Investigation on the Impact of Attention Mechanism in Deep Learning Models for Temperature Prediction

1st Naba Krushna Sabat
Dept. of ECE
NIT, Rourkela
Rourkela, India
nabakrushna_sabat@nitrkl.ac.in

2nd Umesh Chandra Pati
Dept. of ECE
NIT, Rourkela
Rourkela, India
ucpati@nitrkl.ac.in

3rd Santos Kumar Das
Dept. of ECE
NIT, Rourkela
Rourkela, India
dassk@nitrkl.ac.in

***Abstract*—Prediction of the meteorological parameters, such as temperature, humidity, rainfall, wind speed, etc., is a crucial task for industrial and agricultural applications. In recent years deep learning techniques have become more popular for predicting the time series weather data because of their accuracy and promising result. However, adding an attention mechanism in the deep learning model provides more long-term prediction accuracy. This article investigates the potential of attention-based deep learning models for improving the forecasting accuracy of the meteorological parameter temperature. The attention mechanism helps in improving the forecasting accuracy, which is evident from the experimental result analysis in terms of key performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).**

***Index Terms*—Weather forecasting, deep-learning, attention mechanism**

I. Introduction

Weather prediction means the process of estimating the likely state of the atmosphere at a specific date and time. The climatology parameters such as temperature, wind, humidity, precipitation changes abruptly because of natural factors such as solar radiation, volcanic eruption, and human-induced factors like greenhouse gasses, deforestation, pollution, etc. This changes of weather parameters have a significant effect on leaving live [1]. Hence, it is highly essential to measure accurate weather parameter that helps people prepare for severe weather events such as hurricanes, tornadoes, and blizzards, which can save lives and minimize property damage. In agriculture it helps the farmer to plan planting and harvesting schedules, as well as to protect crops from damage due to extreme weather.

Several techniques are adopted to predict the time series data, such as numerical, statistical, machine learning, deep learning, etc. The traditional methods rely on numerical models that use historical data and physical principles to predict future weather patterns. The univariate statistical methods namely auto-regressive integrated moving average (ARIMA), Prophet, Seasonal ARIMA [2], SARIMAX, and multivariate statistical models such as vector auto-regressive (VAR), vector moving average (VMA), and a combination of VAR and VMA known as VARMA are used. However, these methods have limitations in terms of accuracy and scalability because they provide good accuracy on linear input data but are unable to handle non-linear data.

To surpasses the traditional statistical models machine learning models such as Random Forest (RF), Support Vector Machine (SVM), Linear Regression (LR), extreme gradient boost (xgboost), Support Vector Regression (SVR) are used in weather prediction [3]. It gives more accurate result than the statistical method. However this method has limitation to complex input data for long-term forecasting. Recently, with the advancement of technology and computing processes, deep learning models such as Long Short Term Memory (LSTM) [4], stacked LSTM (sLSTM) [5], Gated Recurrent Unit (GRU) [6], Convolutional Neural Network (CNN) [7], Temporal Convolutional Network (TCN) [8], Bi-directional LSTM (Bi-LSTM) [9], Neural Basis Expansion Analysis for Time Series (NBEATS) [10] and Bidirectional GRU (Bi-GRU) [11], are being used for weather prediction, which provide promising results and faster prediction times for long-term forecasting. The models has the ability to analyze larger datasets and identify complex patterns in the data. However, there is ample scope to increase the accuracy of deep learning models. This paper describes the use of attention mechanisms and their impact in deep learning models to increase the accuracy of time-series weather parameters. The four popular deep learning models: LSTM, GRU, Bi-LSTM, and Bi-GRU, are investigated with self-attention mechanism and the results are compared with baseline models.

The rest of the article are structured as follows: Section II discusses the problem statement. The related work is briefly explained in Section III. Section IV describes the methodologies that includes data collection and pre-processing methods, proposed model in detail. The experimental result and analysis of various deep-learning models is presented in Section V. Finally, the article is laid out to conclude in Section VI.

II. Problem Statement

The current study intends to predict one of the meteorological parameter, i.e., temperature, by using the historical time-series temperature data of Dongsi. The time series modeling of the temperature data will be predicted using various widely used models, such as LSTM, GRU, BiLSTM, and BiGRU. In addition, the self-attention mechanism is included to improve prediction accuracy. Finally, this research investigated the impact of the attention mechanism on the deep learning model and aimed to select the best-performing time-series models for predicting Dongsi temperature.

III. Related Work

For time series data prediction, the deep learning method becomes a popular approach because of its ability to automatically extract and learn complex features from raw weather data, making it well-suited for time series weather prediction. This related survey only focus the recent works have been implemented for time series weather data prediction presented in Table I.

TABLE I

Deep learning models used for climatology data prediction

Authors	**Deep learning models used**	**Considering weather parameter**
M. Yu, et al. [12]	LSTM	Temperature
D.Kreuzer, et al. [13]	Convolutional LSTM	Temperature
R. Pandit, et al. [14]	Bi-LSTM	Wind speed
M. Chhetri, et al. [15]	Bilstm-GRU	Rainfall
N. K. Sabat, et al. [16]	NBEATS, GRU, BiGRU, BiLSTM	Temperature
F. Peng, et al. [17]	LSTM-RNN	Temperature, humidity
K Venkatachalam, et al. [18]	T-LSTM	Temperature, pressure, humidity
H. Liu et al. [19]	Convolutional GRU	Wind speed
A. Lawal, et al. [20]	CNN and BiLSTM	Wind speed

IV. Methodologies

Following steps are used in the methodologies to perform this research.

A. Data Collection and Preprocessing

Based on the research scope, the place of meteorological data is located. The raw meteorological data is then transformed to a format that is more conducive to being used by the deep learning model.

1) Research Location: Dongsi, a sub-district situated in Dongcheng District of Beijing, China. This area is located at 39.932°N latitudes and 116.4341°E longitudes with an elevation of 50m and is found at the GPS coordinates $24^{\circ}21'51''N117^{\circ}46'22''E$. Fig. 1 depicts the study area filled with brown, black and white inclined line along with smaller black circle filled with the color green. The country China has placed number of weather stations, one of which is situated in the sub-district Dongsi, Bejing. Subsequently, the temperature parameter of the weather data of Dongsi city is considered for this research work.

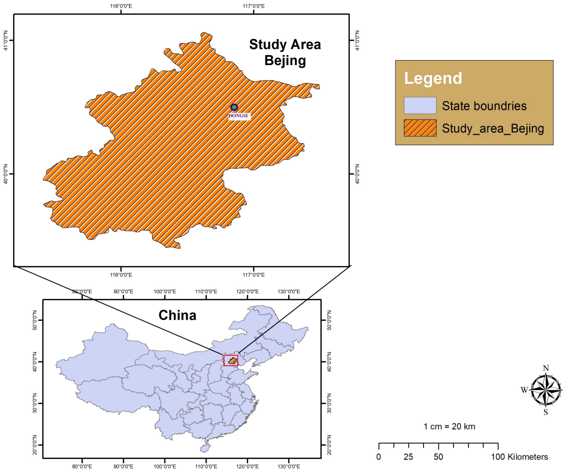

Fig. 1. Research area.

2) Data Preprocessing: This analysis used publicly accessible 2013 to 2017 Dongsi, China data [21]. The hourly temperature data is obtained first, and then it is transformed into daily average temperature data to minimize dimensionality. The neural network needs pre-processed raw data. Preprocessing included data outlier elimination, missing value imputation, data normalisation, etc. Here, Quantile regression method is used to find and remove the data outlier so that the data is not skewed in any way during the modeling process. The median imputation technique is then used to fill in the rest of the gaps in the temperature time series data. In order to train a deep learning-based model smoothly, it is common practise to apply data normalisation to convert the original data from its large-scale variation range into the 0-1 range. Consequently, temperature readings are re-scaled from 0 to 1 using the Min-Max normalisation method. The entire data set is then split into a training set (consisting of 80% of the data), a validation set (10% of the data), and a testing set (10% of the data).

B. Developed Deep Learning Model

In this stage, different widely used deep learning models such as LSTM, GRU, BiLSTM, and BiGRU, are used to model the time-series temperature data of the Dongsi and fed each model's prediction result to the self-attention mechanism for improving the prediction accuracy. The flow of each step for this research is illustrated in the block diagram in Fig. 2.

The baseline models used for the prediction of temperature are briefly explained as follows:

1) LSTM Model: LSTM is a type of RNN model popularly used for processing sequential data. To circumvent the "vanishing gradient" issue seen in conventional RNNs, LSTM models were designed. It has a memory cell and three gates, such as input, output, and forget, to regulate the flow of information.

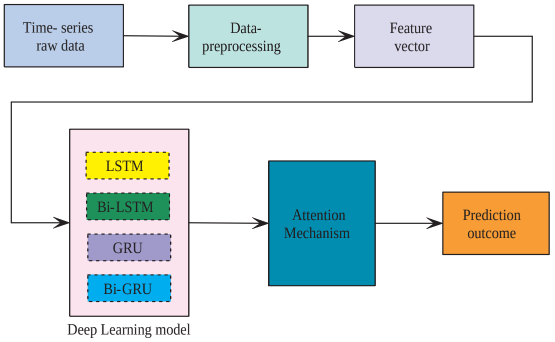

Fig. 2. Block diagram of deep-learning model with attention mechanism for temperature prediction.

The mathematical function working behind the LSTM cell correspond to n^{th} state of input (I_n), output (O_n), forget (F_n), and cell state (C_n) is represented as [4],

$$\begin{aligned} I_n &= \sigma(W^I.[h_{n-1}, X_n]) + b^I \\ F_n &= \sigma(W^F.[h_{n-1}, X_n]) + b^F \\ t_n &= tanh(W^t.[h_{n-1}, X_n]) + b^t \\ O_n &= \sigma(W^O.[h_{n-1}, X_n]) + b^O \\ C_n &= F_n * C_{n-1} + t_n * I_n \end{aligned} \tag{1}$$

Where, W^I, W^F, W^t, W^O and b^I, b^F, b^t, b^O are the weights and bias correspond to input, forget, memory cell and output state of LSTM model respectively [22].

2) Bi-LSTM Model: The Bi-LSTM network more powerful than traditional LSTM networks in processing of time-series data because in Bi-LSTM the information flow in both forward and backward direction. The forward direction processes the input sequence from the first element to the last, and the backward direction processes the input sequence from the last element to the first. This allows the Bi-LSTM network to capture information from both the past and future contexts of a given element in the sequence. The outputs from both the forward and backward directions are concatenated to produce the final output of the network.

3) GRU Model: It is a type of RNN, designed to capture long-term dependencies in sequential data such as time series, text, speech, etc. GRU has two gates, “reset” and “update”, to control the flow of information, which allows it to better handle vanishing gradients compared to basic RNNs. The mathematical formulas guiding the GRU cell’s performance in accordance with the reset (R_n), update (U_n), hidden (h_n) and new hidden $(\tilde{h_n})$ for n^{th} state is expressed as [6],

$$\begin{aligned} R_n &= \sigma(W^R.[h_{n-1}, X_n]) + b^R \\ U_n &= \sigma(W^U.[h_{n-1}, X_n]) + b^U \\ \tilde{h_n} &= tanh(W^h.[R_n * h_{n-1}, X_n]) + b^h \\ h_n &= (1 - U_n) * h_{n-1} + U_n * \tilde{h_n} \end{aligned} \tag{2}$$

where, the parameters W^R, W^U, W^h and b^R, b^U, b^h are the weights of bias of reset, update and hidden states.

4) Bi-GRU Model: The Bi-GRU architecture processes the input sequence in two directions, i.e. forward and backward. The forward and backward hidden states are then combined to capture information from both the past and future context in the input sequence.

5) Self Attention Mechanism: The self-attention mechanism captures the dependencies between different time steps, rather than just modeling the linear relationships between adjacent time steps as in traditional time series models. The attention scores are used to weight the contribution of each time step to the final prediction, allowing the model to capture non-linear relationships between time steps. Basically, in a self-attention mechanism, the input sequence is transformed into three different vectors: query, key, and value. These vectors are computed for each time step in the sequence and then utilised to determine the attention weights. The dot product of the query and the key vectors is used to compute the attention weights. These attention weights are used to compute a weighted sum of the value vectors that represents the information associated with the current time step. This process is repeated for each time step in the input sequence, allowing the model to focus on the most relevant time steps when making predictions.

6) Performance Evolution Criteria: The performance of the models are evaluated using three most popular error metrics criteria such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The error metrics are calculated as follow [16]. The error performance are calculated using the Eq. (3 - 5). The model with the lowest MAE, MSE, RMSE score is considered as the best model for prediction.

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |\beta_t| \tag{3}$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n} (\beta_t)^2 \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (\beta_t)^2} \tag{5}$$

Where, β_t denotes the difference between actual value A_t and predicted value $\widehat{A_t}$, i.e. β_t = A_t - $\widehat{A_t}$. The term n represent the number of data observation.

V. Experimental Results and Analysis

The deep learning models such as LSTM, GRU, attention based LSTM, GRU, Bi-GRU, and Bi-LSTM are executed in google colab platform. Keras, pandas, tensorflow, scikit learn and matplot library of python language are used to design the model. The epochs that are being utilized here have each been set to 100. Both the batch size and optimizer have been set to 64, with Adam as the optimizer. MSE and Rectified Linear unit (ReLU) is used in place of a loss function, and an activation function respectively. Table II outlines the various error calculation parameters that are utilised by all deep learning models. These parameters include MAE, MSE, RMSE score. Fig. [3 - 8] represents the deep learning models LSTM, GRU, attention based LSTM, attention based GRU, attention based BiGRU, and attention based BiLSTM as well as their test predictions. The error performance comparison is shown in the Fig. 9.

TABLE II
PERFORMANCE EVALUATION OF DEEP LEARNING MODEL

Model	MAE	MSE	RMSE
LSTM	2.086	7.402	2.721
GRU	2.011	6.706	2.589
LSTM + Attn. Mechanism	2.359	9.147	3.024
GRU + Attn. Mechanism	1.981	6.583	2.565
BiGRU + Attn. Mechanism	1.942	6.289	2.507
BiLSTM + Attn. Mechanism	**1.845**	**5.679**	**2.383**

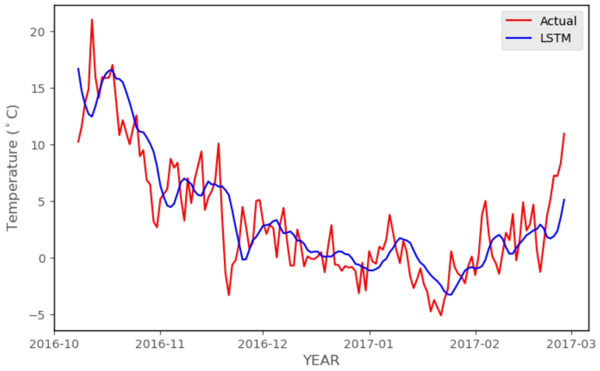

Fig. 3. Prediction analysis of temperature using LSTM model.

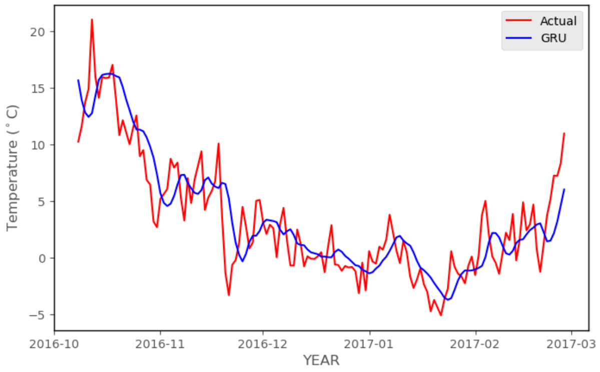

Fig. 4. Prediction analysis of temperature using GRU model.

VI. CONCLUSION

This research examines the effectiveness of the self-attention mechanism in four widely used deep learning models: LSTM, GRU, Bi-LSTM, and Bi-GRU. It has been explored how attention mechanisms improve the efficiency of models. All the models are evaluated and compared using error metrics such as MAE, MSE, and RMSE. The comparative analysis states that the Bi-LSTM model with attention mechanism provides the lowest MAE, MSE, and RMSE values for the study area. Further, a more complex hybrid model, such as a transform model with an attention mechanism, may be utilized to further improve prediction accuracy.

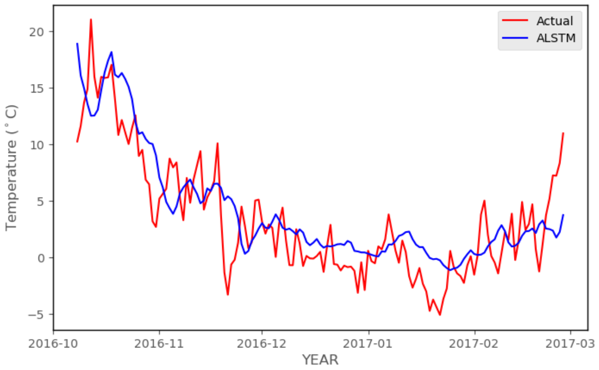

Fig. 5. Prediction analysis of temperature using attention based LSTM model.

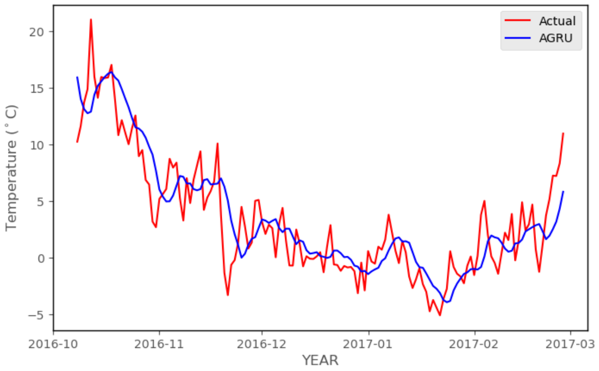

Fig. 6. Prediction analysis of temperature using attention based GRU model.

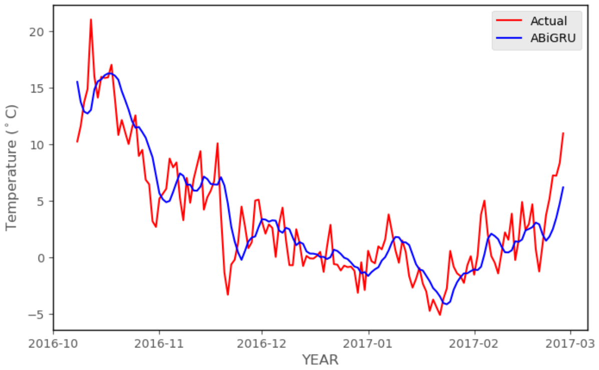

Fig. 7. Prediction analysis of temperature using attention based BiGRU model.

REFERENCES

[1] M. Brennan, T. Hennessy, D. Meredith, and E. Dillon, “Weather, workload and money: determining and evaluating sources of stress for farmers in ireland,” *Journal of agromedicine*, vol. 27, no. 2, pp. 132–142, 2022.

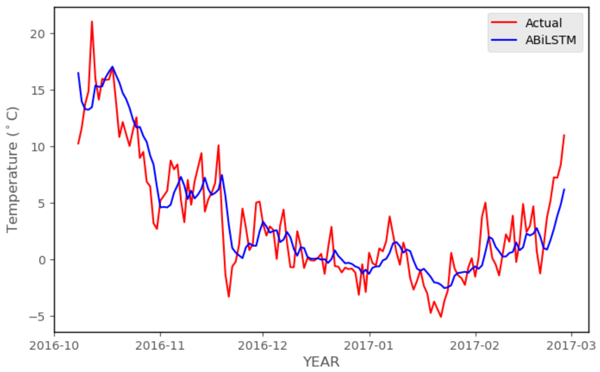

Fig. 8. Prediction analysis of temperature using attention based BiLSTM model.

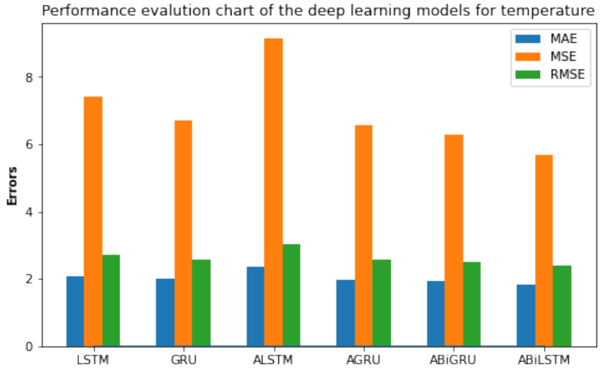

Fig. 9. Comparison of error performance for temperature prediction.

[2] A. P. Kogekar, R. Nayak, and U. C. Pati, "Forecasting of water quality for the river ganga using univariate time-series models," in *2021 8th International Conference on Smart Computing and Communications (ICSCC)*. IEEE, 2021, pp. 52–57.

[3] R. Nayak, M. Tejoyadav, P. Mohanty, and U. C. Pati, "Water quality time-series modeling and forecasting techniques," *Artificial Intelligence of Things for Weather Forecasting and Climatic Behavioral Analysis*, pp. 177–200, 2022.

[4] S. Xingjian, Z. Chen, H. Wang, D. Yeung, and W. Wong, "Woo wc (2015) convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, pp. 802–810.

[5] S. Ghimire, R. C. Deo, H. Wang, M. S. Al-Musaylh, D. Casillas-Pérez, and S. Salcedo-Sanz, "Stacked lstm sequence-to-sequence autoencoder with feature selection for daily solar radiation prediction: a review and new modeling results," *Energies*, vol. 15, no. 3, p. 1061, 2022.

[6] D. Zhang and M. R. Kabuka, "Combining weather condition data to predict traffic flow: a gru-based deep learning approach," *IET Intelligent Transport Systems*, vol. 12, no. 7, pp. 578–585, 2018.

[7] M. Qiu, P. Zhao, K. Zhang, J. Huang, X. Shi, X. Wang, and W. Chu, "A short-term rainfall prediction model using multi-task convolutional neural networks," in *2017 IEEE international conference on data mining (ICDM)*. IEEE, 2017, pp. 395–404.

[8] P. Hewage, M. Trovati, E. Pereira, and A. Behera, "Deep learning-based effective fine-grained weather forecasting model," *Pattern Analysis and Applications*, vol. 24, no. 1, pp. 343–366, 2021.

[9] T. Peng, C. Zhang, J. Zhou, and M. S. Nazir, "An integrated framework of bi-directional long-short term memory (bilstm) based on sine cosine algorithm for hourly solar radiation forecasting," *Energy*, vol. 221, p. 119887, 2021.

[10] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-beats: Neural basis expansion analysis for interpretable time series forecasting," *arXiv preprint arXiv:1905.10437*, 2019.

[11] M. Jaihuni, J. K. Basak, F. Khan, F. G. Okyere, T. Sihalath, A. Bhujel, J. Park, D. H. Lee, and H. T. Kim, "A novel recurrent neural network approach in forecasting short term solar irradiance," *ISA transactions*, vol. 121, pp. 63–74, 2022.

[12] M. Yu, F. Xu, W. Hu, J. Sun, and G. Cervone, "Using long short-term memory (lstm) and internet of things (iot) for localized surface temperature forecasting in an urban environment," *IEEE Access*, vol. 9, pp. 137 406–137 418, 2021.

[13] D. Kreuzer, M. Munz, and S. Schlüter, "Short-term temperature forecasts using a convolutional neural networkan application to different weather stations in germany," *Machine Learning with Applications*, vol. 2, p. 100007, 2020.

[14] R. Pandit, D. Astolfi, A. M. Tang, and D. Infield, "Sequential data-driven long-term weather forecasting models performance comparison for improving offshore operation and maintenance operations," *Energies*, vol. 15, no. 19, p. 7233, 2022.

[15] M. Chhetri, S. Kumar, P. Pratim Roy, and B.-G. Kim, "Deep blstm-gru model for monthly rainfall prediction: A case study of simtokha, bhutan," *Remote sensing*, vol. 12, no. 19, p. 3174, 2020.

[16] N. K. Sabat, R. Nayak, U. C. Pati, and S. Kumar Das, "A comparative analysis of univariate deep learning-based time-series models for temperature forecasting of the bhubaneshwar," in *2022 IEEE 2nd International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, 2022, pp. 1–5.

[17] F.-L. Peng, Y.-K. Qiao, and C. Yang, "A lstm-rnn based intelligent control approach for temperature and humidity environment of urban utility tunnels," *Heliyon*, 2023.

[18] K. Venkatachalam, P. Trojovskỳ, D. Pamucar, N. Bacanin, and V. Simic, "Dwfh: An improved data-driven deep weather forecasting hybrid model using transductive long short term memory (t-lstm)," *Expert Systems with Applications*, vol. 213, p. 119270, 2023.

[19] H. Liu, X. Mi, Y. Li, Z. Duan, and Y. Xu, "Smart wind speed deep learning based multi-step forecasting model using singular spectrum analysis, convolutional gated recurrent unit network and support vector regression," *Renewable Energy*, vol. 143, pp. 842–854, 2019.

[20] A. Lawal, S. Rehman, L. M. Alhems, and M. M. Alam, "Wind speed prediction using hybrid 1d cnn and blstm network," *IEEE Access*, vol. 9, pp. 156 672–156 679, 2021.

[21] S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, and S. X. Chen, "Cautionary tales on air-quality improvement in beijing," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 473, no. 2205, p. 20170457, 2017.

[22] A. P. Kogekar, R. Nayak, and U. C. Pati, "A cnn-bilstm-svr based deep hybrid model for water quality forecasting of the river ganga," in *2021 IEEE 18th India Council International Conference (INDICON)*. IEEE, 2021, pp. 1–6.