JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF AIR POLLUTION USING SUPPORT VECTOR MACHINE

Prathipati Silpa Chaitanya[1*], Galla Devanshi[2] , Vallepu Anusha[3] , Navuluri Anitha[4] ,

Meduri Lahya Priya[5].

[1*]Assistant professor: Vignan's Nirula Institute of Technology and Science for Women

[2,3,4,5] B.Tech Scholar :Vignan's Nirula Institute of Technology and Science for Women.

ABSTRACT:

Air pollution is an international concern with respect to health and environmental issues, and therefore, an efficient monitoring and predictive approach must be designed. This paper presents the utilization of SVM algorithms for air quality prediction based on a robust dataset that features historical pollution records and relevant meteorological variables. The data was sourced from a number of monitoring stations, focusing on key pollutants such as PM2.5, PM10, NO2, and CO. The strengths of SVM in classification and regression were aimed at the development of predictive models that were capable of identifying pollution trends and categorizing air quality levels. These models, across check, utilized performance metrics such as accuracy, precision, and F1 score to establish an overall increased superiority of SVM over the traditionally used methods in regression and classification.

Key words: Air quality prediction, Meteorological variables, Real-time data integration, Hybrid models, Predictive models ,SVM(Support Vector Machine), Historical pollution data.

1.INTRODUCTION:

Air, water, and land are polluted for many reasons and continue to pollute the environment [1]. Among them, polluted air is having the maximum adverse consequence to human health [2]. Breathing in low air quality is harmful to the lungs and heart and arteries [3-5]. At the same time, the contaminants are detrimental to vegetation and all forms of life (Kumar). This kind of air pollution occurs primarily due to the burning of fossil fuels that emit a high amount of sulphur dioxides and nitrogen oxides that combine with the air [6]. Besides that, combustion activities of vehicles and industries also contributes to the increase of air pollution. In order to avert health risks to the environment, the AQI (Air Quality index) has been formulated with components that calculates gas emissions levels associated with pollution of certain areas [7]. From the data on the index, appropriate measures can be put in place to safeguard the health of people.

Some of the less common benefits resulting from the introduction and use of AQI, are most often related to policy-making decision [8, 9]. Municipal government agencies are taking controlled, targeted actions to reduce pollution in specific hotspots where such interventions are likely to be most effective, and this improves the future quality of the environment [10].In view of the complex and multilayered nature of air pollution, the problem is found to be rather complex when it comes to predicting air quality.

A number of such approaches are based on prediction using AI (Artificial Intelligence) methods. In a more recent approach, consisting of improved algorithms, deep neural network networks were the ones to help in modelling the non-linear dependency of data parameters. Similarly, their performance in the analysis of data has remarkably improved as deep learning models provide dependable answers [11]. Likewise, research studies have also applied RFR (Random Forest Regression) and SVR (Support Vector Regression) in [12] for developing the regression model in the prediction of AQI. Emissions of nitrogen oxides in some Italian city – sourced from a publicly available dataset. The performance of both models has been evaluated in terms of coefficient correlation, RMSE, and R2 recalculation and it has been proved that both methods are more efficient in prediction. In the same way in the propose, the research [13] utilized an optimal and hybrid method that combined secondary decomposition and an AI optimization algorithm. In order to evaluate the prediction model, AQI data collected from China from the year 2016 to the year 2018 has been used. Samples are obtained from wavelet decomposition with high frequency, low frequency, and variational mode, and sample entropy is used to smoothen the decomposition process.

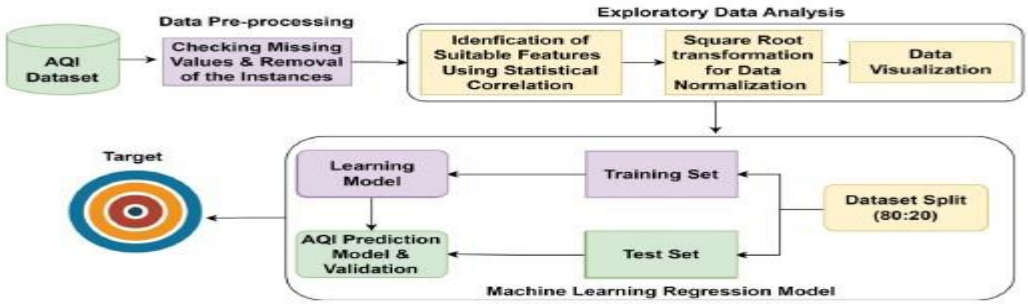

Fig-1 Data preprocessing and model building

LSTM (Long Short Term Memory) based NN (Neural Network) has been used for the prediction of AQI [14]. Also, the BAT has been employed for the optimization of parameters of the least-square SVM, taking into account the factors of air pollutants. It has been outcomes parks so efficient in capturing the properties of AQI data. Further, the assessment of the forecast about the level of pollutants on an hour's basis has been done using ANN in the city of Iran for the year 2009 up to 2010. The algorithmic implementation of the [19] movements of this child neurons was therefore made strong in a hidden layer first and then went forward into a single output layer final performance. The six pollutants that have been included are O3 PM 10 SO2 CO NO2 and PM 2.5 in [20, 21] measures from R2 for AQI are essential in all three due areas RMSE value compared attained more effective air quality forecasts which helped more optimal decision making involving air pollution. The future forecasting of air quality helps to limit the pollution load most especially [22, 23]. Similar, eminent work [24] has built several site quality prediction frameworks leveraging the DL model applying site clustering techniques.

2.LITERATURE SURVEY :

The support vector machine (SVM), M5P model tree and artificial neural network (ANN) are used to forecast pollution, the result shows that the best forecasting performance when different features are used in multivariate modelling with the M5P algorithm over other models. Real time sensor data of downtown Toronto, Canada is used by Wang et al. (2020) [1] and they have applied Land Use Regression (LUR), Artificial Neural Network (ANN) and gradient boost machine learning algorithm to forecast. the study by Chen et al. (2022) [2] uses 3-dimensional variation approach to integrate multi-source PM2.5 data from satellite and ground sensor data Wuhan, Xiang yang and Yichang (China). Six common pollutants (PM2.5, PM10, SO2, NO2, O3, and CO) and pollution levels are forecasted using an automated air quality forecasting method based on machine learning (Ke et al., 2022). [3] LSTM has shown a quick convergence and reduce

training cycle with good accuracy (Navares and Aznarte, 2020). Athira et al. use recurrent neural network (RNN), LSTM and Gated Recurrent Unit (GRU) for PM10 forecasting based on the time series from AirNet data (V. et al., 2018). [4-5] In Bekkar et al. (2021) deep learning algorithms like LSTM, Bi-directional LSTM, GRU, Bidirectional GRU, convolutional neural network (CNN), and a hybrid CNN-LSTM models were examined for predicting PM2.5 concentration. Beijing's AQI analysis and AQI forecast in the range from January 2019 to November 2021 is presented in Liu and You (2022) [6]. It uses an hourly data of major air pollutants namely PM2.5, PM10, and O3 and average to generate daily data, for further analysis and prediction. The ARIMA and LSTM were used for short-term prediction of Beijing's air quality data.

Population density, industrial activity, agricultural practices, thermal power plants, energy sectors, automotive industries, and transportation all affect air pollution differently [11-12](Ravindra, 2019; Ravindra et al., 2020). Several important factors determine air pollution in a given area including particulate matter, gaseous pollutants and metrological factors, which require AQI estimation by a large number of government and non-governmental organizations worldwide [13-14](Bao and Zhang, 2020; Li et al., 2020). Greenhouse gases (GHG) are primarily responsible for global warming, and these GHGs have an impact on plant-soil interactions in addition to climate change that have significant negative impacts on agriculture, the environment and also economy [15] (Malhi et al., 2021). Based on information from 2010 to 2019, the World Health Organization (WHO) published a report on worldwide air quality in 2022. This study examined a wide range of air pollutants mentioned above and discovered that PM2.5 was rising globally based on studies of 6743 cities in 117 different nations, causing 1.7 million yearly mortalities in India alone. The top 20 cities with the highest air pollution were 18 Indian cities, illustrating the seriousness of India's air pollution and adverse health effects in the years to come (Gurjar et al., 2016; Guttikunda et al., 2014) [16-17]. Due to the city's low height and seasonal temperature inversion, air pollution becomes a significant issue over the winter (Sumiya et al., 2023). The rapid development of technologies in smart cities leads to different kinds of pollution in the environment. More specifically, air pollution has increased due to transportation and the excretion of smoke in manufacturing industries [21]. SVR (Support Vector Regression) and RFR (Random Forest Regression) has used for building the regression model for predicting the AQI. Despite the performance of the model, SVR model is not suitable for processing huge number of samples [22]. The development of IoT (Internet of Things) in the prediction of quality measures in the air has assisted in improving the quality of life of humans. In the same way, the recommended research [23] has utilized sensors for the monitoring process and predicts the emission gas level of concentration, which affects the purity of air. The initial process has been handled with sensors, values stored in a cloud-based server have been measured using LSTM, and obtained results have been compared with SVR.

3.PROPOSED METHODOLOGY

To build a model for prediction of air we have used a dataset which contains 16 features and one label. We have used some gases percentages in air to predict whether the air is polluted or not. Gases like PM2.5,PM10,NO,NO2,NOx,NH3,CO,SO2,O3,Benzene,Toluene, Xylene. This model will predict whether the air pollution is severe, poor, very poor, moderate or good. We built a model using SVM(Support Vector Machine).

3.1 Support Vector Machine(SVM)

Support Vector Machines is supervised machine learning algorithm. It is mostly used for classification as well as Regression. The main purpose of SVM is to find the optimal hyperplane that separates different classes in a dataset. The distance between the hyperplane and the nearest data points from each class is known as support vectors. SVM has the ability to handle both linearly separable and non-linearly separable data. If the data cannot be separated by a straight line ,SVM utilizes kernel functions to transform the input space into higher-dimensional space. SVM is memory efficient because it uses only the support vectors in the decision-making process. It will not be effected by the noisy or overfitting issues.

3.2 Working of SVM:

1. Collect and preprocess data.

2. Select suitable kernel (linear, polynomial, RBF, sigmoid).

3. Find optimal hyperplane using optimization techniques.

4. Use trained model to classify new data points and predict the output.

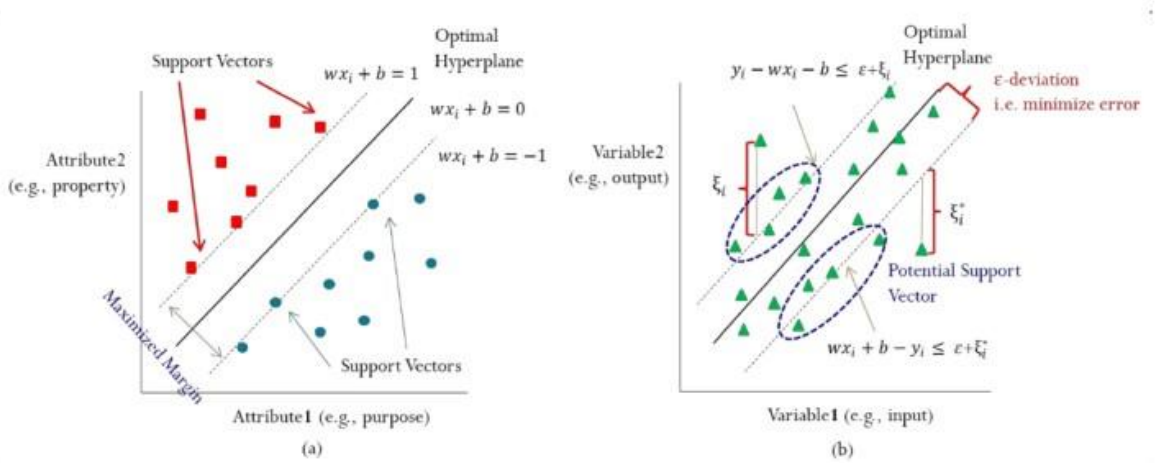

Fig-2 Overview of SVM algorithm: (a) SVM for classification;(b)SVM for regression.

3.3 ALGORITHM

```
from sklearn. svm import SVC

# Training SVM model

svm = SVC(kernel='linear', C=1)

svm.fit(X_train, y_train)

# Model prediction

y_pred = svm.predict(X_test)
```

Firstly, we should import necessary libraries from the scikit learn i.e SupportVectorClassifier(SVC). Initialize SVM classifier with:

C=1: Regularization parameter (trade-off between margin and misclassifications).

We should train the model by dividing it into training and testing sets. By using ” svm.predict” predict the output of the given dataset. In this way we used SVM for predicting the air pollution in the environment.

4.RESULT ANALYSIS:

The model is built using SVM model, as it is giving high accuracy. We have used many algorithms like KNN, SVM, Naïve Bayes, among them SVM got high accuracy. We can observe this in the below given three graphs.

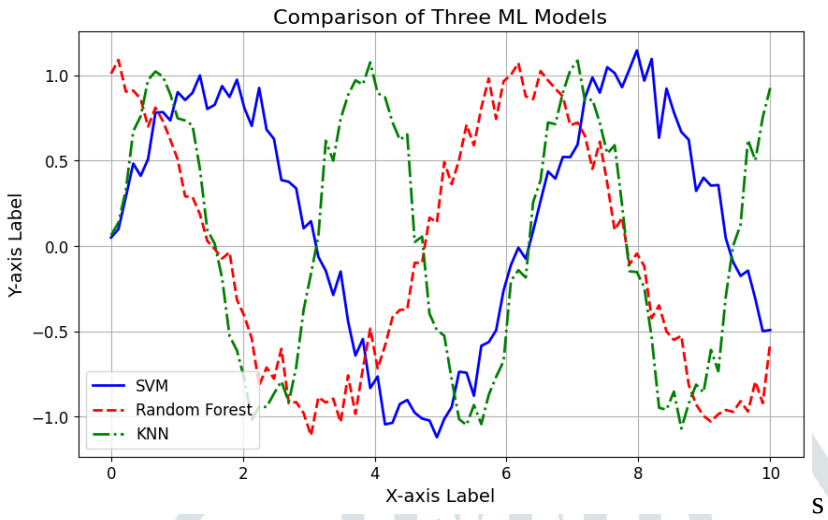

Fig-3 Comparison of Accuracy of three ML model

A comparison of how three machine learning algorithms are performing for three points on the x-axis: SVM, Random Forest and KNN. An initial y-axis metric oscillates between -1 and 1. The blue solid line is the SVM which evolves with fluctuations with significant peaks and troughs in the range. The red dashed line of Random Forest is quite more fluctuating higher values in the middle region. The green dashed-dot line for KNN has large fluctuations but still appears to give similar results for Random Forest about the maximum locations and other overall trends. The graph visually depicts the varying behaviors and the changing patterns of the three models over the range on this plot along the x-axis.

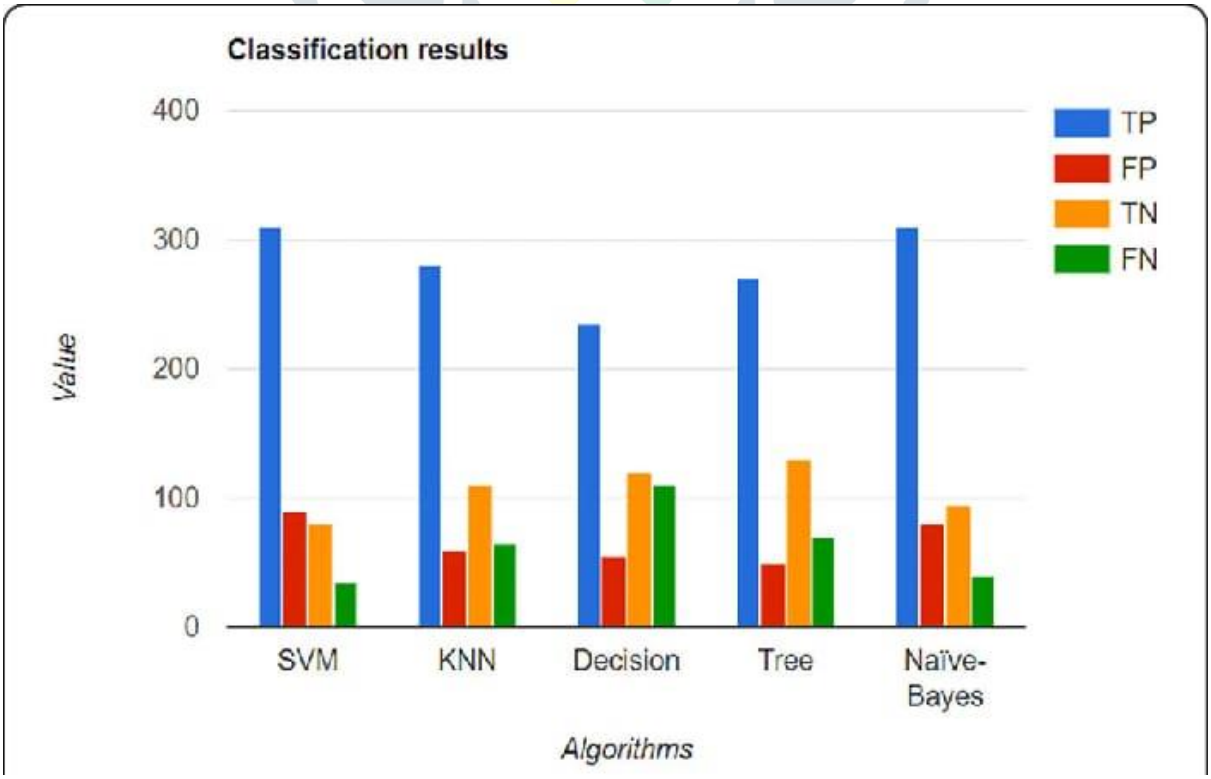

Fig-4 Bar Graph on the Accuracy of KNN , SVM, Decision tree, Navie Bayes.

The graph shows how four machine learning algorithms, SVM, KNN, Decision Tree, and Naive Bayes, compare their performances regarding True Positives, False Positives, True Negatives, and False Negatives. Here, the performance for True Positives is the highest, meaning high accuracy in the right classification of instances as positive; an almost balanced False Positives and False Negatives ratio; and the highest amount of True Negatives. Although high True Positives are achieved with KNN, False Positives and False Negatives are a little higher, which suggests that the model is more likely to misclassify its results. Decision Tree performs fairly well with good True Positives but equally negative False Positives and True Negatives just like KNN. True Positives reach high values in Naive Bayes, almost similar to SVM, but False Positives, False Negatives, and True Negatives are significantly low, which reflects that though it does a great job as a classifier for positive classification the negative classification might not be very sharp. SVM and Naive Bayes generally perform better positive classification whereas KNN and Decision Tree show more erroneous classification.

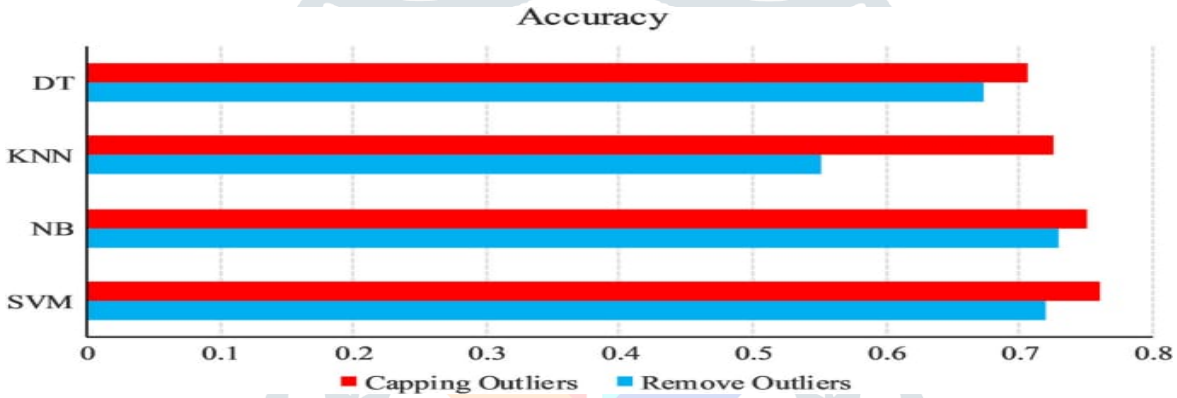

Fig-5 Comparing the accuracy of Four Algorithms.

A comparison graph between four different models that are being used: Decision Tree, K-Nearest Neighbors, Naive Bayes, and Support Vector Machine while applying outlier-handling techniques in two different methods: capping outliers (in red), and removal of outliers (in blue). The data, for the most part, shows that capping outliers generates higher accuracy than removing them for Decision Tree, Naive Bayes, and Support Vector Machine models. On the other hand, K-Nearest Neighbors goes against the inverse trend with very strong improvements in accuracy when removing outliers compared with capping. This tends to suggest that while outlier capping does improve the performance of most models generally, K-Nearest Neighbors is more outlier-sensitive and does significantly better when these are removed. Graph depicts why the choice of outlier handling techniques will have to depend on the specific nature of the machine learning model.

5.CONCLUSION

This promises SVM in modeling air pollution very much toward environmental analysis and prediction. It has the ability to model and predict the level of air pollution in complex and high-dimensional data based on varying environmental parameters like particulate matter, nitrogen dioxide, carbon monoxide, and meteorological data. With promising results in SVM outcomes, further optimization still lies ahead with the selection of appropriate features, model tuning, and incorporating larger data sets that may improve generalization. Finally, SVM technique has provided an effective means to solve problems surrounding air pollution, leading towards better environmental outcomes as well as further public health interventions.

References:

[1] Arepalli, P. G., Naik, K. J., & Amgoth, J. (2024). An IoT-based water quality classification framework for aqua-ponds through water and environmental variables using CGTFN model. International Journal of Environmental Research, 18(4), Article 73. https://doi.org/10.1007/s41742-024-00625-2

[2] & Pavani, V. (2020). Accurate identification and detection of outliers in networks using group random forest methodology. Journal of Critical Reviews, 7(6), 381-384. https://doi.org/10.31838/jcr.07.06.67

[3] Rao, B. T., Patibandla, R. S. M. L., Narayana, V. L., & Gopi, A. P. (2021). Medical data supervised learning ontologies for accurate data analysis. In Semantic Web for Effective Healthcare Systems (pp. 249-267). https://doi.org/10.1002/9781119764175.ch11

[4] Challa, R., YAMPARALA, R., KANUMALLI, S. S., & KUMAR, K. S. (2020, November). Advanced patient's medication monitoring system with ardunio UNO and NODEMCU. In *2020 4th International conference on electronics, communication and aerospace technology (ICECA)* (pp. 942-945). IEEE.

[5] Kanumalli, S. S., Chinta, A., & Chandra Murty, P. S. R. (2019). Isolation of Wormhole Attackers in IOV Using WPWP Packet. *Revue d'Intelligence Artificielle*, *33*(1)

[6] Kanumalli, S. S., Ch, A., & Murty, P. S. R. C. (2018). Advances in Modelling and Analysis B. *Journal homepage: http://iieta. org/Journals/AMA/AMA_B*, *61*(1), 5-8.

[7] Kosaraju, Chaitanya, et al. "A model for analysis of diseases based on nutrition deficiency using random forest." 2022 7th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2022.

[8] Chaitanya, Kosaraju, and Gnanasekaran Dhanabalan. "Secure Route Detection with Multi Level Trust Evaluation Model Using Replicated Auditor Node for Extended Packet Delivery Rate in WSN." Revue d'Intelligence Artificielle 37.4 (2023).

[9] Chaitanya, Kosaraju, et al. "Risk Stratification for Stroke Using Attention Transformer Model." 2024 2nd International Conference on Disruptive Technologies (ICDT). IEEE, 2024.

[10] Wang, A., Xu, J., Tu, R., Saleh, M., Hatzopoulou, M., 2020. Potential of machine learning for prediction of traffic related air pollution. Transp. Res. DTransp.Environ.88,102599. http://dx.doi.org/10.1016/j.trd.2020.102599.

[11] Chen, L., Mao, F., Hong, J., Zang, L., Chen, J., Zhang, Y., Gan, Y., Gong, W., Xu, H., 2022. Improving PM2.5 predictions during COVID-19 lockdown by assimilating multi-source observations and adjusting emissions. Environ. Pollut. 297, 118783. http://dx.doi.org/10.1016/j.envpol.2021.118783.

[12] Ke, H., Gong, S., He, J., Zhang, L., Cui, B., Wang, Y., Mo, J., Zhou, Y., Zhang, H., 2022. Development and application of an automated air quality forecasting system based on machine learning. Sci. Total Environ. 806, 151204. http://dx.doi.org/10. 1016/j.scitotenv.2021.151204. Kletting, P., Glatting, G., 2009. Model selection for time-activity curves: the corrected Akaike information criterion and the F-test. Z. Medizinische Physik 19 (3), 200–206.

[13] V., A., P., G., R., V., P., S.K., 2018. DeepAirNet: Applying recurrent networks for air quality prediction. Procedia Comput. Sci. 132, 1394–1403. http://dx.doi.org/10. 1016/j.procs.2018.05.068.

[14] S. Du, T. Li, Y. Yang, S.-J. Horng, Deep air quality forecasting using hybrid deep learning framework, IEEE Trans. Knowl. Data Eng. 33 (2019) 2412–2424. [3] H. Huang, X. Wei, Y. Zhou, An overview on twin support vector regression, Neurocomputing 490 (2022) 80–92.

[15] S. Zhu, X. Lian, H. Liu, J. Hu, Y. Wang, J. Che, Daily air quality index forecasting with hybrid models: a case in China, Environ. Pollut. 231 (2017) 1232–1244.

[16] Z. Zhang, S. Zhang, Modeling air quality PM2. 5 forecasting using deep sparse attention-based transformer networks, Int. J. Environ. Sci. Technol. (2023) 1–16.

[17] S. Zhang, X. Zhang, X. Zhao, J. Fang, M. Niu, Z. Zhao, J. Yu, Q. Tian, MTDAN: A Lightweight Multi-Scale Temporal Difference Attention Networks for Automated Video Depression Detection, IEEE Trans. Affect. Comput. (2023), https://doi.org/ 10.1109/TAFFC.2023.3312263

[18] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, X. Zhao, Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects, Expert Syst. Appl. (2023) 121692.

[19] H. Fan, X. Zhang, Y. Xu, J. Fang, S. Zhang, X. Zhao, J. Yu, Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals, Inf. Fusion (2023) 102161.

[20] Ravindra, K., 2019. Emission of black carbon from rural households kitchens and assessment of lifetime excess cancer risk in villages of North India. Environ. Int. 122, 201–212. https://doi.org/10.1016/J.ENVINT.2018.11.008.

[21] Ravindra, K., Singh, T., Pandey, V., Mor, S., 2020. Air pollution trend in Chandigarh city situated in Indo-Gangetic Plains: understanding seasonality and impact of mitigation strategies. Sci. Total Environ. 729, 138717 https://doi.org/10.1016/J. SCITOTENV.2020.138717.

[22] Bao, R., Zhang, A., 2020. Does lockdown reduce air pollution? Evidence from 44 cities in northern China. Sci. Total Environ. 731, 139052 https://doi.org/10.1016/J. SCITOTENV.2020.139052.

[23] Malhi, G.S., Kaur, M., Kaushik, P., 2021. Impact of climate change on agriculture and its mitigation strategies: a review, 2021 Sustain. Times 13, 1318. https://doi.org/ 10.3390/SU13031318. Page 1318 13.

[24] Li, L., Li, Q., Huang, L., Wang, Q., Zhu, A., Xu, J., Liu, Ziyi, Li, H., Shi, L., Li, R., Azari, M., Wang, Y., Zhang, X., Liu, Zhiqiang, Zhu, Y., Zhang, K., Xue, S., Ooi, M.C.G., Zhang, D., Chan, A., 2020. Air quality changes during the COVID-19 lockdown over the Yangtze River Delta Region: an insight into the impact of human activity pattern changes on air pollution variation. Sci. TotalEnviron.732,139282https://doi.org/10.1016/J.SCITOTENV.2020.139282.

[25] Gurjar, B.R., Ravindra, K., Nagpure, A.S., 2016. Air pollution trends over Indian megacities and their local-to-global implications. Atmos. Environ. 142, 475–495. https://doi.org/10.1016/J.ATMOSENV.2016.06.030.

[26] Guttikunda, S.K., Goel, R., Pant, P., 2014. Nature of air pollution, emission sources, and management in the Indian cities. Atmos. Environ. 95, 501–510. https://doi.org/ 10.1016/J.ATMOSENV.2014.07.006.

[27] Aarthi C, Ramya VJ, Falkowski-Gilski P, Divakarachari PBJS. Balanced spider monkey optimization with Bi-LSTM for sustainable air quality prediction. Sustainability. 2023;15(2):1637.

[28] Liu B, et al. A sequence-to-sequence air quality predictor based on the n-step recurrent prediction. IEEE Access. 2019;7:43331–45. 33.

[29] Wu Q, Lin H. Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network. Sustain. 2019;50:101657.

[30] . Lu H, Ma XJC. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere. 2020;249: 126169.

[31] Liu H, Li Q, Yu D, Gu YJAS. Air quality index and air pollutant concentration prediction based on machine learning algorithms. Appl Sci. 2019;9(19):4069.

[32] Parkavi P and Rathi S. LSTM Based Air Quality Prediction Model In Smart Cities. IEEE international conference on big data (big data) 2021.

[33] Kalapanidas, E.; Avouris, N. Short-term air quality prediction using a case-based classifier. Environ. Model. Softw. 2001, 16, 263–272.

[34] Corani, G. Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning. Ecol. Model. 2005, 185, 513–529.

[35] Ni, X.Y.; Huang, H.; Du, W.P. Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data. Atmos. Environ. 2017, 150, 146–161.

[36] Athanasiadis, I.N.; Kaburlasos, V.G.; Mitkas, P.A.; Petridis, V. Applying machine learning techniques on air quality data for real-time decision support. In Proceedings of the First international NAISO Symposium on Information Technologies in Environmental Engineering (ITEE'2003), Gdansk, Poland, 24–27 June 2003.

[37] Kurt, A.; Oktay, A.B. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Syst. Appl.2010, 37, 7986–7992.

[38] Bekkar, A., Hssina, B., Douzi, S., Douzi, K., 2021. Air-pollution prediction in smart city, deep learning approach. J. Big Data 8 (1), http://dx.doi.org/10.1186/s40537-021- 00548-1.

[39] Liu, T., You, S., 2022. Analysis and forecast of Beijing's air quality index based on ARIMA model and neural network model. Atmosphere 13 (4), 512. http://dx.doi. org/10.3390/atmos13040512.

[40] Challa, R., YAMPARALA, R., KANUMALLI, S. S., & KUMAR, K. S. (2020, November). Advanced patient's medication monitoring system with ardunio UNO and NODEMCU. In *2020 4th International conference on electronics, communication and aerospace technology (ICECA)* (pp. 942-945). IEEE.

[41] Kanumalli, S. S., Chinta, A., & Chandra Murty, P. S. R. (2019). Isolation of Wormhole Attackers in IOV Using WPWP Packet. *Revue d'Intelligence Artificielle*, *33*(1)

[42] Kanumalli, S. S., Ch, A., & Murty, P. S. R. C. (2018). Advances in Modelling and Analysis B. *Journal homepage: http://iieta. org/Journals/AMA/AMA_B*, *61*(1), 5-8.

[43] Kosaraju, Chaitanya, et al. "A model for analysis of diseases based on nutrition deficiency using random forest." 2022 7th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2022.

[44] Sravanthi, G. L., Devi, M. V., Sandeep, K. S., Naresh, A., & Gopi, A. P. (2020). An efficient classifier using machine learning technique for individual action identification. International Journal of Advanced Computer Science and Applications, 11(6), 513-520. DOI: 10.14569/IJACSA.2020.0110664