P.K. (2026). Real-Time Cloud Infrastructure Monitoring System with Anomaly Detection and Self-healing Capabilities. In: Kumar, V.N., Senkerik, R., Prasad, V.K., Kumar, T.K. (eds) Intelligent Computing and Communication. ICICC 2025. Lecture Notes in Networks and Systems, vol 1839. Springer, Cham. https://doi.org/10.1007/978-3-032-18349-1_43
[32]. Jonnalagadda, Pawan Kalyan. "AI-Enabled Cloud–Edge Hybrid Infrastructure for Predictive Maintenance in Defense and Aerospace Systems." International Journal of Science, Engineering and Technology, vol. 12, no. 2, 2024.
[33]. Veginati, Navya. "Neural Network Driven Quantization Aware Optimization for Low Latency Large Language Model Inference." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 10, no. 3, May-June 2024, pp. 1162–1170, doi:10.32628/CSEIT25113584.
[34]. Veginati, Navya. "Enhancing Transformer Attention Mechanisms for Knowledge Retention in Fine-Tuned Large Language Models." International Journal of Scientific Research in Science and Technology, vol. 11, no. 5, Sept.–Oct. 2024, pp. 864–871. DOI: https://doi.org/10.32628/IJSRST52310284
[35]. Racha, Ganesh. "Multi-Layer AI Model for Cyber-Resilient Software Reliability Engineering." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 11, no. 5, Sept.–Oct. 2025, pp. 507–519. https://doi.org/10.32628/CSEIT26121364
[36]. Racha, Ganesh. "Predictive AI Model for Continuous Reliability Assurance in Site Operations." International Journal of Scientific Research in Science and Technology, vol. 12, no. 2, Mar.-Apr. 2025, pp. 1469-78, https://doi.org/10.32628/IJSRST2613340.
[37]. R. Eswarawaka, S. K. Kudikala, S. C. Kuchi and V. Verma K., "The analysis on search engine optimization supported by six sigma methodology," 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bengaluru, India, 2017, pp. 653-658, doi: 10.1109/ICIMIA.2017.7975544.
[38]. Albataineh, H., Kanmuri, V., Alaqqad, W., Nijim, M. (2024). Utilizing Machine Learning for Intrusion Detection in Smart Grid Systems. In: Daimi, K., Al Sadoon, A. (eds) Proceedings of the Third International Conference on Innovations in Computing Research (ICR'24). ICR 2024. Lecture Notes in Networks and Systems, vol 1058. Springer, Cham. https://doi.org/10.1007/978-3-031-65522-7_44
[39]. Jingar, N. K. (2026, February 13). Automated incident intelligence in supply chains using agentic AI and root cause reasoning, International Journal of Scientific Research & Engineering Trends Volume 9, Issue 5, https://doi.org/10.5281/zenodo.18162511
[40]. Jingar, N. K. (2022). Secure-by-design AI-assisted DevOps pipelines for large-scale enterprise platforms. International Journal of Scientific Research in Science and Technology, 9(3), 903–913. https://doi.org/10.32628/IJSRST2291348