2. LITERATURE REVIEW

In terms of air quality monitoring and prediction, many attempts are being made in the last few years to avoid health hazards and provide ease of living in urban areas. So, the following sub-sections discuss the work done so far in air quality monitoring and air quality prediction. The sub section discusses the major challenges in the domain at the end.

2.1. REVIEW OF EXISTING AIR QUALITY MONITORING SYSTEMS

Various researchers are aggressively engaged in the air quality monitoring domain because of the concern to the critical issue. Li and his team [24] have examined an approach of analyzing monitoring data generated by fixed monitoring stations network every hour laid over Taiwan, the network run by environment protection administration of Republic of China. The Taiwan Air Quality Monitoring Network(TAQMN) comprises 71 air quality monitoring stations of Taiwan that automatically gather air quality parameters every 1 hour. The positions of the stations are distributed over eight different regions of the island. The stations collect primary air pollutants, including carbon monoxide and particulate matter. Data acquisition is implemented using web spiders using Java to retrieve particulate matter 10 data from the EPA website automatically. The data source was comprised of 71-time series, which were stored in a local repository. They carried out the cluster analysis of air pollution data using multi-scale wavelet transformation and a self-organizing map neural network, a visual tool for rendering the data. The motive of such a study is to break down a large area into various small homogeneous regions. The proposed approach does not present any air quality monitoring framework used for parameter collection and monitoring.

Reisinger and his team have demonstrated with differential optical absorption spectroscopy (DOAS)-based device for measuring air pollutants [25]. DOAS permits the determination of the quantity of gas concentration in the atmosphere by assessing the characteristics of the absorption of the particles of the trace gases over an identified length absorption path. The sensors utilized in this approach have a very long optical route, around 100 m to 20 km, at an interval of 20 minutes.

Richards and his team developed a high-throughput network of sensors for real-time air quality monitoring in an urban area using GUSTO sensing technology [26]. Their proposed monitoring system measured several air pollutants precisely in part per billion Levels at small intervals of 2s. Their work was an enhancement of the DOAS based project of the Reisinger et al. team. GUSTO technology and DOAS technology based projects utilized the optical sensors based wired network. The complete setup of the proposed systems was based on wired infrastructure due to the unattainability of wireless sensor networks and the IoT technologies in that era.

In developing countries like India, Central Pollution Control Board is governing the nationwide program, National Air Quality Monitoring Programme(NAMP) [27] for air quality monitoring purposes. This program has a monitoring network comprising 779 fixed monitoring stations across different cities of India. The government of India is boosting and backing projects for air pollution, seeing at the worrying conditions in cities such as Delhi, Bangalore, and Ahmedabad. Though, the reachability of the program has not covered many of the urban and suburban areas of the country.

Due to the developments in wireless sensor networks, sensing machinery, and the IoT, fellow researchers have attempted various approaches to develop air quality monitoring systems. Wireless sensor networks improved the activity of monitoring phenomena heavily. Al-Haija et al. [28] used Arduino microcontroller and generic sensor MQ-2. MQ-2 is the widely available generic sensor, sensitive to multiple gases working based on the semiconductor-based sensing principle. They utilize the unit to measure carbon monoxide and Liquid petroleum gas(LPG) inside the king Faisal university building.

Kularatna and Sudantha [29] implemented an air pollution monitoring system using semiconductor sensors. Sensors were connected to smart transducer interface module-STIM, and network capable application processor(NCAP) was developed on a personal computer. Both types of research [28, 29] have presented the system, implemented with the microcontroller, and utilized general-purpose gas sensors. The implemented system had not provided any feature for reporting the monitoring results in near real-time to the remote repository or server.

Francis Tsaw et al. [30] developed a volatile organic toxicants sensor device to detect volatile organic compounds (VOCs). The sensor was designed using polymer modified tuning for the sensing instrument. They developed a circuit to evaluate the resonant frequencies of fork sensors. The research includes a novel design for making a portable sensor unit that can send the data to mobile over Bluetooth interface.

Authors [31] developed a wireless sensor component to measure indoor air pollutants in home networking. They used a humidity sensor, temperature sensor, and dust sensor in their experiments. The module is easy to install in the home. The monitoring parameters were transferred to a personal computer using an RF transmitter. Wireless sensor networks along with gas sensors had been used for many air quality monitoring systems. Most researches explored the implementation of middleware or networking layer, but Authors [32] focused on designing and developing sensor boards suited for air quality monitoring systems.

Authors [33] undertook the study to measure particulate matter 2.5 inside and outside the truck cab. The study took samples at the spot from the truck cab, which is manually analysed in labs to derive actual values of the pollutants.

OpenSense [34] is a project conducted jointly by EPFL, ETH, Switzerland, and IBM. They explored the feasibility of deploying sensors on the roof-top of automobiles to benefit from available public transport vehicles in creating a broad network of mobile air quality data-gathering sites. The project was considering many alternatives and challenges in forming a community sensing network. However, the exact implementation status and details are not available in depth.

The inAir [35] was a project carried out in the USA to measure and visualize indoor air quality. In this project, they developed a portable device with a visualization monitor placed at home. The device can render the current air quality parameters like particulate matter, carbon monoxide, and Volatile Organic Compounds. The device can work as an indoor portable monitor for those air pollutants. However, the pollutant data are not stored or logged in a repository for analysis purposes.

Postolache et al. [36] demonstrated a network based on Wi-Fi connectivity for indoor air quality monitoring purposes. They used Wi-Fi repeaters for relaying the data to a distant place. Their monitoring system included the processing of sensed pollutant data based on neural networks to recognize certain air pollution events and abnormal operation or malfunctioning of sensors. The individual sensing node transmitted the sensed data using TCP/IP as a communication protocol to the central processing unit. The control unit (laptop) performed the logging and storing of the data stream, processing the received data stream, and publishing it to the web interface through a LabVIEW web server. However, their system was used in an enclosed environment of the laboratory with a single deployment.

Teco EnvBoard [37] presented a generic platform for air pollutant assessment. Teco EnvBoard was developed to work as a sensing platform in air quality monitoring-related researches. The board was carrying some gas sensors along with temperature and humidity sensors. The EnvBoard was supported with a battery for energy consumption and a USB interface. The USB interface could be used to recharge the batteries. So the device was able to work as a standalone site of pollutant collection. The board could be connected to a host device like an Android phone using Bluetooth. The data collection could be triggered via command from the phone also. The data is sent via Bluetooth for storing at some host device and supported with storage on a microSD card, which can be accessed to fetch data later.

Young-Jin Jung et al. [38] designed and developed an air quality monitoring system. They installed ten routers and sensors for temperature, humidity, dust, carbon monoxide, wind direction, wind speed, and altitude. The sensed data were transmitted through a wired OGC network.

Gianfranco Manes et al. [39] developed system for volatile Organic Compound monitoring in the petrochemical plant. The sensing nodes are embedded with the meteorological sensors and volatile organic compound sensors. They used GSM cellular network for relaying data.

SocialCops is one of the data intelligence companies situated in India. They implemented one project for monitor the air quality in the city, Delhi. The project was composed of sensors along with the GPRS and GPS shield. They transmitted the sensed data using the GPRS network [40].

Richard Honicky et al. [41] developed a data acquisition module system with a pollution sensor for carbon Monoxide and ozone measurement. The study was conducted in Ghana for two weeks utilizing six sites. The commodity sensors used in experiments bridged the gap between sampling and analysis by providing real-time air pollutant (carbon monoxide and ozone) data. The research was carried out cooperatively by Intel and UC Berkeley. The sensors for collecting air pollutants were attached to cell phones carrying GPS cellular networks. The study presented several challenges considering the quality of sensed data from connected sensing units, like location coverage and calibration of sensors for accuracy.

Siva V. Girish [42] et al. demonstrated indoor air quality monitoring system. The monitoring system was consisted of a unit for sensing that empowered remote monitoring of air pollutant carbon dioxide, temperature, and relative humidity. They utilized simple mail transfer protocol(SMTP) for remote monitoring. They used a signal processor (CC3100) that was interfaced on the top of an MSP430F5529 - a low-power controller. Their proposed monitoring system also consisted of a notification service with SMS- a short messaging service to the mobile using SIM900 GSM shield/module.

Devarakonda et al. [43] proposed two sensing models in their research work. The first sensing model was developed for deploying on existing public vehicles like buses, cabs, and trams that travel on fixed routes. They designed a sensing box that contained sensors interfaced with a microcontroller board and a GPS module for this model. They used two sensors for measuring carbon monoxide and particulate matter. The board could be powered with the public transport vehicle power supply. The second sensing model was for deployment in the air quality-aware drivers' car. This model was connected to the smartphone of the driver via Bluetooth. They used Arduino Mega 128 microcontroller board. They used SIM52183G/GPRS shield attached to the Arduino board, and this shield was supported with AT command for data conversation over HyperText Transfer Protocol.

A.R. Al-Ali et al. [44] developed a wireless distributed air quality monitoring system. They designed Mobile – DAQ, a mobile Data Acquiring Unit that consists of a microcontroller interfaced with carbon monoxide and nitrogen dioxide sensors. The general packet radio service (GPRS) modem was used as a communication device to transmit air pollutant data.

Liang Zhao et al. [45] proposed a system for indoor air quality detection that consisted of multiple interfaces for communication. They used GPRS, Modbus, and WiFi networks that can be utilized for wired communications, short-range wireless communication, and remote transmission over GPRS to the cloud node or web interface.

CommonSense [46] is a joint project carried out by the University of California and Intel research. They developed hardware that can be embedded with gas sensors along with temperature and humidity sensors. The Data collected by sensors were gathered and processed locally by the core module. The hardware was facilitated with an 802.15.4 interface to connect to the low-power network. The underlying BlueSmiRF Bluetooth module was able to send data to the mobile phone for visualization purposes. The sensor pollutants are sent to a database server for logging and future analysis using Cinterion GPRS radio. With the advancement of technology, GPRS and GSM cellular technology is considered outdated technology.

Authors [47] developed a software platform, and framework for mobile phoned that can be utilized by mobile phone owners to take part in sensor data collecting drives. The software was written in Symbian C++ targeting the operating system Symbian, popular during that duration, and the language followed ANSI C++ standards.

Kim et al. [48] analysed and examined the challenges, platform and infrastructure, processing of sensing data, and issues of designing and implementing a real-time indoor air quality monitoring system. The sensing node comprised an MPU MSP430 microcontroller and the RF chip CC2500, along with few gas sensors. In their network implementation, the sensor node that was part of the sensor network cloud sent the fetched parameters to the sink point (node) using the WSN. The data logger could keep the track of the parameters coming to the sink node, and the sink point could send the data to the middleware. The air quality monitoring data were logged in the middleware and provided to the remote cloud-based platform. The sink node for the sensor node cloud contained the communication unit. The communication unit was containing a microprocessor and low-power RF chip CC2500. The sink node was sending data to the middleware via Serial Communication Interface (SCI). They also proposed a sampling algorithm for energy-saving purposes. The proposed approach is not embedded because the sensing module was connected directly to a computer over the wired interface.

Advancement in sensing technology had empowered the development of tiny low-cost sensors and modules to measure air pollution. Deploying many sensors over a particular region can supplement a costly conventional network of air quality monitoring stations. Community Air Sensor Network (CAIRSENSE) [49] project was carried out in the United States to evaluate emerging low-cost sensors performance in suburban areas. The project deployed low-cost air pollution sensor nodes at air monitoring sites over a surrounding 2-kilometer area in the southeast United States. The nodes were using Arduino Mega 2560 microprocessor with a battery-operated power supply. The data were transmitted to the repository using 900 MHz unidirectional wireless communication via ZigBee network protocol. Castell et al. [50] a developed system was comprising of twenty-four nodes utilizing low-cost sensors. The outcomes were matched with the reference observations in a laboratory environment from CEN, a European standardization organization in Norway.

Authors [51] created a wireless mesh network for communication purposes. The sensor nodes contained sensors with the microprocessor. Once powered up, the mesh nodes can search for other nodes in a particular range for building a mesh network. They used HotPort3200 wireless node for wireless transmission in the 2.4 GHz or 5 GHz frequency range.

Authors [52] developed a system to control indoor air quality using an Arduino Uno controller and sensors. The sensing node powered up with a 5v battery and used the ZigBee s2 module for data transfer to the computer. Moheddine Benammar et al. [53] came up with a modular IoT-based platform for monitoring indoor air quality. The sensor node comprises sensors interfaced to the Gas Pro sensor board and a processor from waspmot. The sensor nodes sent their data to the gateway installed in the same premise using the XBee PRO S2 radio module. The gateway was set up using Raspberry Pi2 model B. The data were posted to the web interface using HTTP from the gateway. Lambebo et al. [54], in their work developed a near real-time temperature and air pollutant monitoring system using a wireless sensor network. The gateway node was designed using low-power hardware with Arduino Uno controller and radio transmission module XBee IEEE 802.15.4 for ZigBee messaging. The task of the gateway was to relay the collected data to the server and gathering data recorded at all the individual sensor nodes. The server end was facilitated with data logging and a web interface for rendering purposes.

Authors [55] used temperature and humidity sensors along with electrolyte-type sensors for carbon dioxide measurement during their research. They used the Zigbee network for transmitting and displaying the data over the computer. Wireless application coding for the functioning of the ZigBee end modules and ZigBee coordinating module is written and developed in nesC.

Tsang et al. [56] developed a ZigBee mesh network for indoor air quality monitoring systems. They implemented an energy reduction scheme with ZigBee wireless sensor network with lower latency and higher throughput. However, the researchers have simulated the WSN for the purpose of indoor air quality monitoring without implementing and operating embedded devices. They designed the scenario of simulation using network-dedicated software OPNET.

Abraham and his team in 2014 have proposed the indoor air quality monitoring system to measure CO and CO2 using the ZigBee module [57]. The experiments were not conducted and tested outdoor environment, in which case ZigBee network setup is required. Kumar et al., with their team, have reported the monitoring system of air pollution parameters [58]. Ferdoush and his team developed ZigBee based prototype with temperature reporting using a raspberry pie board in their experiments [59]. Bacco et al. deployed a fixed and portable sensing node to monitor the carbon monoxide, particulate matter, and unburned hydrocarbons in Italy. They used ZigBee and GPRS base communication during their experiments [60]. In all these proposed methods, the ZigBee-based modules are utilized to complete the communication protocol. The ZigBee modules have a distance limitation while they are relaying or forwarding the packets. The communication based on the ZigBee transition needs more ZigBee devices/nodes to deliver the messages over a longer distance. The ZigBee-based solution is also one of the costly alternatives.

Sharma and his team have implemented the gas sensor-based embedded system. The CNT gas sensor was specially designed and developed. They used the MSP430 controller for interfacing of gas sensors to measure ammonia. However, the researchers did not implement any method for data transmission [61].

Tiwari et al. have demonstrated a system for measuring methane, temperature, and humidity. The Raspberry Pi development board was used to transmit the sensed data to the local webserver [62]. This approach was implemented at research lab of Bits Pilani in Rajasthan in India.

Marques et al. have developed an air quality monitoring system for ambient assisted living. The work was designed to monitor LPG propane with the use of MQ6 general-purpose sensor. The sensing node was getting connected to the gateway node using a radio connection. The gateway node was connected to the end node- laptop via USB and communicated through the serial port. The end node was running an application for data management purposes which was implemented in Java language. The application could post the notification to social media[63].

Authors [64], in their study, proposed IoT based solution for real-time indoor air quality monitoring named e-nose. The proposed system was containing gas sensors along with temperature and humidity sensors. Their proposed system e-nose was produced with open source and low-cost sensors. The sensors were interfaced with the 32-bit ESP32 Wi-Fi controller. The controller was sending data to the IoT-based Blynk platform for recording at the server. It was shown that poor indoor air quality was related to and affected the number of residents to the home and typical emissions because of cooking and cleaning. The deployment was done at a single site and is suitable for indoor monitoring purposes.

Hong-di He et al., in their proposed work, have shown a method to forecast the PM2.5 by using the available and existing data of the nearby air quality monitoring stations [65]. Their proposed work did not involve any direct measurement methods for monitoring the air quality parameters.

Zakariya et al. [66] developed and demonstrated the air quality monitoring system. The proposed method was to integrate a low-cost MQ-135 air quality sensor sensitive to multiple gases, temperature sensors, and relative humidity sensors. The sensors interfaced to the single-board mini-computer - Raspberry Pi 2. They utilized the power of the Raspberry Pi 2 processor to store and visualize data on available IoT platforms. However, the Raspberry Pi is not an economical alternative to suffice the need for many node requirements in air quality monitoring tasks.

A wireless mobile air pollution monitoring application was designed by Dhingra et al. [67]. They used a general-purpose sensor for carbon monoxide and methane measurement connected to the Arduino. Arduino connects to the Wi-Fi and sends data using the HTTP service to the cloud. J. Huang and his team developed the sensing node to report the air quality inside the car. The sensing node was connected to the mobile app using Bluetooth, and the parameters can be sent to the cloud from the mobile app developed [68]. S. Sun and his team [69] explained the benefits of long-term indoor air quality monitoring. The sensing node was transmitting the data to IOTA Tangle through the MAM communication protocol. All these implementations [67-69] utilized the cloud-based approaches in reporting the sensed data. The implemented systems were not tested to evaluate the reliability and accuracy either with indoor or outdoor deployment. Moreover, general-purpose gas sensors were utilized during experiments in place of specialized sensors or sensor modules, which can directly calculate the pollutant parameters such as PM2.5 and PM10 and provide more insights about air quality monitoring.

Challenges in air quality parameters monitoring:

- Many efforts are made for monitoring quality of air using GPS and GPRS based approaches for communicating the fetched parameters to the end-user side. These approaches are now considered to be outdated with the advancement of technology.
- Wireless sensor-based approaches are dominant in recent years due to their applicability in transmission to the central repository over the Wi-Fi network established. WSN based method is mainly implemented through ZigBee-based network setup. However, the approach suffers from the cost demand of such network setup due to the high number of message forwarding ZigBee devices to cope with the limitation of transmission range.
- With the advancement of IoT-based platforms and cloud-based integration support, IoT-based solutions can take advantage of such platforms. Many efforts are made for monitoring the air quality using IoT in recent years, as discussed in the literature review. Though the air quality monitoring using the IoT remained the open area of research because of the inherent issues of the IoT field, which can be described as below:

- Complex architecture design (design changes as per the criteria like deploying location, data storage and monitoring location, etc.)
- No standardization of protocol (Still IoT is under development, communication protocols are also under development)
- Cost is also one of the significant parameters to consider, especially air quality monitoring systems looking at the vast number of monitoring sites(node deployment) requirements.
- Less memory for processing (trade-off between processing vs. cost of the processor)
- Power usage requirement (air quality monitoring at the remote site is supported with battery backup, in such cases, power consumption is a big issue)
- Interfacing of sensors
- reliable delivery
- Authentication of streaming nodes

2.2. REVIEW OF EXISTING AIR QUALITY PARAMETERS PREDICTION APPROACHES

Numerous air quality prediction methodologies have been presented by researchers, which can be classified as statistics based methods, ML (machine learning) based approaches, and, recently, approaches based on the deep learning. The statistics-based method comprises principal component analysis(PCA), Coefficient analysis, linear and non-linear regression-based model, and interpolation-based implementations.

Authors [70] investigated a few feasible methods for estimating the air pollutant concentration (particulate matter) by utilizing existing available pollutants' data at adjacent monitoring stations as an alternative to the actual measurement. They used Spearman correlation and cluster analysis to disclose similar behaviour in Shanghai's particulate matter monitoring network. They also applied linear stepwise regression and the nonlinear support vector regression to predict the pollutant concentration at an intended station in terms of the pollutant values available at adjacent stations. Authors [71,72], in their experiments, tried to estimate the relation between particulate matter or specific pollutants and other pollutants along with meteorological elements collected at the same sampling events. They applied multiple linear regressions in their experiments. K. R. Baker et al. [73] developed a nonlinear regression-based model to approximate the average formed annually for particulate matter released from a single source. They developed separate models for primary and secondary pollutants. The model was also given emission rates in tons/year and the distance between receptor to source as variables for estimation purposes.

Research on human health effects of long-term exposure to air pollutants has played a critical job in modern health impact assessments. The exposure evaluation for epidemiological studies of long-term disclosure to air pollution is a tough challenge due to extensive small-scale spatial variation. Land use regression models were used increasingly over some time in the past. Land-use regression comprised of monitoring air pollutants at usually 20–100 places, extended over the area of study, and the development of stochastic models using the predictor variables typically accessed by geographic information systems (GIS) [74]. The important predictor variables incorporate various types of traffic representations, population density, climate, land use, physical geography (e.g., altitude). Land-use regression approaches have been applied effectively for modeling yearly mean concentrations of nitrogen dioxide, nitrogen oxide, particulate matter, and volatile organic compounds in various settings, including North American and European cities. The performance of the LUR approach in urban and sub-urban areas was proven to be better or alike to other geostatistical methods such as kriging and dispersion models.

The ARIMA- autoregressive integrated moving average method is an efficient and extensively studied approach to predicting the time-series data. The method was first projected in [75], ARIMA gained high acceptance because of its inherent statistics based characteristics, flexibility, and adaptableness to denote a variety of processes. Over the period, as the worry of air quality and living quality in suburban and urban regions has risen, a statistics-based approach such as ARIMA started to predict air pollutants levels.

Authors [76] demonstrated using the time series model to predict the index of air quality or pollution in the city Shah Alam of Selangor. They used the data that consisted of the seventy monthly air pollution index observations of six consecutive years, made accessible by the department of environment of Selangor. They implemented the ARIMA and integrated the long-memory model ARFIMA in their study. The performance evaluators were the mean absolute error, root mean square error, and mean absolute percentage error. Air quality indices, defined by various agencies, were used to grade the ambient air quality all over the globe. The air quality index computed based on the suspended particulate matter(SPM), Respirable suspended particulate matter(RSPM), nitrogen dioxide, and sulfur dioxide for the city Chandigarh, India using twenty-four hours' data were forecasted [77]. They developed three different time series models: ARIMA, ARFIMA, and Holt and Winters (HW) smoothing methods. They predicted the air quality index concerning the major responsible air pollutant RSPM. They applied several model assessment criteria like Mean Absolute Error, Mean Absolute Percentage Error, Root Mean Square Error, and Bias adjusted Akaike's information criterion (AICC). Their conducted experiments demonstrated that the ARFIMA was the more suitable model for prediction purposes. Author [78], in the experiments, demonstrated the performance of ARIMA, and the performance had been compared against a Holt exponential smoothing model to forecast the air quality index day to day values [78].

These methods suffer due to a lack of ability to model non-linear and multivariate types of data. With the availability of a vast volume of historical data for analysis and the requirement for high accuracy prediction performance in various scientific fields, machine learning has drawn interest in stating them as an appropriate solution compared to the other classical statistics-based approaches for time-series prediction. Machine Learning models have been broadly utilized to predict air quality. The machine learning-based methods consisting of fuzzy methods [79, 80], genetic algorithm-based methods [81], and support vector [82, 80] based implementations.

Recent development in neural network-based prediction approaches has presented better forecasting accuracy, outperforming the conventional statistics and machine learning approaches in various domains. In their study, Sahin et al. [83] demonstrated that the correlation between forecasted and the actual observed data was relatively high for all air pollutants using the CNN model. They found that the CNN-based approach was able to predict sulfur dioxide concentrations better compared to pm10 concentrations. Moreover, the prediction model was performing better for the winter than the summer period. Their conducted experiments demonstrated that precise forecasting of missing air pollutants could be achieved using the CNN based approach over the data from Istanbul, which contained unrecorded observations

Archontoula Chaloulakou et al. [84] have implemented artificial neural networks and multiple linear regression-based approaches to predict the PM10 air pollutant. The pollutant data were covering two years' period data for the city Athens of Greece. They used one-third of pollutant data from the available data set as the test data set. The experiments showed that ANN could give satisfactory prediction results as per the necessity if appropriately trained.

Azid et al. [85] found that the artificial neural network approach can be appropriate. They applied successfully as one of the tools for problem-solving for atmospheric management in a better way. They proposed a model based on ANN and compared the results with principal component analysis(PCA) to predict air pollutant index in peninsular Malaysia.

S.TikheShruti et al. [86], in their study, applied two algorithms based on soft computing. Algorithms were: artificial neural network (ANN) and genetic programming (GP) for forecasting the purpose of future air pollutant concentration levels such as sulfur and nitrogen dioxides and particulate matters data of five years for the Pune city of Maharashtra. Their conducted experiments showed that the genetic programming-based model was better compared to the ANN-based model.

Recently the deep learning methods include recurrent neural network (RNN), and Long Short Term Memory (LSTM) based neural network model [87, 88]. The LSTM based network performed more effectively than the RNN based network model due to the gated cell-based mechanisms in the LSTM unit [88, 89]. Jiachen Zhao and his team [90], in their research, used historical air quality data and meteorological data to forecast particulate matter 2.5 contamination of a particular monitoring station over 48 hours. Their prediction was multistep ahead in time for a prediction. The proposed data-driven long short-term memory fully connected (LSTM-FC) network model for prediction using deep learning. The meteorological data contained seven items: timestamp, relative humidity, temperature, pressure, wind direction, and wind speed. The proposed work was comprised of two types of experiments. The first one was a temporal simulator to predict the local variation of particulate matter concentration. The second one was based on the combination of spatial data to understand the influence and spatial dependency of nearby monitoring station data on the central (targeted) station data. The data were collected from the Beijing network of air quality monitoring stations. They used Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to show the experiments' performance. The results were compared with a simple artificial neural network setup.

Li et al. [91] presented the long short term memory network extended (LSTME), for forecasting air pollutants that intrinsically focused spatiotemporal correlations. As per the authors, LSTM layers were used for fetching inherent useful features automatically from historical and auxiliary data. They combined the time stamp and meteorological data for performance enhancement purposes. They collected particulate matter data of Beijing city for experiments. The performance was compared to the autoregressive moving average(ARMA), support vector regression(SVR), and simple LSTM.

In their study, Soh et al. [92] proposed a spatial-temporal deep neural network (ST-DNN) model to predict air pollution. Their proposed model utilized a long short-term memory network to fetch temporal characteristics, exploited KNN and artificial neural network (ANN) to extract spatial information, and combined it with the convolution neural network for conveying the land features. Z. Qi et al. [93] mainly utilized recurrent neural networks to present DAL's deep learning network for air quality prediction. The DAL comprised the selection of features and the semi-supervised training into the various learning layers of the proposed deep learning network. This model utilized various features into consideration in the prediction process.

Authors [94] proposed the short-term prediction model with the deep learning approach for particulate matter concentration prediction. The model was using Beijing PM 2.5 data set available in the UCI repository. They presented a convolution bidirectional gated recurrent unit(CBGRU). The CBGRU was combining 1D convents and a bidirectional gated recurrent unit network. They demonstrated the pm2.5 multistep ahead predictions. The 1D convents learned local features and reduced dimensionality on five input variables using the pooling layer. The low dimensional feature sequence output was given to bidirectional GRU networks to learn the time dependency relation between those sequences. The model was stacked further. The primary research factor of the work was utilizing the local feature extraction capacity of convents as preprocessing step to the bidirectional GRU network. They fine-tuned the performance by hyperparameter setting and compared it with support vector regression, simple GRU, and RNN. They used root mean square error, mean absolute error, and symmetric mean absolute percentage error as evaluation matrices during experiments.

Sookyung Kim et al. [95] proposed the long short-term memory network model to forecast the hourly concentration of the fine dust particles based on historical pollutant data, dust concentration, and climate variables in the surrounding atmosphere at the targeted location in Seoul. They trained and tested individual LSTM models for 25 districts of Seoul. They used the mean square error(MSE) for the evaluation of their proposed model.

Dongming Qin [96] et al. proposed a deep learning approach for air pollutant prediction based on the vast amount of historical environmental data. Their approach presented the method effectively to integrate enormous data (big data) and deep learning. Their proposed model used a convolution neural network (CNN) as the first layer for feature extraction from input data. The output of the last pooling layer from CNN was given as input to the long short-term memory layer. The Long Short Term Memory network was successfully adding the time series forecast function to their proposed model. The root means square error (RMSE) and correlation coefficient (Corr) were used as the performance evaluation matrices. They presented the benefits of their proposed model by comparing it to the numerical model and RNN model.

Ziyue Guan and R. O. Sinnot [97] conducted experiments using several machine-learning approaches to predict the particulate matter 2.5 concentrations. They collected the air pollutant data from the environment protection agency’s website for the Melbourne city of Australia. They applied linear regression, artificial neural networks, and long short-term memory network approaches in the prediction. The LSTM based approach was performing better compared to the two others.

Thanongsak Xayasouk and his team [98] demonstrated a model for predicting particulate matter using long short-term memory and a deep autoencoder(DAE) based model. They used hourly particulate matter data from twenty-five stations in Seoul in South Korea. Their model experimented with tuning and setting of various hyperparameters. They compared the performance with a simple LSTM model using the root mean square error metric.

Saba Gul and his team [99] researched the usage of a recurrent neural network as a prediction framework for the hazardous level in Lahore and Beijing. They classified the historical hazard pollutant data into six categories: good, moderate, and unhealthy for sensitive people, ill, very ill, and hazardous. Their proposed deep learning framework was comprised of three different layers. First was a single LSTM layer and two dense layers following the first layer. Out of these two, one was a hidden layer, and one was an output layer. Their work was related to classification rather than time series prediction. The proposed network model’s performance was evaluated using cross-entropy and accuracy matrices. Their approach performed well due to the LSTM’s ability to model temporal data.

Yue-Shan Chang and his team [100] used the aggregated LSTM model in their proposed prediction framework. The model used three sets of data and aggregated predictive features from those three sets to predict PM2.5. The first set utilized the data from the local station, which contained 17 attributes along with PM2.5. The second data set of PM2.5 and PM10 was derived from nearby stations to realize the effect of the nearby pollutant on local pollutants. The third dataset was derived from the industrial zone to consider the impact of external sources. The model consisted of 2 types of categories mixing and aggregation training. The mixing class trained the independent model to generate predictive features. The aggregation category training made it possible to synchronize the predictive features of sub-models to provide a final composite dataset. The model was utilizing the effect of other (other than PM2.5) pollutants from local stations and the impact of nearby and external sources PM2.5 and PM10 on the prediction of local stations' PM2.5.

Jun Ma and his team [101] proposed the transfer learning-based LSTM network model. The model improved the prediction accuracy of air pollutants, specifically for newly established monitoring stations for which the amount of historical data was limited. The model was trained with the use of existing station data. The trained model's first few layers were frozen, and the upcoming new layers were fine-tuned with the help of data from the new station. So the model could acquire information from current(existing) station data and could transfer the spatial learned knowledge to the new station for improving the accuracy of the new station.

Verma and his team [102] used a bi-directional LSTM based neural network for air pollution monitoring. The work was based on the classification of severity category rather than time-series pollutant level prediction in the future. The bi-directional approach employed can learn features in both forward and backward directions to optimize the LSTM performance further. Weitian Tong et al. [103] utilized the bidirectional LSTM network for spatiotemporal-based interpolation of particulate matter PM2.5 concentrations. The model emphasized both spatial and temporal factors.

Authors [104] used a flexible dropout layer to automatically adjust the dropout rate along with widow size for specific intervals and found good results by adding such layers. The model performs better than other alternatives (without flexible dropout) such as GRU and LSTM.

Challenges in air quality parameters prediction:

- Numerous air quality prediction methodologies have been presented by researchers using statistical methods. The statistics-based approach comprises principal component analysis(PCA), Coefficient analysis, linear and non-linear regression-based models including ARIMA and ARFIMA, and interpolation-based implementations. These methods suffer due to a lack of ability to model non-linear and multivariate types of data.

- In recent years, many papers have been published on predicting air pollution parameters using different models based on classes of neural networks, autoencoder, and deep learning. Most articles explain the models used with a brief outline of the mathematical foundations and schematics for the architectures. However, the results in the individual publications are not comparable and not compared due to their own implementation of algorithms and lack of benchmark standardize tools or algorithms.
- Feedback networks like RNN can utilize the availability of internal memory to remember the input state with the concept of order in time. Feed forward network cannot recall past observation apart from training. The output of the last step is given as input to the current step in RNN. So, the information is traversed through a loop using the recurrent connection to influence the past. RNN based approaches in deep learning approaches are more suitable for time series or temporal data prediction than CNN (Feed forward only). The RNN faces two significant challenges: exploding gradients and vanishing gradients while predicting the longer sequences.
- The deep learning-based model also suffers from overfitting-related issues.
- LSTM based approach is one of the popular choices in air quality parameter prediction, as discussed in the literature review. There are several approaches addressed in the study to improve the LSTM performance further. Still, there is scope for improvement of the performance of LSTM.
- Bidirectional LSTM is one of the approaches to improve the performance of LSTM. In the field of air quality prediction, one approach of using bidirectional LSTM [101] is available to enhance the performance of LSTM. Still, the work is done for label classification where the label is various severity categories rather than actual sequence to sequence (moving window based) value prediction. The Bidirectional model is not employed for the sequence to sequence time series prediction of air quality. Moreover, the performance of such a model depends on how the two layers are merged. Still, the bidirectional LSTM based implementation has the scope of improvement with the methodology employed for constructing the prediction model and fine-tuning of various hyperparameters.