Study and comparison of machine learning models for air quality prediction

Leila ABBAD
Electrical Engineering
and Industrial Computing
ENST Higher School
Algiers, Algeria
leila.abbad@enst.dz

Djallel BRAHMIA
Electrical Engineering
and Industrial Computing
ENST Higher School
Algiers, Algeria
d_brahmia@enst.dz

Mohamed Nadir CHERFIA
Electrical Engineering
and Industrial Computing
ENST Higher School
Algiers, Algeria
m_cherfia@enst.dz

Abstract—In the last several decades and as a result of various kinds of man-made activities, industrialization and human urbanization, the atmospheric environment pollution became a real threat to the human's health. The particles with a diameter of less than 2.5μm, one of the most harmful pollutants present in the air as it causes diseases in the respiratory system as well as cardiovascular ones. Consequently, it is beneficial to predict the PM2.5 concentrations with high accuracy in order to alert people to make the right decision in order to fix the situation and improve the air quality especially in environments where it is essential. The prediction of the PM2.5 concentration have to pass throw a pre-processing stage then fed to the multiple models by passing a data chunk of twelve days to get the prediction for the next day. In this article, a comparative study between different Artificial Intelligence predictions models is presented: Bi-LSTM, Time Distributed Convolutional Neural Network (CNN), and an hybrid model combining both CNN and Bi-LSTM. For this purpose, several architectures were used for the different models: Multi Inputs – Multi Outputs, Multi Inputs - Single Output and the univariate. The CNN extracts the internal spatial correlation between variables and the Bi-LSTM extracts the temporal patterns, the hybridization process proposed of those two models with the multiple Inputs -Single Output architecture gave us the most accurate results.

I. Introduction

The growth in the proportion of the world's urban population demonstrates that more and more people are moving to larger cities. The United Nations (UN) expects the urban population to be approximately 68% in 2050 [1] . The increasing urbanization and industrialization results in many problems: logistics, health care and air quality. In an effort to solve these issues, and to enhance the quality of life of citizens, the smart city approach has been created by embedding information and communication technologies (ICT), as well as fixed/mobile sensors, which are being installed within the city to monitor real-time human behavior. This concept is becoming an infinite source of urban data. In these last couple of decades, common occurrence of smog with the rise of industrialization has pushed the environment pollution to its peak. This means that the pollution is more serious now than ever before. One of the harmful kinds of pollutants is a small particulate matter with a size of 2.5μm or less, known also as PM2.5. Such a particle brings with it grave health damage. The World Health Organization WHO reported that almost 90% of people are breathing polluted air that is beyond the limits of the WHO air quality guidelines and this is causing a lot of respiratory trouble [2] [3],also short-term exposure of a couple of hours to several weeks to PM2.5 can be a cause of death and cardiovascular diseases effects [4]. The classical statistical models have largely been applied to solve air quality prediction challenges. Those approaches are based on the concept of learning from historical data. Some of the notable statistical methods that have been used for air quality forecasting are the autoregressive moving average ARMA [5] , the autoregressive integrated moving average ARIMA [5] and seasonal autoregressive integrated moving average SARIMA [5]. But, with the growing volume and complexity of the data obtained, such methods can no longer satisfy the real demand due to the time taken to learn. As artificial intelligence and big data develop, the use of predictive methods based on machine learning technologies is becoming more popular Artificial neural networks (ANNs) that model the complex non-linear relationships between air pollutant levels and atmospheric parameters [6]. Different ANN structures have been developed to perform air pollution prediction in various areas of study. As the data has grown and the application has gotten more sophisticated, the network built for data analysis is no longer a single network model, but a more complex hybrid network. For example: LUAN et al [7]has proposed a text classification model referred with NA-CNN-LSTM or NA-CNN-COIF-LSTM, which has no activation function in CNN,Kan Chen [8]work on attention based deep learning architecture for the visual question answering task,Inchoon Yeo1 Yunsoo Choi1 Yannic Lops1 Alqamah Sayeed1 [9] released a research proposing a deep learning model that integrates a convolutional neural network and a gated recurrent unit with groups of neighboring stations to accurately predict PM2.5 concentrations at 25 stations in Seoul, South Korea.. However, the weak accuracy and long prediction time of existing methods could not meet the demand of PM2.5 daily life prediction. Due to the complex formation

of PM2.5, the high accuracy and efficiency requirement for prediction, and the challenging of stability in the deep learning network model, it is essential to develop a more efficient model to continuously predict the concentration of PM2.5. In this article, we propose an hybrid CNN-LSTM model to predict the PM2.5 concentration of the next day (24 hours). In order to test which of the alternative models is more efficient, four models including univariate LSTM, multivariate LSTM, univariate CNN-LSTM and multivariate CNN-LSTM, are analyzed and compared. Finally, To evaluate the models, two metrics are adopted, the mean absolute error (MAE) and the root mean square error (RMSE).

The paper is organized as follows: in Section II, the methodologies are described, in which CNN and LSTM are presented in detail and the hybrid model CNN-LSTM for predicting PM2.5 concentration is proposed. In section III, the data preprocessing is completed. In Section IV, the results are presented and discussed. Finally, we conclude our work in Section V .

II. MATHODOLOGY

Deep learning is a Branch of Machine Learning which is focusing on making accurate data-based decisions with huge neural network models that can make those kinds of decisions. Deep learning is especially appropriate in the face of big, highly complex datasets [10]. Deep learning neural networks are good at learning complex random input-output mapping automatically and handling multiple inputs and outputs. These are extremely promising and powerful features in time series forecasting, particularly for problems with complex and nonlinear dependencies, multi-valued inputs, and multi-step forecasts. Those features, together with newer neural network capabilities, have the potential to be very advantageous, however, including the automatic feature recognition offered by convolutional neural networks and the integrated support for sequential data in recurrent neural networks [11].

Time series forecasting is a difficult challenge. Contrary to more simple classification and regression problems, time series issues bring additional complexity of temporal order or dependency among the observations. This can be hard since specialized data processing is necessary during model fitting and evaluation,This time processing structure can also support the modeling, by providing extra context, like trends and seasonality, that can be exploited to increase the power of the model [12].

A. CNN Model

Deep convolutional neural network (CNN) is a particular class of neural networks, which has demonstrated excellent results in various computer vision and image processing related challenges. Some of the exciting application areas of the CNN are image classification and segmentation, object detection, video processing, natural language processing and speech recognition. The powerful learning capacity of deep CNN is principally due to the use of multiple feature extraction stages that can automatically learn features from the data,the CNN topology itself consists of multiple learning levels that are composed of a combination of convolutional layers, nonlinear processing units, and sub-sampling layers. The CNN is a multi-layer feed-forward network, in which each layer, using a set of convolutional kernels, makes multiple transformations. The convolution process provides useful feature extraction from the locally correlated group of data items [12].

CNN is set to become the most powerful deep learning method, and its network structures consist of 1D,2D and 3D CNN [13], the 1D CNN is principally employed for processing sequential data [14], the 2D CNN is commonly used in image and text recognition and the 3D CNN is mostly used for both medical image and video data analysis [15].

The ability of CNNs to learn and then automatically extract features from raw input data can be extended to time series forecasting problems. A sequence of data observations can be handled as a one dimensional image that a CNN model can interpret and distill to its most significant features [11]. Therefore, the time-distributed wrapper is a layer which can be appended to each time slot of the input. Each input must be at least 3 dimensions, and the dimension of index 1 of the input will be considered as the time dimension. Accordingly the time distribute CNN 1D set to be adopted in this paper .The full process of the CNN 1D is illustrated in the following figure 1 .

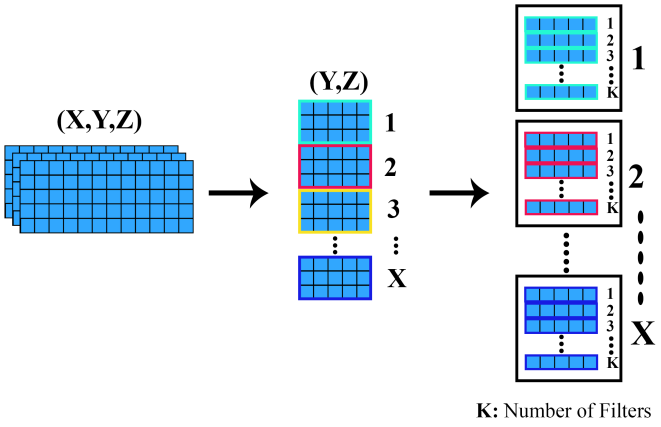


Fig. 1. A time distributed CNN 1D Layer

B. LSTM MODEL

Recurrent neural networks (RNNs) are very dynamic systems; they have an inner state at every time step. This is achieved through circular interconnections between neurons in the top and bottom layers as well as through optional self-feedback connections. Such feedback connections permit RNNs to propagate data from earlier events to current steps of processing. In this manner, RNNs

build a memory of time series events [16]. The original paper introducing the standard LSTM cell was published in 1997,a simple RNN cell was enhanced by appending a memory block that is controlled by input and output multiplicative gates [17] as illustrated in the following figure 2 .

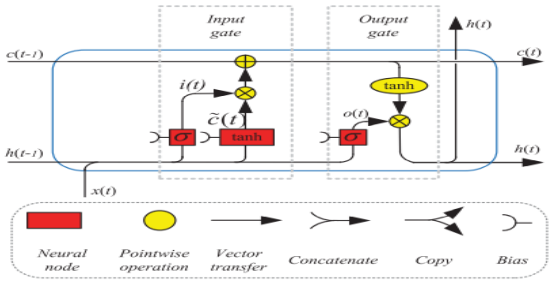


Fig. 2. The Original LSTM architecture [18]

The input gate is capable of blocking irrelevant data from being stored in the cell state, and the output gate decides what information can be output based on the cell state [18]. the original lstm cells were modified by the implementation of a forget gate in the cell,the forget gate can decide which information is going to be dropped from the cell state [19] as illustrated in the figure 3

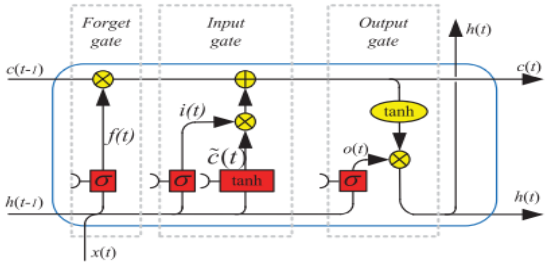


Fig. 3. The LSTM architecture with a forgetting gate [9]

Since the previous LSTM cell gates do not have direct links to the cell state, there is a loss of essential information which affects the network performance. In order to solve this problem, a " peephole " connection is implemented in the LSTM cells [19] as shown in fig 4 :

C. Bidirectional recurrent neural networks

The fundamental concept of bidirectional recurrent neural networks (BRNNs) is to introduce each sequence of forward and backward learning into two separate recurrent networks, both of them connected into the same output layer [20]. In the classical recurrent neural network model and LSTM model, only forward propagation of information is supported, so the state at time t only depends on the given information preceding time t. In an effort to make sure that each instant contains the context infor- mation, the BiLSTM model, combining

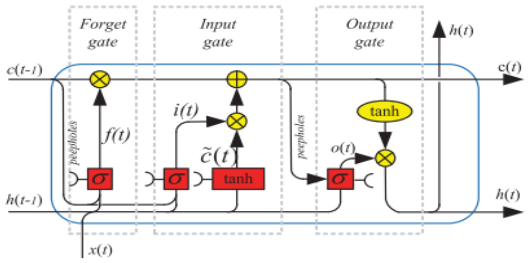


Fig. 4. LSTM architecture [9]

bidirectional recurrent neural network (BiRNN) models and LSTM based units, is implemented to capture the contextual information [21]

D. THE HYBRID CNN-LSTM MODEL

In this section, a hybrid CNN-LSTM model is developed employing a combination of CNN and LSTM shown bellow in fig 6 and its architecture detailed in Fig 9 to enhance the forecast accuracy. The proposed model has multivariate time series data as input and single time series with several steps as outputs, the process of which is illustrated in Fig 5.

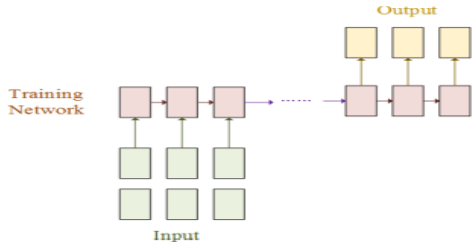


Fig. 5. A simple architecture of the hybrid CNN-LSTM model.

The CNN is adopted for feature extraction, specifically, three one-dimensional convolutional layers are constructed. In order to treat the data in the format that is required by the LSTM, a Flatten layer is added. Overfitting is a frequent phenomenon in deep neural networks (DNNs) and many solutions exist. Of all the solutions, dropout is one of the simplest and most performant. Dropout refer to the point where during the DNN training process, a cell is temporally dropped out of the network depending on a certain probability [23].

In order to prevent overfitting, a Dropout Layer is introduced, the output of which is plugged into the LSTM layer for prediction and finally connected to an Fully Connected Layer.

III. DATA SOURCE AND PREPROCESSING

The chosen dataset https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data) in this

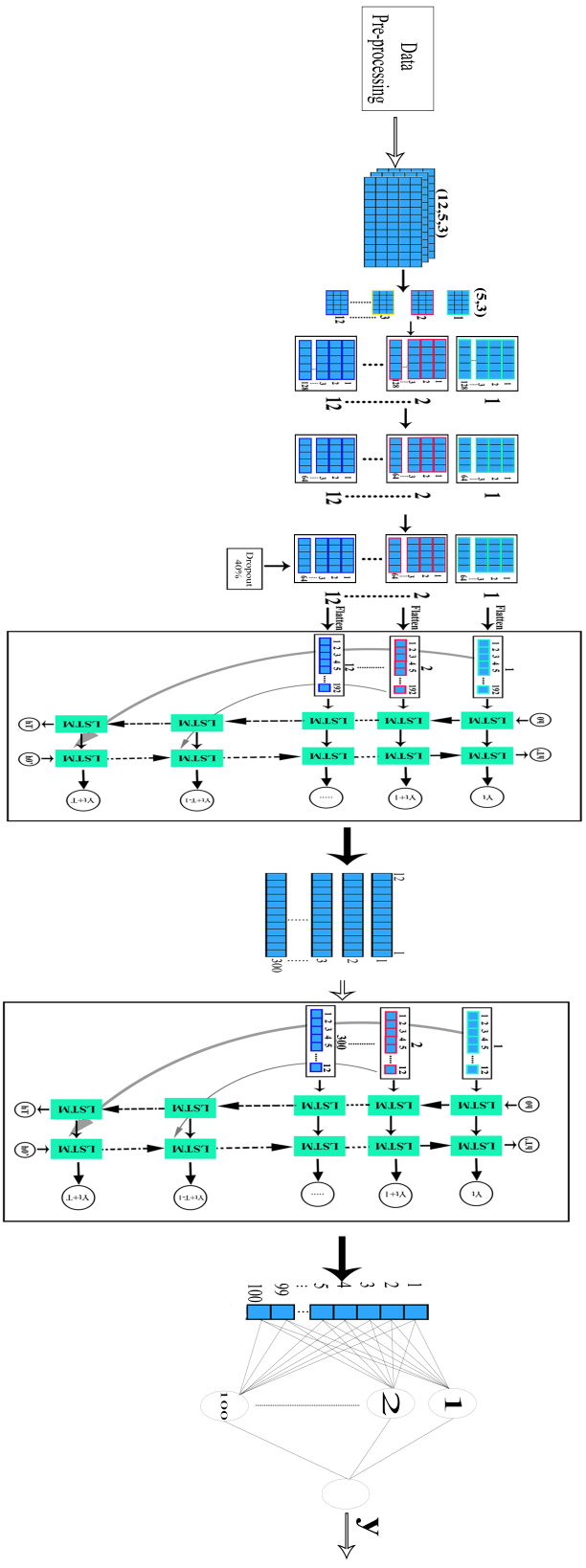


Fig. 6. Network structure of the proposed hybrid CNN-LSTM model

document contains the hourly records of six main air pollutants and six meteorological parameters at multiple stations in Beijing, and we choose to work with the data from Aotizhongxin station in the following study. This Data-set contains 35064 records with multi features, including PM2.5 concentration, PM10 concentration, SO2 concentration, NO2 concentration, CO concentration, O3 concentration, temperature, pressure, dew point temperature, precipitation, Wind direction and wind speed. Therefore, this data-set contains many missing values due to some uncontrollable reasons,so we need first fill in the missing values by using interpolation. Then, the wind direction is coded and converted into numerical values. Then, we analyze the PM2.5 Data Then, the PM2.5 data is analyzed in the following table I

TABLE I
statistical description of PM2.5 concentration values

mean	standard deviation	min	25%	50%	75%	max
82.54	81.95	3.00	22.00	58.00	114.00	898.00

Now an examination of the correlation among the data is made in order to determine the significant parameters using the Pearson correlation method, which is a measurement of the strength of the linear relationship that exists between two items. It ranges from -1 to 1, where a value of -1 signifies a fully negative linear correlation, 0 is no correlation, and + 1 means a fully positive correlation [24] .

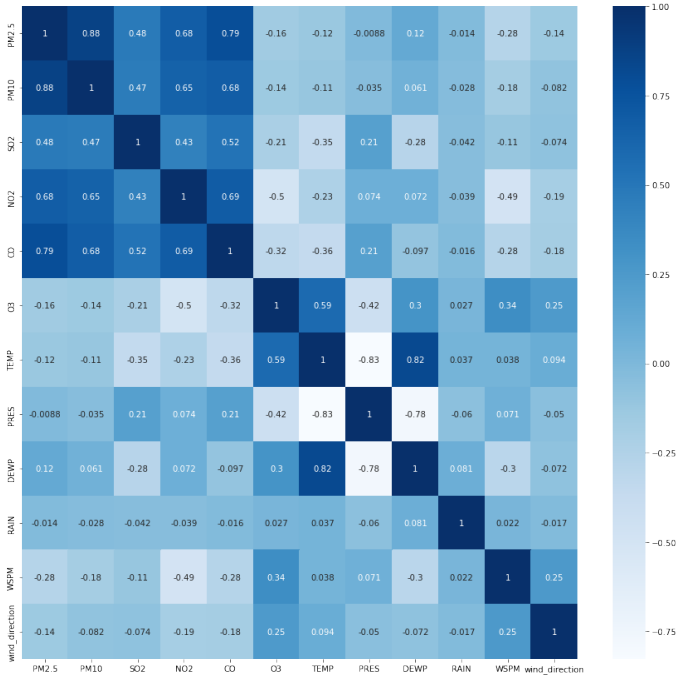


Fig. 7. Correlation between different parameters of our data

Since time series data can exhibit a variety of patterns, it is often useful to decompose a time series into multiple components, each representing an underlying pattern category [25]. In order to finalize the study, time series are decomposed to investigate year-to-year components such as trend, seasonality and residuals, figure 8represents the result.

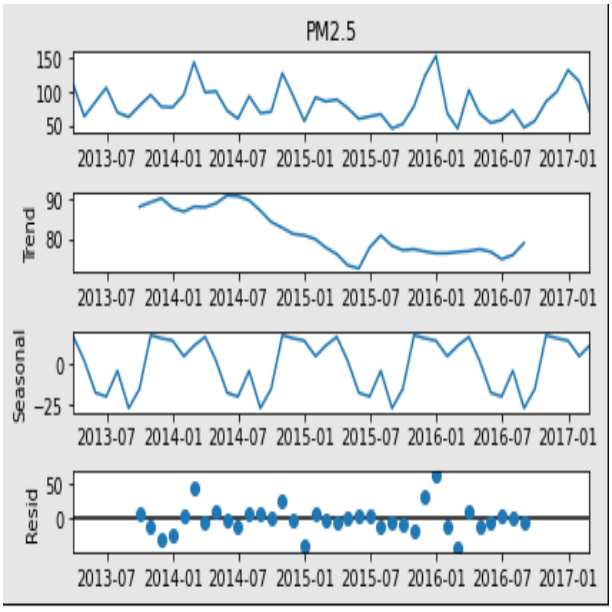


Fig. 8. year-to-year components

After analyzing data, it is noticeable that there is a significant seasonal and trend pattern, and also many parameters which have a low correlation with the target variable. Therefore , we decided to keep only the data which have a strong correlation with pm2. 5 and then we will make three channels out of it: the first one will contain the difference between each day and the previous day to give an overview of the trend to our model, the second will contain the difference between the current value and the same value from the same day of the previous year, as well as the last and main one will contain the current value which will be combined with the second channel to give our model an overview of the seasonal pattern; after all of this, data is reshaped in a 12-day time window. So with the 12-day time frame and the three channels, a good prediction of the next day's PM25 concentrations will be gotten as a result.

To increase the prediction accuracy, data are normalized using the Min-Max normalization methods described in the following equation 1

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

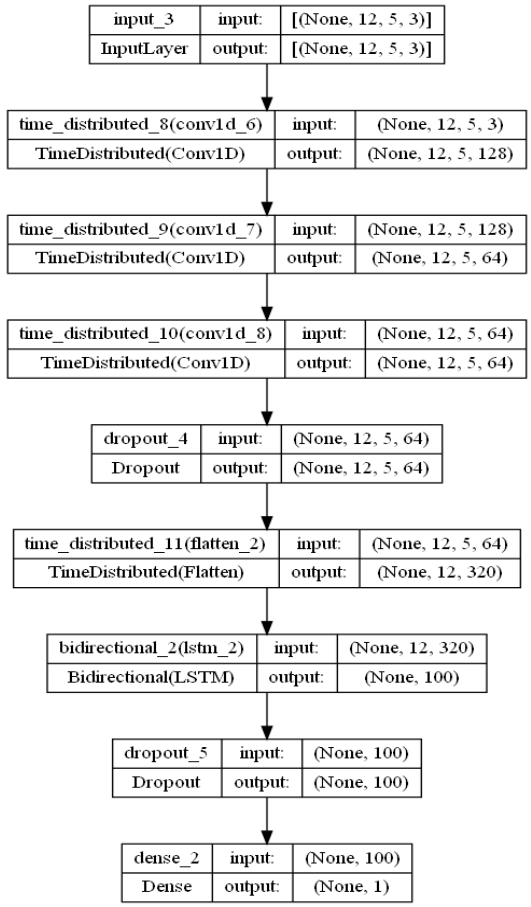


Fig. 9. Internal network structure of the proposed hybrid CNN-LSTM model generated by anaconda platform.

And finally before fitting the model , data is splited to 80% for the training and 20% for the testing, figure 10 illustrate the loss and validation loss functions .

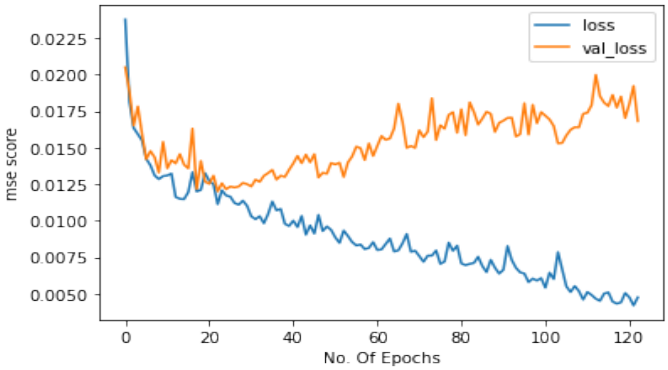


Fig. 10. Loss function

IV. RESULTS AND FINDINGS

In order to evaluate the performance of these models, two indicators were employed, the mean absolute error

(MAE) and the root mean square error (RMSE) defined by the equations 2 and 3.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{y_i} - y_i\right| \tag{2}$$

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(\widehat{y_i} - y_i\right)^2} \tag{3}$$

$\widehat{y_i}$ and y_i indicate the predicted value and true value respectively. Table II shows the RMSE of these nine models over more then 270day , and table III shows the MAE for the same data.

TABLE II
The RMSE of experimental results Over more then 270days

	CNN	Bi-LSTM	CNN-biLSTM
Mi-Mo	59.26	58.66	56.62
Mi-So	57.61	58.66	55.71
Univariate	60.63	60.20	59.65

TABLE III
The MAE of experimental results Over more then 270days

	CNN	Bi-LSTM	CNN-biLSTM
Mi-Mo	40.91	40.09	39.51
Mi-So	41.47	40.03	39.34
Univariate	42.39	42.57	41.52

The figure 11 shows a comparison of true values and predicted values of our model .

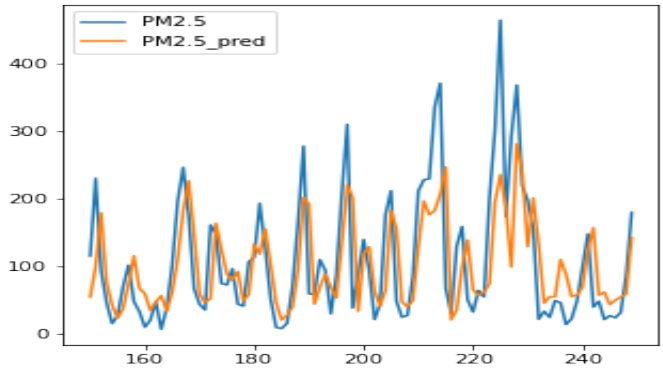


Fig. 11. Comparison of true values and predicted values of our model

Table III show that the MAE of multivariate Multi input single output CNN-LSTM model the value of 39.51 is the minimum. Moreover, the MAE values of multivariate models are obviously lower than that of univariate models, and the MAE values of CNN-BiLSTM models are lower than that of LSTM and CNN models.
Similarly, it is shown in table II that the RMSE of multivariate CNN-LSTM model is the lowest with a value of 55.71. The RMSE values of multivariate models are obviously lower than that of univariate models, and the RMSE values of both the MIMO and MISO CNN-LSTM models are slower than that of LSTM and CNN models. Therefore, according to the previous results, some conclusions on the technical aspects can be summarized: All air quality data patterns in Beijing have a certain time periodicity. However, after a thorough analysis of the PM2.5 concentration, it is clear that there is an annual periodicity in the time series, which is obvious, So it is necessary to use a time frame with 3 channels of air quality data as an input for the next day's PM2.5 concentration forecast.
There are multiple air quality data-related features that can influence the accuracy and performance of PM2.5 prediction, so it is critical to adopt data-driven approaches to identify key properties
Almost all of the models come with their own advantages and disadvantages, so it is important to provide a hybrid CNN-BiLSTM model for predicting PM2.5 concentration, in which CNN is employed to extract related features map from the existing air quality features and then LSTM is chosen to perform predictions.

V. CONCLUSION

The air quality patterns have periodicity thing that was obvious while analyzing the air quality data in Beijing. A hybrid CNN-LSTM deep learning network is being proposed based on convolutional neural network and recurrent neural network to make prediction of PM2.5 concentration in Beijing. The advantages of convolutional neural network for extracting features and recurrent neural network for handling time series data are used to enhance the accuracy of air quality prediction. Due to the periodic nature of the air quality data, the air quality-related feature values with a time frame of 3 channels of the previous 12 days data used as an input, and the next day's PM2.5 concentration predicted and got in the output. The prediction process consists of the following steps:
First, the data set was treated, normalized and then divided into training data (the first 80% of the data) and test data (the remaining 20% of the data). Second, the training data was fed to the proposed hybrid CNN-LSTM model. Third, the predicted values were compared with the actual values. Finally, the model's performance was evaluated by two indicators, which are MAE and RMSE. The models were compared and analyzed, and the results showed that the MISO models perform better than the other models considering the air quality parameters we got, the MAE and the RMSE of which are lower than the others. The multivariate model should be selected if the amount of data is significant and has multiple features, while a univariate model can be considered if it has only one feature.
In the near future we are planning on using a better calculator machine in order to work with hourly data to

allow the model to identify a better pattern of it, to reduce the data's variance and avoid the unpredictable peaks which means that the accuracy of the predicted data will improve and give a much better results. In order to make this work more beneficial we are considering adding the self attention mechanism and encoder decoder architecture to make this comparative study more general.

References

[1] https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html.

[2] AILSHIRE, Jennifer A. et CRIMMINS, Eileen M. Fine particulate matter air pollution and cognitive function among older US adults. American journal of epidemiology, 2014, vol. 180, no 4, p. 359-366.

[3] PÖSCHL, Ulrich. Atmospheric aerosols: composition, transformation, climate and health effects. Angewandte Chemie International Edition, 2005, vol. 44, no 46, p. 7520-7540.

[4] DU, Yixing, XU, Xiaohan, CHU, Ming, et al. Air particulate matter and cardiovascular disease: the epidemiological, biomedical and clinical evidence. Journal of thoracic disease, 2016, vol. 8, no 1, p. E8.

[5] HYNDMAN, Rob J. et ATHANASOPOULOS, George. Forecasting: principles and practice. OTexts, 2018.

[6] KELLEHER, John D. Deep learning. MIT press, 2019.

[7] LUAN, Yuandong et LIN, Shaofu. Research on text classification based on CNN and LSTM. In : 2019 IEEE international conference on artificial intelligence and computer applications (ICAICA). IEEE, 2019. p. 352-355.

[8] CHEN, Kan, WANG, Jiang, CHEN, Liang-Chieh, et al. Abc-cnn: An attention based convolutional neural network for visual question answering. arXiv preprint arXiv:1511.05960, 2015.

[9] Yeo, I., Choi, Y., Lops, Y. et al. Efficient PM2.5 forecasting using geographical correlation based on integrated deep learning algorithms. Neural Comput Applic 33, 15073–15089 (2021).

[10] KELLEHER, John D. Deep learning. MIT press, 2019.

[11] BROWNLEE, Jason. Deep learning for time series forecasting: predict the future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery, 2018.

[12] HAQUE, Md Amaan, VERMA, Abhishek, ALEX, John Sahaya Rani, et al. Experimental evaluation of CNN architecture for speech recognition. In : First international conference on sustainable technologies for computational intelligence. Springer, Singapore, 2020. p. 507-514.

[13] ZHAO, Jianfeng, MAO, Xia, et CHEN, Lijiang. Speech emotion recognition using deep 1D 2D CNN LSTM networks. Biomedical signal processing and control, 2019, vol. 47, p. 312-323.

[14] ABDELJABER, Osama, AVCI, Onur, KIRANYAZ, Serkan, et al. Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. Journal of Sound and Vibration, 2017, vol. 388, p. 154-170.

[15] SHIN, Hoo-Chang, ROTH, Holger R., GAO, Mingchen, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging, 2016, vol. 35, no 5, p. 1285-1298.

[16] ELMAN, Jeffrey L. Finding structure in time. Cognitive science, 1990, vol. 14, no 2, p. 179-211.

[17] HOCHREITER, Sepp et SCHMIDHUBER, Jürgen. Long short-term memory. Neural computation, 1997, vol. 9, no 8, p. 1735-1780.

[18] YU, Yong, SI, Xiaosheng, HU, Changhua, et al. A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 2019, vol. 31, no 7, p. 1235-1270.

[19] GERS, Felix A., SCHMIDHUBER, Jürgen, et CUMMINS, Fred. Learning to forget: Continual prediction with LSTM. Neural computation, 2000, vol. 12, no 10, p. 2451-2471.

[20] SØNDERBY, Søren Kaae et WINTHER, Ole. Protein secondary structure prediction with long short term memory networks. arXiv preprint arXiv:1412.7828, 2014.

[21] XU, Guixian, MENG, Yueting, QIU, Xiaoyu, et al. Sentiment analysis of comment texts based on BiLSTM. Ieee Access, 2019, vol. 7, p. 51522-51532.

[22] KHAN, Asifullah, SOHAIL, Anabia, ZAHOORA, Umme, et al. A survey of the recent architectures of deep convolutional neural networks. Artificial intelligence review, 2020, vol. 53, no 8, p. 5455-5516.

[23] SRIVASTAVA, Nitish, HINTON, Geoffrey, KRIZHEVSKY, Alex, et al. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 2014, vol. 15, no 1, p. 1929-1958.

[24] BENESTY, Jacob, CHEN, Jingdong, HUANG, Yiteng, et al. Pearson correlation coefficient. In : Noise reduction in speech processing. Springer, Berlin, Heidelberg, 2009. p. 1-4.

[25] HYNDMAN, Rob J. et ATHANASOPOULOS, George. Forecasting: principles and practice. OTexts, 2018.