How to Cite:
Kumbhar, V. S., Sohi, S. S., Jayaram, V., Sreelekshmy, P. G., Shukla, S. K., & Abhilash, K. S. (2022). Hybrid artificial neural network algorithm for air pollution estimation. *International Journal of Health Sciences*, *6*(S5), 2094–2106. https://doi.org/10.53730/ijhs.v6nS5.9080

Hybrid artificial neural network algorithm for air pollution estimation

Vijayalaxmi S. Kumbhar
Assistant Professor, PCET's Pimpri Chinchwad College of Engineering and Research, Ravet
Email: vijayalaxmi.kumbar@pccoer.in

Shaminder Singh Sohi
Assistant Professor, Chandigarh University, Gharuan, Mohali, Punjab
Email: Shaminder.e12325@cumail.in

Jayaram V
Research Scholar, Dept of Mechanical Engineering, Noorul Islam Center for Higher Education
Tamilnadu
Email: jayaramvijayan@gmail.com

Sreelekshmy Pillai G
Associate Professor In Civil Engineering, NSS College Of Engineering, Palakkad
Email: sreelekshmypillai@gmail.com

Dr. Surendra Kumar Shukla
Associate Professor, Department of Computer Science & Engineering, Graphic Era Deemed to be University, Dehradun, Uttarakhand, India, 248002
Email: surendrakshukla21@gmail.com

Dr. Abhilash. KS
Managing Director, EduCorp Centre for Research and Advanced Studies Pvt. Ltd. Thiruvananthapuram, Kerala
Email: dr.abhilashks@gmail.com

***Abstract*---**In recent years, airborne broadcasting has grown more prevalent in cities. Air quality degradation is a severe air pollution issue that exists daily. To forecast the amount of pollutants, Artificial Neural Network (ANN) and Linear Vector Quantization (LVQ) techniques were utilized. The data set dimensions are defined by the pre-processing procedure and the feature extraction mechanism. The ANN model predicts categorization concentration, allowing the LVQ model to classify direct situations with greater accuracy using explanatory factors. The ANN+LVQ model outperformed other

Manuscript submitted: 18 Feb 2022, Manuscript revised: 27 April 2022, Accepted for publication: 9 June 2022

technologies in terms of classification accuracy. The raw data was cleaned to improve the accuracy of the prediction algorithms. The pollutants discovered in the collection are NO_2, NO_x, O_3, Benzene, Xylene, NH_3, CO, SO_2, PM_{10}, NO, and Toluene. The performance of the recommendation and forecast models were tested in this study using two datasets in two distinct experiments. In urban, rural, and industrial settings, the proposed ANN model is successful in detecting air quality and predicting pollution levels. The ANN-LVQ model obtained 90% percent sensitivity, 97.59% accuracy, and 99.46% specificity with 2.43% error rate. The suggested model's accuracy is much greater than that of other current research models.

***Keywords*---**Air Quality Estimation, Hybrid Model, Artificial Neural Network, Pollutants, Linear Vector Quantization.

Introduction

Airborne diseases have become more common in cities in recent years. Smog caused by waste products such as Nitric Oxide (NO), Hydrocarbons, and Carbon Monoxide (CO), as well as synthetic compounds made from biological sources, has a negative impact on the environment. Artificial Neural Network (ANN) models have gained popularity in recent years as a way of determining and forecasting environmental air pollution. Chronic illnesses are caused by the worsening quality of the air in metropolitan areas. The use of an ANN model does not demand a thorough understanding of the dynamic connection between levels of air pollution and other explanatory factors. In latest years, the public has had more access to powerful and less difficult computer tools for the construction and deployment of ANN and their training algorithms. The prediction system for the Index of Air Quality (IAQ) helps to smart environments, where sophisticated sensor technologies may be employed to produce healthy living circumstances for occupants. To estimate air pollution, most environmental scientists employ ANN models. Between one input and output layer, ANN preserves a set of hidden layers. The layer that isn't visible analyses the data for the next layer then sends the results to the output layer of the previous layer. Several Machine Learning (ML) and ANN techniques have been introduced to resolve issues on air pollution prediction and forecasting [1]. ANN algorithms are useful in building such models because they can capture the nonlinear behavior of complex atmospheric systems. Using data from monitoring stations, the Air Quality Index (AQI) may be generated by concentrating on certain air pollutants for a set period of time [2]. The technique for converting air toxin concentration to an AQI differs due to the efficacy of various air toxins. AQI is a valuable tool for assessing air quality on a regular basis and is thus used to raise public awareness in both rural and urban areas [3].

New approaches in ANN are needed to explore the quality of the air in real frames and analyze elusive patterns based on the data acquired [4]. Air pollution is caused by high levels of air pollutants, which harm the environment by causing solid, liquid, and gaseous particles to remain in the atmosphere. Those variations in the biosphere have a detrimental impact on human health, limit visibility, and

disrupt the biosphere's natural equilibrium. A pollutant or poison is a particle or gas that occurs in the atmosphere and has a harmful influence on human health [5]. Air pollution has a variety of consequences on living things, including human and animal health, as well as the environment as a whole [6]. Human health and the environment, including animal life, are affected by various geographical locations, global temperature changes, and environmental variations. Air pollution has a variety of effects on life on Earth. Polluted air has a negative influence on all important industries, including health, agriculture, and the economy.

Feature selection is a method of selecting features from a set of options. The levels/amounts of ten different air contaminants were indicated using Multilayer Perceptron (MLP) [7]. The input data layer is exposed to the outside world to detect all associated signals. The gathered information is fed into a single concealed layer. The building of a logical model is automated using a neural network, which is an information-examination approach. The goal of Neural Network is to create scalable computer programs that can be modified as new data becomes available. The diverse neural networks offer a great deal of promise to use in a variety of sectors, including climate, medical, education, gaming, robotics, and much more. One of the most significant features of machine learning is the capacity of computers to learn rapidly and efficiently. There are several calculations that aid in the creation of Neural Network devices and strategies. Chauhan *et al* [8] implemented Convolutional Neural Network (CNN) model to observe the trends in air quality using the future forecast modeling. The first stage of this method focused on data preparation and analysis, while the second step is used to validate the model. Models have been used to appropriately categorize the data. The general notion is implemented using the scripting language and Python programming. Wang *et al* [9] proposed Chi-Square Test (CT) with Long Short-Term Memory (LSTM) to predict AQI. An LSTM network with hidden neurons used in the new method can perform long-term memory functions. According to the experimental data, the CT-LSTM technique can better reduce AQI prediction error.

Wardana *et al* [10] proposed a 1D CNN-LSTM network with 4 post-quantization approaches in the Tensor Flow framework. Model validity is preserved when dynamic range and float16 quantization are used, however latency is not much reduced. In terms of model size and latency, full integer quantization outperformed competitor TFLite models, albeit at the cost of model accuracy. For Raspberry Pi 3 Model B+ (RPi3B+) and Raspberry Pi 4 Model B edge devices, the proposed approach was created (RPi4B). The Raspberry Pi 4's more capable CPU resulted in lower latency. Zhang *et al* [11] fused multiple machine learning models for predicting the air quality. The focus feature group contains historical meteorological data, statistical data, date information, and polynomial fluctuations. Using the Light Gradient Boosting Machine (LGBM) model, 500 most important characteristics were choose from the supplemental data and feed them into the Gradient Boosting Decision Tree (GBDT) and LGBM models. They filtered the most significant 300 attributes using the eXtreme Gradient Boosting (XGBoost) model and fed them into the three prediction models.

Sharma *et al* [12] proposed Recurrent Neural Network (RNN-LSTM) model for predicting the AQI of a location. The RMSE value of AQI is calculated using the ARIMA model. The time-series data of various pollutant concentration levels was fed into a RNN. Zhou *et al* [13] proposed a Deep Multi-output LSTM (DM-LSTM) model that was incorporated with three deep learning algorithms. Regional multi-step forward air quality forecasts were improved using a DM-LSTM. The DM-LSTM model has the potential to dramatically reduce time-lag phenomena and solve the problem of over fitting. DM-LSTM was able to successfully capture the heterogeneities in various air pollutant-producing systems. Nawahda *et al* [14] proposed a fuzzy classifier with quantification algorithm for solving the air quality monitoring problem in e-noses. All of the variables around e-nose sensors should be considered in the identification and quantification models. A confidence evaluation is performed to assess the accuracy of all classifier predictions. Only samples with a high level of confidence were sent to the third step, where quantification was performed. The efficacy of the characteristics utilized in e-nose classification has a big impact on its success. 16 MOX sensors from the FCQM type were used in this research. Bai *et al* [15] analyzed the air pollution prediction methods and classified using statistical forecasting techniques, artificial intelligence methodologies, and numerical forecasting techniques. In addition to traditional methodologies, hybrid approaches were applied to enhance air pollution forecasting. Artificial intelligence technologies like the neural network approach boosted forecasting and computing skills even with non-linear and unbalanced information.

Methodology

ANN is a mathematical model that aids in the organization and/or functionality of a biological Neural Network (NN). The model frequently complicates the input and output linkages or employs them to neutralize data patterns. Neurons are commonly divided into three levels. A group of neurons at the input level receives input from a user agent. The output level of a neuron is the level at which it provides data to the user agent. The hidden layers are located between the two levels. The architecture of ANN is illustrated Figure 1.

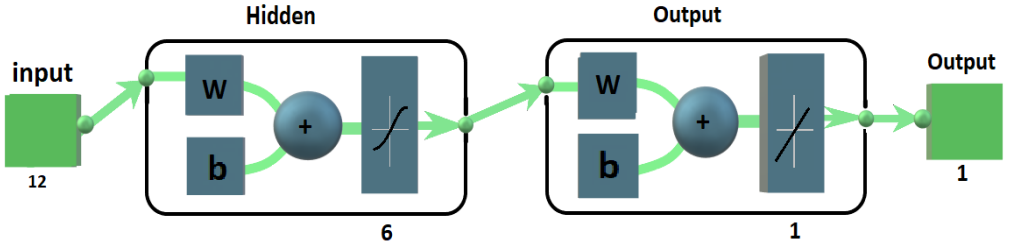

Figure 1: Architecture of ANN

The neurons in the hidden layer are shared by all neurons and do not directly interact with the user agent. Each neuron has the power to alter NN-carrying processing in a specific way. Although the output and input levels are crucial, the layers can also serve as entry and exit points. The NN employs numerous epochs to select the classified categories as training progresses. In this recommended

strategy, weighting factors are specified to use the LVQ methodology for prediction. The input and output levels, respectively, are made up of nodes pertaining to input and output variables. Data is transported between these tiers via weighted links. The flowchart of proposed ANN-LVQ classifier is illustrated in Figure 2.

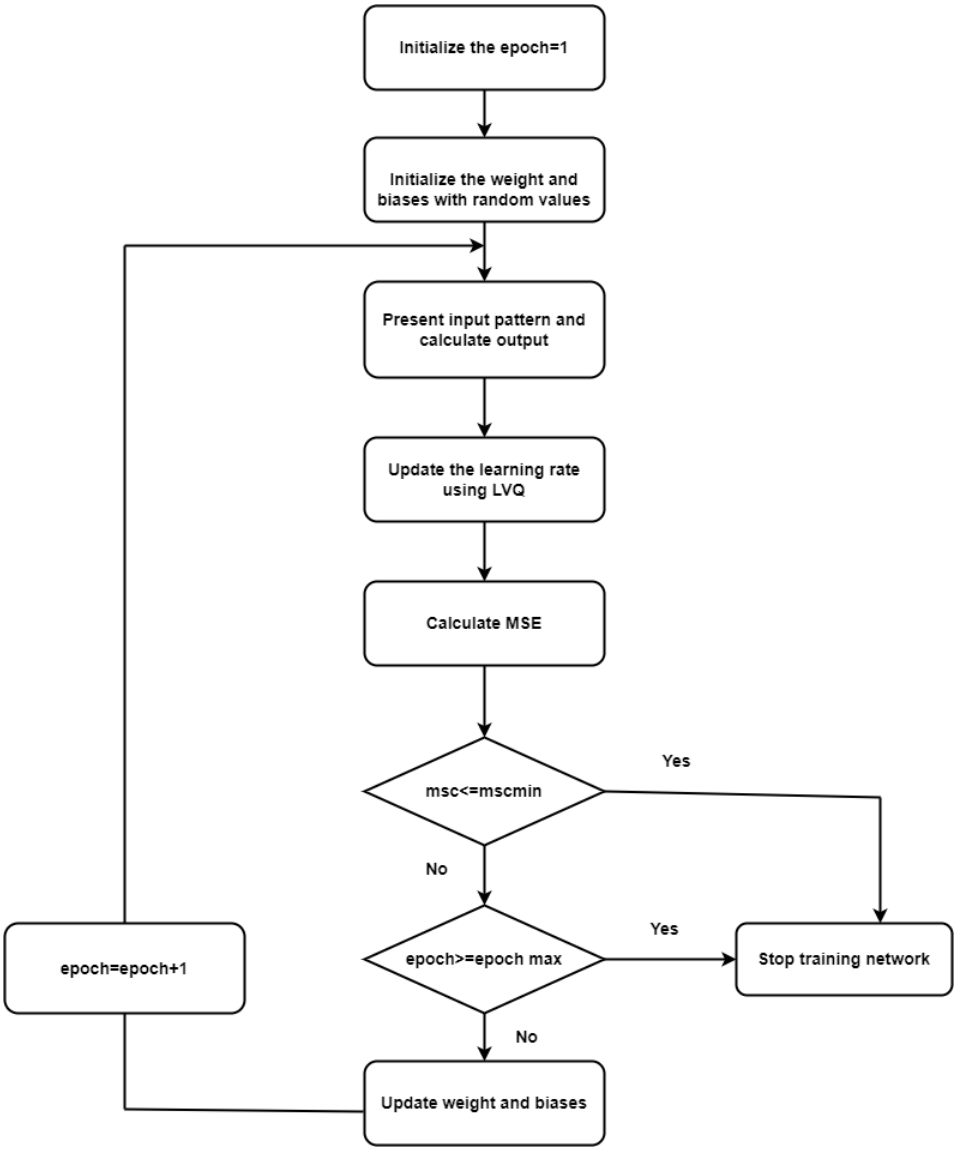

Figure 2: Flowchart of Proposed ANN-LVQ

The NN employs numerous epochs to select classified categories as training progresses. Training determines the connection weights. The count of training cycles starts with random numbers being added to the weight. In this technique, the weighting parameters are set using the LVQ methodology for prediction. A

back propagation computation is included in the ANN and is represented in Eqn 1.

$$x_i = \sum_{j-1}^{n} \Delta W_{i.jX_j} \tag{1}$$

Where, the number of inputs is defined as n, the weight of the connection between nodes i and j is represented as ΔW, meanwhile, input from node j is defined as X. In this equation, the LVQ technique is used to choose ΔW.

LVQ is a controlled learning system based on the scale of vectors and a regulated form of ANN for classification. Select a positive value for the learning speed parameter (α = 0.9) and set the initial synaptic weight for minor random values with intervals of [0, 1]. Alpha learning speeds were varied from 0.1 to 0.9 to achieve maximum efficiency the alpha must be 0.9. Next, the input vector x_i was started from the random data space. The LVQ will travel in the direction xi if the class label x and the weight vector ΔW are near. If the labels for class x have differing average values, the weight vector ΔW will shift away from x_i. Assuming that the weight vector in parallel $\Delta(t)$ is close to the input vector x_i , the mathematical equation representation is as follows for ΔW Eqn. 2.

$$\Delta(t) = \arg\min \| X_i - \Delta W(t) \| \tag{2}$$

Let CW stand for the class of $\Delta(t)$, and Cx_i for the class of x_i. If both classes are similar, then CW = Cx_i, the weight vector $\Delta(t)$ is modified and ANN parameters are written in Eqn. 3.

$$\Delta(t+1) = W(t) + \eta(t)(x_i - W(t)) \tag{3}$$

The learning rate of the adaptation technique is given by (t). If both classes are different, then $CW \neq Cx_i$, then ANN is denoted by Eqn. 4. Based on the condition, either Eqn. 2 or Eqn. 3 can be used to update the weighting function in ANN. Using the weighting function, the output of the predicted value is determined by Eqn. 4.

$$y_i = x_i W(t+1) \tag{4}$$

The output y_i predicts the AQI from the input pollutants based on six classes namely good, moderate, poor, satisfactory, severe and very poor. The process flow of proposed AQI prediction system is depicted in Figure 3.

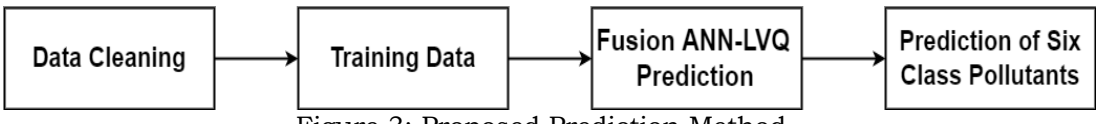

Figure 3: Proposed Prediction Method

The ANN-LVQ is a neural network with a single feed architecture consisting of an output and input unit. The AQI classification that results is solely determined by the reserves as among effort vectors. The struggle layer places dual participating vectors in the same class if the two input vectors are relatively close. The dataset put aside for this concept during the pre-processing step was used to test the hybrid NN model illustrated in Figure 4.

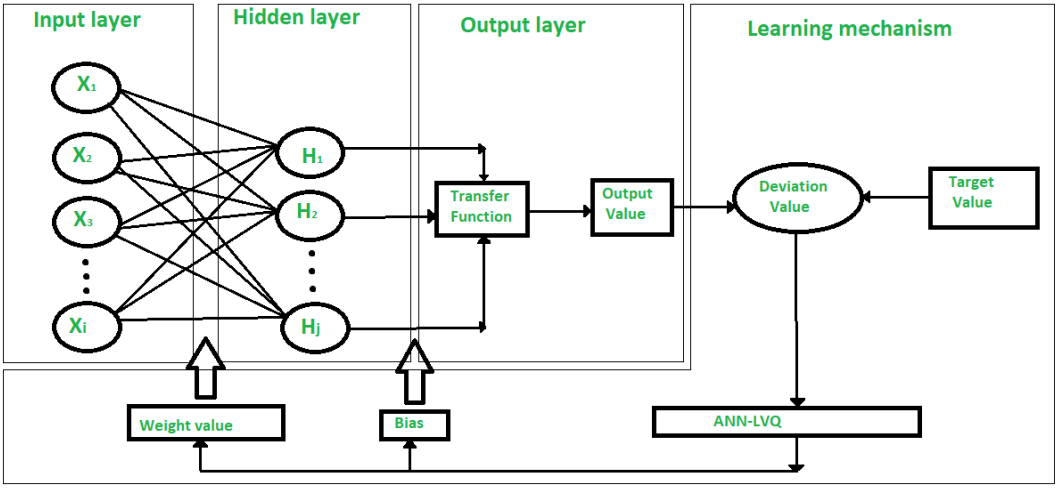

Figure 4: Proposed Hybrid ANN and LVQ Model

Dataset Description

This research utilized the Kaggle data set. The data set includes air quality data from numerous monitoring stations around India on an hourly and daily basis, as well as AQI data for twenty cities and 230 monitoring stations. Amaravati, Ahmedabad, Aizawl, Amritsar, Bhopal, Bengaluru, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulum, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kolkata, Kochi, Lucknow, Mumbai, Patna, Shillong, Thiruvananthapuram, Talcher and Visakhapatnam are among the cities taken into consideration. To increase the accuracy of the prediction approaches, the raw data were treated with a data cleaning procedure. The data set for present study includes 12 pollutants, namely Nitrogen Oxide (NO), Nitrogen dioxide (NO_2), Nitric Oxide (NO_x), Carbon monoxide (CO), Ammonia (NH_3), Particulate Matter (PM2.5, PM10), Sulphur Dioxide (SO_2), Benzene (C_6H_6), Xylene (C_8H_{10}), Ozone (O_3) and Toluene (C_7H_8).

Data Cleaning

The raw data was cleaned to ensure that it was appropriate for the descriptive phase of the final analysis. Any errors in the dataset that may have affected the prediction model were detected and deleted using the data cleaning procedure. Missing data, noise-related data, and imbalanced data were among the mistakes. The columns with null values have been eliminated, and the values have been normalized. As a result, a dataset that had undergone data cleaning was error-

free and could be used as input for the proposed model. The two most crucial phases in data cleaning are feature extraction and pre-processing.

Pre-processing

There would be a significant amount of information records and attribute numbers in the input data collection. The number of characteristics is determined by the number of dimensions, and the data set dimension must be defined by a pre-processing procedure. If the data set dimension is defined, the procedure uses feature extraction to read the data point value. The approach treats the observed missing data point as noisy input and eliminates the data point from the data collection. The flow chart of preprocessing is depicted in Figure 5.

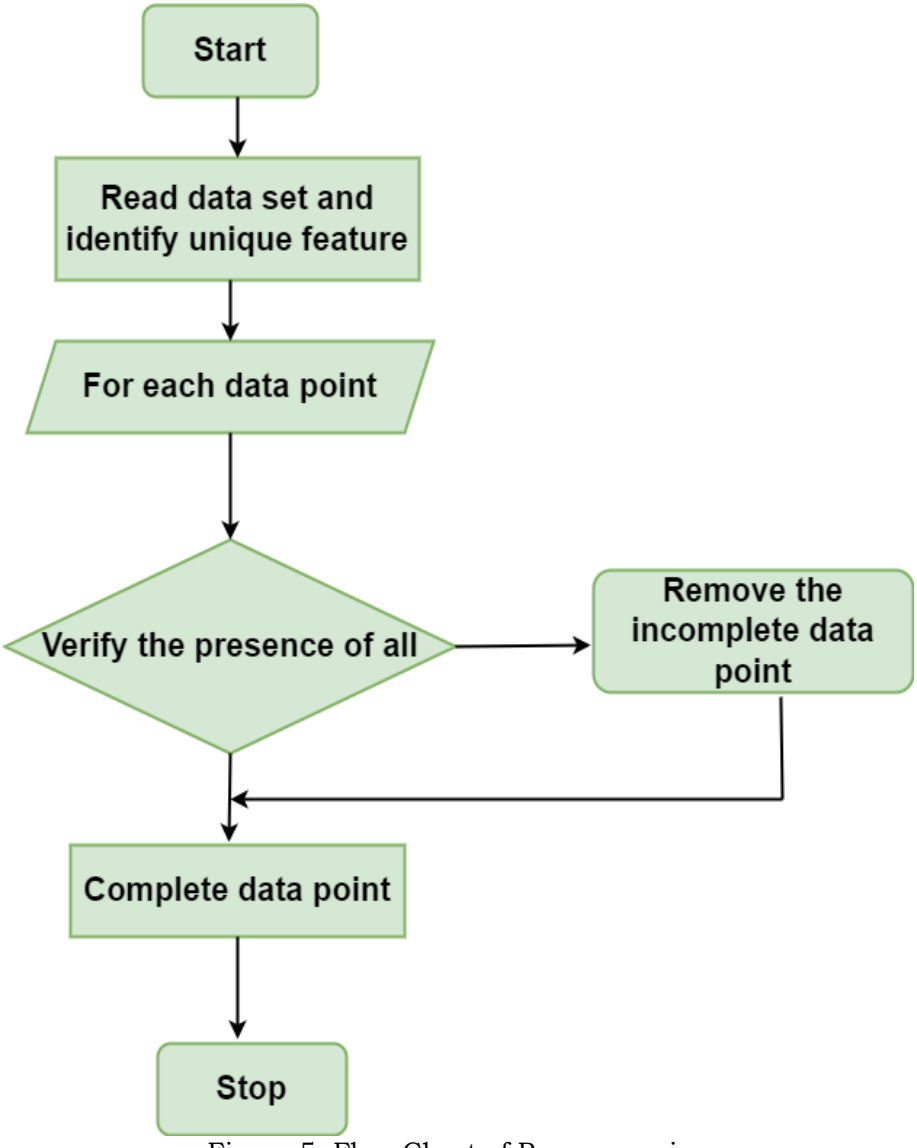

Figure 5: Flow Chart of Pre-processing

Feature Extraction

Figure 6 shows the process flow of feature extraction. Feature extraction is a technique for removing needed characteristics from the stage of input results. The data set is divided into three sections: testing, training, and validation. Instead of remembering an outcome, the training dataset is used to see if the model has extended to other datasets. Cross-validation is a technique that involves splitting datasets into subgroups numerous times. The model has evaluated each data point to eliminate bias, and the variance has been computed several times.

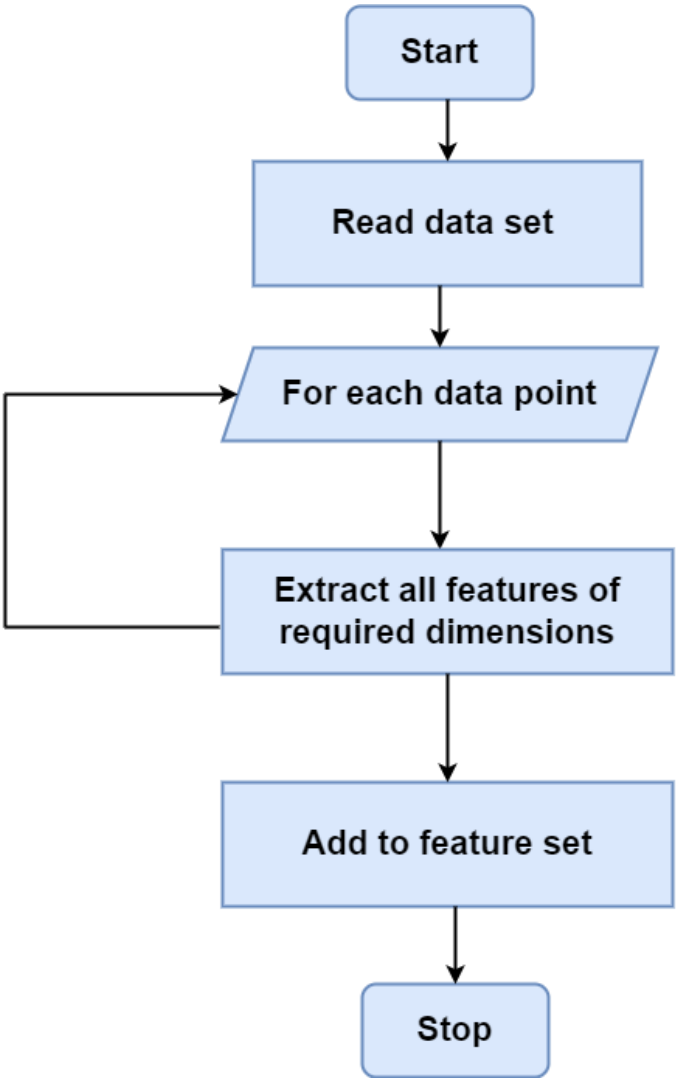

Figure 6: Flow Chart of Feature Extraction

Experimental Results

Two separate studies used two datasets to validate the prediction and recommendation algorithm's performance. As a consequence, the algorithms were created in MATLAB Version 18.a on a system with a 1.8GHz Pentium IV CPU. Sensitivity, Accuracy, false omission rate, specificity, false discovery rate, and error rate were all used to validate the data. The accuracy rate is the percentage of correct predictions in a given number of predictions, whereas the error rate is the percentage of incorrect guesses in a given number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{5}$$

$$Error\,Rate = \frac{FP + FN}{TP + TN + FP + FN} \times 100 \tag{6}$$

The terms sensitivity and specificity are used to understand the link between the system's input and output variables, as well as to verify the model's resilience in the face of uncertainty. Sensitivity represents the True Positive (TP) rate, while specificity represents the True Negative (TN) rate.

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \times 100 \tag{8}$$

The ratio of expected False Positive (FP) values to the total number of true and FP values is known as the False Positive Rate (FPR).

$$FPR = \frac{FP}{TP + FP} \times 100 \tag{9}$$

FOR refers to the ratio of the predicted False Negative (FN) values to the total number of true and false negative values.

$$FNR = \frac{FN}{TN + FN} \times 100 \tag{11}$$

Table 1: Performance of ANN- LVQ

	Parameters (%)					
Methodology	Accuracy	Sensitivity	Specificity	Error Rate	FPR	FOR
Adaptive ANN	57. 82	60.49	70.56	42.20	75.75	5.12
Traditional ANN	90.08	81.92	92.42	9.94	12.68	7.08
ANN-LVQ	97.59	90	99.46	2.43	.13	0.08

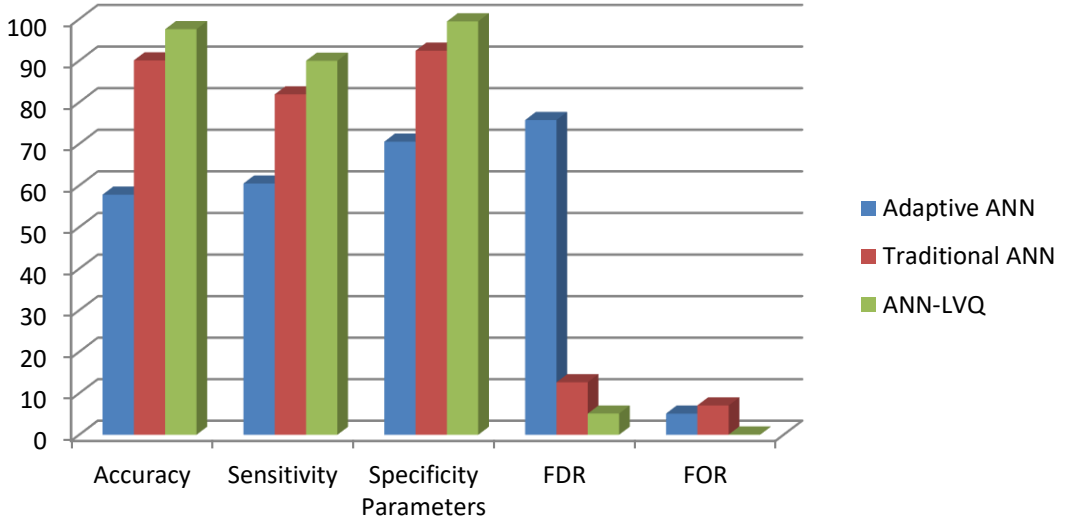

Figure 7: Graphical Illustration Classifier Performance of ANN-LVQ

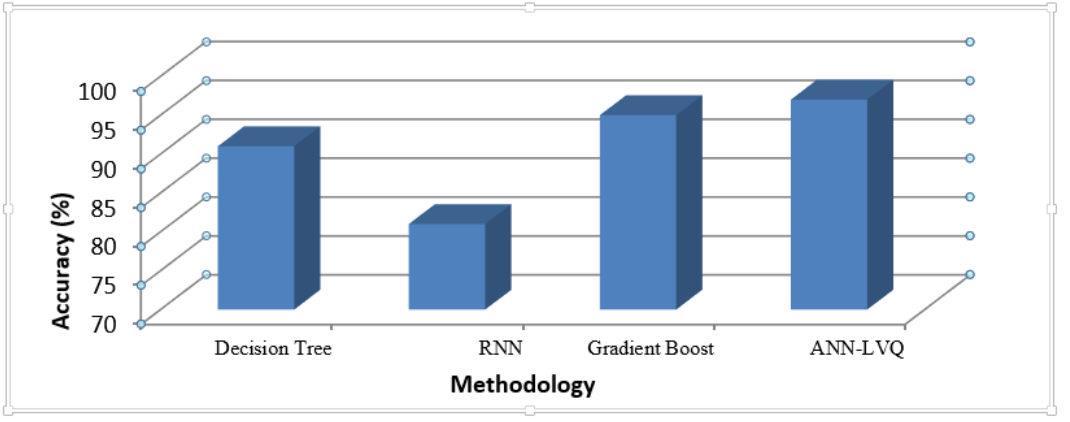

Figure 8: Comparison of the Classifier Accuracy

According to the findings, the adaptive ANN model performed worse than the other two models. The adaptive ANN model, for example, had an accuracy of only 57.82% and a 42.20% error rate. Furthermore, the regular feed-forward ANN outperformed the adaptive ANN in terms of lowest error rate, FNR, and FPR. The adaptive ANN model only had 60.49% sensitivity and 70.56% specificity, whereas the conventional ANN model had 81.92% sensitivity and 92.42% specificity. The

standard ANN model, on the other hand, was projected to have a 7.07% error rate and a long calculation time. LVQ was chosen as the ANN model's weighting function in this investigation. This model has 97.59% accuracy, 0.08 percent FNR, 2.43% error rate, 90% sensitivity and 99.46% specificity. The accuracy of the proposed model (ANN –LQV) was compared to that of the other models developed by various researchers in Figure 8.

The right input-output ratio was determined using a variety of selection methods. To begin, the technique sorted the variables based on specified indications and eliminated the factors that were not as relevant. Based on the comparison of the ANN-LVQ technology with the other technologies in Table 2 demonstrated that ANN-LVQ scored the highest accuracy (97%) than that of the Decision Tree and Naive-based, RNN and Gradient descent algorithms because the ANN model predicts the concentration of classification so that the LVQ model classifies the direct conditions with higher accuracy based on explanatory variables.

Conclusion

Air quality prediction is a complex endeavor due to its dynamic nature. Air quality, particularly in metropolitan areas, is gradually altering. The prediction system for the Index of Air Quality (IAQ) helps to smart environments, where sophisticated sensor technologies may be employed to produce healthy living circumstances for occupants. There are a few air toxins that are harmful to human health such as Sulphur Dioxide (SO2), Carbon Monoxide (CO), Respirable Suspended Particulate Matter (RSPM), Nitrogen Dioxide (NO2), Particulate Matter (PM) (2.5 and 10), and Ozone (O3). High levels of air pollution have detrimental health effects, such as an increased risk of asthma, giddiness, heart failure, and other diseases. All major industries, including health, agriculture, and the economy, are negatively impacted by air pollution. In addition to human health, air pollution has a range of environmental consequences, including acid rain, haze, eutrophication, effects on animals, forest loss, ozone depletion, and global climate change. Using data from monitoring stations, the AQI may be generated by concentrating on certain air pollutants for a set period of time. Depending on the toxin, the Air Quality Index (AQI) fluctuates. The ANN model is effective in detecting air quality and predicting pollution levels in rural, urban, and industrial areas. According to the findings, the ANN-LVQ model has 90 percent sensitivity, 97.59% accuracy, and 99.46% specificity with a 2.43% error rate. The suggested model's accuracy is much greater than that of other recent research models. As a consequence, data from additional smart cities might be analyzed using this model for air pollution studies.

References

1. F. Bre, J. M. Gimenez, and V. D. Fachinotti, “Prediction of wind pressure coefficients on building surfaces using artificial neural networks,” Energy Build., 2018, doi: 10.1016/j.enbuild.2017.11.045.
2. N. H. A. Rahman, M. H. Lee, M. T. Latif, and Suhartono, “Forecasting of air pollution index with artificial neural network,” J. Teknol. (Sciences Eng., 2013, doi: 10.11113/jt.v63.1913.

3. A. Azid, H. Juahir, M. T. Latif, S. M. Zain, and M. R. Osman, “Feed-Forward Artificial Neural Network Model for Air Pollutant Index Prediction in the Southern Region of Peninsular Malaysia,” J. Environ. Prot. (Irvine,. Calif)., vol. 04, no. 12, pp. 1–10, 2013, doi: 10.4236/jep.2013.412a1001.
4. “Vapnik, V. (2013). "Introduction to Artificial Neural Network Theory", The nature of statistical learning
5. A. Alimissis, K. Philippopoulos, C. G. Tzanis, and D. Deligiorgi, “Spatial estimation of urban air pollution with the use of artificial neural network models,” Atmos. Environ., vol. 191, pp. 205–213, 2018, doi: 10.1016/j.atmosenv.2018.07.058.
6. P. A. Rahman, A. A. Panchenko, and A. M. Safarov, “Using neural networks for prediction of air pollution index in industrial city,” 2017, doi: 10.1088/1755- 1315/87/4/042016.
7. A. Challoner, F. Pilla, L. Gill, G. Adamkiewicz, and M. P. Fabian, “Prediction of Indoor Air Exposure from Outdoor Air Quality Using an Artificial Neural Network Model for Inner City Commercial Buildings,” mdpi.com, 2015, doi: 10.3390/ijerph121214975.
8. Chauhan, R., Kaur, H., & Alankar, B. (2021). Air quality forecast using convolutional neural network for sustainable development in urban environments. *Sustainable Cities and Society*, *75*, 103239.
9. Wang, J., Li, J., Wang, X., Wang, J., & Huang, M. (2021). Air quality prediction using CT-LSTM. *Neural Computing and Applications*, *33*(10), 4779-4792.
10. Wardana, I. N. K., Gardner, J. W., & Fahmy, S. A. (2021). Optimising deep learning at the edge for accurate hourly air quality prediction. *Sensors*, *21*(4), 1064.
11. Zhang, Y., Zhang, R., Ma, Q., Wang, Y., Wang, Q., Huang, Z., & Huang, L. (2020). A feature selection and multi-model fusion-based approach of predicting air quality. *ISA transactions*, *100*, 210-220.
12. Sharma, A., Mitra, A., Sharma, S., & Roy, S. (2018, October). Estimation of air quality index from seasonal trends using deep neural network. In *International Conference on Artificial Neural Networks* (pp. 511-521). Springer, Cham.
13. Zhou, Y., Chang, F. J., Chang, L. C., Kao, I. F., & Wang, Y. S. (2019). Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. *Journal of cleaner production*, *209*, 134-145.
14. Nawahda, A., & Zhong, J. Classification with Quantification for Air Quality Monitoring.
15. Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air pollution forecasts: An overview. *International journal of environmental research and public health*, *15*(4), 780.