International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Ambient Air Pollution Forecasting System using Deep Neural Networks

Geethika Jujjavarapu, Siddhartha Duggirala, Anulekha Kavutarapu, Ravikishan Surapaneni

Abstract: ***Air pollution is a major problem that has been recognized throughout the world. Harmful impacts of air contamination include hypersensitive reactions such as throat irritation, itchy eyes, nose, and some other serious problems. In recent years, the number of fatalities occurred due to air pollution has been increasing dramatically. In this paper, various air pollutants such as Carbon Monoxide, Methane or natural gas, LPG, and air quality at different places of city are measured using sensors. Further, the detected values are then used in the prediction of future values. The evolution of deep neural networks and Internet of Things made this possible to detect and forecast the concentration of pollutants underlying in the air. We use a special module called pyFirmata firmware which is used to connect the Arduino with python and upload the data into csv file on Jupyter Notebook. Here, the data collected is univariate i.e. it varies with only time. Though there are many statistical models to predict time series datasets such as ARIMA, their efficiency is low. Deep Neural Networks works well for predicting univariate as well as time series datasets. Hence, the Keras sequential model is employed to predict the hourly future values of air pollutants based on previous readings. The final results of prediction are compared with the actual values and error is calculated. As a result, the level of air pollutants at a particular hour can be predicted. The concentration of air pollutants in coming years, month or week helps us to reduce its concentration to lesser than the harmful or toxic range.***

Keywords: Air Pollution, Deep Neural Networks, Keras Sequential Model, pyFirmata, Univariate Data.

I. INTRODUCTION

Air contamination is a significant reason for heart attack, chronic bronchitis, lung malignancy, intense respiratory diseases, and intensifies asthma. The deterioration of human health due to the increasing rate of air pollutants mainly CO, methane necessitates the prediction of air pollution. Numerous climatic factors, for example, wind course, temperature, atmospheric pressure, and moistness impact the concentration of toxins in the air..

Manuscript received on March 15, 2020.
Revised Manuscript received on March 24, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Geethika Jujjavarapu* , Computer Science program at VR Siddhartha Engineering College. affiliated to JNTUK University, Kakinada.
Siddhartha Duggirala, Bachelor of Technology in Computer Science and Engineering at VR Siddhartha Engineering College afflicted to JNTUK University, Kakinada.
Anulekha Kavutarapu , B.Tech final year in Computer Science and Engineering at VR Siddhartha Engineering College affiliated to JNTUK University, Kakinada.
Mr. Ravikishan Surapaneni , Associate Professor, CSE department at VR Siddhartha Engineering College

Moreover, The Future Level Of Pollution Can Be Determined Based On Previous Values. Deep Learning Neural Networks Are Proficient In Spontaneous Learning And Feature Extraction From Raw Information
This element of neural networks can be utilized for time series prediction problems, where models are originated legitimately from the raw information without the direct requirement to transform the data using standardization and normalization or differencing to make the data stationary. In this paper we proposed a model which predicts the air pollutants on the hourly basis.

A. Motivation

The knowledge of the pattern of contamination levels is imperative to identify ahead of time, so vital prudent steps need to be considered to keep humans from unsafe impacts of prolonged exposure to contamination. In this way, the expectation of poison levels is one of the most significant issues for common organizations and is likewise significant in climate checking.

As indicated by WHO information, an expected 42 lakh individuals died due to the impacts of air contamination in India[1]. Some 3.8 million unexpected losses every year are credited to outside (surrounding) air contamination. At any rate 130 million individuals inhale demeanor of value that is multiple times or increasingly over the WHO standards; twelve of 24 urban areas found on earth with the most noteworthy yearly degrees of air contamination are in India. A uniform scale of 2 million people in India is an unanticipated loss due to air pollution[2].

Along these lines, foreseeing the degree of air quality is a major undertaking in deciding the air quality. In this model, we surveyed the dataset at various locations in Vijayawada. We utilized this dataset to prepare a profound neural system to anticipate the future values.

B. Objective and Scope

The fundamental objective of the project is to provide a platform that monitors the parameters and help to create better and pollution free future life. This paper centers around the ambient air pollutant forecasting system. In order to forecast, the pollutants must be detected, and this is possible with the help of some sensors as well as Arduino Uno module. This module helps to connect the sensors together for detection. Furthermore, it focuses on refined modeling for anticipating hourly air contamination concentrations based on historical air pollution data and the hour at which the concentration is detected.

II. RELATED WORK

In this section, we discuss about some conventional methods and techniques in the literature employed to predict air pollution.

Till now, different air pollution predicting procedures are proposed, and chiefly characterized into statistical approaches, the deep machine learning techniques, and the shallow learning methods. Statistical approaches are Linear regression, Logistic regression, Time series [3], K-nearest neighbor algorithm [4], etc. In these methods, the accuracy is limited due to their impotence to predict multivariate and non-linear data. Traditional machine learning approaches include decision trees, Support Vector Machine(SVM)[5], Naive Bayes, Multilayer Perceptron(MLP),artificial neural networks [6], etc. For the forecast of PM2.5, the effectiveness of attribute selection by random forest was explored by Shamsodinni. He claimed that the model performed better in comparison to Feedforward neural network. Akram and Ramin [7] discovered improvement in forecast of PM2.5 and air quality index by combining decision trees and Feedforward neural network. The data is predicted based on the meteorological dataset with previous values excluding the weather conditions at which forecast is done.

Lately, AI techniques have shown their attainability in non-linear patterns. We can see various types of implementations in air quality prediction[8]. Karpinnen evaluated the performances of artificial neural network , linear regression and an ensemble model to predict the levels of PM10 and NO2. It was deduced that the artificial neural networks exhibit better performance when compared with other two models, mainly in forecasting PM10[9]. However, many machine learning models are implemented for prediction , some areas are not considered where there is more pollution.

The adequacy of Recurrent Neural Networks(RNN) is indicated in managing time-series data . In order to forecast the data in future the present values are needed. BRNNs may partially acquire this by delaying the yield to include future knowledge[10]. Theoretically, an immense broaden may be implemented, however, the forecast results dribble if the extend is too enormous. Simultaneously, the output is delayed using some partitions method which is implemented proficiently in order to get well-built outcomes for continuous information. Moreover, two distinct networks, are incorporated for every input and the results are integrated using geometric or arithmetic mean to obtain the final forecast. It is difficult for having better result because the training of the network is based on indistinguishable knowledge which may be considered as a biased model.

Deep learning models are considered as an eminent model for exhibiting good performance on complex learning tasks. The solutions for the prediction of deep learning model can be found in numerous literatures. A variant of deep learning, LSTM, is a modified recurrent neural network (RNN) for supervised temporal difference learning. LSTM consists of loops that retain historical events for greater utilization of its input. LSTM is utilized to solve various sequence problems such as stock price prediction, machine interpretation, traffic stream forecast, text generation, text recognition in videos, and speech recognition etc. It exhibits great efficiency in all these areas that model temporal data with association of input variables to target out.

III. SYSTEM ARCHITECTURE

The system architecture is comprised into two parts. They are:

1. Detection of air pollutants
2. Prediction of air pollutants

A. Detection of Air Pollutants

In this stage, the data is gathered from multiple sensors such as MQ-135, MQ-7, MQ-4, and MQ-2. Initially, these gas sensors are connected to an Arduino Uno microcontroller at specific pins. The ground pin of all the sensors are connected to GND of Arduino and Vcc is connected to 5V on Arduino. Fig.1 shows the detection phase of air pollution. The code required for each sensor is written IDE and compiled to fix any errors. The compiled code is uploaded to microcontroller which in turn gets stored in the respective sensors. The sensors in response detects the gases present in the atmosphere. The analog values obtained are first converted to ppm and can be seen on serial monitor of Arduino IDE.

- The MQ 135 sensor is suitable for detecting or measuring of air quality. It is used to detect the gases such as CO2, NOx, Benzene, Alcohol, NH3, Smoke. It is connected to A0 pin of Arduino Uno microcontroller.
- The MQ 7 sensor which is used to detect concentrations of carbon monoxide present in the air. It is connected to A1 analog pin of Arduino Uno.
- The MQ 4 sensor is used in the detection of methane and natural gas. It is connected to A2 pin of Arduino Uno.
- The MQ2 sensor is used in detecting the leakage of gas. It is connected to A3 analog pin. All these sensors are connected via breadboard to make a compact circuit.

The permissible limits of pollutants are shown in Table I

Table I: Safe Level Of Pollutants

Pollutant	Safe limit
CO2	<350 ppm
CO	<70 ppm
Methane	<50 ppm
Gas	<1000 ppm

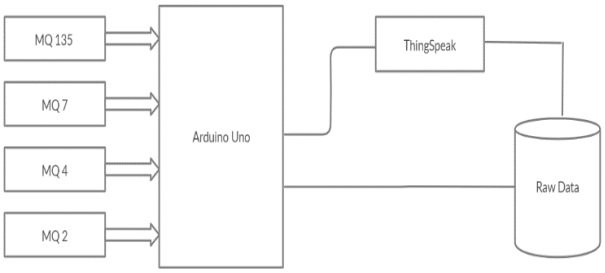

Fig 1. Detection of air pollution

B. Prediction of Air Pollutants

The detected data is uploaded to ThingSpeak, a cloud platform especially designed for IoT.

The dataset required for prediction can be obtained either by exporting data from ThingSpeak and importing as a csv file or by directly reading serial input from Arduino in python. The later can be achieved by uploading StandardFirmata code into Arduino and importing pyFirmata module in python.

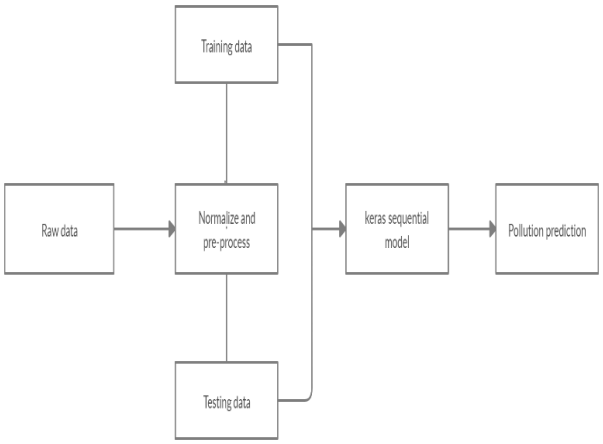

Fig 2. Prediction of air pollution

The proposed system consists of following steps:
1.Initially the gas sensors are calibrated and connected to Arduino, detects the air pollutants in the atmosphere.
2.The data is collected from Arduino to Jupyter Notebook using pyFirmata module.
3.The values read using the sensors are appended to dataset(csv file).
4.The dataset is pre-processed, and inputs are selected.
5.Split the dataset into training and testing data.
6.Train the Keras Sequential model using training dataset.
7.The model is tested to predict the pollutants after specific period.
8.Analyzing the difference between actual and predicted values.
The overall methodology of the proposed system is explained in the next section.

IV. MACHINE LEARNING APPROACH

Common data which is in a raw format can be handled by standard machine learning methods. The word Deep Learning (DL) has risen with the development of feature learning. Using Deep learning techniques, the detection and classification of normal data can be automatically discovered. It has contributed significant development in solving problems related to AI for long time with the help of machine learning algorithms. It additionally comes up with better classification and representation of time-series analysis contrasted to approaches when preprocessed and trained appropriately.

In machine learning, artificial neural networks (ANN) comprise of nodes that are actuated through weighted associations with previously dynamic neurons.

Deep Neural Network (DNN) is like the composition of ANN but signals and processes data more proficiently. It consists of latent layers that are placed between input and output layers that determines weight of layers between input and final layers.

Time series data analysis has been researched by analysts for many years. Conventional models do not have the ability to model linear data with high precision rate. Deep Learning is combined with real-time data for the analysis of time series prediction because time series data is complex, multidimensional and noisy.

We discuss about the dataset collection, its processing along with prediction mode below.

A. Dataset Collection

For prediction purposes, we compiled the real time dataset of air pollution including the concentrations of air quality, CO, methane and gas.

The dataset collected comprised of following attributes:

- Created time
- Concentration of air quality(in ppm)
- Concentration of CO(in ppm)
- Concentration of methane(in ppm)
- Concentration of gas(in ppm)

The values read from Arduino are appended to the dataset and imported as csv file or the dataset is exported from ThingSpeak platform.

B. Dataset Preprocessing

Initially, the dataset is grouped on hourly basis by calculating the mean of air pollutants at a specific hour. By doing this, we can predict the level of air pollutants at a particular hour of day. The window transformation is applied to a dataset of nested elements to produce a dataset of nested windows. Then the flat map is used to ensure that the shape of dataset remains constant i.e. to flatten the chunks in dataset into elements in a dataset.

The dataset is filled with a buffer of elements and the elements are randomly sampled from this buffer, thus, restoring the chosen elements with latest elements. Finally, the dataset is divided into batches of consecutive elements.

C. Initializing the Model

In this paper, we have used a sequential model. The Sequential provides easy approach to construct a model. It allows to construct layer one after another. The weight of every layer is associated to the following layer.

The add function is used to append layers at the end of model. Here we added three hidden layers and a target layer. The number of neurons in each hidden layer is 250. Dense attribute is conventional type of layer where, the nodes in the preceding layer is associated to its next layer nodes.

An activation function permits models to assess non-linear associations. The activation function used is Rectified Linear Activation (ReLU). Fig.3 shows the graph of ReLU activation function.

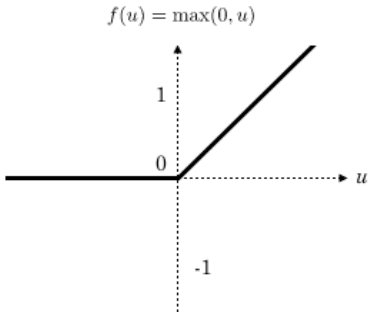

Fig 3. Rectified Linear activation function

The hyperparameters of the sequential model and its values are shown in Table II.

TABLE II: Hyperparameters Of Proposed Model

Hyperparameters	Values
Input layer	1
Output layer	1
Hidden layer	3
Epoch	100
Batch size	2

D. Training and Testing the Model

The train dataset and test data comprise of 62.5% and 37.5% of total data respectively. The model is trained with by passing it in the neural network for 100 times (epochs=100). After training the model, it is tested using test dataset to forecast the concentration of pollutants for a period of 6 hours.

E. Compiling the Model

We will be using 'SGD' as our optimizer. It controls the learning rate of various parameters and reduces the error function. Mean squared error is used for error or validation loss function. It calculates the average of squared difference between the predicted and original measures as follows:

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

where n represents number of data points, y_i represents actual values, and $\tilde{y}_i$ represents predicted values

V. EXPERIMENTAL RESULTS

In this proposed model, we obtained the results using deep neural networks. Root mean square error metric is employed to compare predicted results with the actual values. RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

We predict the air quality values for the following 6 hours from specified point of period. The below Table III consists of RMSE values we calculated for different air pollutants.

Table III: Keras Sequential Model Results

Air pollutant	RMSE
Air quality	2.55
CO	3.33
Methane	4.46
Gas	2.27

The following Fig 4(a), 4(b), 4(c), and 4(d) represents the variation between actual and predicted values of air pollutants we considered in this paper.

Using these residual plots, the level of air pollutants at any particular hour of day can be predicted based on previous values.

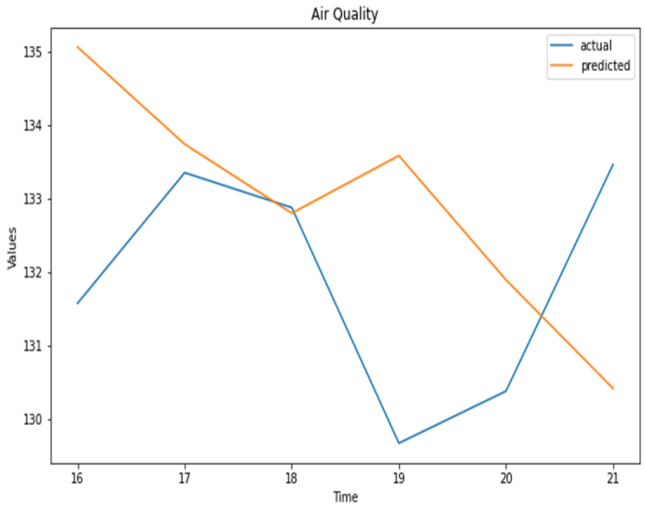

Fig 4(a). Air quality residual plot

Retrieval Number: F9425038620/2020©BEIESP
DOI:10.35940/ijrte.F9425.038620
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

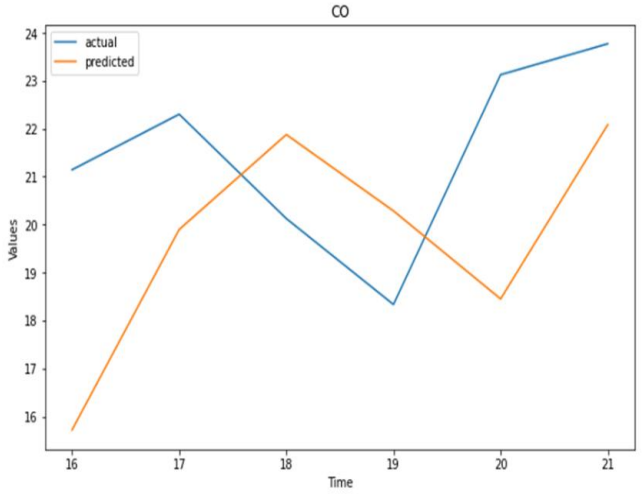

Fig 4(b). CO residual plot

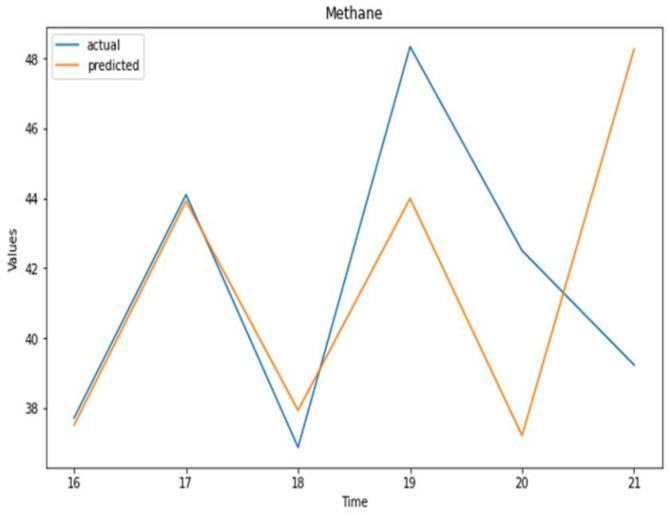

Fig 4(c). Methane residual plot

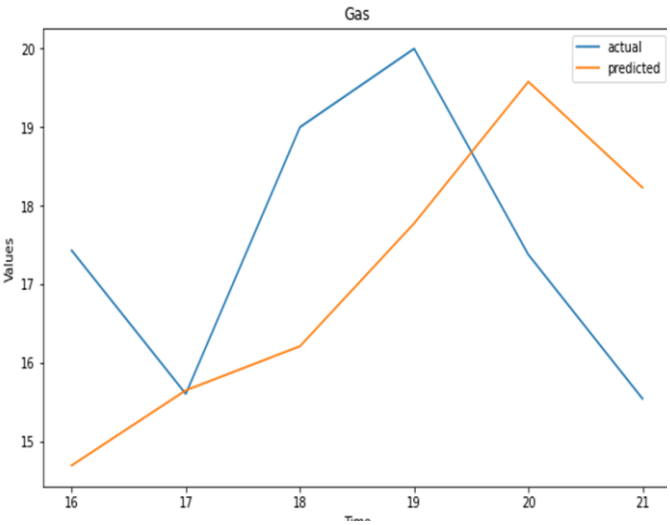

Fig 4(d). Gas residual plot

VI. CONCLUSION AND FUTURE SCOPE

In this literature, we implemented deep neural network with keras sequential model to predict the air quality, CO, methane, and gas. This model works well for time series datasets. The validation loss of the model for air quality is 3.12, CO is 5.05, methane is 4.31, and gas is 4.80. The dataset used in this model is for a short duration; it limits the model capacity. Hence, use of dataset for longer durations with negligible time gaps is recommended for further improvement. In future, we can employ several climatic factors such as measure of temperature, humidity, dew etc to elevate the accuracy of model.

REFERENCES

1. "Ambient air pollution - a major threat to health and climate," *World Health Organization Global Ambient Air Quality Database*, 2018.
2. "Air pollution in India", *State of Global Air 2019* [Online] Available: https://www.stateofglobalair.org/ , 2019.
3. Radmila Janković, Marijana Ćosović , Alessia Amelio, "Time Series Prediction of Air Pollutants : A Case Study for Serbia, Bosnia and Herzegovina and Italy", 2019.
4. Yang Rui-jun, Ding Dan-feng, Yan Feng, "Application of Improved KNN Algorithm in Air Quality Assessment" :*HPCCT 2019: Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference*, Pages 108–112, June 2019
5. Wei Wang, Wei-guo Shen, Bin Chen, "Air Quality Index Forecasting Based on SVM and Moments", 2018
6. Sheen Mclean Cabaneros, John Kaiser Calautit, Ben Richard Hughes, "A review of artificial neural network models for ambient air pollution prediction", *Environmental Modelling & Software*, Volume 119, Pages 285-304, September 2019.
7. Akram Jamal, Ramin Nabizadeh Nodehi, "Predicting Air Quality Index Based On Meteorological Data: A Comparison Of Regression Analysis", *Artificial Neural Networks And Decision Tree*, 2017.
8. Frank J. Kelly and Julia C. Fussell, "Air pollution and public health: emerging hazards and improved understanding of risk", *Environ Geochem Health*, 2015.
9. Jaakko , Leena, Juhani, Stephen, "Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations", *Atmospheric Environment*. Volume 37, Issue 32, October 2003.
10. J. Rene Beulah, K. Mahesh Babu, "Air Quality Prediction based on Supervised Machine Learning Methods", *International Journal of Innovative Tech and Exploring Engg.* , Vol 8, Pages 206-212, 2019.

AUTHORS PROFILE

Geethika Jujjavarapu is currently pursuing Computer Science program at VR Siddhartha Engineering College. affiliated to JNTUK University, Kakinada. She will complete her under graduation in 2020 with a B.Tech in Computer Science. She has interest in the fields of Data Science, Machine Learning, Cyber Security, and Deep Learning.

Siddhartha Duggirala is currently pursuing final year Bachelor of Technology in Computer Science and Engineering at VR Siddhartha Engineering College afflicted to JNTUK University, Kakinada. He will be graduating with B.Tech degree in 2020. He has interest in the fields of Operating Systems, Microprocessors and Internet of Things.

Anulekha Kavutarapu is currently pursuing B.Tech final year in Computer Science and Engineering at VR Siddhartha Engineering College affiliated to JNTUK University, Kakinada. She will complete her under graduation with B.Tech degree in 2020. Her major interests include Database Management, Web Technologies and Information Security.

Mr. Ravikishan Surapaneni is currently working as an Associate Professor, CSE department at VR Siddhartha Engineering College He has more than twenty years of teaching experience. His research excellency in the area of data analytics published more than 20 papers in various reputed journals.