Journal of Mines, Metals and Fuels, *73(10): 3073-3083; 2025. DOI: 10.18311/jmmf/2025/49500*

Online ISSN : 0022-2755

Journal of Mines, Metals and Fuels

Contents available at: *www.informaticsjournals.co.in/index.php/jmmf*

Deep Learning and IoT based Air Quality Monitoring and Prediction System

Ajay Kumar[1], Pallavi S. Bangare[2], E. Jayanthi[3], Manisha P. Mali[4], Venkatesan Hariram[5], Mohan Awasthy[6], F. Sammy[7] and Abhishek Raghuvanshi[8*]

[1]School of CSE, IILM University, Greater Noida - 201306, Uttar Pradesh, India
[2]Sinhgad Academy of Engineering, Savitribai Phule Pune University, Pune - 411041, Maharashtra, India
[3]Department of Computer Science and Engineering,Presidency University, Bangalore - 560089, Karnataka, India
[4]Department of Computer Science and Engineering, Vishwakarma Institute of Information Technology, Pune - 411048, Maharashtra, India
[5]Department of Practice of Medicine, Vinayaka Mission's Homoeopathic Medical College and Hospital, Vinayaka Mission's Research Foundation Deemed to be University, Salem - 636308, Tamil Nadu, India
[6]Bharati Vidyapeeth Deemed to be University Department of Engineering and Technology Navi Mumbai - 400614, Maharashtra, India
[7]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram - 522502, Andhra Pradesh, India
[8]Department of Computer Science and Engineering, Mahakal Institute of Technology, Ujjain - 456006, Madhya Pradesh, India; abhishek14482@gmail.com

Abstract

The air quality is quite poor due to the high quantities of particulate matter in the surrounding environment. Incomplete coal combustion, automobile emissions, and particulate matter (PM2.5) emissions are all potential causes of air pollution that may have an impact on human health. Carbon Monoxide (CO) is another gas produced mostly by incomplete combustion of fuel. Vehicle engines are responsible for 80% of carbon monoxide emissions in urban areas. Smog is an inescapable component of city life, especially in big metropolitan areas. This paper introduces an IoT and deep learning-based architecture for air quality monitoring and forecasting. In this framework, the input data set is created by gathering air quality-related data from various sensors and IoT devices. The raw input data set is then reduced into features using the principal component analysis approach. This strategy is also useful for coping with noisy and missing data in the data collection. The classification model is created utilising convolution, neural network, recurrent neural network, and LSTM deep learning approaches. The classification model is trained and tested on a preprocessed data set. Performance of various models is evaluated using metrics such as R2, RMSE, and MSE. The air quality index is also anticipated in the conclusion.

Major Findings: The finding shows that the *CNN consistently outperforms the other models, achieving the lowest MSE and RMSE values and the highest R^2 scores. This implies that CNN is the most effective model for predicting air pollution levels.*

Keywords: *Air Quality Monitoring, Convolution Neural Network, Internet of Things, Principal Component Analysis, Root Mean Square Error, Sensors*

**Author for correspondence*

1.0 Introduction

Due to escalating pollution from automobiles and industrial facilities, environmental degradation is intensifying as urban areas expand in size and population. While air is vital for life, it seems that most individuals are either aware of the detrimental quality of the air or ignorant of the harm caused by pollution. Although several kinds of pollution exist globally, such as water, soil, thermal, and noise pollution, air pollution is especially perilous due to its potential to induce numerous health ailments. Currently, about 90% of the global population is subjected to air pollution, resulting in 7 million fatalities, as reported by the World Health Organisation. Pollution significantly exacerbates heart disease, stroke, asthma, allergies, and cancer, especially lung cancer. Present global air pollution issues, including ozone depletion, illustrate the effects of this pollution on both human health and the ecology. Contemporary technology and innovations in the current industrial revolution have rendered air pollution one of the most urgent challenges of our day[1].

The expanding economy, increased industrialisation, and the proliferation of industry zones have all rendered air quality and pollution monitoring a significant societal concern. The atmosphere is contaminated with various gaseous, liquid, and solid contaminants, adversely affecting humans, plants, and animals worldwide. Air pollution causes sinus infections, asthma, allergies, toxic dust exposure, neurological risks, and several other health complications. Researchers face considerable challenges in assessing and regulating air quality. The US Environmental Protection Agency states that indoor air pollution is 100 times more severe than outside air pollution. Despite the fact that the majority of individuals already allocate 80% to 90% of their time indoors, the interior air quality exerts a more immediate impact on health than outside air quality. Indoor air pollutants are a thousand times more likely to enter the lungs than those found outside, potentially leading to vertigo, chemical sensitivity, sick building syndrome, and several other health issues. Controlling indoor air quality is essential since safety measures can mitigate hazards[2]. To provide optimal air quality, indoor air quality monitoring must be both efficient and effective.

In the era of Industry 4.0, when issues such as industrial automation, intelligent pollution monitoring, and smart transportation are prioritised, technologies like the Internet of Things and artificial intelligence have assisted several industrial sectors. The emergence of real-time data analytics, automation, and continuous networking has fundamentally altered the operations of traditional industries. Moreover, several enterprises have lately announced significant real-time monitoring opportunities facilitated by the Internet of Things (IoT) and cloud computing. As a result, some researchers have considered integrating these technologies with air quality monitoring systems. The inclusion of IoT design will enable real-time air quality monitoring. Nonetheless, it constituted the primary focus of the investigation. The combination of these new technologies with wireless sensor networks can significantly enhance air quality by automatically sensing, collecting, evaluating, and visualising data[3].

The air quality is severely concerning due to the high levels of particle matter in the surrounding surroundings. Additional sources of air pollution that may affect human health are incomplete coal combustion, vehicular emissions, and Particulate Matter (PM2.5) emissions. Carbon monoxide (CO) is another gas mostly generated by the incomplete combustion of fuel. Vehicle engines account for eighty percent of carbon monoxide emissions in metropolitan areas. In urban environments, particularly in large metropolitan regions, smog is an unavoidable characteristic of city living. Smog mostly consists of ground-level ozone, generated by the chemical interaction among carbon monoxide, Volatile Organic Compounds (VOCs), and solar radiation. Consequently, carbon monoxide is the primary contaminant in urban haze. In metropolitan areas, Particulate Matter (PM) levels between 2.5 and 10 elevate the risk of cardiovascular illness, pulmonary damage, subclinical atherosclerosis, and heightened mortality from circulatory conditions. Carbon monoxide has become a significant air pollutant in some areas. The epidermis is susceptible to harm from extended exposure. Acute exposure to particulate matter has been linked to heart attacks, strokes, and hospitalisations. Numerous poisonings and fatalities have been associated with carbon monoxide in the environment globally. Dizziness, tremors, nausea, and emesis are among the documented symptoms of carbon monoxide poisoning; nonetheless, these symptoms are quite common. A positive Babinski sign, tremors, cervical rigidity, and fluctuations in awareness are all neurological indications. These diseases

provoke abnormal reflex responses in the neurological system. Children are especially susceptible to the hazards of carbon monoxide poisoning. In some circumstances, carbon monoxide concentrations may escalate to perilous levels, potentially resulting in unconsciousness or fatality. Over fifty percent of the issues related to illnesses are acknowledged by individuals in impoverished countries. Environmental issues in emerging nations such as India are escalating rapidly, leading to an increase in air quality index monitoring. The air quality in metropolitan areas and cities like as Ahmedabad, Kolkata, and Delhi is compromised due to fuel emissions from industrial activities and vehicles. Delhi residents may face heightened risk of severe respiratory diseases, cancer-related symptoms, and other respiratory emergencies owing to the city's air quality index exceeding 1,000. Individuals in rapidly urbanising, densely populated cities are at a heightened risk of mortality due to air pollution-related factors[4].

Most common sources of air pollution are shown in Figure 1.

Primary Pollutants: These pollutants are emitted directly from a source, which may be either natural (such as volcanic eruptions or wildfires) or anthropogenic (such as carbon monoxide from automobiles). Primary pollutants include ammonia (NH_3), sulphur dioxide, and particulate matter.

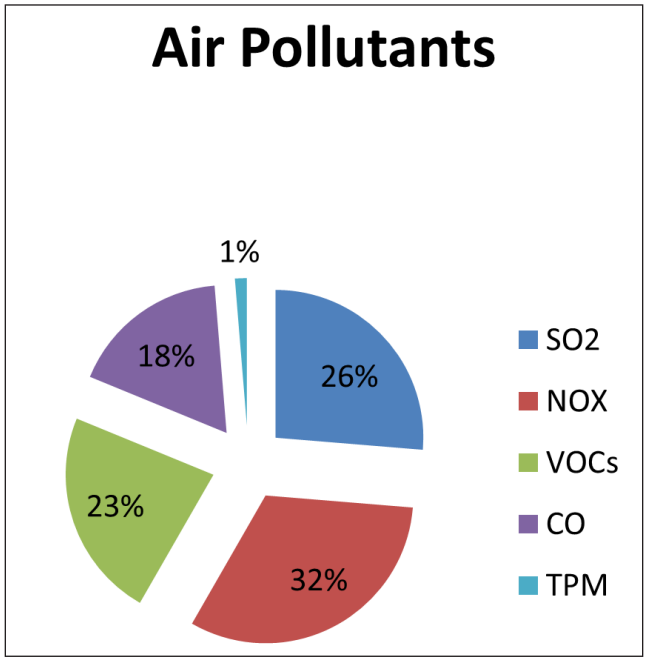


Figure 1. Most common sources of air pollution.

Secondary Pollutants: These pollutants are toxins that interact with each other or react with fundamental atmospheric toxins to generate them. Supplementary poisons, conversely, are not produced directly. At ground level, ozone is a notable contaminant. It is essential to recognise that several pollutants are categorised as both primary and secondary toxins. Secondary poisons are classified into three distinct categories: Certain solid particulate toxins are produced when main gaseous pollutants interact with photochemical smog. Consequently, smog may arise from severely contaminated air. For instance, elevated coal burning in a singular area produces smoke and sulphur dioxide, culminating in a conspicuous smog cloud. Ground-level ozone is formed by the reaction of Nitric Oxide (NOx) with Volatile Organic Compounds (VOCs). Ozone (O_3) is a crucial element of the troposphere. It is a crucial element of several stratospheric areas, including the ozone layer. Likewise, NOx and VOCs interact to produce peroxyacetyl nitrate ($C_2H_3NO_5$), a secondary contributor to air pollution.

The AQI is a widely utilised instrument for assessing air quality levels in municipal waste management. Numerous approaches have been developed. Numerous scientists and researchers involved in previous AQI decision-making committees have determined that there is no one optimal approach to address every conceivable scenario. The calculation of AQI employs several aggregation functions to determine the levels of different pollutants. A primary objective of the proposed Air Quality Index (AQI) applications is to assist individuals in identifying locations with poor air quality and to provide information on its severity, sources, and the presence of either singular or many pollutants.

An increasing number of "smart cities" are emerging globally due to the proliferation of the Internet of Things (IoT). The government employs the Internet of Things to deliver various public-sector services to people. A novel technology enabling connectivity with physical objects is the Internet of Things, or IoT. Via the internet, ordinary objects or devices outfitted with sensors, hardware, electronics, and software may interact autonomously and assign identities to one another, independent of human intervention. The Internet of Things (IoT) has the potential to induce significant transformations in the annals of computing, notwithstanding its status as a nascent technology. The Internet of Things (IoT) has boundless

opportunities for several sectors, although mankind now comprehends just a little portion of these applications. Significant operational transformations are occurring in urban areas at now. Utilising the Internet of Things (IoT), these pioneering smart city models aim to improve the quality of life for those in their area. Historically, environmental or governmental organisations have mostly established permanent, fixed monitoring stations to assess air pollution levels or other air quality indicators. The primary benefits of these pollution monitoring stations are their dependable readings and their capacity to evaluate the accessibility of various air contaminants. This type of monitoring enables the extrapolation of contamination levels over prolonged durations at specific sites or regions.

Challenges associated with utilising permanent sites for air quality or pollution monitoring encompass exorbitant setup and maintenance expenses, with inadequate spatial data resolution, perhaps resulting in erroneous assessments over the extensive research area. Therefore, it is essential to devise techniques and systems that allow electronics clustered in confined spaces to monitor and react to variations in ambient air pollution. Once the sensors are established and synchronised, it is quite viable to obtain data from the environment through the Internet of Things. Comprehensive research is now being conducted on air sensing sensor systems that utilise scattered infrared light for absorption, electrochemical reactions, resistance measurement, and photoionisation.

The significant advancement in technology over recent years is astounding. The AI paradigm, wherein computers autonomously make judgements instead of according to preordained rules, is increasingly infiltrating all facets of society. Machine learning has transformed the focus of all organisations within its domain, from startups to industry stalwarts. Air pollution, because to its immediate impact on humans, has garnered significant attention from several specialists and has emerged as a critical worldwide issue. To safeguard human health from the severe issues associated with air pollution, individuals are acquiring enhanced techniques for evaluating air quality data. A multitude of professionals is engaged in current air quality monitoring and forecasting to facilitate the implementation of government initiatives concerning climate change and air pollution, therefore delivering precise data to support critical decision-making. The method via which AI systems collect data from sensors and acquire autonomous behaviour is referred to as machine learning. One rationale for monitoring air quality indicators is the potential to utilise machine learning techniques. The following systems use the four devices examined herein: decision tree, random forest, naive Bayes, and k-nearest neighbour algorithms. Although other scholars have proposed alternate algorithms for incorporation into the recommended system, a comparison review of their performance against a single research using identical data and circumstances to assess all four algorithms remains absent.

Machine learning algorithms analyse customer data to establish correlations with behaviour over time. This enables the team to align product development with client preferences and enhance promotional initiatives[5].

This article describes an IoT and deep learning-based platform for air quality monitoring and forecasting. In this framework, the input data set is created by gathering air quality-related data from various sensors and IoT devices. The raw input data set is then reduced into features using the principal component analysis approach. This strategy is also useful for coping with noisy and missing data in the data collection. The classification model is created utilising convolution, neural network, recurrent neural network, and LSTM deep learning approaches. The classification model is trained and tested on a preprocessed data set. Performance of various models is evaluated using metrics such as R2, RMSE, and MSE. The air quality index is also anticipated in the conclusion. Section 2 presents work related to air quality monitoring. Section three presents proposed methodology, section 4 presents results and discussion and section 5 contains conclusion and future work.

2.0 Literature Review

The study used regression neural networks to improve air quality prediction by reducing modelling errors compared to typical full-range models. This was achieved by separating data into sub-ranges for distinct contaminants, which improved air pollution analysis and forecasting. Two air monitoring stations in Poland were used to forecast air quality in this study, and the authors compared the prediction errors of the sub-range models to those of a single full-range regression model. The precision of air quality models is enhanced by using sub-range models. When compared to traditional regression

techniques, artificial neural networks significantly cut down on prediction errors[6].

Using a vanilla Recurrent Neural Network (RNN), the authors of suggested predicting the air quality index for several tiers of Indian cities. Long Short-Term Memory (LSTM) and vanilla RNN with fractional derivatives provide similar results. The suggested approach improves upon previous methods of AQI prediction by making use of fractional derivatives[7]. When compared to LSTMs, RNNs using fractional gradient-based backpropagation perform as well, if not better. Utilising historical AQI data from five Indian cities, the study employs a vanilla Recurrent Neural Network (RNN) enhanced with fractional derivatives to forecast air quality. The AQI is more accurately predicted by RNNs that are based on fractional gradients. Findings for AQI and gas prediction that are on par with LSTM.

The correlation coefficient between PM2.5 and NO_2 in 2011 is 0.7036, indicating a favourable relationship. A logistic regression analysis of the current air pollution state yielded a result of 0.9430. Researching the characteristics of regression models as a research guide and updating data connected to air pollution research according to the dataset, Wang *et al*., utilised regression analysis on top of Beijing's air quality data from 2017 to 2021[8].

A second approach to air pollutant concentration prediction using the Kalman-attention-LSTM model has been suggested in this study[9]. For the O_3 pollutant, the second forecast using the Kalman-attention-LSTM model outperforms the competition. Results for the remaining five pollutants (CO, SO_2, NO_2, PM10, PM2.5, and NO_2) similarly show improved accuracy and reduced mistakes in their predictions. Its performance is superior to that of other conventional models. This research employs a two-step procedure to forecast the geographical and temporal concentration of pollutants with great accuracy. Predicting pollution levels is much easier with a spatiotemporal approach. Current categories used to measure air quality in big cities are acceptable for the suggested technique, according to the study's findings[10]. Commercial air monitoring stations in Peninsular Malaysia were evaluated for their capacity to enhance the prediction of the Air Pollution Index (API) using Principal Component Analysis (PCA) and Artificial Neural Network (ANN). This study's ANN prediction technique outperformed the Replaced Principal Component Score (RPCs) as an input factor for API prediction[11].

To forecast the concentration of PM10 air pollution in Tehran, it is recommended to use a neural network back propagation model and a genetic algorithm. The performance of a network is negatively affected by major changes that occur over an extended duration. Conversely, shorter intervals were shown to produce more consistent findings than longer ones[12-14]. In this work, a hybrid model was utilised to enhance ANN and forecast the average daily PM2.5 pollution concentration. The model included both air mass trajectory and wavelet transform. Daily weather predictions were generated by the model using pollution forecast variables and a multi-layer perceptron back propagation neural network (BPNN-MLP). It is suggested that the spatial model and wavelet transform be used for air quality prediction since the findings shown a considerable improvement in the forecast accuracy of PM2.5. Machine learning models that use meteorological datasets and performance metrics like MAE, MAPE, and RMSE to forecast PM2.5 pollution include XGBoost, AdaBoost, ridge, lasso, linear regression, random forest, KNN, and XG. Therefore, when it comes to forecasting PM2.5 pollution, XGBoost and AdaBoost, random forests, and KNN are generally more trustworthy models. When compared to the existing models, the proposed ones are superior at predicting PM2.5 pollution with much lower error rates[15].

The A dual model was employed to forecast air quality, utilising an ANN model in addition to an ARIMAX (Autoregressive Integrated Moving Average with Explanatory) prototype. The capacity to anticipate the targeted pollution limit exceedance can be enhanced with the use of such traffic management systems and techniques if the predicted concentration is moderately higher than the typical level. Simulating modelled concentrations of air pollutants from LES in urban streets is done using regression models. It is shown that, on new independent data, log-linear regression often gives the best and most robust results. When it comes to replicating concentrations of air pollutants, the log-linear regression is superior to the dummy model. To estimate pollutant concentrations, regression models might be used instead of LES simulations.

We calculated the study area's Air Pollution Index (API) using neural networks and classification prediction models for spatial and temporal models[16]. If you know the linear connection between the predictor and the

predictor, you may use the ANN model to make accurate and consistent forecasts about air pollution in industrial locations. Used a number of regression models available in the sklearn package to make AQI predictions[17]. Among these models were ridge regression, XGBoost, KNN, ElasticNet, decision trees, additional trees, neural network regression, LASSO, and neural network regression. One important metric for performance is Root-Mean-Squared Error (RMSE), and the extra trees model had the lowest RMSE and most accuracy of the bunch.

This study examined the use of artificial neural network modelling and multilinear regression to forecast ozone concentrations in connection to meteorological variables like temperature, relative humidity, wind speed and direction, and relative humidity. When compared to linear regression, ANN models provide more accurate predictions of ozone concentrations. The accuracy of the ANN model was significantly improved by optimising the hidden neurones.

Using Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest as examples, this article explores supervised methods for machine learning that may be used to make predictions about air quality based on meteorological factors. It offers a more effective method for predicting air quality. In order to make predictions, it makes use of LR, SVM, DT, or RF. Air quality predictions made by machine learning outperformed those by more conventional means. When it comes to predicting AQI, the Random Forest algorithm performs admirably[18].

To demonstrate how distributed deep learning models may considerably enhance training time and prediction accuracy, this study employs a convolutional bi-directional Long Short-Term Memory autoencoder model trained utilising a distributed architecture to forecast the concentration of air quality particles. In terms of accuracy, the suggested model surpasses eleven baseline models. Both training time and resource usage are decreased by the distributed strategy[19]

Research on statistical models and ML for filling in data gaps and eliminating background noise was conducted[20-23]. In a study that compared four methods for detecting air pollution, the researchers found that the NARX method performed the best. The other three methods were ANN, Geographically Weighted Regression (GWR), and Support Vector Regression (SVR). While the spline approach was used for daily and weekly predictions, the fourier series was used to construct monthly weather forecasts.

3.0 Methods

In this section, an Internet of Things (IoT) and deep learning-based framework for monitoring and predicting air quality is presented. Figure 2 illustrates this structure in its entirety. Within this framework, the input data set is created by collecting data relating to air quality via the utilisation of a variety of sensors and Internet of Things devices. After that, the method of principal component analysis is used in order to limit the number of features that are present in the raw input data set. Additionally, this method is helpful in dealing with the noisy data and missing data that are included in the data collection. LSTM deep learning, convolution, neural networks, and recurrent neural networks are the approaches that are used in the construction of the classification model. Training and testing of the classification model are carried out with the help of the data set that has been preprocessed. R2, RMSE, and MSE are some of the characteristics that are used to evaluate the performance of different models. In addition, the air quality index is forecasted at the conclusion.

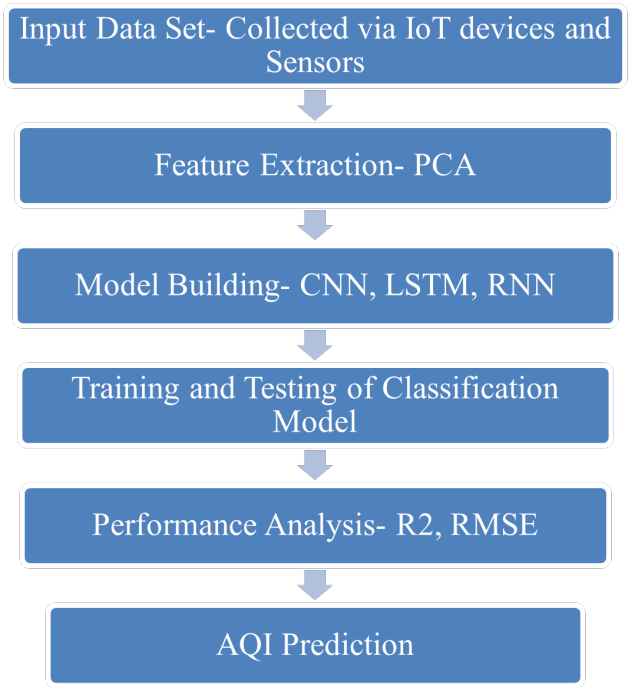


Figure 2. Deep learning and IoT based air quality monitoring and prediction system.

Principal component analysis technique is used to reduce features from the raw input data set. This technique also helps in dealing with the noisy and missing data in the data set. A specific category of RNNs, termed Long Short-Term Memory, have the ability to learn sequences with extended temporal dependencies. They are particularly designed to avert issues associated with long-term reliance. Memory encoding is a technique that entails progressively committing extensive sequences to memory. In forecasting procedures using sequential or time-series data, a recurrent neural network, a type of artificial neural network, is employed. Typical applications of these deep learning layers encompass automatic image labelling, neural machine translation, and natural language processing, all of which are vulnerable to sequential or temporal inaccuracies. These layers employ contemporary voice-assisted technologies such as Siri, Google Assistant, and Alexa to enhance user experience. The primary distinction between RNNs and CNNs is that RNNs possess a memory that modifies current inputs and outputs by incorporating all previous data from earlier inputs. The training technique for this network remains unchanged. In contrast to traditional neural networks, which regard inputs and outputs as entirely distinct, RNNs produce outputs that are contingent upon both the inputs and their configurations. Another distinguishing characteristic of RNNs is that their parameters are uniformly shared across all levels of the network. Recurrent neural networks employ same weight parameters over all layers, whereas feedforward networks assign distinct weights to each node. In recent years, RNNs have achieved significant success across several applications, including language modelling, translation, image annotation, audio recognition, and sequence generation. Initially, we will analyse one individual RNN node. We can manually modify the activation function if it accommodates the input from the preceding phase and the current state Xt, integrating with Tanh as the initial function.

Periodic verification of the most recent data may be necessary to accomplish the current task. In any case, we do not frequently encounter this issue. Extended commands or terms may lead to information loss in conventional RNN networks. This is attributable to the methodological approach, which just accounts for the most recent data available at the node and cannot maintain extended setups. The term "vanishing gradients" refers to this problem. During training, a Recurrent NeuralNetwork (RNN) adjusts its weights according to gradients calculated at each time step or iteration. The relative slope will be calculated as minimal if the influence of the prior order on the plane is negligible. Secondly, while managing extensive orders, it is apparent that a slight incline of the preceding layer yields minimal contextual weight[24]. This network's independence from earlier input leads to problems with short-term memory. Recurrent Neural Networks (RNNs) constitute a fundamental network architecture that underpins several deep learning frameworks. Recurrent networks may include both feedback mechanisms to prior layers and just feed-forward connections. This delineates the distinctions between recurrent networks and conventional multilayer networks. This feedback enables the RNN to monitor previous inputs and model problems as they develop. We will thereafter analyse a conventional topology referred to as LSTM, which is one of the several architectural styles observed in RNNs. In networks, feedback serves as a vital difference. This may stem from output layers, hidden layers, or a mix thereof. The architecture of a recurrent neural network is seen in Figure 3.

A Convolutional Neural Network (CNN) comprising many layers is stimulated in an animal's visual cortex[25]. This method is most appropriate for applications involving image dissemination. Architects usually utilised handwritten credits at that period. Similar to deep networks, subsequent coatings assimilate the input's inherent features and utilise them to integrate previously identified components (such as edges) from preceding layers. A convolutional layer receives a receptive field

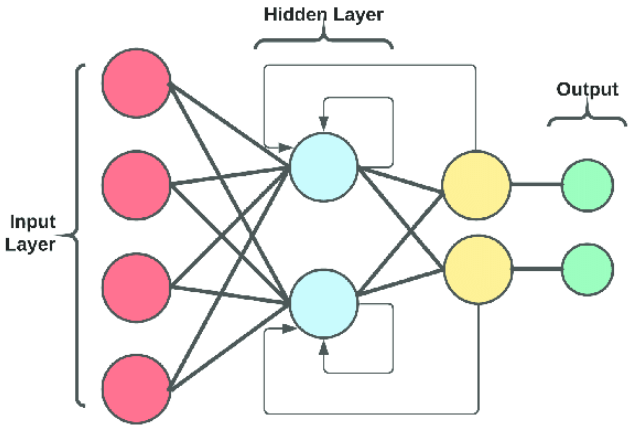


Figure 3. architecture of a recurrent neural network.

to extract fundamental information from the input image. This constitutes one of the layers in the LeNet-CNN architecture employed for feature extraction and reorganisation. Subsequently, the pooling phase ensues. Down sampling is an effective strategy that reduces the dimensionality of collected primitives while preserving the most relevant data, typically using max pooling. Subsequent to further complexities and integration processes, fully linked multi-layer perceptrons are produced. A network of nodes—one for each detected number—that identifies image properties represents the last layer of the system's segmentation. The network can acquire new talents through backpropagation. Figure 4 shows a convolutional neural network model.

Deep Learning (DL) neural networks have discovered novel applications with the incorporation of convolutional, pooling, and specialised sorting layers, among others. Convolutional Neural Networks (CNNs) possess several applications outside image processing, including natural language processing and video recognition. Applications encompass natural language processing, image recognition, and video analysis. In recent years, LSTM has gained prominence as an RNN architecture for many applications. LSTMs are found in everyday items; for instance, LSTM technology enables

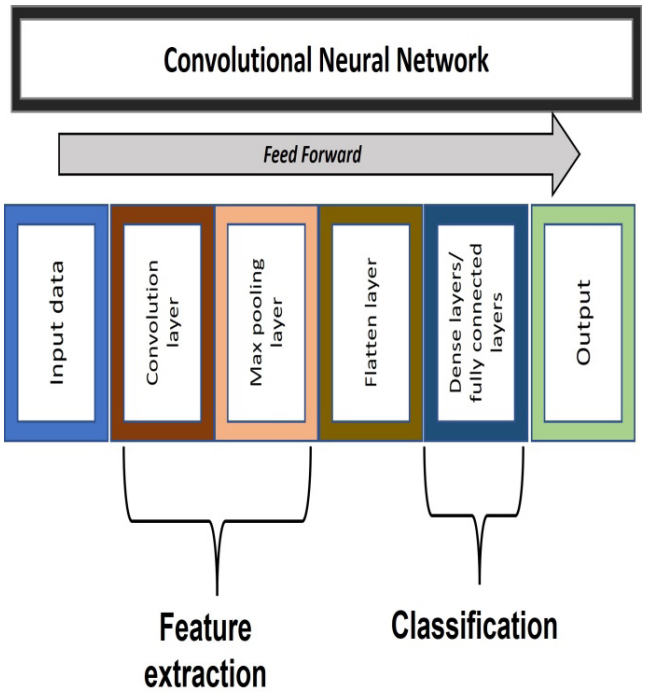


Figure 4. Convolutional neural networks model.

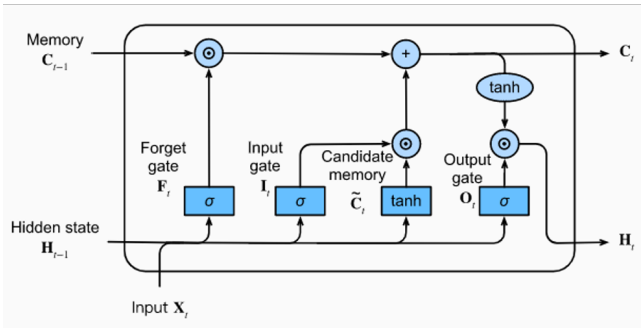


Figure 5. Basic LSTM model.

IBM Watson to do groundbreaking conversational voice recognition. Rather of employing the conventional neuron-based neural network architecture, LSTM utilises the principle of retention cells. A memory cell can retain values for either a short or extended duration, dependent upon the input. Consequently, the cell may consider pertinent factors in conjunction with the final calculation.

Three gates govern the ingress and egress of the LSTM holding cell, which is tasked with processing evidence[26]. New reports are regulated to enter memory by access gates. For a cell to acquire new information, it must first discard old information, a process governed by ignore gates. The output gate also regulates the use of the cell's stored data at its output. The gate weight is further recorded in the cell. These weights are refined by a training procedure, often Backpropagation by Time (BPTT), due to inaccuracies in the system output. Figure 5 shows a simple LSTM network.

4.0 Result Analysis

The **Air Quality Dataset** by **fedesoriano** is a comprehensive collection of air pollution measurements, specifically designed for machine learning and time series analysis[27]. The dataset comprises hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an air quality chemical multisensor device. These sensors measure concentrations of various air pollutants, including:

- **CO (Carbon Monoxide)**
- **NMHC (Non-Methane Hydrocarbons)**
- **C6H6 (Benzene)**
- **NOx (Nitrogen Oxides)**
- **NO2 (Nitrogen Dioxide)**

In addition to pollutant concentrations, the dataset includes meteorological data such as temperature, relative humidity, and absolute humidity.

The original input data set is reduced in features using the principal component analysis technique. This method is also beneficial in addressing the chaotic and absent data in the data set. The convolutional neural network, recurrent neural network, and LSTM deep learning techniques are employed to construct the classification model. The preprocessed data set is employed to train and evaluate the classification model. The performance of a variety of models is evaluated based on parameters such as R2, RMSE, and MSE. The air quality index is also predicted at the conclusion.

For performance comparison, R square, Mean Squared Error and Root mean square error parameters are used.

- R Square (R2): R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  Where
  y_i: actual value
  $\hat{y}_i$: predicted value
  $\bar{y}$: mean of actual values
  n: number of data points
- Mean Squared Error (MSE): Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals. Mathematical representation for MSE is given below.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Root Mean Squared Error (RMSE): The Root Mean Square Error (RMSE) is defined as the square root of the mean square error. One of the key benefits of RMSE is its differentiability, making it suitable for use as a loss function. Additionally, RMSE is always a non-negative value. Its implementation closely resembles that of the Mean Square Error (MSE).

The performance metrics, including R^2, MSE, and RMSE, for different machine learning models, including RNN, LSTM, and CNN, are compared in Table 1. and Figure 6, 7 and 8. CNN consistently outperforms the other models, achieving the lowest MSE and RMSE values and the highest R^2 scores. This implies that CNN is the most effective model for predicting air pollution levels.

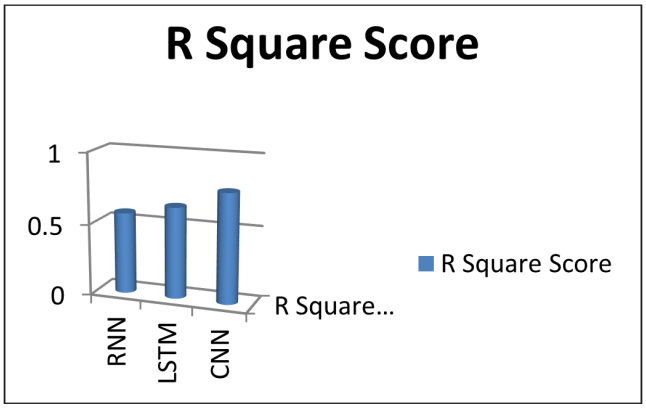


Figure 6. Performance comparison of R^2 of CNN, LSTM and RNN for air quality prediction.

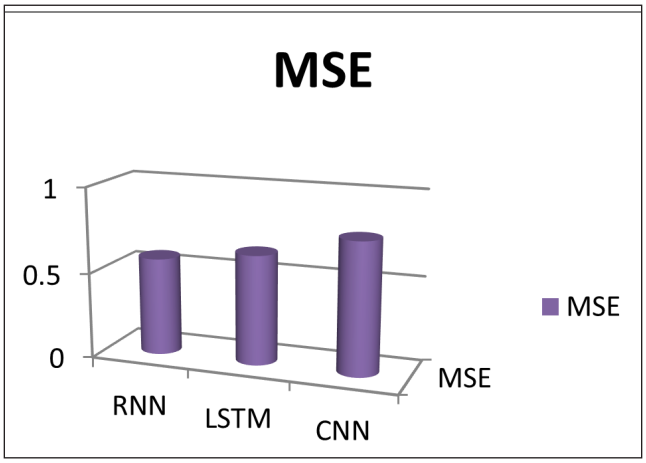


Figure 7. Performance comparison of MSE of CNN, LSTM and RNN for air quality prediction.

Table 1 compares the performance of three models-RNN, LSTM, and CNN-for air quality prediction in terms of three evaluation metrics R 2 score, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). From the models, the best overall performance by the CNN with the R^2 score of 0.77, which means higher correlation between predicted target and actual target. It also logs the minimum MSE (2587.14) and RMSE (50.86) indicating less deviation and error in the prediction. The intermediate LSTM model performs fair

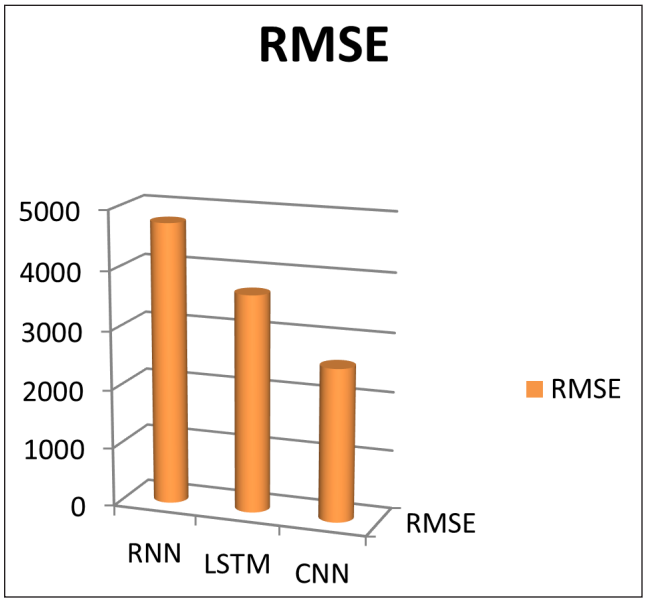


Figure 8. Performance comparison of MSE of CNN, LSTM and RNN for air quality prediction

Table 1. R^2, MSE, and RMSE of CNN, LSTM and RNN for air quality prediction

Model	R Squared Score	MSE	RMSE
RNN	0.57	4756.21	68.96
LSTM	0.64	3675.34	60.62
CNN	0.77	2587.14	50.86

with an R^2 of 0.64, MSE of 3675.34, and RMSE of 60.62 being better than the RNN and worse than the CNN. The worst performance is observed in the RNN (R^2 = 0.57) which also yields the highest errors (MSE = 4756.21, RMSE = 68.96) among these three models indicating the possible lack of capturing the temporal dynamics of air quality data of RNN relative to the other architectures (LSTM and CNN).

5.0 Conclusion

This paper talks about an architecture for tracking and predicting air quality that is built on the Internet of Things and deep learning. In this system, the input data set is made by collecting data from different monitors and Internet of Things devices about the quality of the air. The principal component analysis method is then used to break down the raw data set into features. This method can also be used to deal with noise or missing data during the data collection. Deep learning techniques like convolution, neural networks, recurrent neural networks, and LSTM are used to make the classification model. A set of data that has already been handled is used to train and test the classification model. Metrics like R2, RMSE, and MSE are used to judge the performance of different models. In the end, the air quality score is also expected. CNN always does better than the other models; it has the lowest MSE and RMSE numbers and the best R^2 scores. This means that CNN is the best tool for figuring out how much pollution is in the air.

6.0 References

1. Gowda V, Varadaraju VD, Bharath N. Review of air quality monitoring: Case study of Peenya industrial area. Int J Eng Technol Manag Sci. 2023; 7(1):186-93. https://doi.org/10.46647/ijetms.2023.v07i01.026
2. Keswani A, Akselrod H, Anenberg SC. Health and clinical impacts of air pollution and linkages with climate change. Hardin CC, editor. NEJM Evidence. 2022; 1(7). https://doi.org/10.1056/EVIDra2200068 PMid:38319260.
3. Zhu D, Cai C, Yang T, Zhou X. A Machine learning approach for air quality prediction: Model regularization and optimization. Big Data and Cognitive Computing. 2018; 2(1):5. https://doi.org/10.3390/bdcc2010005
4. Kabir S, Islam RU, Hossain MS, Andersson K. An integrated approach of belief rule base and convolutional neural network to monitor air quality in Shanghai. Expert Systems with Applications. 2022; 206:117905. https://doi.org/10.1016/j.eswa.2022.117905
5. Méndez M, Merayo MG, Núñez M. Machine learning algorithms to forecast air quality: A survey. Artificial intelligence Review. 2023. https://doi.org/10.1007/s10462-023-10424-4 PMid:36820441 PMCid:PMC9933038.
6. Hoffman S, Filak M, Jasiński R. Air quality modeling with the use of regression neural networks. Int J Environ Res Public Health. 2022; 19(24):16494. https://doi.org/10.3390/ijerph192416494 PMid:36554373 PMCid:PMC9779138.
7. Arora S, Singh NSS, Singh D, Shrivastava RR, Mathur T, Tiwari K, *et al.*, Air quality prediction using the fractional gradient-based recurrent neural network. Computational Intelligence and Neuroscience. 2022; 2022:1-14. https://doi.org/10.1155/2022/9755422 PMid:36531923 PMCid:PMC9757944.

8. Karthikeyan S, Moses ML, Perarasi T, Raju N. Machine learning based air pollution monitoring and predicting cause for pollution. IOP Conference series earth and environmental science. 2022; 1125(1):012020. https://doi.org/10.1088/1755-1315/1125/1/012020
9. Zhou H, Wang T, Zhao H, Wang Z. Updated prediction of air quality based on Kalman-Attention-LSTM network. Sustainability. 2022; 15(1):356. https://doi.org/10.3390/su15010356
10. Russo A, Lind PG, Raischel F, Trigo R, Mendes M. Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales. Atmospheric Pollution Research. 2015; 6(3):540-9. https://doi.org/10.5094/apr.2015.060
11. Shen C. A comparison of principal components using TPCA and nonstationary principal component analysis on daily air-pollutant concentration series. Physica a Statistical Mechanics and Its Applications. 2016; 467:453-64. https://doi.org/10.1016/j.physa.2016.09.014
12. Mazuruse G, Nyagadza B, Tumbure A, Makoni T, Muvuti A. Algorithmic optimization for efficient air quality prediction models through machine learning: A case study of Shillong city in India. Next Research. 2025; 100346. https://doi.org/10.1016/j.nexres.2025.100346
13. Alyousifi Y, Othman M, Almohammedi AA. A novel stochastic fuzzy time series forecasting model based on a new partition method. IEEE Access. 2021; 9:80236-52. https://doi.org/10.1109/access.2021.3084048
14. Zhang S, Li X, Li Y, Mei J. Prediction of Urban PM2.5 Concentration based on wavelet neural network. 2018 Chinese Control and Decision Conference; 2018. p. 5514-9. https://doi.org/10.1109/CCDC.2018.8408092 PMid:29615796 PMCid:PMC5882958.
15. Catalano M, Galatioto F, Bell M, Namdeo A, Bergantino AS. Improving the prediction of air pollution peak episodes generated by urban transport networks. Environmental Science and Policy. 2016; 60:69-83. https://doi.org/10.1016/j.envsci.2016.03.008
16. Rahman NHA, Lee MH, Latif MT, Suhartono S. Forecasting of air pollution index with artificial neural network. Journal Teknologi. 2013; 63(2). https://doi.org/10.11113/jt.v63.1913
17. Mahanta S, Ramakrishnudu T, Jha RR, Tailor N. Urban air quality prediction using regression analysis. TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON); 2019. p. 1118-23. https://doi.org/10.1109/tencon.2019.8929517
18. Simu S, Turkar V, Martires R, Asolkar V, Monteiro S, Fernandes V, *et al*., Air pollution prediction using machine learning. 2021 IEEE Bombay Section Signature Conference (IBSSC); 2020. p. 231-6. https://doi.org/10.1109/ibssc51096.2020.9332184 PMid:32190864.
19. Mengara AGM, Kim Y, Yoo Y, Ahn J. Distributed deep features extraction model for air quality forecasting. Sustainability. 2020; 12(19):8014. https://doi.org/10.3390/su12198014
20. Dun M, Xu Z, Chen Y, Wu L. Short-term air quality prediction based on fractional grey linear regression and support vector machine. Mathematical Problems in Engineering. 2020; 2020:1-13. https://doi.org/10.1155/2020/8914501
21. Alfasanah Z, Niam MZH, Wardiani S, Ahsan M, Lee MH. Monitoring air quality index with EWMA and individual charts using XGBoost and SVR Residuals. MethodsX. 2024; 14:103107. https://doi.org/10.1016/j.mex.2024.103107 PMid:39802429 PMCid:PMC11721863.
22. Jahanshahi A, Jahanianfard D, Mostafaie A, Kamali M. An Auto Regressive Integrated Moving Average (ARIMA) model for prediction of energy consumption by household sector in Euro area. AIMS Energy. 2019; 7(2):151-64. https://doi.org/10.3934/energy.2019.2.151
23. Malik S, Ullah I, Kim DH, Lee K. Heuristic and statistical prediction algorithms survey for smart environments. J Inf Process Syst. 2020; 16(5):1196-1213. https://doi.org/10.3745/JIPS.04.0191
24. Pande CB, Kushwaha NL, Alawi OA, Sammen SS, Sidek LM, Yaseen ZM, *et al*. Daily scale air quality index forecasting using bidirectional recurrent neural networks: Case study of Delhi, India. Environmental Pollution. 2024; 351:124040. https://doi.org/10.1016/j.envpol.2024.124040 PMid:38685551
25. Chauhan R, Kaur H, Alankar B. Air quality forecast using convolutional neural network for sustainable development in urban environments. Sustainable Cities and Society. 2021; 75:103239. https://doi.org/10.1016/j.scs.2021.103239
26. Naresh G, Indira B. Air pollution prediction using multivariate LSTM deep learning model. Int J Intell Syst Appl Eng. 2023; 12(8s):211-20.
27. Air quality dataset [Internet]. Kaggle; 2020. Available from: https://www.kaggle.com/datasets/fedesoriano/air-quality-data-set