A picture containing logo

Description automatically generated

**K. E. Society’s  
Rajarambapu Institute of Technology, Rajaramnagar  
An Autonomous Institute**

**SYNOPSIS OF UNDERGRADUATE RESEARCH**

|  |  |
| --- | --- |
| **Name of program:** | B. Tech (Information Technology) |
| **Name of student:** | Akash Ashok Mahajan (1954012) |
| **Date of registration:** | December 2021 |
| **Name of guide:** | Dr. A. C. Adamuthe |

**Proposed Title:**

Air quality forecasting using statistical and machine learning methods

1. **Introduction and literature review**

Growing industrialization and urbanization has threatened the quality of the air we breathe. Government authorities (for example, Government of India) has established air quality monitoring systems in regions where chances of airborne breathing diseases are high such as in Mumbai, Delhi, Kanpur, Kolkata etc. These monitoring devices are passive and real-time hence there is little or no way to predict the next hour’s air quality which is potentially beneficial for government to spread awareness or take important decisions associated with public interest. Air quality can be forecasted based on historical data collected from the air quality monitoring systems. Various techniques like statistical analysis and machine learning can be applied to this available data to predict the air quality index for certain cities around the India and other countries.

Time series forecasting and time series analysis both are used in variety of applications in the field of transport, retail, education, finance etc. Similar work has been proposed in [1], [2], and [3]. For literature survey, we study the different applications of time series forecasting such as crop yield prediction [2], gold price forecasting [3], load forecasting [4], stock prices prediction [5, 6], Sales forecasting [7] etc. The air quality forecasting [8, 9] is also studied. Most of the paper proposes direct implementation of statistical models like ARIMA, ARMAX, Principal Component Analysis (PCA). Some papers compare deep learning and machine learning methods with statistical models (Table 1).

Table 1 References with proposed or compared methods

|  |  |  |
| --- | --- | --- |
| **Reference** | **Paper title** | **Method proposed/compared** |
| [2] | An applied time series forecasting model for yield prediction of agricultural crop | Seasonal Adaptive ARIMA |
| [3] | A CNN–LSTM model for gold price time-series forecasting | CNN, LSTM |
| [4] | Day-ahead building-level load forecasts using deep learning vs. traditional time-series techniques | CNN, RNN, Seasonal ARIMAX. |
| [5] | Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices | ARIMA |
| [6] | The Use of Artificial Neural Networks in the Analysis and Prediction of Stock Prices | Feedforward MLP |
| [7] | Sales forecasting using time series and neural networks | ARIMA with interventions, Back propagation neural network model |
| [8] | Neural networks and periodic components used in air quality forecasting | Self-organizing maps, multilayer perceptron |
| [9] | Three improved neural network models for air quality forecasting | Adaptive Radial Basis Function (ARBF) network, Principle Component Analysis (PCA) |

1. **Problem definition**

**Compositions and properties of air**

Air is lightweight, odorless, and colorless element of nature. It is a mixture of various gases which makes up the atmosphere of earth. Figure 1 illustrates the concentration of various gases in the air. Nitrogen holds the largest part of air followed by Oxygen, Argon, other gases, and Carbon Dioxide.

Figure 1 Concentration of various gases in air

**Major air pollutants**

According to World Health Organization (WHO) Particulate Matter (Sulphate, Ammonia, Nitrates, Black Carbon etc.), Nitrogen Dioxide, Sulphur Dioxide, and Ozone are major contributors to both indoor and outdoor air pollution. The WHO has also provided the threshold for presence of these elements in air.

**Air Quality Index**

Air Quality Index (AQI) is the measure (See Figure 2) to report the compositions of pollutants in air. The concentration of the pollutants is calculated physically (By traditional methods or by smart air quality monitoring systems) and reported.

In India, most of the metropolitan areas has a smart air quality monitoring system installed in certain areas of city. The sensors transmit the data collected to the central repository where it is stored and shown on website [*https://app.cpcbccr.com/AQI\_India\_Iframe/*](https://app.cpcbccr.com/AQI_India_Iframe/)

Table

Description automatically generated

Figure 2 Air Quality Index measures

In proposed research work, we define our problem as a time series analysis problem where we forecast the future data based on the historical learning.

1. The sensors on an air quality monitoring system transmits the data about air quality in certain area of a city. It consists of concentration data of Particulate Matter (PM) PM10, PM2.5 and other air pollutants in an ambient (outdoor) air.
2. Every instance of the data is transmitted with a timestamp associated to it indicating at what time these readings was captured. Based on this information we calculate the Air Quality Index (AQI).
3. The previously calculated AQI can be used to predict the future AQI using a time series analysis method or algorithm.
4. Sometimes sensors may transmit faulty measurements because of physical conditions or malfunctioning. These kinds of errors can be either reduced or removed completely to avoid incorrect and faulty forecasting.

Graphical user interface, application

Description automatically generated

Figure 3 Dataset snapshot

The dataset contains 16 different columns, however, not all parameters affect the Air Quality Index. Pollutants like PM10, PM2.5, NO2, SO2, CO2, O3 etc. contribute majorly.

1. **Methods used for forecasting**

In this section, we will learn about few widely used methods for forecasting the values for a given dataset. We study two different sets of methods namely statistical methods and machine learning methods.

* 1. **Statistical methods**
     1. **Moving Average [10]**

Moving average (also known as Moving Mean) is a method of time series analysis. In this method, we take fixed size subset of data items from a given large dataset. Then, for each subset we calculate a mean or average value. It is called moving average because initially we take a fixed size subset and calculate its mean. On next iteration, we add next data items to this subset and remove first data items and again calculate new average.

This method is useful in smoothing the data when there are lot of fluctuations over the time.

* + 1. **Weighted Moving Average [11]**

Weighted Moving Average (WMA) is like Moving Average (MA), but it assigns different weights to different data points, whereas MA assigns equal weights to each datapoint in a given sample window. We calculate the WMA with help of a N-day period. The latest data point assigned the highest weight and the oldest one is assigned the lowest weight.

The advantage of WMA is that it will assign more weight to the recent readings. Hence, we can have more understanding of the recent trend in data.

* + 1. **Exponential Smoothing [12]**

Weighted Moving Average gives weights to the data points in decreasing order where difference between weights between two points is same. On other hand, Exponential Smoothing assigns exponentially decreasing weights to each of the data point in given sample window.

* + 1. **Autoregressive Moving Average (ARMA) [13]**

ARMA is formed from two different method: Autoregressive (AR) and Moving Average (MA). The AR part is associated with regressing over the past values with order ‘p’ written as *AR(p)*. The MA part is associated with modelling the error with order ‘q’ written as *MA(q)*. The whole term is written as *ARMA(p, q)*.

* + 1. **Autoregressive Integrated Moving Average (ARIMA) [14]**

ARIMA model is also called box-Jenkins model. It consists of Autoregressive, Integrated, and Moving Average models. We represent ARIMA as ARIMA(p, d, q). p, d, and q are the orders of the autoregression, the difference, and the average respectively. Box-Jenkins method uses the ARIMA as baseline which assumes that past outcomes can affect the future outcomes.

* + 1. **Linear Prediction [15]**

As name suggest, Linear prediction uses a linear function to predict the outcomes. The input to the function is usually the past outcomes. It includes the regression algorithms namely linear regression, polynomial regression etc.

* + 1. **Autoregressive Integrated Moving Average with Exogenous Inputs (ARIMAX) [16]**

The ARIMAX is extension to ARIMA model. The ‘X’ in ARIMAX suggests adding a new exogeneous variable in the analysis to support prediction of the expected variable. The ‘X’ can be any variable of our choice from the dataset. ARMAX on other hand, can be said as a special case of ARIMAX with order of ‘I’ factor is 0.

* 1. **Machine Learning methods**
     1. **K-Nearest Neighbor Regression (KNN) [17]**

KNN is generally used for interpreting the neighboring ‘K’ data points for a given data point and assigning the numeric value to output based on the nature of neighbors. It uses distance functions to calculate the similarities.

* + 1. **Support Vector Regression (SVR) [15]**

Support Vector Regression is the component of Support Vector Machine (SVM). The SVM is widely used classification algorithm which tries to find a hyperplane in a n-dimensional space. SVR algorithm involves decision boundary, ‘+d’ and ‘-d’ from the hyperplane above and below it respectively. Only those points within these decision boundaries are considered for regression.

* + 1. **Classification and Regression Trees (CART) [18]**

CART is a decision tree-based algorithm. As its name suggest, CART is used for both classification and regression problems. The data is partitioned and fit onto a simple regression model.

* + 1. **Recurrent Neural Network (RNN) [19]**

RNN is special kind of neural network which preserves the output provided by one layer and later provide it to input layer to predict new output. They are called recurrent because sending output to input layer and then getting output is done several times.

* + 1. **Long Short-Term Memory (LSTM) [20]**

RNN faces exploding or vanishing gradient problem, which is covered by more sophisticated model, LSTM.

Deep Learning and optimization of current methods have been applied to forecast time series [3, 4, 20].

1. **Proposed work**
   1. **Objectives**

Following are the objectives of proposed work:

* Literature review of statistical methods and Machine Learning methods for solving the problem of forecasting the sensor measurements.
* Implementation of the statistical methods like exponential smoothing.
* Implementation of learning-based algorithms such as neural network and deep learning methods.
* Comparison of collected results and choosing the best performing model.
  1. **Possible outputs**

In this research work, we are going to design a forecasting model which can forecast the measurements (readings) of the sensors of an air quality monitoring system. We will compare statistical and machine learning approaches and get following outputs from the study:

1. Comparative analysis report of various approaches used for forecasting which are based on either statistics or Machine Learning methods.
2. Appropriate model for forecasting of sensor measurements in a smart air quality monitoring system.
3. Produce analysis of the data associated with sensor measurements in smart air quality monitoring system with help of statistical methods, machine learning methods, or both.
4. **Required facilities**

To carry out the study and research work, facilities that will be required are as below:

* Programming languages: Python and R programming language.
* Tools: PyCharm, Anaconda, Python and R programming libraries etc.
* Standard computing hardware with GPU capabilities is sufficient.

1. **Expected date for completion of work:** May 2022
2. **Approximate Expenditure:** *NA*

**References**

[1] De Vito, S., Massera, E., Piga, M., Martinotto, L., & Di Francia, G. (2008). On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. Sensors and Actuators B: Chemical, 129(2), 750-757.

[2] Thiyagarajan, K., Kodagoda, S., Van Nguyen, L., & Ranasinghe, R. (2018). Sensor failure detection and faulty data accommodation approach for instrumented wastewater infrastructures. IEEE Access, 6, 56562-56574.

[3] Freeman, B. S., Taylor, G., Gharabaghi, B., & Thé, J. (2018). Forecasting air quality time series using deep learning. Journal of the Air & Waste Management Association, 68(8), 866-886.

[4] Reddy, P., & Sureshbabu, A. (2019, June). An applied time series forecasting model for yield prediction of agricultural crop. In International Conference on Soft Computing and Signal Processing (pp. 177-187). Springer, Singapore.

[5] Livieris, I. E., Pintelas, E., & Pintelas, P. (2020). A CNN–LSTM model for gold price time-series forecasting. Neural computing and applications, 32(23), 17351-17360.

[6] Cai, M., Pipattanasomporn, M., & Rahman, S. (2019). Day-ahead building-level load forecasts using deep learning vs. traditional time-series techniques. Applied energy, 236, 1078-1088.

[7] Mondal, P., Shit, L., & Goswami, S. (2014). Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. International Journal of Computer Science, Engineering and Applications, 4(2), 13.

[8] de Oliveira, F. A., Zárate, L. E., de Azevedo Reis, M., & Nobre, C. N. (2011, October). The use of artificial neural networks in the analysis and prediction of stock prices. In 2011 IEEE International Conference on Systems, Man, and Cybernetics (pp. 2151-2155). IEEE.

[9] Ansuj, A. P., Camargo, M. E., Radharamanan, R., & Petry, D. G. (1996). Sales forecasting using time series and neural networks. Computers & Industrial Engineering, 31(1-2), 421-424.

[10] Kolehmainen, M., Martikainen, H., & Ruuskanen, J. (2001). Neural networks and periodic components used in air quality forecasting. Atmospheric Environment, 35(5), 815-825.

[11] Wang, W., Xu, Z., & Lu, J. W. (2003). Three improved neural network models for air quality forecasting. Engineering Computations.

[12] Kumar, A. K., Lakshmi, A. S., & Rao, P. J. N. (2020, December). Moving average method-based air pollution monitoring system using IoT platform. In Journal of Physics: Conference Series (Vol. 1706, No. 1, p. 012078). IOP Publishing.

[13] Zhuang, Y., Chen, L., Wang, X. S., & Lian, J. (2007, June). A weighted moving average-based approach for cleaning sensor data. In 27th International Conference on Distributed Computing Systems (ICDCS'07) (pp. 38-38). IEEE.

[14] Brown, R. G., & Meyer, R. F. (1961). The fundamental theorem of exponential smoothing. Operations Research, 9(5), 673-685.

[15] Benjamin, M. A., Rigby, R. A., & Stasinopoulos, D. M. (2003). Generalized autoregressive moving average models. Journal of the American Statistical association, 98(461), 214-223.

[16] Nelson, B. K. (1998). Time series analysis using autoregressive integrated moving average (ARIMA) models. Academic emergency medicine, 5(7), 739-744.

[17] Lin, K., Lin, Q., Zhou, C., & Yao, J. (2007, August). Time series prediction based on linear regression and SVR. In Third International Conference on Natural Computation (ICNC 2007) (Vol. 1, pp. 688-691). IEEE.

[18] Peter, Ď., & Silvia, P. (2012, September). ARIMA vs. ARIMAX–which approach is better to analyze and forecast macroeconomic time series. In Proceedings of 30th International Conference Mathematical Methods in Economics (Vol. 2, pp. 136-140).

[19] Ban, T., Zhang, R., Pang, S., Sarrafzadeh, A., & Inoue, D. (2013, November). Referential knn regression for financial time series forecasting. In International Conference on Neural Information Processing (pp. 601-608). Springer, Berlin, Heidelberg.

[20] Lewis, R. J. (2000, May). An introduction to classification and regression tree (CART) analysis. In Annual meeting of the society for academic emergency medicine in San Francisco, California (Vol. 14).

[21] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.

[22] Yadav, A., Jha, C. K., & Sharan, A. (2020). Optimizing LSTM for time series prediction in Indian stock market. Procedia Computer Science, 167, 2091-2100.

|  |  |
| --- | --- |
| Date:  Place: Rajaramnagar | Akash Ashok Mahajan  Student (Roll No.: 1954012) |
| **Dr. Amol C. Adamuthe**  Guide, Department of IT,  RIT, Rajaramnagar. | **Dr. Sandip A. Thorat**  Head of Department, Department of IT,  RIT, Rajaramnagar |