Utkal University Journal of Computing & Communications,
Vol. 2, Issue. 1 (2024) *e-ISSN:2584-203X*

Air Quality Prediction Using ANN and RNN - Comparative Study

Ebin Joy[1], Akhil Raveendran Pillai[1]
[1]Department of AI & Automation, University West, Trollhattan

ebin.joy@student.hv.se,
akhil.raveendran-pillai@student.hv.se

Abstract: This work aims to predict Absolute Humidity (AH) using Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN). The dataset is first preprocessed by removing columns that contain unused, null, or negative values. Noise symbols are then stripped from the remaining data, which is subsequently converted to a floating-point format to facilitate processing. Finally, feature scaling is applied to enhance model performance. The dataset includes different sensor values like carbon monoxide (CO), benzene (C6H6), nitrogen oxides (NOx), non-methane hydrocarbons, and temperature, all of which influence AH — a key measure for weather forecasting and environmental monitoring. Both ANN and RNN models are used to capture complex relationships and patterns in the data. The final step involves evaluating both models using performance metrics like Mean Absolute Error (MAE) and Mean Squared Error (MSE) to select the one that performs the best.

Keywords: RNN, ANN, MSE, MAE

1. Introduction

The Air Quality UCI Dataset is one of the critical sources of data to determine the specifics of air quality. The data set contains multiple sensor values such as carbon monoxide (CO), benzene (C6H6), nitrogen oxides (NOx), non-methane hydrocarbons, and absolute humidity (AH), which is a key variable for measuring air quality. Absolute humidity (AH) is the measure of water vapor in the air and is influenced by factors such as temperature, pollution levels, and weather conditions. This work aims to clean the data and prepare for the modeling, train, and tune the Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN), compare the model efficiency metrics.

This paper is structured as follows: Section 2 reviews existing works, Section 3 contains preliminary data analysis. The proposed model is introduce in Section 4. Section 5 discuss the model performances and future work. Finally, Section 6 conclude with a highlight of the research result and future scope.

2. Literature Survay

The objective of this research is to predict Absolute Humidity (AH) using deep learning models, specifically Artificial

✉ ebin.joy@student.hv.se

Neural Networks (ANN) and Recurrent Neural Networks (RNN) [1]. To achieve this, the dataset needs to be preprocessed and cleaned thoroughly before being used for training. The requirements are categorized into functional and non-functional areas [11].

Recent advances in machine learning have significantly enhanced the capability to predict air quality by leveraging both ANNs and RNNs. Early work laid the foundation by applying conventional neural network models to forecast air pollution levels, while recent studies have shifted toward more sophisticated architectures, such as long short-term memory (LSTM) networks and hybrid models, to capture temporal and spatial dependencies in the data.

Initial comparative studies, such as the one by Ghosh et al. (2007) [4], explored various ANN configurations for air quality prediction, establishing a baseline for performance and highlighting the potential of neural networks in this domain. These foundational works provided insights into how feature selection and network architecture could influence prediction accuracy, setting the stage for further enhancements with more advanced models.

The progression from traditional ANNs to recurrent architectures marks a significant evolution in air quality forecasting. For example, studies by Ghosh et al. (2019, 2022) [1, 3] and Hossain et al. (2018) [2] introduced LSTM and other RNN variants that can effectively capture long-term temporal dependencies inherent in air quality time series. Hossain et al. (2018) specifically demonstrated the utility of backpropagation through time in training these models, while Ghosh et al. (2022) extended these approaches with LSTM networks optimized for real-world data streams.

Further improvements in model performance are evident in the work of Ghosh et al. (2022) [5], where an optimized RNN framework was developed to address both the non-linear nature and the dynamic behavior of pollutant concentrations. The use of adaptive architectures in these studies illustrates the ongoing trend toward leveraging deep learning to improve prediction accuracy in complex environments.

Recognizing the importance of spatial relationships in urban air quality, recent research has also integrated graph-based approaches with recurrent networks. Le (2023) [6] and Xu et al. (2023) [7] propose spatiotemporal graph convolutional and dynamic graph neural networks that adapt edge attributes, thereby capturing the interdependencies between different geographical locations. These studies represent a new frontier in the field, as they combine the strengths of spatial modeling with the temporal learning capabilities of RNNs.

Similarly, Han et al. (2020) [8] have explored multi-adversarial spatiotemporal networks that jointly predict air quality and weather parameters, indicating a move toward more holistic and interconnected prediction systems. Hybrid models, such as those presented by Bui et al. (2018) [9] and Zhang et al. (2020) [10], combine convolutional neural networks (CNNs) with LSTM or Bi-LSTM frameworks to leverage both local feature extraction and sequential learning, thereby enhancing overall model robustness.

Beyond the aforementioned RNN-based strategies, additional recent studies provide further insights into the comparative performance of different deep learning architectures. For instance, Li et al. (2020) [12] introduced a hybrid deep learning model that effectively balances temporal forecasting with spatial heterogeneity, while Zhang et al. (2021) [13] demonstrated the efficacy of CNN-based approaches in predicting urban air quality in Beijing. Kumar and Singh (2022) [14] expanded on these ideas by applying ensemble learning techniques, which have shown promise in improving the stability and accuracy of predictions in smart city applications.

Nguyen and Tran (2022) [15] offer a detailed comparative analysis of deep learning models across different regions in Southeast Asia, highlighting the challenges and potential of transferring models across varying environmental contexts. Further, Garcia and Lopez (2023) [16] and Patel and Mehta (2023) [17] emphasize the benefits of deep neural network approaches and spatiotemporal learning in capturing the fine-grained dynamics of pollutants, suggesting that the integration of real-time data streams with advanced modeling techniques can lead to substantial

improvements in forecasting performance.

The surveyed literature indicates a clear evolution from traditional ANN-based methods to more advanced RNN architectures and hybrid models that integrate spatial and temporal data. While early studies established the feasibility of using neural networks for air quality prediction, recent research has focused on optimizing these models to handle the complex, non-linear, and spatiotemporal nature of air pollution data. This body of work underscores the importance of comparative studies—like the current paper—to evaluate the trade-offs between ANN and RNN approaches, thereby guiding future research toward more robust and accurate prediction systems.

3. Data Analysis

3.1. Data Processing

I. Preprocessing: The dataset must be cleaned by removing unused columns, dealing with null values, and handling any negative values to ensure the quality and consistency of the input data.

II. Noise Removal: It is important to remove any irrelevant symbols or noise from the data that might interfere with model training. The dataset has CO(GT), C6H6(GT), T, RH, AH have noise symbols which would be difficult for model to parse.

III. Feature Scaling: CO levels are labeled in parts per million (ppm) whereas NOx concentrations are labelled in parts per billion (ppb) and temperature is in degrees Celsius. The values of the corresponding feature can be within a wide range.

IV. Datatype Conversion: Datatype Conversion is achieving the right data type for all features in the dataset. Consistency of input data is therefore very important when it comes to feeding data into the models in any of the machine learning workflows. All the sensor data are converted to Float 64

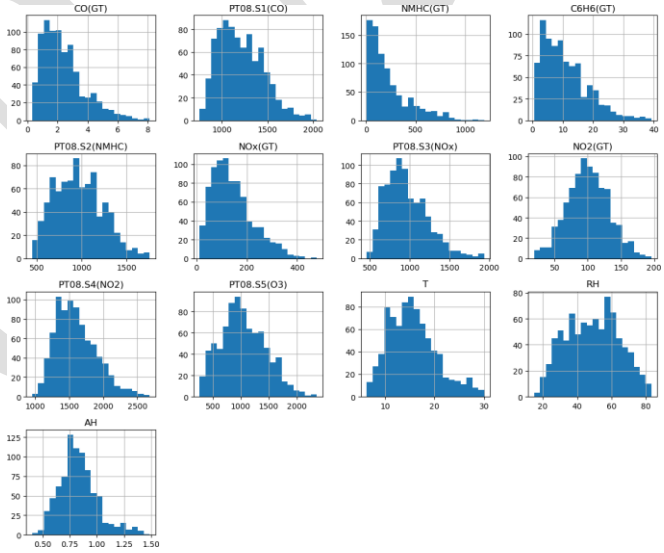

Fig 1. Histogram of Dataset

V. Histogram: The histograms in the Fig 1 illustrate how the data values are distributed. For example, as for the CO and NOx, the pollution levels are normalized and there are only several extreme values. We also

notice that temperature and humidity values are much less skewed towards the extremities and are centered more closely around the middle. The Absolute Humidity (AH) is mainly slightly below the average value. These graphs give us an idea of how the plotted data looks on the graphs as whether most of the values are clustered or if there are some extraordinarily high or low values that need special consideration.

3.2. Correlation Heat map

The heatmap shown in Figure 2 distinguishes from the previous one because it illustrates the dependence of various air quality indicators (CO, NOx, temperature) on each other. Each number reflects how closely tied the two measurements are. If the number is closer to 1 they have either a direct positive correlation while numbers closer to -1 reflect a negative correlation. For instance, movement up or down of the curve of the CO and C6H6 are in the positive direction, implying that if one of them rises, the other also will rise. Interestingly, AH has some correlation with CO and NO2 but not as close as the previous two parameters. This tells us how various parameters in the air interact in relation to the other.

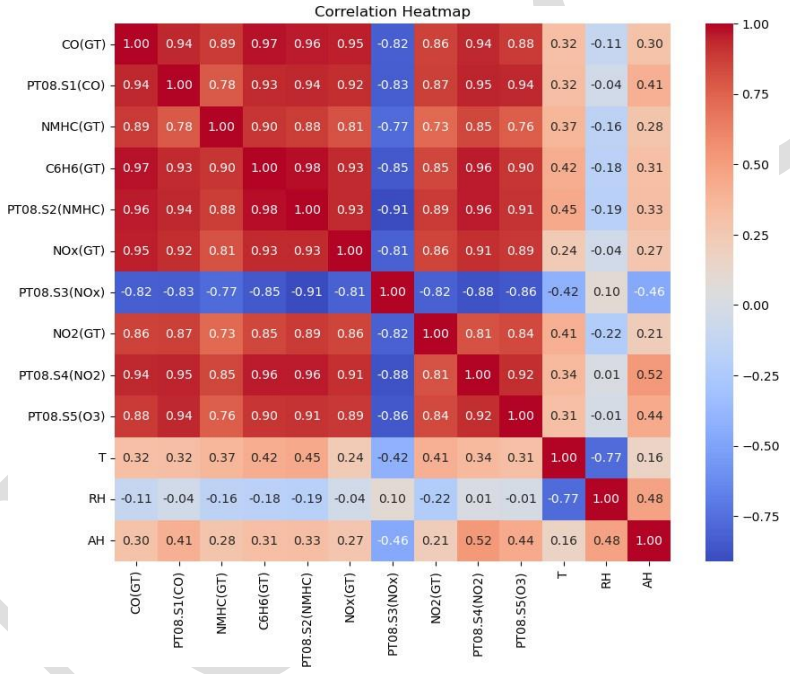

Fig 2. Correlation Heat Map of the Used Dataset

4. Proposed Model

The process of model development involves creating, implementing, and training several models to address the target problem which in this case is determining the absolute Humidity (AH) of the air. In this work, we focus on two powerful machine learning architectures: In the two subcategories of Neural Networks, there are Artificial Neural Networks (ANN) as well as Recurrent Neural Networks (RNN).

4.1. Artificial Neural Network (ANN)

Artificial Neural Network is a feedforward model with two hidden layers added to the prediction tasks using ReLU activation and drop-out layers to avoid overfitting and an output layer for regression tasks. This is because the target variable is predicted by training the model on the pre-processed data and the early-stopping was used during the

training. This work utilizes 128 neurons and ReLU activation for input features and 64 neurons with ReLU activation for hidden layer and single neuron for final prediction. The model is trained with the Adam optimizer which help in tuning the learning rate during the training process while using Mean Squared Error (MSE) as the loss function as its a regression algorithm. Moreover, to avoid over-training the early stopping technique is used to ensure the model does not continue for more epochs when the validation loss does not improve for the next 10 epochs, and the weights are replaced with the best weights for the model. Training of the model is done under a maximum of 50 epochs with a fixed batch size of 32. ANN models are perfect for capturing non-linear relations between inputs and outputs which is crucial particularly when using predictors of AH that depict environmental factors. Capable of handling multi-dimensional data, such as multiple features simultaneously. ANN is a flexible model that can learn complex patterns in data, especially in cases where the relationship between input features and the target variable is complex.

4.2. Recurrent Neural Network

Recurrent Neural Network (RNN) captures the temporal dependencies in the data such as reshaping the input features into sequences, enabling the RNN to process time-series data which makes it suitable for predicting Absoule Humidity (AH). Here we use 128 units of neuron and ReLU activation function and return sequence true which full output is passed to the next layer. The dropout rate is 0.2 to prevent overfitting. The second layer SimpleRNN layer uses 64 units of neurons with ReLU activation. The final layer is the Dense layer with unit=1 and activation=linear which gives us the predicted AH value.

5. Performance Evaluation

The performance of the Artificial Neural Network (ANN) and Recurrent Neural Network (RNN) models is evaluated based on two key metrics: and Mean Absolute Error (MAE). These metrics give information about how well Absolute Humidities (AH) are indicated by the models as against the actual values.

5.1. ANN Performance

The performance of ANN like Loss curve and True value are shown in the Figure 3 and 4. The ANN model reached to standard loss of 0.0011 and the MAE reached up to 0.0249. In the loss value, the smaller the better, it indicates that the model is close to the target value, where loss value show the mean squared error between the output and the real values. MAE stands for mean absolute error which reflects the mean of the absolute differences between the predicted and actual values equal to 0.0249 that says that on average the model predicts that AH will be 0.025 units off. This indicates that the proposed model of ANN is quite efficient and can make reasonable estimates for AH.

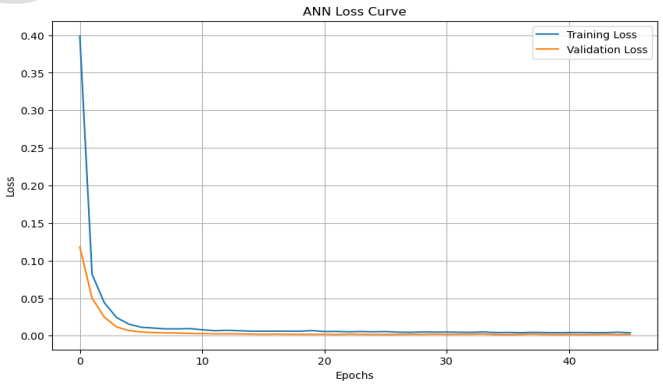

Fig 3. ANN Loss Curve

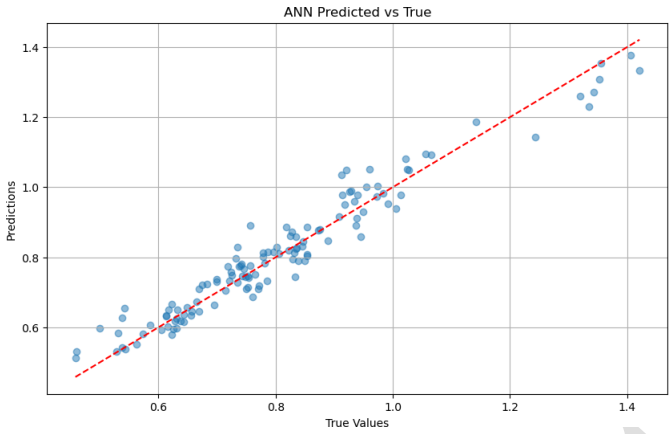

Fig. 4. ANN Predicted VS True value

5.2. RNN Performance

A similar performance of RNN are shown in the Figure 4 and 5. The RNN Model observed a loss of 0.0028 and MAE of 0.0415. As has been highlighted earlier, the loss is a bit higher than that of ANN, which means that the RNN is not very good at predicting values of y; The MAE of 0.0415 means that the average error made by the RNN is higher than that of the ANN. This means although the RNN can aptly identify patterns within the time series data, its prediction accuracy of AH is lower than that of the ANN.

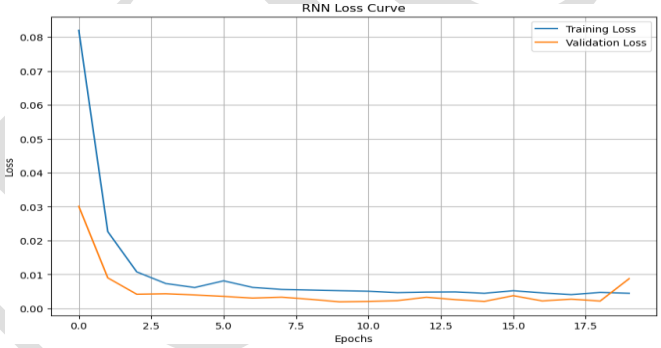

Fig 5. RNN Loss Curve

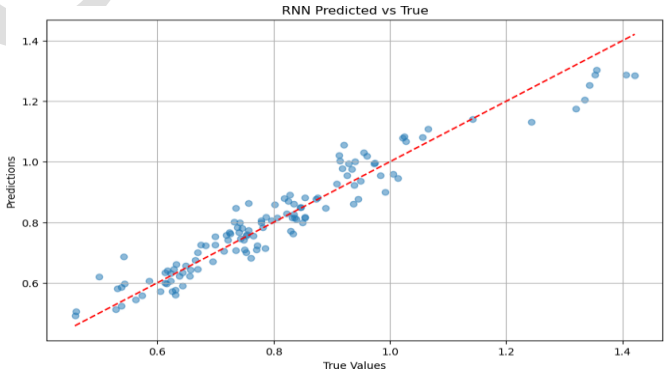

Fig 6. RNN Predicted Vs True Value

It is observed that ANN performed more than RNN in terms of prediction accuracy. RNN outperformed because temporal dependencies were not strong.

5.3. Future Plan Enhancing with Real-Time Applications

- In developing the data pipeline, an automated system is implemented to continuously ingest live air quality data from IoT devices. For real-time streaming predictions, optimized models are deployed on cloud platforms such as AWS or Azure, ensuring prompt and scalable forecasting.
- The Advanced Model Architectures like the exploration of (LSTM) Long Short Term Memory models and to be more accurate in predicting the sequences. LSTM, with its ability to capture long-term dependencies, will mark a qualitative improvement in the quality of temporal predictions. Use an ensemble approach that combines the features of ANN with those of RNNs (or LSTM) in order to enhance strength and generalization.
- The user interface development such as web or mobile applications, which designs a user interface where stakeholders can interact with real-time predictions and insights from air quality models. Visualization Tools to Install dynamic charts and graphs as a means of reporting alerts and trends over time on critical air quality thresholds.
- Host models and data pipelines on scalable cloud infrastructure to accommodate the much larger data volumes in the cloud deployment and the model updates by Implement a feedback mechanism to regularly retrain models with updated data, ensuring the system remains adaptive to evolving environmental conditions.

To improve the model's ability to capture long-term dependencies in data we can use Long Short Term memory (LSTM) and advanced RNN that address vanishing and gradient problems.

6. Conclusion

The study explored and compared the capabilities of Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN) in predicting Absolute Humidity (AH), a critical variable in assessing air quality and environmental conditions. The work aimed to determine which model could better handle the complexity of AH prediction, leveraging their respective strengths in capturing non-linear relationships (ANN) and temporal dependencies (RNN). The study clearly shows that ANN excels in comparison to RNN in predictive accuracy as well as efficiency. ANN model has provided a Mean Absolute Error (MAE) of 0.0249 and a loss value of 0.0011, which is an indicator of its good ability to find and classify complex patterns in the dataset. The architecture used, which includes two hidden layers with ReLU activation, dropout layers to reduce overfitting, and an Adam optimizer for dynamic learning rate adjustment, has played a critical role in the success of the model. The model thus applies these features effectively to generate accurate predictions with lower deviation from actual AH values.

The RNN model exhibited a higher MAE on 0.0415 and a loss of 0.0028 making it less accurate than the first two models, RNN by itself is powerful in obtaining those temporal dependencies. In this particular dataset, however, it did not perform to expectations relative to ANN's performance. This is due to the fact that the dataset bears a low temporal correlation, thereby negating a lot of RNN's advantages of sequentially processing information. It may also be due to hurdles like the vanishing gradient problem in the training of RNNs, which may have compromised its ability to optimally converge. The visual comparisons between predicted and experimental results with respect to AH

values which go a long way in validating the supremacy of ANN. Almost all the ANN predictions are faithful to the true values while RNN predictions show considerable deviations especially for extreme or outlier data points. Again, the loss curves for both models strongly support the above summarization with ANN converging faster, and more stably, at a lower error, when compared to RNN.

References

1. Raheja, S., & Malik, S. (2022). Prediction of air quality using LSTM recurrent neural network. International Journal of Software Innovation (IJSI), 10(1), 1-16.

2. Septiawan, W. M., & Endah, S. N. (2018, October). Suitable recurrent neural network for air quality prediction with backpropagation through time. In 2018 2nd international conference on informatics and computational sciences (ICICoS) (pp. 1-6). IEEE.
3. Cho, K., Lee, B. Y., Kwon, M., & Kim, S. (2019). Air quality prediction using a deep neural network model. Journal of Korean Society for Atmospheric Environment, 35(2), 214-225.
4. Barai, S. V., Dikshit, A. K., & Sharma, S. (2007). Neural network models for air quality prediction: a comparative study. In Soft Computing in Industrial Applications: Recent Trends (pp. 290-305). Springer Berlin Heidelberg.
5. Waseem, K. H., Mushtaq, H., Abid, F., Abu-Mahfouz, A. M., Shaikh, A., Turan, M., & Rasheed, J. (2022). Forecasting of air quality using an optimized recurrent neural network. Processes, 10(10), 2117.
6. Le, V. D. (2023). Spatiotemporal graph convolutional recurrent neural network model for citywide air pollution forecasting. arXiv preprint arXiv:2304.12630.
7. Xu, J., Wang, S., Ying, N., Xiao, X., Zhang, J., Cheng, Y., ... & Zhang, G. (2023). Dynamic Graph Neural Network with Adaptive Edge Attributes for Air Quality Predictions. arXiv preprint arXiv:2302.09977.
8. Han, J., Liu, H., Zhu, H., Xiong, H., & Dou, D. (2021, May). Joint air quality and weather prediction based on multi-adversarial spatiotemporal networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 5, pp. 4081-4089).
9. Bui, T. C., Le, V. D., & Cha, S. K. (2018). A deep learning approach for forecasting air pollution in South Korea using LSTM. arXiv preprint arXiv:1804.07891.
10. Zhu, X., Zou, F., & Li, S. (2024). Enhancing Air Quality Prediction with an Adaptive PSO-Optimized CNN-Bi-LSTM Model. Applied Sciences, 14(13), 5787.