A NOVEL AIR QUALITY PREDICTION SYSTEM WITH LONG TERM STORAGE WITH HYPER PARAMETER TUNING

G.Kalaivani, Research Scholar
Department of Computer Science
School of Computing Sciences
Vels Institute of Science, Technology Advanced Studies (VISTAS)
Pallavaram, Chennai, India
kalaivanikamalakannan@gmail.com

S. Kamalakkannan, Associate Professor
Department of Information Technology
School of Computing Sciences
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Pallavaram, Chennai, India
kannan.scs@velsuniv.ac.in

P.Kavitha, Research Scholar
Department of Computer Science
School of Computing Sciences
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Pallavaram, Chennai, India
pkavikamal@gmail.com

Abstract— Air pollution is the most important contributor to a variety of serious health problems as well as weather transformation. Air quality interpreters are needed to design human activity at a particular environmental location and lessen the adverse pollution detrimental impacts. The authors have examined the problem in difficulty of forecasting the classification of pollutant concentrations in the future and have proposed a novel method based on Long Short-Term Memory (LSTM) model. This is a Deep Neural Network (DNN) model that is identified to work well with consecutive prediction difficulties. Through the data obtained, this approach produces a prediction model that reliably forecasts the Air Quality Index (AQI) from Central Pollution Control Board (CPCB) website. LSTM is experimented eight number of times to choose the best possible function using Python's LeakyRelu package to assess hyperparameter optimizations. The comparative analysis of accuracy metrics such as R Squared, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are measured for different models. The proposed model discovered R Squared is more than 1, indicating that the hyper parameter tuned model is the best fit model, based on the concureent experiments. This dataset has the highest prediction accuracy and less prediction loss.

Keywords: Air pollution, Central Pollution Control Board, long short-term memory, Deep Neural Networks, Root squared, accuracy, Air Quality Index

I. Introduction

Air pollution [1] is a primary hazard for coronary heart related diseases, stroke, continual disruptive respiratory disorder, lung cancer, critical respirational contaminations, and exacerbating asthma. The great charge of boom in PM2 five (particles), predicting air pollutants is important for improving public health management. The human respiratory system may easily take in particles with a width of less than 2.0 microns (PM2.5) [2]. The existing studies show that levels of pollutants additionally depend on meteorological elements together with wind velocity, temperature, barometric pressure, and wind path. In addition, pollutant concentrations are inherently time-dependent, and future sequences of pollutant levels will depend on past pollutant degree sequences, meteorological elements, and visitors' conditions. LSTM model is a deep neural model that is recognized to work properly with sequential prediction

responsibilities. This paper proposes an LSTM-based result for predicting future air impurity absorptions. It is essential to recognize earlier the fashion of the pollutant's degrees in order that important preventive actions can be occupied to save ourselves from the damaging consequences of long-time acquaintance to pollutants. Thus, the forecast of contaminant degrees is one of the maximum essential issues for municipal corporations and is likewise essential in climate change observation.

According to data from the World Health Organization (WHO) [3], 1.6 million people are predicted to die in India in 2012 as a result of air pollution. At least a hundred and forty million human beings breathe air of high-satisfactory this is ten instances or extra above the WHO protection boundary; 13 of the globe's twenty cities with the best yearly phases of air pollutants are in India. Some of the most polluted cities in the world are located in India, including Ahmedabad, New Delhi, Lucknow, and Patna. About 2 million Indians die prematurely each year as a result of air pollution [4]. The commonplace air pollutants present are particulate subjects (PM2.5 and PM10), carbon monoxide (CO), ozone (O3), nitrogen dioxide (NO2) and lead (Pb). The Government of India (Ministry of Environmental, forest, and climate change) has mounted devices in many cities and these sensors can discover the attention of these contaminants on the exact environmental zones. These device size statistics is made widely accessible via CPCB. Numerous pollutants listed above, Particulate Matter (PM) 2.5 is important. Its thickness is less than

2.5 microns and can permeate profound into the bronchial tube of the human respirational scheme. Therefore, forecasting PM2.5 level is an important task in influential air quality. This paper examines datasets from the different cities such as Ahmedabad, Patna, Delhi etc., from CPCB website. The customisation of this dataset is done to train the LSTM network and forecast imminent sequences of pollution concentrations of this regions. In this work comparative analysis of the predictive function of LSTM eight number of times recursively. Moreover, Hyperparameter adjustment procedure is executed to Identify the optimal hyperparameter tuning for every LSTM machine learning model. Finally, comparative analysis of traditional LSTM with fine-tuned LSTM are analysed using R squared and Accuracy for the air quality prediction. The remaining work is divided into four sections: section 2 discusses about the literature review, section 3 discusses methodologies, algorithms and implementation details section 4 depicts outcomes and performance evaluation, and section 5 discusses conclusion and future research.

II. Relevant Works

In current years, several ML approaches have been anticipated to solve the problem of air pollution prediction. This section introduces and analyses some of the most important tasks in this area. In this section, the most applicable and today's results proposed inside the works that employs Deep Learning (DL) methods for air quality is stated. Ameer et al. [5], stated that DL has base stacked LSTM model for predicting future air pollutants awareness. Neural Network Auto Regressive (NNAR) method used to predict PM2.5 absorption and contemporary a predictive model for forecasting PM2.5 only for the subsequent hour. The dissertation also delivers a comparison study of predictive performance of the preservative form of Holt-Winters technique, Auto Regressive Integrated Movement Average (ARIMA) model and NNAR model. Taking into account characteristics such as weather factors, pollutants (NO2, CO, PM2.5, etc.), generation of festival / holiday / traffic information of data specified at the time they forecast the contaminants next one hour, next six hours and next twelve concentrations time for the region of Agra and Delhi. P. Yuan et al., [6] have analysed the contamination of the city and plotted it conferring to the considered geographic area. The authors used Apache Spark to analyse data from 2008 to 2014. In addition, authors have linked the forecast precision of logistic regression with the Naive Bayes (NB) set of rules. The originate NB predicts information more precisely than other ML systems used to classify indefinite air quality predictions. Although this paper shows good outcomes in relations of Apache Spark dispensation time, this process is not suitable for actual time series forecast. Hoque et al., [7], the author deals with the forecast of air contaminants such as ozone, Particulate Matter (PM2.5), and sulphur dioxide. This paper practice optimization and regulation methods to predict air contaminant levels the subsequent day. The author uses the dataset from two stations to predict the value. One station forecasts O3 and SO2 values, and the other station provides O3 and PM2.5 values and have modelled the data based on comparison and cast-off linear regression for federation. Root Mean Squared Error (RMSE) was the criterion used. This limitation of work arises from a linear regression model that cannot predict or handle unexpected events. In addition, the study uses only data from two stations, restraining its generalization.

Elgeldawi et al. [8], stated that a deep neural network for prediction PM2.5 absorption for the subsequent twenty-three hours from the specified time straight away. The paper uses a trained non-stacked LSTM model information on the attentiveness of contaminants and other keys Works with preceding time phases. In this work, developed a solitary layer LSTM system longer recall for predicting improved outcomes. The author selects only the functions needed to moderate and progress overfitting predicted performance. The anticipated deep neural networks complexity of the network calculation is found to be low and excellent. Kratzert et al. [9], analyzed day-to-day air contamination forecasts from 74 towns in China were investigated using ML techniques. Five dissimilar organization methods were implemented to predict the results, using a dissimilar feature set than the Decision Tree (DT) model. They functioned on feature assortment techniques, which displayed that DNN had a low rate of conjunction limitation. Borovkova et al., [10], the authors present a study that proposes a procedure that exhibits improved predictive power by increasing R2 and decreasing RMSE when run on Hong Kong dataset. Extreme Learning Machines (ELMs) have been shown to perform well in relation to accuracy, generalization, and forcefulness. No important difference was found among the prediction accuracy of the individual models. ELM performed greatest on predictive pointers such as R2 and RMSE. The author attained a training time of 95 RMSE and 0.07 seconds [11]. The prediction of AQI, and its impact on health was presented by authors Mohit et al., [12]. The authors carried out a DT approach and NB J48 for classification. The outcomes they acquired showed that DT algorithm performs with 91.9978% accuracy. However, there are numerous limitations with this research, along with the issue that the dataset used became restricted. Moreover, the selection tree strategies are carried out poorly over non-stop variables and may have troubles with overfitting. Abimannan et al., [13] stated the classification of

AQI dataset prediction. In their portraits, the authors employed K-means set of rules; again, in this research the dataset used become restrained. Further problems stand up when trying to expect destiny values, a weak point in K-means techniques. Another examines for forecasting air pollution in Canada makes use of a Multilayer Perceptron Neural Network (MLPNN) [14]. The authors deal with the difficulty of air best prediction and model accuracy. However, the quantity of data used in the study is confined and the computational cost for seasonally updating of the version is huge.

III. Proposed Methodology

This section comprises of method used in this research to predict air quality. The dataset is collected from Kaggle dataset. The dataset includes information on air quality and the AQI at the hourly and daily levels from numerous stations located throughout various Indian cities. The dataset consists of 12 features with 29,544 instances from 23 different Indian cities. Table 1 shows a sample dataset containing observations from 2015 to 2020 [15] for a particular Vishakhapatnam city among 23 different cities. Analysis of some major air pollutants, namely PM2.5, PM10, NO2, CO, SO2, O3, and AQI predictions are data variables for cities and dates. The seven measurements used to calculate the AQI are PM2.5, PM10, SO2, NOx, NH3, CO, and O3. If there are at least 16 values, the average value from the previous 24 hours is utilized for PM2.5, PM10, SO2, NOx, and NH3. For CO and O3, the highest value during the previous 8 hours is used. Every measurement is transformed into a Sub-Index based on pre-established groups. Due to a shortage of data points or a lack of measuring, measurements are occasionally unavailable. The highest Sub-Index is the Final AQI, provided that at least three out of the seven components and at least one each of PM2.5 and PM10 are present. The LSTM [16] algorithm is used to predict AQI with concurrent executions of the LSTM model and finally fine tune the model through hyperparameter tuning. The basic work flow of the proposed system is shown in figure 1. After collecting data from CPCB website, data pre-processing such as removing punctuations, reducing the repeated words, assigning tokens etc., are carried out. Then air quality prediction is evaluated through ML and Deep Learning (DL) algorithms. The best fit model LSTM and Dense Net is found by recurrently evaluating the LSTM model eight number of times. The hyperparameter tuning [17] is carried out in three parameters unit size, activation function and batch size. Through error and trail method, LSTM will be checked concurrently. The final output will be checked with comparative analysis of fine-tuned and untuned models.

Table 1 Sample AQI for the Visakhapatnam city

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
29521	Visakhapatnam	2020-06-22	33.17	108.22	5.58	42.45	27.06	13.70	0.73	13.65	34.85	3.99	10.24	2.32	95.0	Acceptable
29522	Visakhapatnam	2020-06-23	25.40	83.38	2.76	34.09	19.92	13.13	0.54	10.40	43.27	2.88	12.03	1.33	100.0	Acceptable
29523	Visakhapatnam	2020-06-24	34.36	90.90	1.22	23.38	13.12	14.45	0.56	10.92	35.12	2.99	3.15	1.60	86.0	Acceptable
29524	Visakhapatnam	2020-06-25	13.45	58.54	2.30	21.60	13.09	12.27	0.41	8.19	29.38	1.28	5.64	0.92	77.0	Acceptable
29525	Visakhapatnam	2020-06-26	7.63	32.27	5.91	23.27	17.19	11.15	0.46	6.87	19.90	1.45	5.37	1.45	47.0	Good
29526	Visakhapatnam	2020-06-27	15.02	50.94	7.68	25.06	19.54	12.47	0.47	8.55	23.30	2.24	12.07	0.73	41.0	Good

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
29527	Visakhapatnam	2020-06-28	24.38	74.09	3.42	26.06	16.53	11.99	0.52	12.72	30.14	0.74	2.21	0.38	70.0	Acceptable
29528	Visakhapatnam	2020-06-29	22.91	65.73	3.45	29.53	18.33	10.71	0.48	8.42	30.96	0.01	0.01	0.00	68.0	Acceptable
29529	Visakhapatnam	2020-06-30	16.64	49.97	4.05	29.26	18.80	10.03	0.52	9.84	28.30	0.00	0.00	0.00	54.0	Acceptable
29530	Visakhapatnam	2020-07-01	15.00	66.00	0.40	26.85	14.05	5.20	0.59	2.10	17.05	NaN	NaN	NaN	50.0	Good

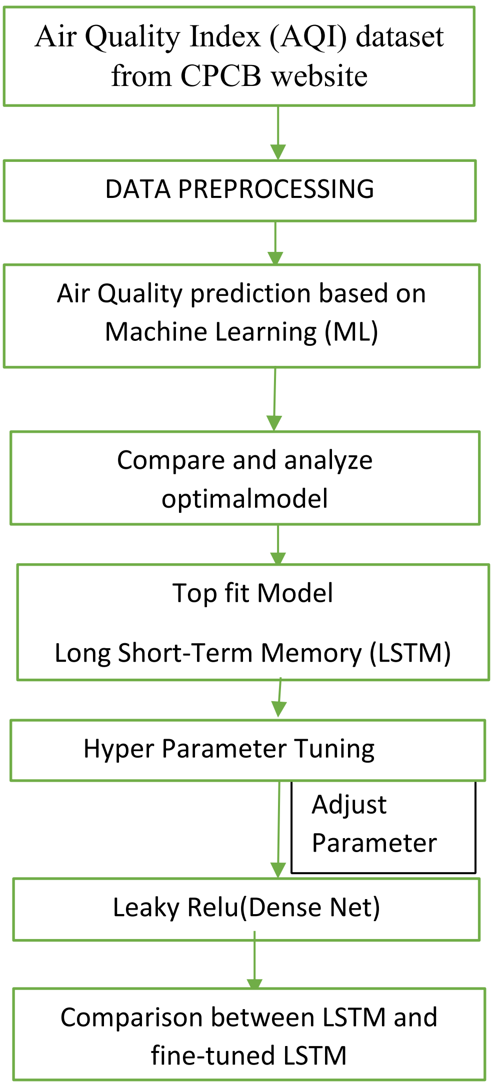


Figure 1 Proposed Process Flow

a. LSTM and DenseNet

DenseNet (Y.Liu et al., 2018) is the most used Deep Learning Models under progress in current years. Familiarized by former researchers on DenseNet connection properties for each layer and feature map all consequent layers. The subsequent layer obtains the input function maps from all layers The DenseNet design flow comprises of tightly connected blocks through the transition layer. Figure 2 shows the basic dense net architecture. When other DL algorithms connect functions Map amongst layers by accumulation and use DenseNet current layer and chain amongst layers. When the l th level receives an input feature map from every previous layer as input X0, X1, ……., Xl-1 shown in eqn(1).

$$Xl = Hl([X0, X1, \ldots, Xl-1])\text{-------}(1)$$

Where [X0, X1, ..., Xl-1] is feature map from layers 0, 1, ..., l-1. The main advantage of Dense net has overcome associated to other disappearances gradient backpropagation, feature reprocess, and number of parameters not deeper than other parameters learning model.

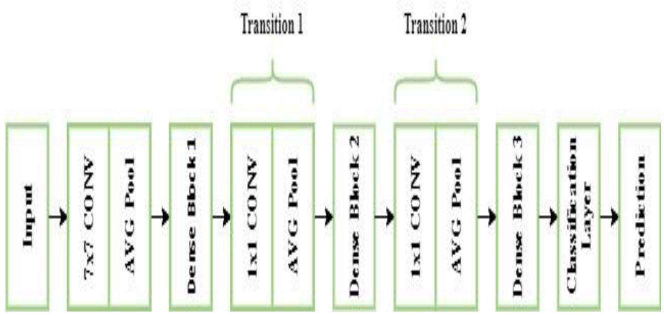


Figure 2 Basic DenseNet Design (Y.Liu et al., 2018)

The disadvantages of DenseNet are its concentration of layers so it requires reckoning training time. Another disadvantage is the large amount of memory usage for the following reasons: Inefficiency was introduced by concatenating function maps Model. One way to rectify these difficulties is to optimize with a batch strategy method which is LSTM model (Athira et al., 2018) incorporating with DenseNet. Figure 3 shows batch processing method. The batch processing system uses the following ratios of adaptive batch sizes, activation function through training to achieve quicker training time. The implementation of batch processing is hyperparameter adjustment requirement. Hyperparameters used in stacking strategies The methods are stack size and batch size. Batch size activation function is part of hyperparameter DL. Hyperparameters need to be adjusted to get the best batch size [19] and activation function to apply to batch processing approach. This paper describes adjusting DenseNet hyperparameter to get the optimal stack size, activation function and unit range.

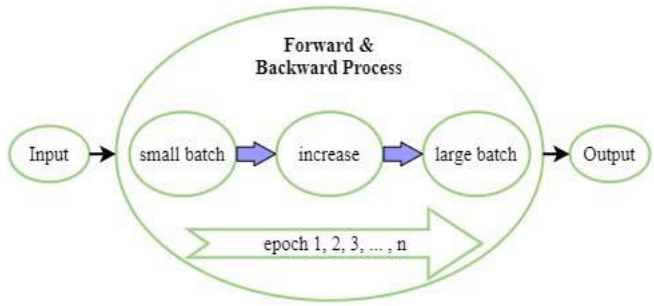


Figure 3 Batch Processing Method

Figure 4 shows the process steps of the implemented process incorporating LSTM with dense net architecture. In this research, LSTM method with DenseNet is taken for experimental. To get the stack size and learning rate values through trial-and-error manual usage of parameters. An evaluation will be made for each selected candidate by recurrent analysis of LSTM eight number of times. These values determine the finest series of values in a selected candidate. Experiments are based on hyperparameters batch size, activation function and unit size.

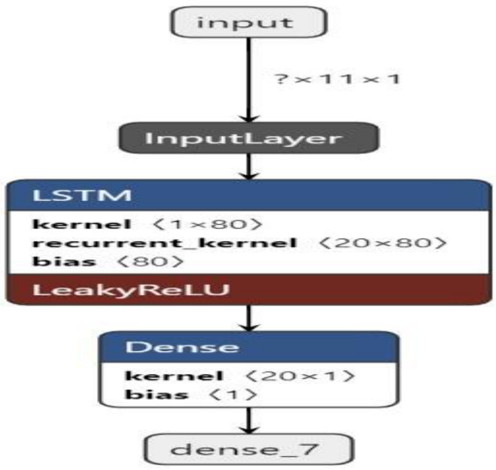


Figure 4 Proposed LSTM (Athira et al., 2018) and DenseNet Architecture

ALGORITHM: Fine-tuned LSTM model

Input parameter: unit size, activation code and batch size

1 Start

2 Tuning the model with tuned_lstm(units, act_func, batch): model = Sequential() model.add(LSTM(units, input_shape=in_dim, activation=act_func))

3 DenseNet model is added using model.add(Dense(out_dim))

4 Comparing the error, loss and validation model.compile(loss="mse", optimizer="adam")

5 Predicting the best fit model through model.fit(train_X, y_train, epochs = 100, batch_size = batch, validation_data=(test_X, y_test)) y_pred = model.predict(test_X)

6 Print the accuracy by return (model, acc_scores(y_test, y_pred)

7 End

The proposed LSTM incorporating with DenseNet algorithm are shown in algorithm. The three input parameters are unit size, activation code and batch size are formed. The DenseNet method evaluate error, loss and validation with accuracy metrics [20] Mean Squared Error (MSE). The prediction of model is determined through error and trial manual method. LSTM model is then recurrently run eight number of times and found the LSTM model which was run at 5th time is the best fit model.

IV. Results

The hyperparameter tuning with LSTM are compared and analysed for air quality prediction in this section. First, the accuracy of LSTM ML model [21] in classifying air quality are analysed using the LSTM model hyperparameters through the default settings. The value is provided by the Python library package [22]. Figure 5 shows the AQI of various cities. X axis is the dimension of air pollution in top cities in terms of seconds (s) and Y-axis measured in terms of year and month. The sequential model which has LSTM incorporated with DenseNet train and nontrained parameters are shown in figure 6. The total parameters trained are 1781and non-trainable parameters are null. The output of the preceding step is used as input in the current step of the LSTM. This addressed the problem of long-term dependence of RNNs, where RNNs cannot forecast words stowed in long-term memory, but can make more precise predictions built on current information. As the gap length increases, the RNN works less efficiently. By default, LSTMs can retain information for a long time. It is used for processing, forecasting, and classifying time series data.

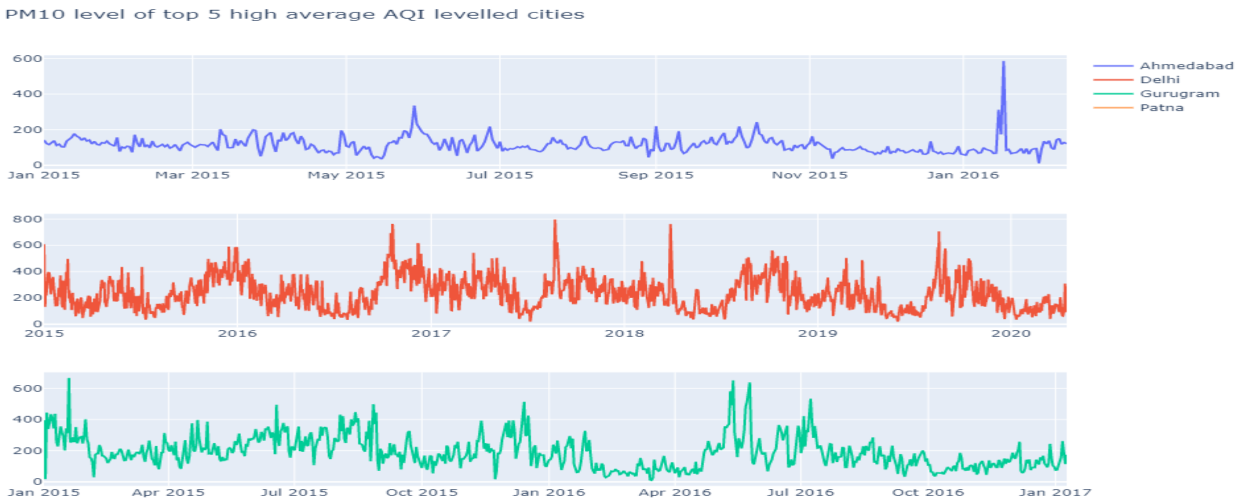
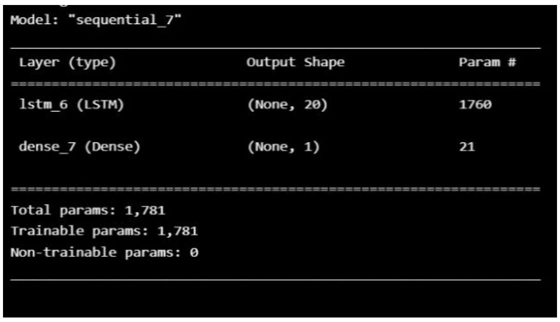


Figure 5 AQI (PM 10, PM2.5, CO, NH3, AQI, O3, NO2) levels of top 5 high average AQI leveled cities



```
Model: "sequential_7"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_6 (LSTM)               (None, 20)                1760

 dense_7 (Dense)             (None, 1)                 21

=================================================================
Total params: 1,781
Trainable params: 1,781
Non-trainable params: 0
_________________________________________________________________
```

Figure 6 Sequential Model (LSTM+DenseNet) Parameters

Figure 7 shows the plot of the accuracy of LSTM classifiers which has used trial and error for 8 number of times.

Table 2 shows the accuracy measures such as R squared, RMSE, and MAE, which are depicted in figure 8. The Mean Absolute Error (MAE) is a metric for measuring the difference in error between two observations. The MAE was displayed in seconds. When compared to other models, LSTM has a high level of accuracy. Predicted and observed timings, tracking, and starting time comparisons, and measuring the alternative measurement methods with x axis measured in terms of models and y axis is measured in terms of seconds (s). The square root of MSE is RMSE, while the term seconds denotes the square root of R squared (s).

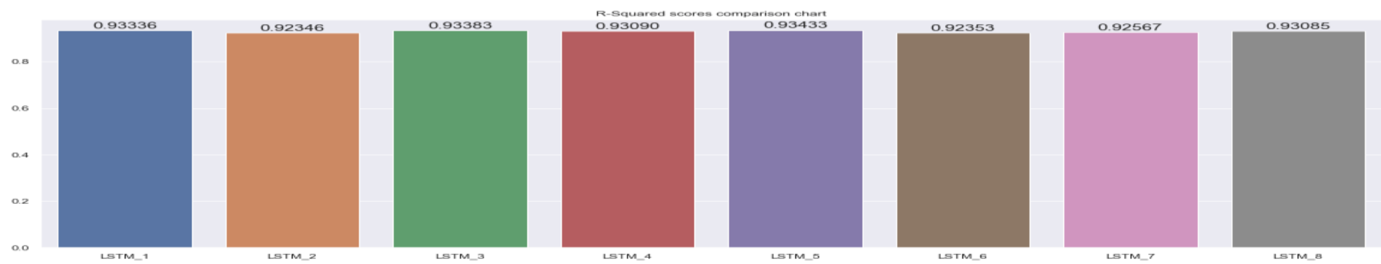


Figure 7 Comparison of R Squared Metrics in Recurrent LSTM Models

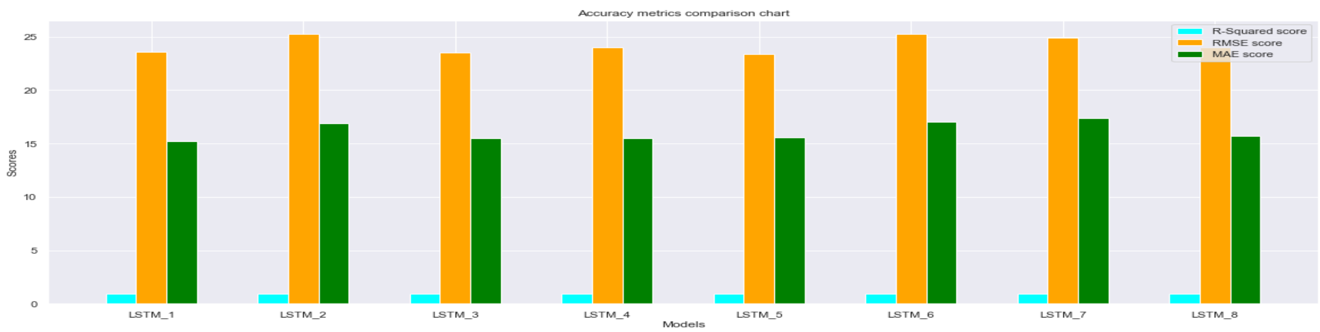


Figure 8 Accuracy Metrics (RSquared, RMSE and MAE) Comparison for Eight Recurrent Trial and Error LSTM Model

Table 2 Comparative Analysis of Dense Net, LSTM, Fine-tuned LSTM

Accuracy Metrics	Dense Net	LSTM	Fine-Tuned LSTM
R-Squared	0.9299	0.9321	0.9343
RMSE	24.1764	23.7882	23.4069
MAE	15.5572	15.4386	15.5977

LSTM runs concurrently through Leaky Relu python package. R Squared is less than 1 which is measured using proposed hyperparameter tuning technique. These parameters outperforms the existing system with the difference of 0.0021. Figure 9 shows the comparative analysis of accuracy metrics x and y coordinates are measured in terms of seconds(s). This shows that, fine tuning the LSTM model increases by R Squared value in milli seconds. Figure 10 shows the comparative analysis of R squared with default LSTM model and fine-tuned LSTM model. Matching the model hyperparameters to the test set means that the hyperparameters can overfit the test set. Estimating performance using the same test set would be an overestimate. Accuracy and R squared of these models sorted out overfitted model. Estimation error in the predicted model is calculated through LeakyRelu activation function. Dropout is used for prediction and training as argument. During prediction mean and standard deviation are evaluated. Test data is given as T with 1000 initially to estimate uncertainty. In this case, the accuracy of the model is higher and prediction loss is low. Binary loss entropy is the prediction loss which means prediction algorithm undergo loss when predicting the real label as 0 or 1. The prediction loss is lesser in this hyperparameter tuned model.

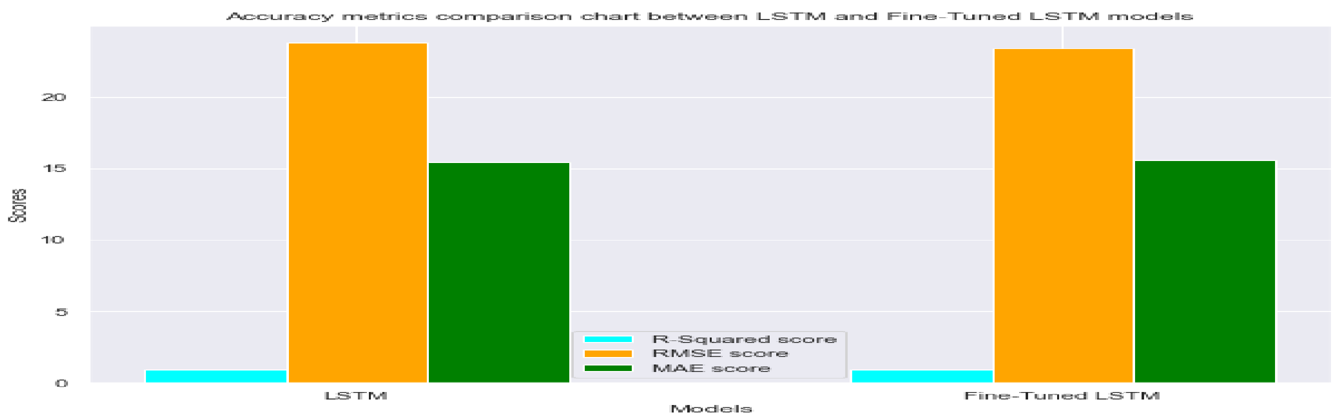


Figure 9 Comparison of Accuracy Metrics for LSTM and Fine-tuned LSTM

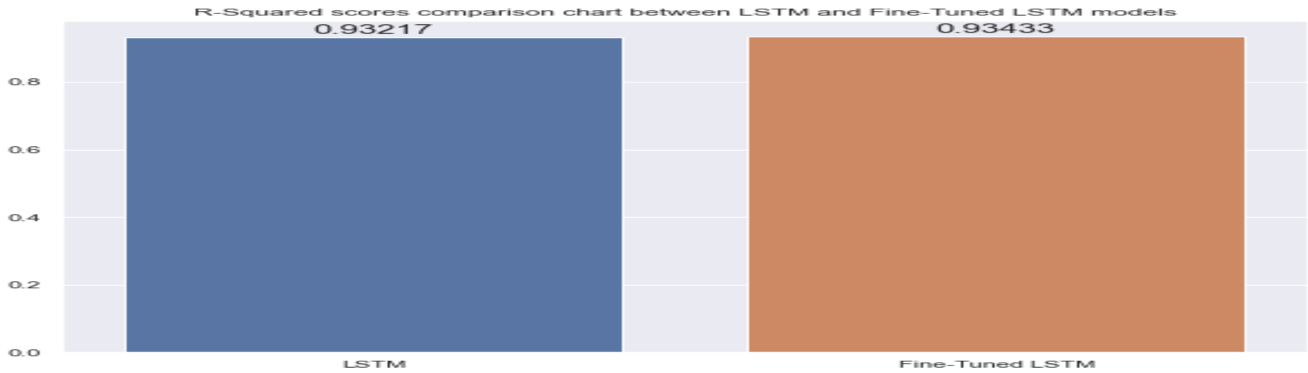


Figure 10 Comparison of R squared for LSTM and Fine-tuned LSTM

V. Conclusion and Future Scope

Forecasting air quality can aid in a variety of civil choices as well as safeguarding inhabitants' health. In this case, a novel system is presented combining LSTM with DenseNet in the research to predict air quality concentrations and memory. The most suitable candidates characteristics extracted from a real-time dataset for training purposes coefficients of connection for the specified geographical region are determined. The proposed model has proved that memory usage and rapid analyses of LSTM model is higher through several experiments using data collected from CPCB AQI for 23 different cities in India. The trial-and-error method is used in LSTM model 8 number of times and LSTM _5 is the best fit model since R Squared is more than 1 and RMSE is less than 1.2. The future work can be done many numbers of areas not specified to single place. The future research can be done in many other algorithms and can focus on much less memory usage.

References

[1] A. Nugroho and H. Suhartanto, "Hyper-Parameter Tuning based on Random Search for DenseNet Optimization," 7th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), 2020, pp. 96-99,doi: 0.1109/ICITACEE50144.2020.9239164.

[2] D. Srivastava, Y. Singh and A. Sahoo, "Auto Tuning of RNN Hyper-parameters using Cuckoo Search Algorithm," Twelfth International Conference on Contemporary Computing (IC3), 2019, pp. 1-5, doi: 10.1109/IC3.2019.8844900.

[3] D. Ao, Z. Cui and D. Gu, "Hybrid model of Air Quality Prediction Using K-Means Clustering and Deep Neural Network", Chinese Control Conference (CCC), 2019, pp. 8416-8421, doi: 10.23919/ChiCC.2019.8865861.

[4] P. Singh, T. L. Narasimhan and C. S. Lakshminarayanan, "DeepAir: Air Quality Prediction using Deep Neural Network," TENCON- IEEE Region 10 Conference (TENCON), 2019, pp. 869-873, doi: 10.1109/TENCON.2019.8929470.

[5] Ameer, Saba & Shah, Munam & Khan, Abid & Song, Houbing & Maple, Carsten & Islam, Saif & Asghar, Muhammad, "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities," in IEEE Access, 2019, vol. 7, pp. 128325-128338, doi: 10.1109/ACCESS.2019.2925082.

[6] P. Yuan, Y. Mei, Y. Zhong, Y. Xia and L. Fang, "A Hybrid Deep Learning Model for Predicting PM2.5"', 7th International Conference on Intelligent Computing and Signal Processing (ICSP), 2022, pp. 274-278, doi: 10.1109/ICSP54964.2022.9778520.

[7] Hoque, Kazi Ekramul & Aljamaan, Hamoud.''Impact of Hyperparameter Tuning on Machine Learning Models in Stock Price Forecasting'', IEEE Access Journals and magazines, 2020, vol. 9, pp. 163815-163830, 2021, doi: 10.1109/ACCESS.2021.3134138.

[8] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis," *Informatics*, vol. 8, no. 4, p. 79, Nov. 2021, doi: 10.3390/informatics8040079.

[9] Kratzert, Frederik & Klotz, Daniel & Brenner, Claire & Karsten, Schulz & Herrnegger, Mathew. ''Rainfall-Runoff modelling using Long-Short-Term-Memory (LSTM) networks''. Hydrology and Earth System Sciences Discussions, 2018, pp. 1-26. doi: 10.5194/hess-2018-247.

[10] Borovkova, Svetlana & Tsiamas, Ioannis.'' An ensemble of LSTM neural networks for high-frequency stock market classification'', Wiley Journal of Forecasting,2018, Vol. 38, doi: https://doi.org/10.1002/for.2585.

[11] Khalid, Rabiya & Javaid, Nadeem,'' A Survey on Hyperparameters Optimization Algorithms of Forecasting Models in Smart Grid'', Sustainable Cities and Society Journal,2020, Vol. 61, doi: 10.1016/j.scs.2020.102275.

[12] Mohit Bansal, Anirudh Aggarwal, Tanishq Verma and Apoorvi Sood, '' Air Quality Index Prediction of Delhi using LSTM'', International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2019, Vol. 8, Issue 5.