International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Comparative Analysis of LSTM, BILSTM and ARIMA for Time Series Forecasting on 116 years of Temperature and Rainfall Data from Pakistan

Asif Khokhar[1], Shahnawaz Talpur[2], Mohsin A. Memon[3]
[1 3] Department of Software Engineering, Mehran University of Engineering and technology Jamshoro Sindh, Pakistan
[2]Department of Computer System Engineering, Mehran University of Engineering and technology Jamshoro Sindh, Pakistan

ARTICLEINFO

Article History:

Accepted: 01 April 2023
Published: 15 April 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
350-357

ABSTRACT

Numerous aspects of human life, including agriculture, transportation, and health, are significantly influenced by weather, both economically and socially. Rain has an impact on landslides, floods, and other natural disasters. We are motivated to create a model for comprehending and forecasting rain in order to provide advanced warning in a spectrum of areas such as transport, agriculture, and so on because of the numerous consequences that rain and temperature have on human survival. In this study, a dataset for temperature and rainfall for Pakistan for 116 years is used. Comparative analysis of ARIMA, LSTM and BILSTM is performed. For this study, 90% of data is used for training and the 10% for testing. Normalization is also performed to clean data. According to the results, LSTM and BILSTM are better than ARIMA but for specific cases of rainfall, BILSTM performed better than LSTM and for Temperature LSTM outperformed BILSTM.

Keywords : LSTM, BILSTM, ARIMA, Rain Forecasting and Temperature Forecasting.

I. INTRODUCTION

One of the major problems that earth faces today is climate change. Climate change has affected every aspect of the natural environment. Climate change has disturbed the biodiversity, affected the temperature and rainfall. Rain and Temperature has a significant influence on many facets of human existence, both economically and socially, such as agriculture, transportation and health. Rain has an effect on natural catastrophes, including floods and landslides. The many effects of rain and temperature on human existence motivate us to develop a model for understanding and forecasting rain in order to offer early warning in a variety of fields/needs such as transportation, agriculture, and so on.

The availability of water supplies and a number of other areas of life are being impacted by climate change in both industrialized and developing nations. Over the past several decades, study on climate change brought on by global warming has gained significant relevance [11]. Pakistan is one of those developing countries that relies heavily on agriculture for economic means so forecasting Rain and temperature are of immense importance [12].

The increase in applications of machine learning in the field of environment has enabled humans to create intelligent applications for the prediction and forecasting of occurrence of environmental events. Researchers have been working on prediction of different meteorological elements such as rain, temperature and humidity using machine learning [13]. The key techniques for performing prediction of elements like temperature and rainfall are regression analysis and time series analysis [14]. Among the time series analysis the most common algorithms for making such forecasting are ARIMA, LSTM and BILSTM [15]. In this study comparative analysis of ARIMA, LSTM and BILSTM is performed for both temperature and rainfall data from Pakistan ranging from 1901 to 2016. The research questions for this study are:

1. Are deep neural networks better at forecasting time series data than ARIMA for two different datasets?
2. Which Neural network architecture performs better?
3. Are there any trades-offs related to the time for these techniques?

Rest of the paper is arranged as section 2 is literature review where related work is discussed, section 3 is methodology that exhibits how this study is conducted and in section 4, experiments and results of the study are discussed in detail and section 5 is conclusion that shows impact, limitation and future work of this study.

II. RELATED WORKS

Pakistan's economy is based on agriculture and the country is extremely vulnerable to climate change. Climate change is a natural phenomenon that occurs because of greenhouse gas emissions caused by industrialization, deforestation, urbanization, and fuel burning, resulting in changes in temperature and precipitation [1]. Agriculture is Pakistan's largest industry, accounting for roughly 53% of total employment. This sector accounts for 25% of Pakistan's GDP [2]. There are several studies that show the impact of rainfall and temperature on the yield of a crop [3] [4]. So in order to deal with the problem caused by fluctuation in temperature and rainfall it is important to have a mechanism for the reliable forecast of rain and temperature for the years to come. One of the most common elements of climate modeling and agricultural studies is spatial-temporal analysis of rainfall in [5] researchers have used numerous interpolation techniques and found that EBKRP (Empirical Bayesian Kriging Regression Prediction) works best in this case. In [6] researchers have used MCA maximum covariance analysis between temperature and precipitation to find patterns in variability in the environment. In [7] researchers have used the patterns in monsoon rainfall in India to find its relationship on the EL-Nino cycles. In [8] researchers have created MLR Multiple Linear Regression and PCR Principal Component Regression models which are used in rainfall prediction in Pakistan. In [9] research discusses the study of one-day annual-max rainfall for Nawabshah and Hyderabad cities in Sindh, Pakistan from 1961 to 2011. The analysis was conducted using the STATISTA tool for interpolating and predicting rainfall time series. One way to analyse and forecast any perimeter is to perform the trend analysis so in [10] researchers have performed the prediction of rainfall using the trend analysis.

III. METHODOLOGY

Figure 1 shows the methodology diagram for this study. There are four main steps involved in the methodology 1 is Data collection and cleaning, 2 is data analysis, 3 is splitting dataset and 4 is Modeling and evaluation. Let us discuss each of these steps in detail.

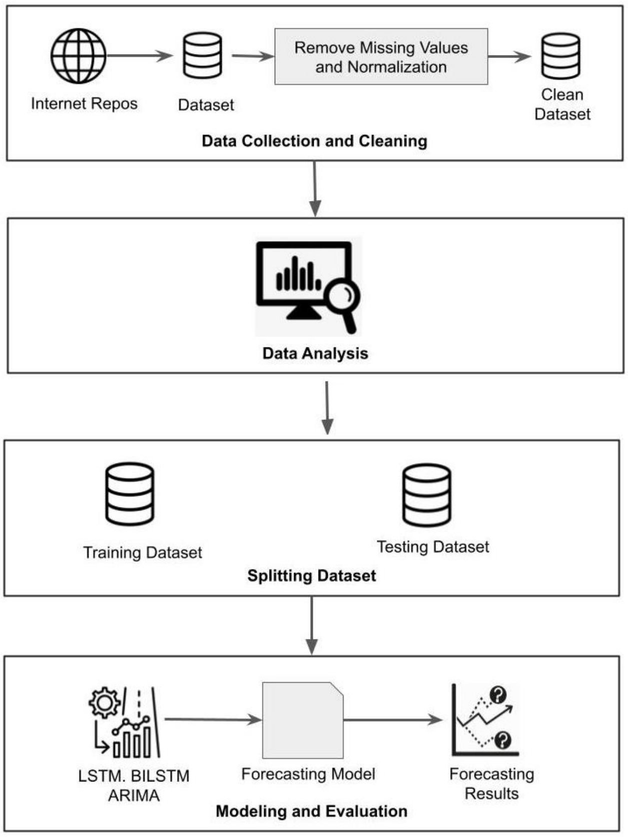

Fig. 1: Methodology Flow Diagram of the study

A. Data Collection and Cleaning:

First step in any machine learning flow is the collection of dataset. In this study two datasets are used, one dataset is used for rainfall forecasting and second is used for temperature forecasting. Both of these datasets are collected from an online repository [15] [16]. Both dataset contains 1392 rows from the year 1901 to 2016. Data is recorded each month. After the collection of both dataset, there is a need for cleaning them for missing values. Both datasets are checked for missing values and if present missing values are replaced by the mean of the column. After the dataset is free of any missing values it is fed to standard scaler for normalization so that the scale of the dataset becomes consistent for the forecasting.

B. Data Analysis:

The datasets are reviewed and we observed from the dataset. Figure 2 shows the monthly analysis of the rainfall and temperature in millimeter and Celsius respectively. According to figure 1, the months with highest rainfall are July and August due to the monsoon season and October and November had the least amount of rain. The hottest months of the year are June, July and August and the coldest month is January.

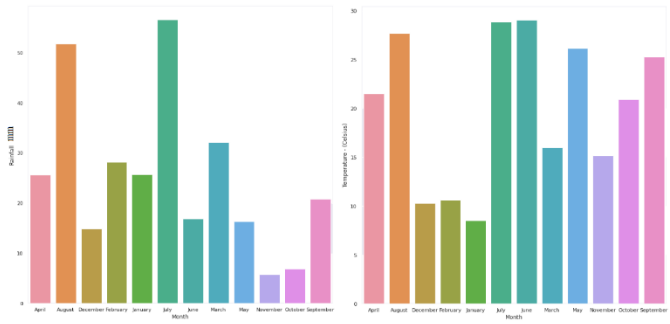

Fig. 2: Monthly Rainfall and temperature analysis.

Figure 3 shows the season based rainfall and temperature. Summer and spring are the seasons with highest rainfall and highest temperature although winter is coldest but least rainfall occurs in autumn. With the increase in temperature chances of rain increases.

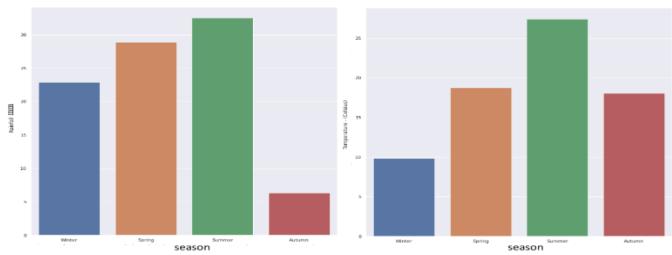

Fig. 3: Seasonally Rainfall and temperature analysis.

Table 1 shows the year with highest rainfall. 1944 and 1956 are the years with highest rainfall of 494 mm and 431 mm respectively. This table shows the top 5 years with highest rainfall. Last in this list is the year 1959 with rainfall of 404 mm.

Table 1: Highest rainfall in years.

5 Highest Rainfall Years	
Year	**Rainfall**
1944	494.37035

1956	431.99396
1994	423.27579
1942	414.70332
1959	404.21003

Table 2 shows the top 5 years with highest average temperature throughout the year. This data is from 1901 to 2016 so all these values lie in this range. 2016 is the hottest year in the dataset followed by 2004. The last on this list is the year 1941 with an average temperature of 20.

Table 2: Highest Temperature in years.

5 Highest Temperature Years	
Year	**Average Temperature**
2016	21.414617
2004	21.051189
2010	20.999923
2006	20.994692
1941	20.984646

C. Data Splitting

To train a model there is need for two types of dataset: one for training and one for testing. In this study, 90% of the dataset is taken for training and 10% of the dataset is taken for validation.

D. Modelling and Evaluation:

Once there is data for training and validation. Modelling and evaluation can be performed. Training set is passed to various algorithms to generate models and these models are evaluated using a testing set.

There are numerous ways to model time series data, but the popular among them are Moving averages and Deep Neural Networks. Among these neural networks the sequential data is conventionally trained using RNN but RNN lacks the ability to remember the previous data, which is crucial for time series forecasting. A special type of RNN is LSTM and BILSTM that has ability to remember previous input that makes it suitable for this use case. For moving averages, ARIMA is considered among the best models for time series forecasting. Discussing them one by one.

LSTM: Long - Short - term Memory Networks, most often referred to as "LSTMs," are a unique class of RNN that can recognize long-term dependencies. They were first presented by Schmidhuber & Hochreiter (1997), and several authors developed and popularized them in subsequent works. They are currently frequently utilized and perform incredibly well when applied to a wide range of issues. The long-term reliance problem is specifically avoided with LSTMs. They do not strain to learn; rather, remembering knowledge for extended periods is their default habit. LSTMs are able to selectively recollect or forget things in this way. There are three different dependencies on the information at a specific cell state.

1. The prior condition of the cell
2. The prior Hidden condition
3. The input for this time interval

Figure 4 shows the LSTM architecture.

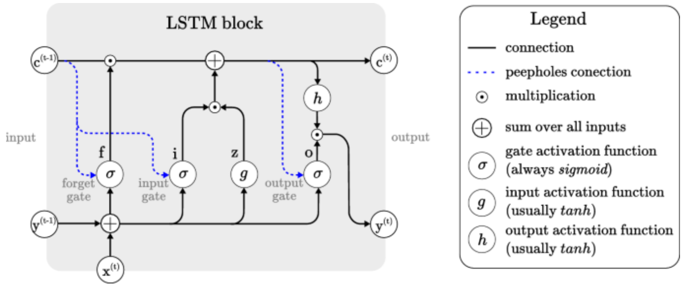

Fig. 4: LSTM Architecture [17]

BILSTM: A RNN used largely for sequence processing is called Bidirectional LSTM (BILSTM). It may use data from both sides and, unlike regular LSTM, the data flows in both directions. In both ends of the sequence, it is a potent tool for modelling the sequential relationships between sequences. In conclusion, BILSTM changes the direction of data flows by adding one extra LSTM layer. It simply implies that in the extra LSTM layer, the input data flows backward. The outputs from the two LSTM layers are then combined in a variety of ways, including average, sum, multiplication, and concatenation. Figure 5 shows the BILSTM architecture.

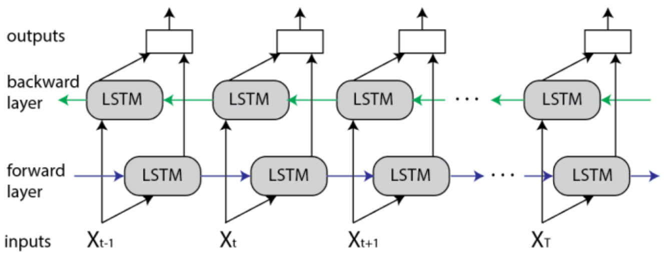

Fig. 5: BILSTM Architecture [18]

ARIMA: The acronym ARIMA, which stands for "Auto Regressive Integrated Moving Average," refers to a family of models that uses a time series' own previous values—specifically, its own delays and lagged prediction errors—to "explain" the time series in order to predict future values. ARIMA models may be used to represent any "non-seasonal" time series that has patterns and is not just random noise. The key elements of ARIMA is:

1. Based on historical values, ARIMA models forecast future values.
2. Lagged moving averages are used by ARIMA to smooth time-series data.
3. They are frequently utilized in quantitative analysis to predict upcoming asset price trends.

IV. RESULTS AND DISCUSSIONS

A. ARIMA Results:

ADF Test: An often-used statistical test to determine whether a particular Series is stationary or not is the (ADF Test) Augmented Dickey Fuller test. When examining the nature of a series, it is one of the statistical tests that is most frequently applied. Null hypothesis is not stationary to reject null hypothesis p values must be less than 0.05. By looking at it can be seen that both of the series are stationary. It is important to check if time is stationary or not, because ARIMA is applied on Stationary time series now it is confirmed that series are stationary. Let's have a look at ARIMA's results.

Table 3. ADF Results

Dataset	Test Statistics	P-Value	Stationary
Rainfall Dataset	-6.695889008362762	4.0007354131483325e-09	Yes
Temperature Dataset	-3.791598778074882	0.002994449601225318	Yes

ARIMA models are chosen based on some criteria in this study, that criteria is AIC, Akaike information criterion, the lesser the value of AIC the better the models are for both of the datasets. The best performing ARIMA models are shown in table 3 with AIC of 10964.602 and 3756.720 for rainfall and temperature dataset respectively.

Table 4: ARIMA results

Dataset	AIC	Time	Model
Rainfall Dataset	10964.602	59.374 seconds	ARIMA(0,0,1)(1,1,0)
Temperature Dataset	3756.720	49.201 seconds	ARIMA(1,0,0)(1,1,0)

Figure 6 shows the RMSE score of the model generated from ARIMA. RMSE is the root mean squared error. The lesser the values of RMSE the better the performance. The RMSE for rainfall model is 16.37 and for temperature model it is 1.14.

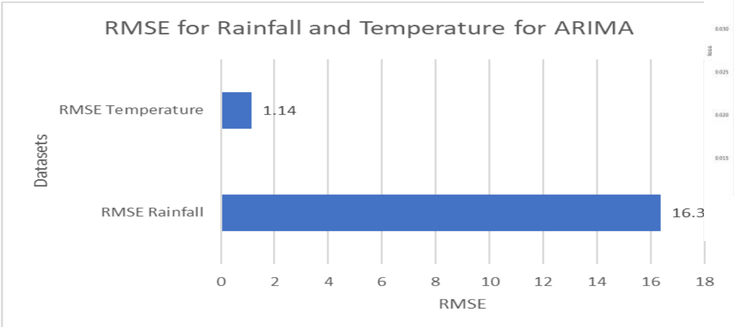

Figure 6: RMSE for ARIMA of Both Datasets

B. LSTM and BILSTM on Rainfall dataset

After ARIMA, LSTM and BILSTM are applied on the data and results of those models are discussed in this section. Figure 7 shows the RMSE results for LSTM and BILSTM models which are trained on 100, 300 and 500 epochs. LSTM has higher RMSE score for all epochs, BILSTM has better performance and lower RMSE for all three 100, 300 and 500 epochs. Best results for both of them are achieved on 500 epochs.

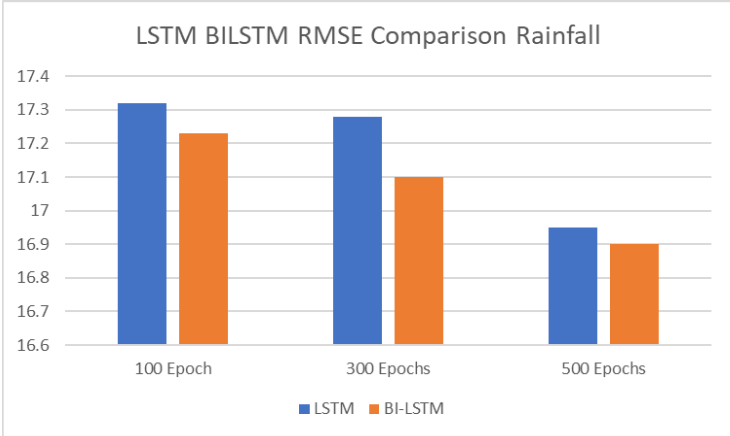

Fig. 7: RMSE for LSTM and BILSTM for Temperature

Figure 8 shows the training and validation curve of LSTM and BILSTM for rainfall dataset at 500 epochs. Curves Show that the BILSTM model converges better than the LSTM model.

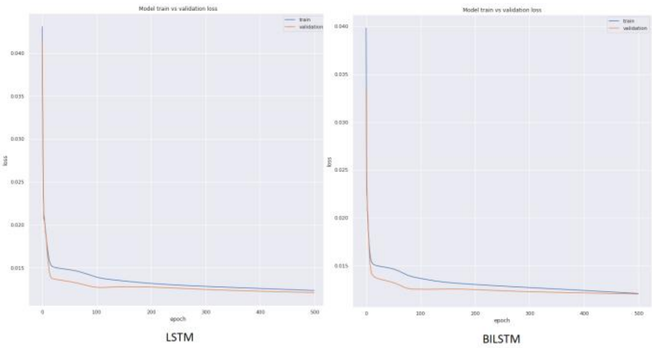

Fig. 8: Training and Validation Loss curve

C. LSTM and BILSTM on Temperature dataset

This section applies the models LSTM and BILSTM to the data and discusses the outcomes.

Figure 8 shows the RMSE results for both LSTM and BILSTM models trained on 100, 300 and 500 epochs. For 100 and epochs BILSTM worked better, for 300 epochs LSTM and BILSTM have the same performance metric and for 500 epoch LSTM performed better than BILSTM.

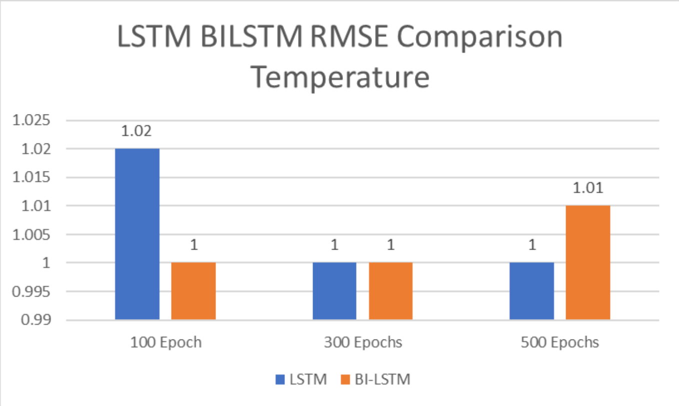

Fig. 9: RMSE for LSTM and BILSTM for Temperature

At 500 epoch due to very minor difference in results both of the loss curve looks almost the same both converging well. Although BILSTM has very minor differences which cause RMSE to increase.

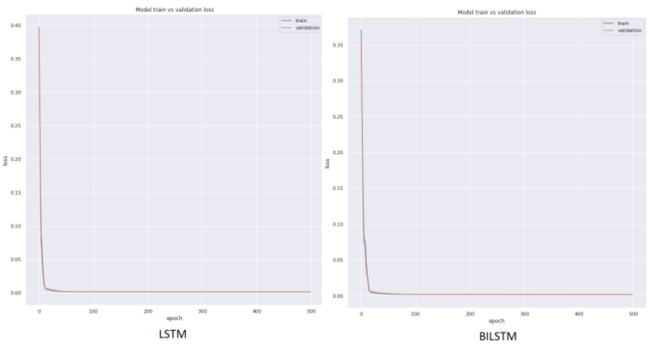

Fig. 10: Training and Validation Loss curve

V. CONCLUSIONS

Climate change is one of the main issues the planet is now facing. Every part of the natural surroundings has been impacted by climate change. The temperature, rain, and biodiversity have all been impacted by climate change. In this atmosphere of uncertainty there is dire need of prediction models for various factors such as temperature and rainfall especially for countries like pakistan where dependence of agriculture is high. In this study ARIMA, LSTM and BILSTM were used to create predictive models for rainfall and temperature. Based on the evaluation LSTM and BILSTM are better than ARIMA and for rainfall BILSTM works better and for Temperature LSTM provided good results.

VI. REFERENCES

[1]. Abbas, Sohail, and Zulfiqar Ali Mayo. "Impact of temperature and rainfall on rice production in Punjab, Pakistan." Environment, Development and Sustainability 23.2 (2021): 1706-1728.

[2]. Imran, Muhammad Ali, et al. "Impact of Climate Smart Agriculture (CSA) practices on cotton production and livelihood of farmers in Punjab, Pakistan." Sustainability 10.6 (2018): 2101.

[3]. Siddiqui, R., Samad, G., Nasir, M., & Jalil, H. H. (2012). The impact of climate change on major agricultural crops: evidence from Punjab, Pakistan. The Pakistan Development Review, 261–274.

[4]. Hanif, U., Syed, S. H., Ahmad, R., Malik, K. A., & Nasir, M. (2010). Economic impact of climate change on the agricultural sector of Punjab [with comments]. The Pakistan Development Review, 49, 771–798.

[5]. Ali, Ghaffar, et al. "Spatial–temporal characterization of rainfall in Pakistan during the past half-century (1961–2020)." Scientific reports 11.1 (2021): 1-15.

[6]. Syed, F.S., Kucharski, F.: Statistically Related Coupled Modes of South Asian Summer Monsoon Interannual Variability in the Tropics. Atmos. Sci. Lett. 17(2), 183–189 (2016)

[7]. Kucharski, Fred and Muhammad Adnan Abid. 2017. "Interannual Variability of the Indian Monsoon and Its Link to ENSO." Oxf. Res. Encycl. Clim. Sci. (November):1–24

[8]. Adnan, M., Rehman, N., Ali, S., Mehmood, S., Mir, K.A., Khan, A.A., Khalid, B.: Prediction of Summer Rainfall in Pakistan from Global Sea-Surface Temperature and Sea-Level Pressure. Weather. 72(3), 76–84 (2017)

[9]. Mahessar, Ali Asghar, et al. "Rainfall Analysis for Hyderabad and Nawabshah, Sindh,

Pakistan." Engineering, Technology & Applied Science Research 10.6 (2020): 6597-6602.

[10]. Falzon, Greg, et al. "ClassifyMe: a field-scouting software for the identification of wildlife in camera trap images." Animals 10.1 (2020): 58.

[11]. Hattermann, F.F.; Krysanova, V.; Gosling, S.; Dankers, R.; Daggupati, P.; Donnelly, C.h.; Flörke, M.; Huang, S.; Motovilov, Yu.; Buda, S.; et al. Cross-scale inter-comparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins. Clim. Change 2017, 141, 561–576.

[12]. Yaqoob, Nusrat, et al. "The effects of agriculture productivity, land intensification, on sustainable economic growth: a panel analysis from Bangladesh, India, and Pakistan Economies." Environmental Science and Pollution Research (2022): 1-9.

[13]. Singh, Nitin, Saurabh Chaturvedi, and Shamim Akhter. "Weather forecasting using machine learning algorithm." 2019 International Conference on Signal Processing and Communication (ICSC). IEEE, 2019.

[14]. Anjali, T., et al. "Temperature prediction using machine learning approaches." 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT). Vol. 1. IEEE, 2019.

[15]. Usmani, Z.-ul-hassan. Pakistan temperature. Kaggle. Retrieved March 1, 2023, from https://www.kaggle.com/datasets/zusmani/pakistan-temperature (2021, March 19)

[16]. Usmani, Z.-ul-hassan. (2021, March 13). Rainfall in Pakistan. Kaggle. Retrieved March 1, 2023, from https://www.kaggle.com/datasets/zusmani/rainfall-in-pakistan

[17]. Van Houdt, Greg, Carlos Mosquera, and Gonzalo Nápoles. "A review on the long short-term memory model." Artificial Intelligence Review 53.8 (2020): 5929-5955.

[18]. Alawneh, Luay, et al. "A comparison of unidirectional and bidirectional lstm networks for human activity recognition." 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). IEEE, 2020.

Cite this article as :

Asif Khokhar, Shahnawaz Talpur, Mohsin A. Memon, "Comparative Analysis of LSTM, BILSTM and ARIMA for Time Series Forecasting on 116 years of Temperature and Rainfall Data from Pakistan", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 2, pp.350-357, March-April-2023. Available at doi : https://doi.org/10.32628/CSEIT2390238

Journal URL : https://ijsrcseit.com/CSEIT2390238