Predictive Analysis on Climate Change of Pakistan using Machine Learning Algorithms

Abdul Wasay, Muhammad Talha Siddiqui

Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology
SZABIST 99 3rd Ave, Block 5 Clifton, Karachi, 75600, Pakistan

Abstract— Climate had been changing negatively throughout the world due to excessive Global warming and weakening of ozone layer and Pakistan situated in Southern Asia is facing its effects immensely in form of extreme temperatures, less rainfall and poorer air quality. The knowledge of how climate change adaptation is an important aspect of existing policy sectors and operations, it is crucial to be sure of timely climatic actions across different levels. The study of climate change has been made difficult due to the lack of relevant data available, the accuracy of the existing methods proving to be ineffective. Introduction of machine learning algorithms to interpret the data and predicting the change in future will provide the researchers with the information required either to prevent it from happening or to create adaptation policies for it. We will be using various models defined in our research onto our data scrapped from different sources and illustrate its usefulness with quantitative analysis. Our interpretations will be providing a predictive analysis for the researchers in analyzing the trends and patterns from the historical patterns to the predicted patterns and take precautionary measures in how to avoid any disastrous situation. We culminate the paper by reflecting on the merits and pitfalls of using ML algorithms, which algorithm proves to be the most efficient and provides with the best accuracy for the resultant values.

Keywords— LSTM (**Long Short-Term Memory**), RNN (**Recurrent Neural Network)**, Linear Regression, Time Series Forecasting, MSE (**Mean Square Error**), MAE (**Mean Absolute Error**).

I. Introduction

A. Background

Climate change has been affecting life for hundreds of years and is expected to continue changing the world in the future. Global warming is at its peak and is expected to increase in the near future having numerous side effects on the environment and Pakistan being a south Asian country is facing the effects of global warming with extreme temperatures, poor air quality, low precipitation hence having various effects on to the environment and people residing in it.

B. Research Question

Since we know climate change is a global issue of our time and we are at a very critical moment to where the climate is headed due to great change in global warming. For us, where is the climate of Pakistan headed in the near future and how will it affect the society? What will be the effects of global warming for climate of Pakistan and when should we prepare for the disasters that are inevitable?

C. Study Purpose

The ML models we will be using will be providing us with the prediction of temperatures, rainfall and greenhouse gases. We will be selecting a particular model with the best accuracy such that the information it provides can be used either to prepare for the inevitable and take precautions before the time arrives or the government should take initiatives for a better environment and control on greenhouse gases.

II. Literature Review

Literature review is a necessity in the project development. It provides with the understanding of the project with its background. It also provided with the detail for following the best practices in development of project. Literature review provided with the details of how other researchers used different techniques to get their result along with analyzing dependencies and feasibility of the project. Literature review also suggests various tools and algorithms for project development.

1. In this study [3], series of monthly mean temperature and monthly total precipitation, the two climatic variables, were analyzed in forty-seven stations across Spanish peninsula. ARIMA models were used to reflect the changes in trends over the period 1940-2013. A study was conducted to identify possible trends from different regions to examine the variations in trend over time. 12-month predictions were made using ARIMA models, with more than 50% of the series modeled by both.

2. In the study [5], India is mentioned where 60% of the citizens relying upon the agriculture. Rain fall prediction is the maximum essential undertaking for predicting early forecast of rainfall. The paper represents easy linear regression approach for the early prediction of rainfall to help farmers for taking suitable decisions on crop yielding. The easy linear regression analysis technique was implemented at the dataset gathered over six years of Connor in Nilagris district from Tamil Nadu state. The experiment and simple linear regression method take advantage of the suitable consequences for the rain fall [2].

3. In this paper [4], the authors applied ANN and CNN in weather forecasting, it shows how their future predictions were solely based on previous input values. As CNN are a form of deep learning technique that can help classify, recognize, and predict trends in climate change and environmental data. In this review, two NN system designs have been selected to be utilized as weather prediction models for vitality utilization. The first one is ANN, which is used several algorithms such as Long Short-Term Memory. And secondly CNN was chosen for different applications. In conclusion this paper compares the accuracy of both CNN and ANN on weather predictions while acknowledging their differences and it was noted that ANN forms have a significantly higher performance than CNN.
4. This paper [2], explains how the climate has changed in the past 100 years as e global average temperature has increased by approximately 0.6°C in addition, global atmospheric CO2 concentrations have risen by nearly 38%, considering these factors and taking global warming as a major indication, climate change may effectively alter ecosystem structure and function. in light of increasing concerns over these ecological issues the authors proposed the use of ANN model as they are relatively accurate when used for short-term predictions. Paper also states that even though global climate is a long-term experimental research ANN's remain a better choice than many traditional methods when dealing with nonlinear problems, and possesses great potential for the study of global climate change and ecological issues. They can solve problems in difficult situations in which measurements are difficult to conduct or when only incomplete data are available. It is also anticipated that ANN will be widely adopted for the study and research of global climate change.
5. In this paper [6] techniques of data mining in the field of Meteorology are put in the spotlight. By using the process of data mining, it is made possible to extract relevant data and determining mutual factors from a data set. This process assisted in predicting outcomes that help in making a suitable decision, by analyzing the data set with respect to various possible perspectives and identifying patterns, relations and correlations about different weather elements including temperature, wind pressure, humidity and rainfall etc. A system is designed using K-Nearest Neighbors (KNN) Algorithm for weather prediction. However, after comparing the predicted results with the real results it was observed that the comparison was very suitable and acceptable.
6. In this paper [7], it is explained that how Long Short-Term Memory (LSTM) is used to predict sea surface temperature (SST). LSTM is a recurrent neural network, and SST is formulated as a time series regression problem. The introduced network architecture is made up of a layer of LSTM and a full-connected dense layer where first layer is used for time series relationship modelling and the other layer is utilized for mapping the outcome of LSTM taken to a final prediction result. The team carried out experiments on the coastal seas of China to reassure the effectiveness and accuracy of the system.

III. METHODOLOGY

A. Workflow Diagram

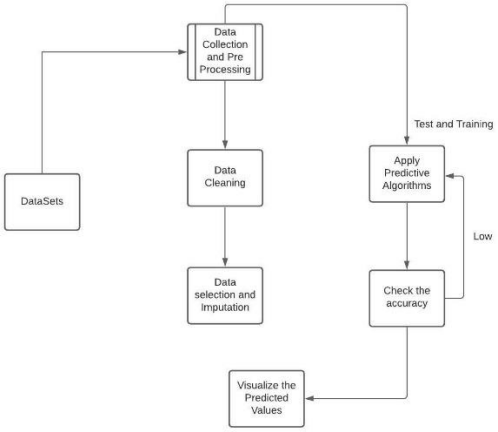

Fig 1. Generalized Workflow of Project

We went through the approach as described in above figure 1 in four phases, Data collection and data preprocessing, data cleaning, data selection according to our requirement and data imputation. The general factors we will be working upon are temperature, rainfall, humidity and greenhouse gases factors. The data is scrapped from different sources and is worked upon our predictive models for the prediction of climate such as day, month or year wise and its adverse effects on the environment. Our ML models will be processing the climatic data of Pakistan consisting of different factors as well as those effecting climate change. Our model will be separately trained for each factor and tested with. Our systems would be tested for efficiency & accuracy in-order to find prediction pattern for the near future. Once we have the predictive analysis of those factors, the predicted data will be used to integrate all the factors into one main predictive analysis model that will provide the overall analysis in change of climate over the years and the prediction for the future.

First step of the project would be data collection, after collection data needs to be processed due to which data goes through several steps usually known as data processing cycle. Below are the steps included in data processing cycle.

B. Data Collection

Huge datasets regarding each and every factor of climate contains immense importance when it comes to predicting the future climate. Data collection is the most time taking task in any project and data has to be in an appropriate condition to be

used. Since Pakistan's climate data is not easily available as there are no such data lakes available, data had to be scrapped from different websites.
Climatic factors such as rainfall and temperature were scrapped from a website with the help of a python script that scrapped the data from the website and saved it in the excel format. Collection of Air quality factors and sea level were available on websites but since all this data was available in year wise, we had to interpolate this data.
The data collected for rainfall, Greenhouse gases, AQI and temperature comprises from 12 cities of Pakistan.
Below is the table with complete Data members.

TABLE I
DATA MEMBERS

S.NO	Attribute	Type	Description
1.	DATE	Datetime	Year, month, day.
2.	AVG	Float	Average temperature
3.	MAX	Float	Maximum temperature
4.	MIN	Float	Minimum temperature
5.	R.H	Float	Relative Humidity
6.	PRCP	Float	Total Precipitation
7.	SLP	Numeric	Mean Sea level
8.	FF	Float	Fossil Fuels emission
9.	NO	Float	Nitrogen Oxide emission
10.	CO2	Float	Carbon Dioxide emission
11.	Meth	Float	Methane emission
12.	Flu Gases	Float	Fluorinated gases
12.	GHG	Float	Total Green House

C. Data Analysis

Data analysis is an important step before moving ahead with the creation of model. Within this process are data cleaning and transformation of data for modeling data.
The scrapped data was in different excel files, so we appended all these files with the help of python script for appending excel files.
Once we had the Complete data we scrapped from the websites, they had many missing values which couldn't be eradicated from the data since it will affect our efficiency and inputting random values would have made our trained model biased. We used interpolation script of python to input the missing values with the help of entire data removing any certainty of biasness. For few factors we had very limited data and for that purpose, we normalized our data and performed over sampling method in python according to our requirement and de-normalized the data for further evaluation.

D. Long Short-Term Memory

Long Short-Term Memory networks are a type of an RNN which are capable of learning long dependencies for a longer period of time. To get the relationship between time series data, adopting LSTM is the best option.

LSTM was first designed by Juergen Schmid Huber's in 1997. It is a kind of recurrent neural network that was designed to model sequences for their long-range dependencies more accurately then a convolutional RNN. It can process sequence of inputs and output in pairs.
Figure .3 shows the structure of a LSTM cell.

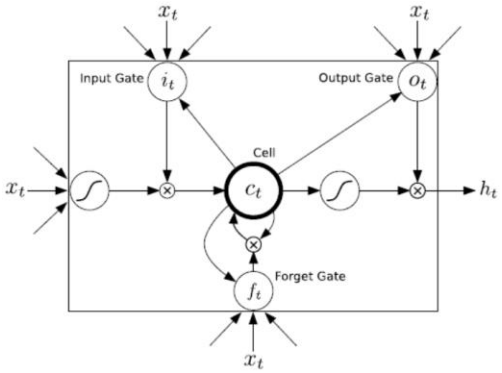

Fig .3 Structure of LSTM cell, [7]

We combined the LSTM with a full connected layer to build a basic LSTM block.
Since there are two basic neural layers in a block, LSTM can capture the temporal relationship. Whilst the output of LSTM layer is a vector, the hidden vector of last time step, for better combination for the output vector, a full-connected layer was used. Fig. 5 shows a full-connected layer.


Fig. 4 A full connected-layer [7]

The architecture of the network is like a cuboid, the x asix stands for latitude while the y axis stands for longitude and the z axis is for the time step. Figure. 6 shows the architecture of the network.

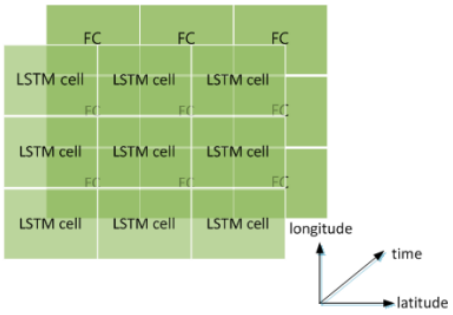

Fig. 5 Network Architecture [7]

IV. Experimental Results and Discussions

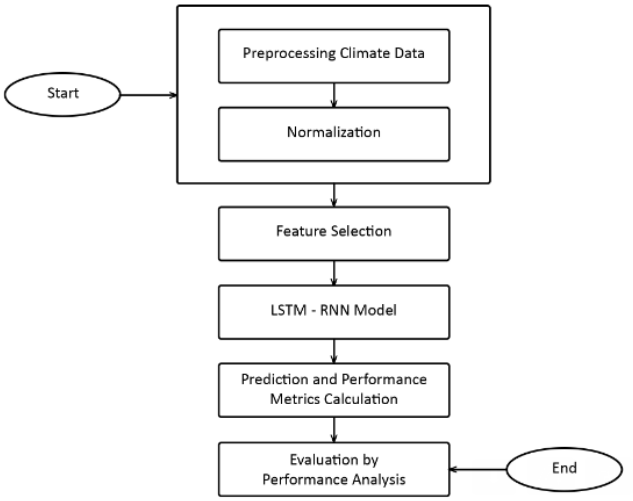

Fig. 6 Generalized LSTM (RNN) working model flowchart.

We used the LSTM recurrent neural network model to predict the maximum and minimum temperature, relative humidity and rainfall pattern based on the historical data. We trained and tested the model using the daily data.

According to these aspects mentioned above, we first designed a simple experiment to determine the critical values using the basic LSTM block to predict average temperature for a particular year.

Once we determined the structure of the network, there are certain things to be determined to train the model. We optimized the model using Adam optimizer. It is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments. This method can speed up the convergence of network training, which can adapt the learning rate to the parameters, performing larger updates for infrequent and smaller updates for frequent parameters

We compared the performance and check the Model validation using Error Calculation.

The (MSE) mean square error and (MAE) mean absolute error is a method of measuring how accurate a predicted system is.

Mean Square Error

$$MSE = \frac{1}{N}\sum_{t-0}^{n}(y_t - \acute{y}_t)^2 \qquad (1)$$

- Y_I = Actual value
- Y_I = Predicted value
- N = Total number of samples.

We predicted the maximum temperature for Karachi for the next 5 years. We only used 10 epochs initially since we built a beginner model and once, we got the results, now we will be focusing upon 100 epochs to get a better result. Below is the Pseudocode for prediction.

1. Import the Libraries required.
2. Import the Dataset file, print its head
3. Append the data with time step and check the Xtrain Xtest shape.
4. Set ip units, lstm units, op units and optimizer to define LSTM Network (A)
5. Normalize the dataset (Di) into values from 0 to 1.
6. Select training window size (tw) and organize Di accordingly
7. **for**
   n epochs and batch size do
   Train the Network (A)
   Check for Epoch loss with time and find MSE and MAE.
   end for
8. Run Predictions using A.
9. Read the predicted values onto excel and save.

Table II explains the test case for Maximum temperature trained on 10 epochs using LSTM model.

TABLE III
Test Case

Test Case ID	Test Case Name	Test Case Summary	MSE	MAE
1.	Max Temp Forecasting	Forecasting maximum temperature using LSTM (RNN)	0.1225275	0.0269616

The figure. 4 shows a graph plotted is to compare between the actual values and the predicted values for the given data.

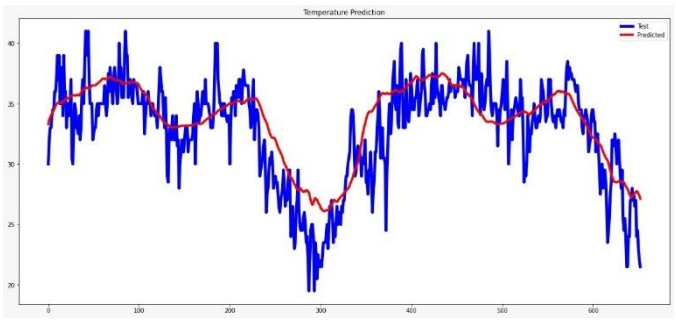

Fig. 7 Comparing actual and predicted values

Figure. 5 shows the 5 years of predicted maximum temperature in the city of Karachi.

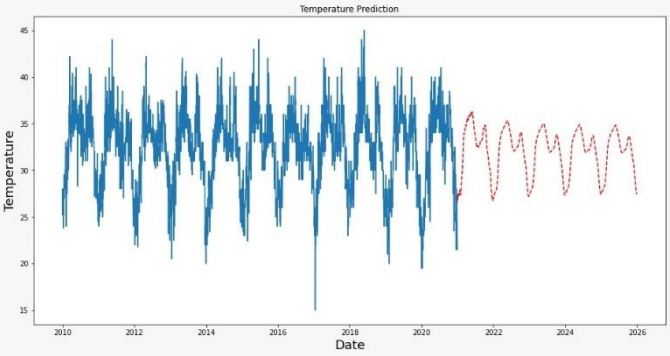

Fig. 8 Five years of predicted maximum temperature

The table III shows the head of the data frame that shows the actual and predicted values.

TABLE IIIII
Actual and Predicted Values

Serial #	Actual	Predicted
0	30.0	29.332
1	32.0	31.221
2	33.0	33.846
3	34.5	34.023
4	32.5	33.889

we continue to train the model adding five years of the predicted values to get the result for the next 5 years, and get a new model for ourselves.
We will repeat this process for our different parameters listed in Table I.

V. Conclusion

Data mining techniques are an efficient method of analyzing and compiling data from various sources. These techniques are helpful for the pre-processing of the data and using ML algorithms for prediction is a better solution then those inefficient ways country's meteorological department opt for.

This paper concludes that using Data mining techniques for data gathering and pre-processing and using LSTM along with Recurrent Neural network for climate change prediction gives a close to accurate result and can be considered as an alternative for previous traditional methods. Comparison shown in this paper between actual and predicted values showcases that the use of these algorithms is of one of the best suited techniques for this application. For this reason, the main dependency in predicting these factors is the unbiased availability of the data.

VI. Future Work

The results of the proposed work are efficient and encouraging for us to explore the domain further by including more factors and extending our scope with disaster prediction as well in relate to the predicted values.

Based on our result we can conclude that an accurate prediction model will help the country in various sectors especially in agriculture to cultivate their farms according to the weather predicted and also in disaster management to take precautionary measures long before the time arrives.

In future, we would like to develop this model further based on different climatic parameters as well as disaster prediction.

REFERENCES

[1] Khan, Z. &. (2014). *Hourly Based Climate Prediction Using Data Mining Techniques by Comprising Entity Demean Algorithm.*

[2] Liu Z L, P. C. (2010). Application of artificial neural networks in global climate change and ecological research. An overview. . *ChineseSci Bull, 2010, 55: 3853.*

[3] Mulomba Mukadi, P. G.-G. (2021). *Time Series Analysis of Climatic Variables in Peninsular Spain. Trends and Forecasting Models for Data between 20th and 21st Centuries.*

[4] S. Kareem, Z. J. (2021). An evaluation of CNN and ANN in prediction weather forecasting: A review. *Sustainable Engineering and Innovation, vol. 3, no. 2, pp. 148-159.*

[5] Sreehari, E. &. (2019). Climate Changes Prediction Using Simple Linear Regression. *Journal of Computational and Theoretical Nanoscience.*

[6] Yousif. (2013). *Weather Prediction System Using KNN Classification Algorithm.*

[7] Muhuri, P. &. (2020). *Using a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to Classify Network Attacks.*

[8] Jitendra Kumar, R. G. (2018). *Long Short Term Memory Recurrent Neural Network (LSTM-RNN) Based Workload Forecasting Model For Cloud Datacenters.*