International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(11s) 2026 pp.577-591
https://doi.org/10.70917/ijcisim-2026-3773
Article

Deep learning assisted data visualisation framework for real time data interpretation and predictive analytics

Sanjay Kumar Gupta[1], Dharmendra Kumar Tripathi[2], Hakim Singh Jat[3], Vinay Joshi[4], Pradnya Vikhar[5]

[1] Rustamji Institute of Technology, BSF Academy, Gwalior, Madhya Pradesh, India.
Email: drsanjaygupta@rjit.ac.in
ORCID: 0009-0002-7862-9493
[2]Rustamji Institute of Technology, BSF Academy, Gwalior, Madhya Pradesh, India.
Email: dharmendratripathi@rjit.ac.in
ORCID: 0009-0000-7999-6768
[3]Rustamji Institute of Technology, BSF Academy, Gwalior, Madhya Pradesh, India.
Email: dr.hsjat@rjit.ac.in
ORCID: 0009-0006-4123-3718
[4]InfinAIverse Pvt. Ltd., Jaipur, Rajasthan.
Email: vinay.joshiii@gmail.com
[5]Rajarshi Shahu College of Engineering, Pune.
Email: pradnyav123@gmail.com
Corresponding Author: Sanjay Kumar Gupta

Abstract: Rapid industrialization and urbanization have significantly increased the importance of air quality prediction for environmental monitoring, public health protection, and intelligent urban management. However, conventional forecasting methods often struggle to capture the complex spatial and temporal characteristics of environmental data, resulting in limited predictive accuracy. This study proposes a deep learning-assisted data visualization framework for real-time air quality interpretation and predictive analytics by integrating Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The Beijing Multi-Site Air Quality Dataset was utilized, followed by data cleaning, missing value imputation, Min–Max normalization, feature representation, and dataset partitioning in a 70:15:15 ratio for training, validation, and testing, respectively. The CNN extracts high-level spatial features, while the LSTM captures temporal dependencies to enhance forecasting performance. Experimental results demonstrate that the proposed hybrid CNN–LSTM model outperforms the standalone CNN and LSTM models, achieving an MSE of 11.38, MAE of 2.41, RMSE of 3.37, and an R^2 value of 0.972. Furthermore, the model achieved an MAE of 2.18, RMSE of 3.05, and an R^2 of 0.981 for 1-hour-ahead forecasting, demonstrating its effectiveness for accurate real-time air quality prediction and intelligent visualization-based decision support.
Keywords: Deep Learning (DL); Hybrid CNN–LSTM; Air Quality Prediction; Predictive Analytics; Intelligent Data Visualization.

1. Introduction

The proliferation of digital technologies such as Internet of Things (IoT) devices, cloud computing, social networks, healthcare systems and industrial automation has resulted in an unprecedented amount of data being created across varied application areas [1]. Structured and unstructured data flows at an enormous rate in an organization and need to be processed efficiently, interpreted and used in decision making. Traditional data visualization techniques are good for presenting historical data but are not suitable for real-time streams or complex multidimensional data [2,3]. They are usually based on static models that offer few analytical tools and do not capture patterns or make accurate forecasts. Thus, there is a growing need for intelligent visualization frameworks that integrate advanced analytics with automatic learning algorithms to enable real-time decision making [4].

Deep Learning (DL) has become a paradigm shift technology that can capture hierarchical features and complex nonlinear relationships from large-scale data sources [5]. CNNs, RNNs, LSTM networks, GRUs, Autoencoders and Transformer-based models have been used with great success in applications ranging from healthcare diagnostics to financial forecasting, industrial monitoring, cybersecurity, transportation and smart city management [6]. Compared to traditional machine learning models, DL models can learn meaningful feature representations automatically from raw data, which is highly beneficial for improving the accuracy of the prediction and decreasing the reliance on human feature engineering. The capabilities of DL, especially its ability to process streaming data continuously and accurately predict insights, make it a perfect fit to build intelligent data visualization systems [7].

Data visualization has been transformed from simple graphing to the inclusion of analytical intelligence to explore, interpret and understand complex data sets more efficiently [8]. Interactive visualization dashboards help spot anomalies, trends, hidden correlations and performance indicators with easy-to-use visual interfaces. These systems can be used to visualize high-dimensional data into meaningful graphs, which makes the analytical process easier and helps to make better decisions. The use of intelligent visualization in various fields like healthcare, industry, finance, environmental monitoring, smart infrastructure, etc. significantly improves the situational awareness, increases the operational efficiency, optimizes resource usage, and allows for quick response to constantly changing data [10]. The overall architecture of a real-time analytics system is depicted in Figure 1, where the data coming from IoT devices are preprocessed and analyzed, and then displayed via visualization tools for supporting the informed decision-making process.

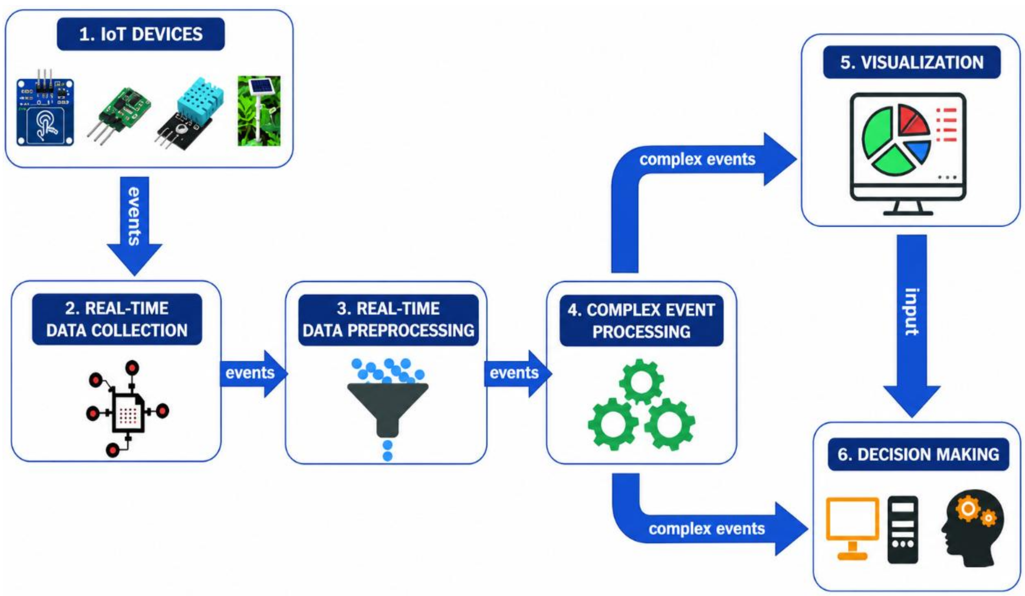


Figure 1. General workflow for real-time data acquisition, preprocessing, intelligent analytics, visualization, and decision-making [11].

DL and data visualization have become a new paradigm in smart visual analytics by coupling automated feature learning with dynamic visual interpretation [12]. DL assisted visualization systems can present prediction outcomes, uncertainty estimates, anomaly alerts, feature importance and temporal behavioural patterns generated through advanced learning models, instead of just displaying information in historical. This integration facilitates real-time monitoring, enhances the interpretability of predictive models, and aids in proactive decision-making, aiding in anticipation of critical events [13]. Predictive maintenance, healthcare diagnostics, financial risk assessment, intelligent transportation, energy management, environmental monitoring, and cybersecurity are only a few examples of fields that require the precise interpretation of rapidly changing data and thus have drawn significant interest from DL powered visual analytics [14].

However, there are a number of challenges remaining in the implementation of effective visualization frameworks using DL techniques. The current solutions are mainly geared towards predictive modeling or visualization in isolation, leading to non-scalable, non-interpretable and non-responsive solutions [15]. In addition, there are still a number of unsolved research challenges related to handling heterogeneous data sources, low latency data processing, making the visualization readable, and being able to explain predictions [16]. The combination of high-performance DL models and interactive visualization platforms demands efficient data preprocessing, feature extraction, real-time inference, and adaptive visualization strategies that can accommodate the evolving needs of users in various application areas [17].

The study aims to develop a data visualization framework for real-time data interpretation and predictive analytics by combining automated data pre-processing, deep feature extraction, predictive modeling, and intelligent visualization within a single architecture to cope with the aforementioned challenges. It uses the power of DL algorithms to harness complex patterns within the constantly flowing data streams and provides user friendly visual dashboards for real-time data monitoring and analysis. The suggested system makes anomaly detection, trend forecasting, predictive analytics and interactive decision support possible by adaptive visualization techniques. Its modular structure allows its seamless integration with different data sources and its scalability in cloud and edge-based environments. The following are the research objectives:

- To develop a deep learning-assisted framework for real-time data visualization and predictive analytics using environmental data.
- To preprocess air quality and meteorological data through cleaning, imputation, normalization, and feature representation for effective model training.
- To design a hybrid CNN–LSTM model for accurate $PM_{2.5}$ prediction by learning spatial and temporal data patterns.
- To assess the performance of the proposed model using MSE, MAE, RMSE, and R^2, and compare it with standalone CNN and LSTM models.
- To provide intelligent visualization of prediction results for real-time air quality monitoring and decision support.

2. Literature of Review

The use of DL methods for predictive analytics with heterogeneous and real time data has been widely studied recently. Sammy et al. (2025) [18] designed a visual analytics pipeline that uses DL algorithms with preprocessing, transfer learning, and CNN architectures such as VGG16, GoogleNet, and ResNet50, with ResNet50 demonstrating the highest classification accuracy on various satellite datasets. Likewise, Selmy et al. (2024) [19] compared the traditional forecasting models (ARIMA and SARIMA) with DL models and showed that LSTM and hybrid CNN-LSTM models predicted the temperature time-series with lower errors, and Saravanan et al. (2023) [20] used LSTM with sentiment analysis to predict the stock market. In addition, Akpan et al. (2026) [21] showed that machine learning algorithms could be usefully applied to the prediction of traffic congestion based on environmental data, while Mahdi et al. (2025) [22] compared various traffic congestion prediction models to be used for smart governance applications, concluding that Artificial Neural Networks (ANNs) could be the most accurate of the models. These studies collectively confirm that DL architectures are key to higher prediction accuracy, scalability, and performance in real-time analysis across a variety of applications. A comparison of the existing data visualization and predictive analytics frameworks using DL across various application fields is shown in Table 1.

Table 1. Comparative Analysis of Existing DL-Based Data Visualization and Predictive Analytics Frameworks

Ref.	Methodology / Models	Dataset / Data Source	Key Findings
Sammy et al. (2025)	CNN (VGG16, GoogleNet, ResNet50), Transformer, Transfer Learning, PSO Optimization	MODIS, Landsat-8, Sentinel-2	ResNet50 achieved the highest classification accuracy (95.2%) with robust real-time environmental monitoring capabilities.
Selmy et al. (2024)	ARIMA, SARIMA, LSTM, CNN-LSTM	Temperature Time-Series Data	CNN-LSTM and LSTM significantly outperformed traditional statistical forecasting models with lower prediction errors.

Rajesh et al. (2025)	Random Forest, Gradient Boosting, XGBoost, LSTM, SHAP	Air quality sensors, satellite, meteorological, demographic data	Developed an interpretable real-time prediction framework with cloud dashboards and health-risk visualization.
Liu et al. (2023)	CNN with Temporal Pooling (VEVF)	Sports Video Data	Proposed an AI-based visualization framework achieving 98.7% accuracy for sports video analysis.
Zhu et al. (2022)	Attention-based Evidential RNN, Edge Computing	Clinical Blood Glucose Dataset	Enabled real-time glucose prediction with mobile visualization and superior prediction accuracy.
Saravanan et al. (2023)	LSTM + Sentiment Analysis (NLP)	Twitter & Facebook Stock Data	Integrated sentiment analysis with LSTM to improve stock market prediction accuracy and visualization.
Tuarob et al. (2021)	Ensemble Learning, Contextual Feature Engineering	News, social media, Financial Data	Combined predictive analytics with interactive visualization, achieving low forecasting error (MAPE = 0.93).
Akpan et al. (2026)	KNN and Machine Learning Models	Air Pollution and Traffic Data	KNN reduced traffic prediction error by over 30% using environmental indicators.
Razali et al. (2022)	Random Forest, JRip	Student Academic Dataset	Achieved up to 100% accuracy and enabled real-time academic performance visualization.
Mahdi et al. (2025)	Linear Regression, Decision Tree, Random Forest, ANN	Open Government Data	ANN achieved the highest predictive accuracy for evidence-based policy and resource planning.

Another area of significant research interest is in combining predictive analytics with interactive data visualization, enabling efficient interpretation and decision making based on the data. Liu et al. [23] introduced an effective visualization framework based on Artificial Intelligence (AI) to analyze sports videos by introducing CNN and Temporal Pooling layers, which resulted in high accuracy and effective visualization of tactical performance. Likewise, Tuarob et al. (2021) [24] introduced a single architecture that combined financial data collection, contextual feature engineering, ensemble learning, and interactive visualization for real-time stock market prediction, which helps investors better understand the market situation. Zhu et al. (2022) [25] also enhanced visualization with smartphone apps and cloud-based dashboards to serve as continuous glucose trajectories and prediction results for healthcare monitoring. Rajesh et al. (2025) [26] added dashboards utilizing cloud technology, mobile alerts, and visual risk mapping for real-time environmental monitoring and health advice generation, while Razali et al. (2022) [27] integrated machine learning and visualization dashboards for learning analytics purposes, allowing for early intervention by way of real-time monitoring of academic performance. While these studies show the increasing significance of predictive analytics plus visualization, most current frameworks are application specific and focus on either predictive performance or visualization. Thus, the need for a comprehensive DL-aided visualization system, which can seamlessly integrate the processing of heterogeneous data, real-time interpretation, interactive visualization, and predictive analytics under a single roof is an important research path.

3. Research Methodology

Figure 2 illustrates the overall workflow of the proposed deep learning-assisted data visualization framework, including data collection, preprocessing, dataset partitioning, hybrid CNN–LSTM model development, prediction, visualization, and real-time decision support for accurate air quality monitoring and predictive analytics.

3.1 Dataset Description

Beijing Multi-Site Air Quality Dataset [28] is a publicly available multivariate time-series dataset with air quality and meteorological observations recorded hourly at 12 air quality monitoring stations, all under the national control, in Beijing, China. This data set covers the time period from March 1, 2013 to February 28, 2017, and contains the following air pollutants: PM2.5, PM10, SO_2, NO_2, CO, O_3, among others, and the following meteorological parameters: temperature, pressure, dew point, wind direction, wind speed, rainfall, humidity, etc. It is observed in large quantities in real-world applications, which makes it well suited for use in DL predictive analytics, real-time data interpretation, and intelligent visualization applications.

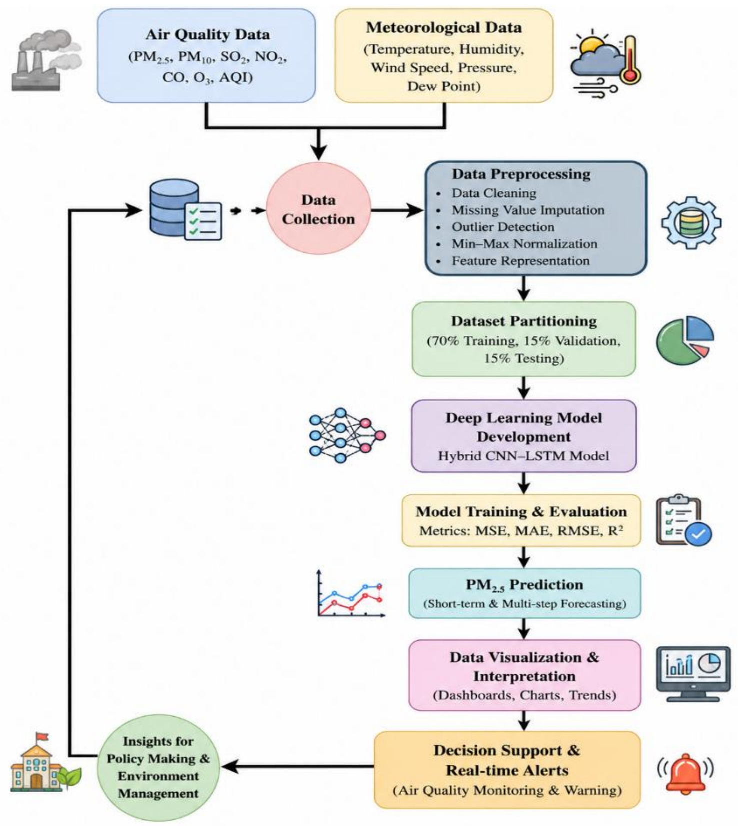


Figure 2. Proposed Methodology

3.2 Data Preprocessing

In the proposed DL assisted data visualization framework, data preprocessing is an important step, and the quality of data input directly affects the accuracy of the prediction and effectiveness of data visualization. The Beijing Multi-Site Air Quality Dataset suffers from missing observations, noise, inconsistencies, and changes in feature scales, as it is hourly air quality measurements from multiple monitoring stations that are collected over multiple years. For these reasons, a thorough preprocessing pipeline is used to preprocess the raw environmental data and convert it into a consistent and learning-friendly format. Preprocessing is a process that includes the following steps: data cleaning, missing values handling, normalization, and partitioning of the data set for model training.

3.2.1 Data Cleaning

The first step is to detect and eliminate inconsistent records, duplicate records, and bad observations that might adversely affect the performance of the models. Sometimes, air quality observations taken at several monitoring stations include partial sensor readings or abnormal readings due to the failure or incorrect operation of the sensors or due to communication problems.

Let

$$D = \{x_1, x_2, x_3, \dots, x_n\}$$

represent the original dataset containing n observations. After removing duplicate and inconsistent records, the cleaned dataset is represented as

$$D_c = D - \{x_{duplicate}, x_{invalid}\}$$

where D_c denotes the cleaned dataset used for further processing.

3.2.1. Missing Value Imputation

The data from environment may be missing due to sensor failure, communication loss, or maintenance. On missing numerical values, no samples are deleted and the values are replaced by mean imputation to maintain data continuity.

For a feature X,

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

where

X_i represents the observed values and N denotes the number of available observations.

Each missing value is replaced by

$$X_{missing} = \bar{X}$$

This helps to reduce the loss of information and ensures that there is statistical consistency across the dataset.

3.2.2. Feature Normalization

The dataset contains variables measured in different units, including pollutant concentrations (µg/m^3), atmospheric pressure (hPa), temperature (°C), humidity (%), and wind speed (m/s). Directly training DL models on features with different numerical ranges may lead to unstable convergence.

Therefore, Min-Max normalization is applied to transform all numerical features into the interval [0,1].

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where

X is the original feature, X_{min} is the minimum value, X_{max} is the maximum value, and X' is the normalized feature.

Normalization helps all features converge toward a shared range and helps prevent features with greater numerical values to dominate the optimization process.

3.2.3. Dataset Partitioning

Once preprocessed, the dataset is split into training, validation, and testing set to guarantee the model's development and assessment. A partition ratio of 70:15:15 is implemented, with 70% of the data used for training the model, 15% for model validation to optimize the hyperparameters, and 15% for model testing to evaluate the model's predictive performance and generalization ability with new test data.

$$D = D_{train} \cup D_{validation} \cup D_{test}$$

3.3 Input Feature Representation

The input feature representation stage is used to convert the preprocessed environmental observations into a structured multivariate feature matrix, to be used for DL-based prediction. The proposed framework provides a characterization of the air quality conditions based on the key air pollutants and meteorological parameters such as $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3, temperature, atmospheric pressure, dew point, wind speed, relative humidity and rainfall. These features describe pollutant behaviour and environmental changes, which allow the model to learn more complex relationships linked to air quality dynamics. Normalization is performed on the selected attributes and arranged in a single feature matrix which is fed into the proposed hybrid CNN–LSTM model. This representation enhances the efficiency of learning and offers complete information for making a correct prediction and interpreting data in real-time.

The input feature vector is represented as

$$F = \{PM_{2.5}, PM_{10}, SO_2, NO_2, CO, O_3, T, P, DP, WS, RH, RF\}$$

where T, P, DP, WS, RH, and RF denote temperature, atmospheric pressure, dew point, wind speed, relative humidity, and rainfall, respectively.

The normalized input feature matrix is expressed as

$$X = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^{n \times m}$$

where n represents the total number of observations and $m$denotes the number of selected input features. The feature matrix X is then fed into the CNN module where the automatic hierarchical feature learning takes place, further temporal modeling is performed by the LSTM network.

3.4 Proposed Hybrid CNN–LSTM Model Development

3.4.1. CNN-Based Spatial Feature Learning

The first stage of the proposed hybrid CNN–LSTM DL system is the CNN used to automatically learn meaningful spatiotemporal information from the multivariate air quality dataset. Conventional machine learning techniques require some handcrafted feature engineering, while CNN learns feature representations hierarchically from the normalized input data directly using convolution and pooling operations. The extracted environmental variables (such as pollutant concentrations and meteorological variables) are presented in feature matrices and fed into the convolutional layers in the proposed framework. Each of these layers contains several learnable filters that detect local correlations and underlying patterns between the input variables, helping to improve the representation of features and eliminate noise and redundancy. Then the feature maps are down-sampled for pooling to reduce the computational complexity and increase the robustness of the model. The high-level features thus extracted are then passed to the LSTM network which learns temporal sequences and predicts future air quality.

The convolution operation performed by CNN is expressed as:

$$F_k(i,j) = \sigma\left(\sum_m \sum_n X\,(i+m, j+n) \cdot W_k(m,n) + b_k\right)$$

where X represents the input feature matrix, W_k denotes the k^{th} convolution kernel, b_k is the bias term, $\sigma(\cdot)$ is the activation function (ReLU), and F_k is the generated feature map.

To reduce feature dimensionality while preserving the most significant information, max-pooling is applied as

$$P(i,j) = \max_{(u,v)\in\Omega} F(u,v)$$

where Ω denotes the pooling window and $P(i,j)$ represents the pooled feature map. The extracted high-level spatial features are then passed to LSTM module, which effectively learns the temporal dependency and enhances the overall predictive performance of the proposed hybrid CNN–LSTM framework.

3.4.2. LSTM-Based Temporal Dependency Learning

The CNN module is followed by a LSTM network to learn temporal dependencies and sequential patterns in the multivariate air quality data. Conventional neural networks have a tendency to lose out on long-term context because of the vanishing gradient problem caused by temporal correlation of the environmental parameters like $PM_{2.5}$, PM_{10}, temperature, humidity and gaseous pollutants. However, LSTM has two drawbacks, which it addresses with a memory cell and three gating mechanisms—Input Gate, Forget Gate, and Output Gate—that selectively preserve or forget information over time. In the proposed hybrid approach, the high-level spatial features extracted by the CNN are processed sequentially by the LSTM network to learn temporal variations and to make accurate forecasts in the future. The LSTM's ability to capture and learn from long-term dependencies between past observations makes it more accurate for prediction and more reliable for real-time predictive analytics. The acquired temporal representations are then used for visualization and decision support in the designed intelligent framework.

The forget gate computes the amount of previous information that should be forgotten from the memory cell as follows:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

where x_t denotes the current input, h_{t-1} is the previous hidden state, W_f represents the weight matrix, b_f is the bias vector, and $\sigma(\cdot)$ is the sigmoid activation function.

The hidden state generated by the LSTM is expressed as

$$h_t = o_t \odot \tanh(C_t)$$

where o_t is the output gate, C_t denotes the updated cell state, $\tanh(\cdot)$ is the hyperbolic tangent activation function, and $\odot$ represents element-wise multiplication. These hidden representations are used to store the temporal information over a long period of time and thereby, the proposed CNN–LSTM model is able to achieve good prediction of the video frame and high analytical accuracy in the real-time system.

3.4.3. Proposed hybrid CNN–LSTM Archiecture

The proposed CNN–LSTM model combines the feature extraction ability of CNN with the temporal learning ability of LSTM networks to accurately predict the air quality in real time. The pre-processed multivariate environmental information is fed into the CNN layers, with convolution and pooling automatically learning high level spatial features and discarding redundant information. These learned feature representations are then fed as temporal inputs to an LSTM network that captures long-term temporal dependencies and historical trends in the time-series data. The proposed architecture takes advantage of both spatial and temporal features for better prediction accuracy than individual CNN or LSTM models. Moreover, the forecast values are delivered to the intelligent visualization module to provide real-time graphical interpretation, anomaly monitoring and decision support.

The hybrid model can be represented as

$$Y = \text{LSTM}(\text{CNN}(X))$$

where X denotes the pre-processed input feature matrix, CNN(·) extracts high-level spatial features, LSTM(·) learns temporal dependencies, and Yrepresents the predicted output used for real-time visualization and predictive analytics. Table 2 presents the hyperparameter configuration adopted for training the proposed hybrid CNN–LSTM predictive analytics model.

Table 2. Hyperparameter Settings of the Proposed Hybrid CNN–LSTM Model

Parameter	**Value**
Convolution Layers	2
Number of Filters	32, 64
Kernel Size	3 × 3
Activation Function	ReLU
LSTM Units	128
Optimizer	Adam
Batch Size	32
Loss Function	Mean Squared Error (MSE)

3.5 Predictive Analytics Module

The predictive analytics module employs the proposed hybrid CNN–LSTM model to predict the future air quality level for the next few hours based on the past air pollutant concentration and weather parameters. The CNN extracts spatial features which are in turn processed by the LSTM network to model the long-term temporal dependencies to make accurate predictions. The trained model continuously processes incoming data streams from the environment, and identifies emerging trends in pollution, detects anomalies, and predicts future air quality levels.

These predictions can help with proactive monitoring, early risk assessment and decision making in relation to the environment. In addition, the forecasted values are sent to the visualization part, which displays them in interactive dashboards, trend charts and alerts, allowing the interpretation of the environmental conditions in real time. The DL integration with the predictive analytics greatly boosts the forecast accuracy and the intelligent monitoring and dynamic visualization within the proposed framework.

The predicted output of the hybrid model is expressed as

$$\hat{Y}_{t+1} = f(X_t; \theta)$$

where $\hat{Y}_{t+1}$ represents the predicted air quality value at the next time step, X_t denotes the input feature vector, $f(\cdot)$ is the trained hybrid CNN–LSTM model, and θ represents the optimized model parameters.

3.6 Intelligent Data Visualization Module

The intelligent data visualization module translates the predictive results of the hybrid CNN–LSTM model into real-time visualizations that can be interpreted and used for decision making. Dynamic dashboards, line charts, heatmaps, and gauge indicators visualize the predicted air quality values, their temporal trends, and the air quality anomaly alerts, allowing real-time monitoring of environmental conditions. The framework combines predictive analytics and visualization, making it easier to interpret the complex multivariate data, and to detect unusual pollution patterns in a timely fashion. As the data streams into the visualization interface, it continually updates the screen with data, allowing for real-time monitoring and increasing situational awareness for environmental management, policy planning, and informed decision making. The integration enhances the ease of access, interpretability, and usability of predictive insights within the proposed framework.

$$V_t = \Phi(\hat{Y}_t, T_t, A_t)$$

where V_t represents the visualization output at time t, $\hat{Y}_t$ denotes the predicted air quality values, T_t represents temporal trend information, A_t indicates anomaly alerts, and $\Phi(\cdot)$ is the visualization mapping function that transforms predictive results into interactive dashboards for real-time interpretation. The visualization module is implemented using interactive dashboard components that continuously update with streaming prediction results.

3.7 Performance Evaluation

The proposed hybrid CNN–LSTM framework is tested through standard regression metrics to check the accuracy of the predictions and reliability of the model. Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Coefficient of Determination (R^2) are used to measure differences between actual and predicted air quality values, and these metrics are employed because air quality prediction is a time-series forecasting problem. The lower values of MAE and RMSE and the higher R^2 value reflect better predictive performance and model generalisation.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | Y_i - \hat{Y}_i |$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

where Y_i is the actual air quality value, $\hat{Y}_i$ is the predicted value, $\bar{Y}$ is the mean of the observed values, and N denotes the total number of test samples.

4. Results and Discussion

4.1 *Comparative Performance Analysis of Deep Learning Models*

Table 3 shows the prediction performance of the proposed hybrid CNN–LSTM model and the single CNN/LSTM model with the conventional regression metrics. The hybrid CNN–LSTM network has the lowest MSE value, which is 11.38, MAE value, which is 2.41, and RMSE value, which is 3.37, suggesting a lower prediction error and better forecasting accuracy. In addition, it has the highest coefficient of determination ($R^2 = 0.972$), which indicates that there is a high correlation between the actual and predicted air quality values. The results show that the combination of CNN and LSTM has a significant impact on the reliability of predictions and the overall performance of the model.

Table 3. Performance Evaluation Metrics of the Proposed Hybrid CNN–LSTM Model

Model	MSE	MAE	RMSE	R^2
CNN	18.72	3.48	4.33	0.941
LSTM	15.86	3.11	3.98	0.953
Proposed CNN–LSTM	**11.38**	**2.41**	**3.37**	**0.972**

Figure 3 shows the comparative performance of CNN, LSTM, and the proposed hybrid CNN–LSTM model based on four evaluation metrics: MSE, MAE, RMSE, and R^2. The hybrid CNN–LSTM model shows the smallest MSE (11.38), MAE (2.41), and RMSE (3.37) which signify that it has significantly less prediction errors than the CNN and LSTM model. Furthermore, it shows the highest R^2 value (0.972), indicating a higher correlation between the concentration of $PM_{2.5}$ predicted and observed. The results indicate that the proposed CNN–LSTM model can achieve better predictions than CNN and LSTM models by minimizing prediction errors, enhancing general forecasting performance, and improving the generalization of the model. The proposed hybrid CNN–LSTM model outperformed the standalone CNN by reducing the MSE by around 39.2%, highlighting the synergy between spatial feature extraction and temporal sequence learning.

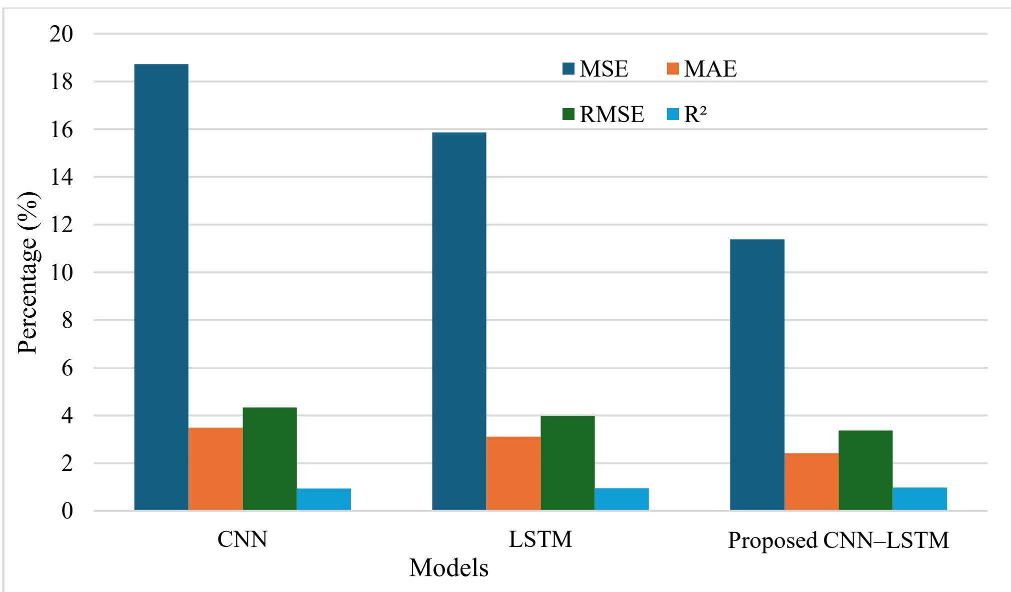
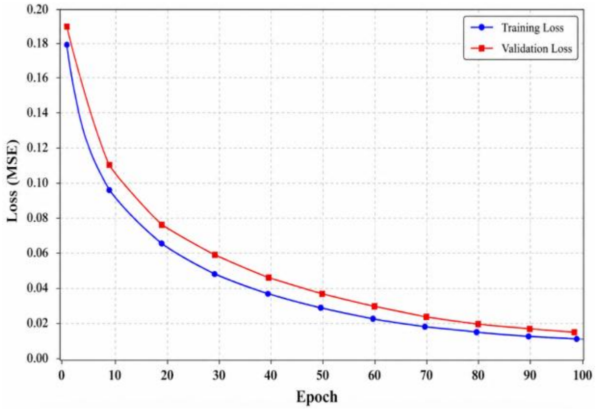


Figure 3. Performance comparison of CNN, LSTM, and the proposed hybrid CNN–LSTM model.

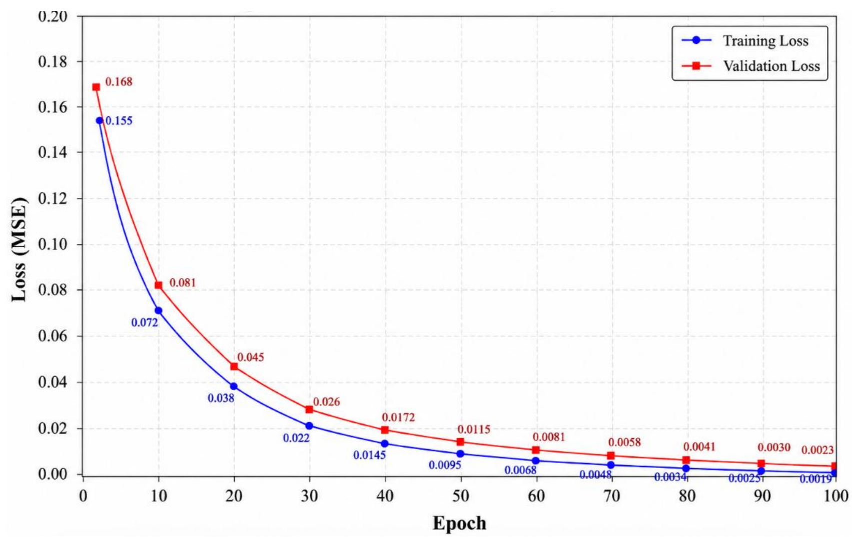
4.2 *Training Performance Analysis*

The CNN training and validation loss curves are shown in Figure 4 in comparison with the LSTM model and the proposed CNN–LSTM model for 100 epochs of training. The loss for all three models steadily decreases, suggesting convergence during training. But, the hybrid CNN–LSTM gives the best convergence speed and least final loss value, where the training loss goes down from 0.155 to 0.0019 and the validation loss goes down from 0.168 to 0.0023. The CNN and LSTM models, on the other hand, converge to more similar loss values. The small difference

between the training and validation curves, consistently, validates the proposed hybrid model as it is well generalized and has high predictive power with low overfitting.



(a) CNN (b) LSTM



(c) Hybrid CNN+LSTM

Figure 4. Training and validation loss curves of (a) CNN, (b) LSTM, and (c) the proposed hybrid CNN–LSTM model.

4.3 Prediction Performance Analysis

The forecasting accuracy of the proposed hybrid CNN–LSTM model at various prediction horizons is shown in Table 4. The model has the best performance for 1 hour ahead forecast with MAE of 2.18, RMSE of 3.05 and R^2 of 0.981. The longer the forecasting time window (3, 6, and 12 hours), the higher the forecasting error, and the lower the R^2 value, which is slightly reduced to 0.963. Even though this is the expected decrease, the model consistently provides high predictive accuracy, and shows strong performance in both short and long-range air quality forecasting.

Table 4. Prediction Performance at Different Forecast Horizons

Forecast Horizon	MAE	RMSE	R^2
1 Hour Ahead	2.18	3.05	0.981
3 Hours Ahead	2.43	3.29	0.976
6 Hours Ahead	2.78	3.71	0.970

12 Hours Ahead	3.16	4.08	0.963

The result of the actual and predicted concentration of $PM_{2.5}$ by the proposed hybrid model CNN-LSTM for test dataset is shown in figure 5. The predicted values are in good agreement with observations, both in terms of abrupt changes and trends, in air quality. The quantitative assessment also shows the effectiveness of the proposed framework, with an MSE of 11.38, MAE of 2.41, RMSE of 3.37 and an R^2 value of 0.972, thus showing very good level of agreement between the predicted and observed $PM_{2.5}$ concentrations. The slight difference between the two curves indicates the validity of the proposed hybrid CNN–LSTM model for real-time air quality forecasting, which offers strong, stable, and highly reliable predictions.

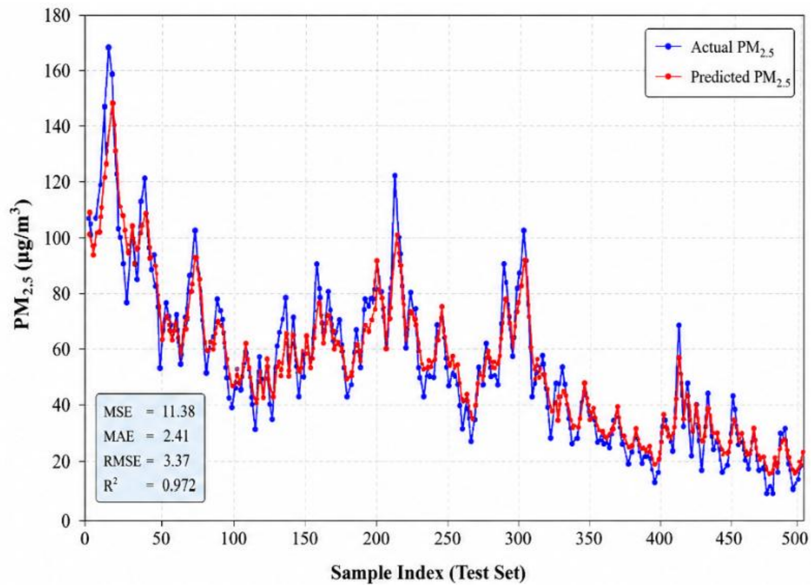


Figure 5. Comparison of actual and predicted $PM_{2.5}$ concentrations using the proposed hybrid CNN–LSTM model.

The proposed hybrid CNN–LSTM model is used to conduct the correlation analysis between the actual and predicted $PM_{2.5}$ concentrations, as shown in Figure 6. The majority of the data points are tightly clustered along the line of best fit , suggesting a high level of agreement between the actual and predicted data. The fitted regression line is very close to the reference line, which indicates high predictive consistency of the model. The quantitative results are also valid to the model with the MSE value of 11.38, MAE value of 2.41, RMSE value of 3.37, and an R^2 value of 0.972. The small dispersion of the data points in the presented figure validates the effectiveness and reliability of the proposed hybrid CNN–LSTM model for real-time $PM_{2.5}$ forecast. (y=x)

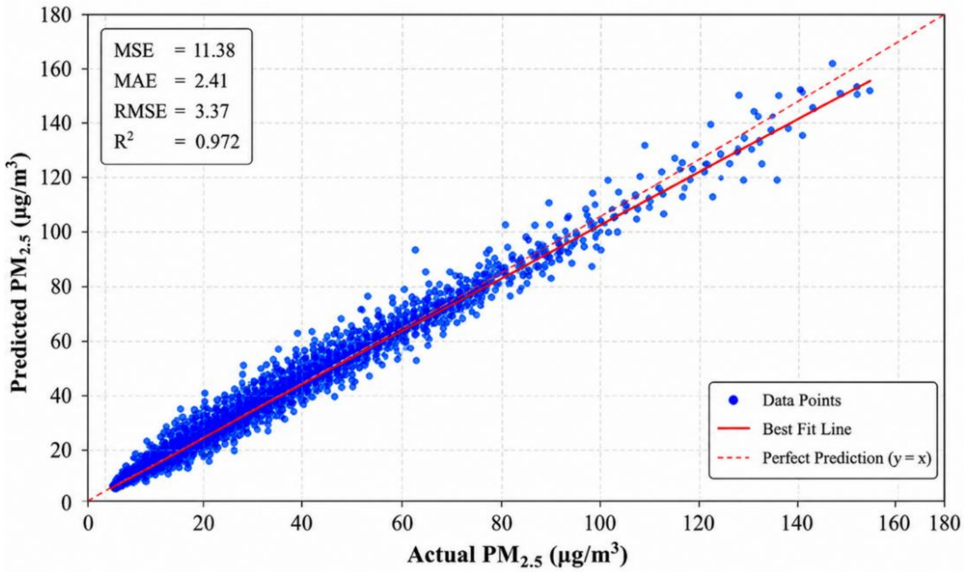


Figure 6. Scatter plot comparing actual and predicted $PM_{2.5}$ concentrations using the proposed hybrid CNN–LSTM model.

4.4 Error Analysis

Table 5 shows the prediction accuracy of the proposed hybrid CNN–LSTM model for various air quality conditions. The model has the greatest accuracy in the good category, with an MAE of 1.82, RMSE of 2.65, and R^2 of 0.981. The prediction errors slowly grow as the air quality worsens from Moderate to Hazardous, while the R^2 value slowly halves itself to 0.961. However, the model is always able to predict $PM_{2.5}$ concentrations with high accuracy for all categories, which shows that the model is robust, reliable, and effective for predicting $PM_{2.5}$ concentrations under different environmental conditions.

Table 5. Performance of the Proposed CNN–LSTM Under Different Air Quality Conditions

Air Quality Category	MAE	RMSE	R^2
Good	1.82	2.65	0.981
Moderate	2.33	3.14	0.975
Unhealthy	2.91	3.82	0.969
Hazardous	3.54	4.43	0.961

Time-series and scatter plot analyses were performed to further validate the forecasting ability of the proposed hybrid CNN–LSTM model, by comparing the $PM_{2.5}$ forecasts with the corresponding ground-truth observations. The results of visual comparison shown in Figure 6 and Figure 7 also support the effectiveness of the proposed model.

The prediction error distribution of the proposed hybrid CNN–LSTM model for $PM_{2.5}$ prediction is shown in the figure 7. The distribution of prediction errors is approximately normally distributed with a mean near zero, suggesting that most prediction errors are small and unbiased. The mean prediction error of 0.21 µg/m³ and median error of 0.15 µg/m³ illustrate a low level of systematic bias between the observed and modeled concentrations of $PM_{2.5}$. Moreover, the standard deviation is 3.31 µg/m³ and the RMSE is 3.37 µg/m³, which demonstrates low variation in the errors and high stability in the prediction. The errors are clustered around zero, which further validates the robustness, consistency, and reliability of the hybrid CNN–LSTM model proposed for real-time air quality prediction.

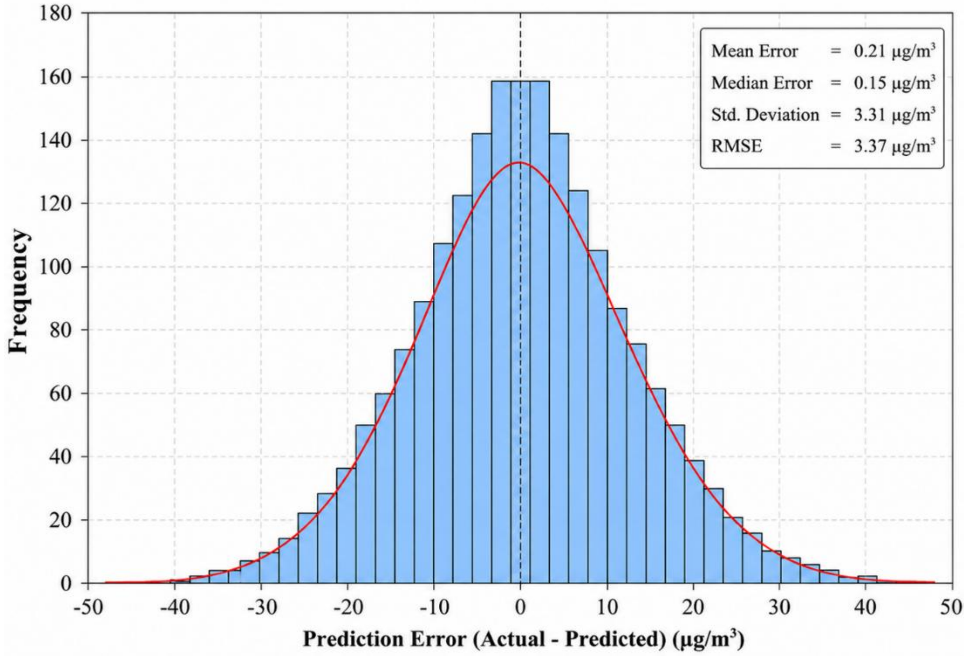


Figure 7. Prediction error distribution of the proposed hybrid CNN–LSTM model for $PM_{2.5}$ forecasting.

5. Conclusion and Future Scope

In this study, the authors proposed a DL-based data visualization system for real-time air quality interpretation and predictive analytics which utilizes CNN and LSTM in a unified hybrid system. The suggested framework involved data preprocessing, feature representation, predictive modeling and intelligent visualization to analyze the multivariate environmental data efficiently. To test the effectiveness of the proposed approach in the real world, the Beijing Multi-Site Air Quality Dataset was used. Experimental results showed that the hybrid CNN+LSTM model outperformed the CNN and LSTM models in all the metrics used and provided good generalization of the model with high prediction accuracy for MSE (11.38), MAE (2.41), RMSE (3.37) and R^2 (0.972). Moreover, the framework was able to provide robust forecasting under various time horizons and various air quality conditions and provide interactive visual analytics for temporal trends of pollution. The prediction error analysis also validated the robustness and stability of the proposed model, which shows it can be used to monitor the environment, provide intelligent decision support and for smart city applications.

Future research will be directed towards incorporating Transformer-based architectures, explainable artificial intelligence (XAI), and edge-enabled IoT platforms for further enhancing prediction accuracy, explainability, and real-time deployment capabilities.

References

1. Bablu, Tarek Aziz, and Mohammad Tanvir Rashid. "Edge computing and its impact on real-time data processing for IoT-driven applications." Journal of advanced computing systems 5, no. 1 (2025): 26-43.
2. Emma, Lawrence. "Big data analytics for real-time insights and strategic business planning." In Proc. Conf. 2024.
3. Sultana, Rebeka. "Artificial intelligence in data visualization: Reviewing dashboard design and interactive analytics for enterprise decision-making." International Journal of Business and Economics Insights 5, no. 3 (2025): 01-29.
4. Uddin, Md Kazi Shahab. "A review of utilizing natural language processing and AI for advanced data visualization in real-time analytics." Global Mainstream Journal 1, no. 4 (2024): 10-62304.
5. Ahmed, Shams Forruque, Md Sakib Bin Alam, Maruf Hassan, Mahtabin Rodela Rozbu, Taoseef Ishtiak, Nazifa Rafa, M. Mofijur, A. B. M. Shawkat Ali, and Amir H. Gandomi. "Deep learning modelling techniques: current progress, applications, advantages, and challenges." Artificial Intelligence Review 56, no. 11 (2023): 13521-13617.
6. Shiri, Farhad Mortezapour, Thinagaran Perumal, Norwati Mustapha, and Raihani Mohamed. "A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU." arXiv preprint arXiv:2305.17473 (2023).
7. Wang, Qianwen, Chen Zhu-Tian, Yong Wang, and Huamin Qu. "A survey on ML4VIS: Applying machine learning advances to data visualization." IEEE transactions on visualization and computer graphics 28, no. 12 (2021): 5134-5153.

8. Shakeel, Hafiz Muhammad, Shamaila Iram, Hussain Al-Aqrabi, Tariq Alsboui, and Richard Hill. "A comprehensive state-of-the-art survey on data visualization tools: Research developments, challenges and future domain specific visualization framework." IEEE Access 10 (2022): 96581-96601.
9. Kobi, Joseph. "Developing dashboard analytics and visualization tools for effective performance management and continuous process improvement." International Journal of Innovative Science and Research Technology (IJISRT) 9, no. 10.38124 (2024).
10. Akter, Momena, and Sai Praveen Kudapa. "A comparative analysis of artificial intelligence-integrated bi dashboards for real-time decision support in operations." International Journal of Scientific Interdisciplinary Research 5, no. 2 (2024): 158-191.
11. Fournier, Fabiana, and Inna Skarbovsky. "Real-time data processing." In Big Data in Bioeconomy: Results from the European DataBio Project, pp. 147-156. Cham: Springer International Publishing, 2021.
12. Afzal, Shehzad, Sohaib Ghani, Mohamad Mazen Hittawe, Sheikh Faisal Rashid, Omar M. Knio, Markus Hadwiger, and Ibrahim Hoteit. "Visualization and visual analytics approaches for image and video datasets: A survey." ACM Transactions on Interactive Intelligent Systems 13, no. 1 (2023): 1-41.
13. Adepoju, ADEBUSAYO HASSANAT, B. L. E. S. S. I. N. G. Austin-Gabriel, O. L. A. D. I. M. E. J. I. Hamza, and A. N. U. O. L. U. W. A. P. O. Collins. "Advancing monitoring and alert systems: A proactive approach to improving reliability in complex data ecosystems." IRE Journals 5, no. 11 (2022): 281-282.
14. Afanaseva, Olga Vladimirovna, Timur Faritovich Tulyakov, and Artur Airatovich Shaimardanov. "Deep Learning-Based Visual Analytics for Efficiency and Safety Optimization in Power Infrastructure." Eng 7, no. 3 (2026): 135.
15. Olayinka, Olalekan Hamed. "Big data integration and real-time analytics for enhancing operational efficiency and market responsiveness." Int J Sci Res Arch 4, no. 1 (2021): 280-96.
16. Fekete, Jean-Daniel, Yifan Hu, Dominik Moritz, Arnab Nandi, Senjuti Basu Roy, Eugene Wu, Nikos Bikakis et al. "Human-Data Interaction, Exploration, and Visualization in the AI Era: Challenges and Opportunities." arXiv preprint arXiv:2603.05542 (2026).
17. Rane, Nitin Liladhar, Mallikarjuna Paramesha, Saurabh P. Choudhary, and Jayesh Rane. "Machine learning and deep learning for big data analytics: A review of methods and applications." Partners Universal International Innovation Journal 2, no. 3 (2024): 172-197.
18. Sammy, F., Thiyagarajan Chettier, Venkata Boyina, Harshala Shingne, Kamal Saluja, Manisha Mali, R. Sugumar, and A. Shobana. "Deep Learning-Driven Visual Analytics Framework for Next-Generation Environmental Monitoring." Journal of Applied Science and Technology Trends (2025): 114-122.
19. Selmy, Hend A., Hoda K. Mohamed, and Walaa Medhat. "A predictive analytics framework for sensor data using time series and deep learning techniques." Neural Computing and Applications 36, no. 11 (2024): 6119-6132.
20. Saravanan, Vijayalakshmi, Laxman Paudel, Pramod Acharya, Praveen Paramasivam, and Anju S. Pillai. "An Efficient LSTM-Based Deep Learning Model for Stock Prediction Analytics and Real-time Visualization." (2023).
21. Akpan, Edidiong E., Oluwatobi Akinlade, Oluwaseyi O. Alabi, Oluwasesan A. David, Oluwaseyi F. Afe, and Sunday Adeola Ajagbe. "Optimizing traffic signal control using machine learning and environmental data." In Artificial Intelligence and Applications, vol. 4, no. 1, pp. 48-56. 2026.
22. AL-Inizi, Mahdi Salah Mahdi. "Enhancing governmental decision-making through predictive analytics with machine learning-based data-driven framework." Babylonian Journal of Machine Learning 2025 (2025): 86-96.
23. Liu, Aijun, Rajendra Prasad Mahapatra, and A. V. R. Mayuri. "Hybrid design for sports data visualization using AI and big data analytics." Complex & Intelligent Systems 9, no. 3 (2023): 2969-2980.
24. Tuarob, Suppawong, Poom Wettayakorn, Ponpat Phetchai, Siripong Traivijitkhun, Sunghoon Lim, Thanapon Noraset, and Tipajin Thaipisutikul. "DAViS: a unified solution for data collection, analyzation, and visualization in real-time stock market prediction." Financial innovation 7, no. 1 (2021): 56.
25. Zhu, Taiyu, Lei Kuang, John Daniels, Pau Herrero, Kezhi Li, and Pantelis Georgiou. "IoMT-enabled real-time blood glucose prediction with deep learning and edge computing." IEEE Internet of Things Journal 10, no. 5 (2022): 3706-3719.
26. Rajesh, M., R. Ganesh Babu, Usha Moorthy, and Sathishkumar Veerappampalayam Easwaramoorthy. "Machine learning-driven framework for realtime air quality assessment and predictive environmental health risk mapping." Scientific Reports 15, no. 1 (2025): 28801.
27. Razali, Mohd Norhisham, Habiel Zakariah, Rozita Hanapi, and Emelia Abdul Rahim. "Predictive model of undergraduate student grading using machine learning for learning analytics." In 2022 4th International Conference on Computer Science and Technologies in Education (CSTE), pp. 260-264. IEEE, 2022.
28. Zhang, S., Guo, B., Dong, A., He, J., Xu, Z., & Chen, S. X. (2017). Cautionary Tales on Air-Quality Improvement in Beijing. Proceedings of the Royal Society A, 473(2205), 20170457.