Global Scientific JOURNALS

GSJ: Volume 12, Issue 1, January 2024, Online: ISSN 2320-9186

www.globalscientificjournal.com

Comparative study between Machine Learning Models and a Deep Neural Network Model for Air Temperature Prediction in Zambia

Gershom Chishala[1,*], Professor Christopher Cheembe[2], Dr. Nyirenda Mayungo[3] Mwenya, T Musonda[4]

1. ICT Department, Kapasa Makasa University, Chinsali, Zambia.gershomchishala8@gmail.com
2. ZCAS University, Zambia. christopher.chembe@zcas.edu.zm
3. Computer Science Department, University of Zambia. mayumbo.nyirenda@cs.unza.zm
4. Computer Science department, Copperbelt University. mtmusonda2016@gmail.com

Abstract:

Accurate temperature forecasting is important and can be used in specific applications such as agriculture, weather prediction, energy management and early warning systems in cases of extreme cold and hot temperatures. This research focuses on leveraging on the capabilities of deep learning to enhance temperature predictions accuracy. Machine learning models for regression were used as baseline models to compare with the outcome of the Deep Neural Network. The DNN model proposed in this research employs a multi-layer neural network which learns relationships and patterns in the historical dataset. The dataset was downloaded from kaggle website and extracted data specific to Zambia. The model performance was evaluated using various metric which included R-Squared, Mean Absolute percentage error and root mean squared error. The results demonstrated that the DNN model moderately outperformed traditional machine learning models on the test data. The baseline machine learning models included both linear and ridge regression, K-Nearest Neighbor, Random forest. The findings from the research can be instrumental in improving the accuracy of weather forecasting which can in turn help in agriculture planning, agriculture adaptation and climate monitoring.

Key Words: Temperature, deep neural network, regression, random forest, Pre-processing

1. Introduction

Atmospheric temperature variation affects a lot of natural systems that eventually impact human life. Consistent hydro power generation, solar energy management, Agriculture planning and weather monitoring all tap into the knowledge of atmospheric temperature variation throughout the year. In recent decades, adverse effects of climate change in general have negatively affected a broad spectrum of human activities and sometimes threating its very existence [1]. A lot of systems have been designed and implemented to help monitor and predict temperature changes.

However, the main constrain has been the accuracy [2] of the systems used to predict the temperature variations which use different features as inputs.

Various machine learning linear regression models have been used to help achieve temperature change forecasting. In order to have good results, regression models should heavily depend on a lot of data with many features [3]. Overfitting is a common problem that linear regression models have [4]. They tend to perform very well on the training data but fails to generalize on the test data or unseen data. This creates a poorly performing model that cannot offer accurate prediction. Similar studies have been conducted howbeit on water surface temperature which had different parameters to consider [1]. Traditional machine learning models like Gradient Boosting Tree, linear regression and Random Forest have been studied and used for example in predicting temperature and humidity in Kuala, Terengganu state, Malaysia [5]. GBT, RF and linear regression models performed less than deep learning models. Many studies have shown that traditional machine learning models give poor performance on nonlinear data like temperature variations at compared to multi-layer perceptions MLP [1].

1.1 Research Problem

This study aims at comparing the results of three traditional machine learning models (Linear Regression, Ridge Regression, K-Nearest Neighbor, and Random Forest) and a multi-layer deep neural network. Since linear models give good results when the number of features is comparatively larger than the number of samples in a dataset [4], we used it in this study as a baseline algorithm since our dataset had fewer features and more sample data. Again, because regression models perform poorly on nonlinear data like the one contained in our dataset, we used them here purely as baseline models [1]. To solve these problems highlighted, we decided to use feed the two features in our dataset into a Deep Neural Network which performs better on nonlinear data. Generally, neural network models perform better than traditional machine learning models and so our proposed DNN model is expected to give us better results compared to the baseline models. Many research problems available have used neural networks for air temperature forecasting but on different geographical target areas from the one explored in this paper.

1.2 Related Work

Use of both machine learning and deep learning in temperature prediction has continued to be researched on available datasets. In [6, 3], three neural network models from Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM) network and a combination of a Convolution Neural Network (CNN) and LSTM were used to compare their accuracy on the John F. Kennedy international airport corpus historical dataset. The combination of LSTM and CNN outperformed the other data sets. Although the study showed that the accuracy increased with an in increase in the complexity of the neural network, they could not get accurate results for more future distant days. LSTM was experimented in predicting surface water temperature of a reservoir.

Other researchers have also experimented with MLP, LSTM and CNN as independent models applying them to high frequency features with daily average, minimum and maximum temperature forecasts [7]. Their research also showed that more frequent daily data gives better predictions than less frequent data. They also confirmed that CNN model outperformed the other two models. In their research they pointed out the need to use loss balancing algorithms for multitasking learning to further improve the results in future studies.

Several machine learning models which included Lasso regression, Decision Trees, Random forest and Convolution Neural Networks with recurrence plots were employed to predict long-term air temperature in Paris in the month of August [8]. The CNN+RP approach handed them better results as compared to the classical CNN and other traditional machine learning models they used. They proposed as future work in which the CNN would be modeled to output more than one target (temperature) variable but rather include others like wind information etc.

A paper has been published in which researchers developed a deep neural network to (DNN) for understanding the relationship between air and surface water temperatures [1]. Their main aim was to understand the long-term relationship between air temperature and surface water temperature. They employed the DNN and transfer learning TL because DNN alone does not perform well on unexperienced data (out-of-range training data). The results showed that the DNN with TL outperformed classical DNN.

In a recent study, 12 different regression machine learning methods were used to predict air temperature in Seoul city, South Korea [9]. XG Boost and Light Boost where the most performing amongst the regression methods used. Among the notable methods are KNN, Lasso and Ridge regression, Support Vector regression, Decision Trees among others. They proposed that other time series algorism like LSTM and ARIMA could be employed to get even better results.

Jaharabi et al. also used machine learning and deep learning algorithms to create models for temperature predictions in major cities of Bangladesh [10]. Models such LSTM, ARIMA, SARIMA, Prophet and RNN were employed to filter out abnormalities, preprocess and predict future trend. Amongst the listed time series models, the best performing models was SARIMA. The performance did not perform well in Delhi because it was affected by the high levels of pollution in the city. They also proposed to use a dataset covering wider span to determine the effect of pollution on temperature prediction.

A variation of temperature prediction is one that involved micro-climate monthly temperature forecasting [11]. In their paper, they proposed a dense neural network to predict measurements of temperature and humidity corresponding to that of sensors located in the green house. The DNN showed a very close correlation between the true values and the predicted value for both temperature and humidity. Using a combination of a number of optimal sensors and the proposed DNN Architecture, they were able to predict the micro-climate change of the green house. They proposed that their work could be extended from monthly data variation to daily data monitoring.

2. Material and Methods

2.1 Dataset

The dataset employed in this study was downloaded from kaggle (**https://www.kaggle.com/datasets/sevgisarac/temperature-change**). The most important features of interest that it had were the month, year and a target of temperature change for each year. This dataset contains the data for most of the counties and so we had to down size it to our geographical location of interest during preprocessing which is Zambia. The data was recorded between 1961 and 2020 covering 59 years. The primary programming language that was used is python together with its libraries that included Keras, Matplotlib, pandas, numpy and skitlearn. Anaconda's jupytor note book was used as integrated development Environment (IDE).

2.2 Regression Techniques

This study explored five (5) different regression methods which included linear regression, ridge regression, KNN, Random Forest and Deep Neural network. Each one of these is explained in the following subsections.

2.2.1 Linear Regression

Linear models make predictions based on the linear function of the input features using a general prediction formula below.

$$\hat{y} = w(o) * x(0) + w(1) * x(1) + \cdots w(p) * x(p) + b \quad (1)$$

For a dataset with a single feature, the equation is reduced to

$$\hat{y} = w(0) * x(0) + b \quad (2)$$

Here $w(0)$ is the slope and b is the y-axis offset while $x(o)$ is the single feature. The formula for the linear regression score is given below. The problem that linear regression has is that it has no hyper parameters that can be used for regularization.

The score is given by the formula below.

$$MSE = \frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_i) \quad (3)$$

Where: -

n is the number of data points.

y_i is the actual value of the dependent for the $i - th$ data point.

y_i is the predicted value of the dependent variable for the $i - th$ data point.

2.2.2 Ridge Regression

Ridge regression is also a linear model but results in magnitudes of coefficients that are as small as possible. What this means is that the coefficients will be very close to zero. With ridge regression, each feature has as little effect on the outcome as possible. This enables the model to be regularized in order to avoid overfitting.

The formula to measure the score in ridge regression is

$$J(\beta) = RSS + \alpha\sum_{j=1}^{P}\beta_j^2 \quad (4)$$

Where

- $J(B)$ is the objective function
- RSS is the residual sum of squares, which measures the difference between the observed and predicted value.
- B_j represents the coefficients of the regression model and
- The last term consisting alpha and sum represents the regularization term

2.2.3 K-Nearest Neighbor (KNN)

The most basic form of KNN considers only one closet training data in prediction consideration. Instead of taking only one point (neighbor), we can instead take an arbitrary number of neighbors. Although KNN was initially developed for classification problems, we can also use it for regression problems. In KNN regression, the target variable is predicted by averaging the data samples from the nearest neighbors. The equation for KNN regression is defined as below:

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i(x) \qquad (5)$$

The determining factor for KNN performance is the choice of the number of k-nearest neighbors. If the number of k-nn neighbors is too low then the chances of over fitting will be high. Conversely if k-nn number is too low, the model will likely result into under fitting. Under fitting results in many outliers on the prediction. Hence the number for knn needs be iteratively chosen.

2.2.4 Random Forest

Random Forest (RF) is a collection of individual decision trees working independent of each but in parallel with other trees [9]. RF makes use a bagging technique (and sometimes pasting techniques). Despite each tree doing a fair job of predicting, each one of them overfits to a certain extent [4]. This overfitting can then be reduced by averaging their results (prediction). Random forest ensures each tree has a different result by injecting randomness in each tree. The final obtained result is gotten from the average of N tress as shown in equation 6 below [9].

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \qquad (6)$$

2.2.5 Deep Neural Networks (DNN)

A Deep Neural Network is a feed forward artificial neural network multi-layer perceptron with three basic layers, which are the input, hidden and output layers [9]. The neural network constantly updates hyper parameters like weights and bias through back propagation [12]. The first layer or the input layer collects data from the features related to the output target. The hidden layer is a hierarchy of neurons stack on top of each other. The task of the output layer is to output the target. The output layer make use of the activation function with output of the previous layer as the input of the next layer as given below [9, 12].

$$y^k = \sigma(w^k y^{k-1} + b^k) \tag{7}$$

The regression calculation of a DNN is given blow.

$$y_t = a_0 + \sum_{j=1}^{n} a_j f\left(\sum_{i=1}^{m} \beta_{ij} y_{t-j} + \beta_{aj}\right) + \varepsilon_1 \tag{8}$$

The generic structure of the DNN is shown in figure 1 below

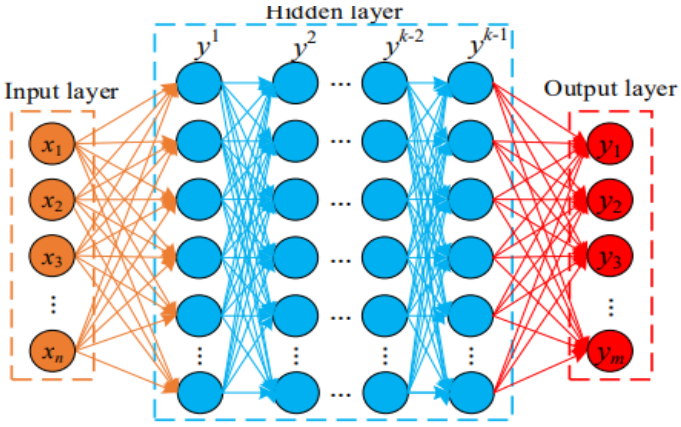

Figure 1:shows a flow chart of the Deep Neural Network

3.0 Results and Discussion

3.1 Evaluation Criteria

The R^2 is the metric that has been used to evaluate the performance of baseline regression models which are later compared to the classical DNN model. The R^2 metric is defined as

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \tag{9}$$

K-fold cross validation is a resampling technique used to evaluate ML-based regression methods' performance independently from the dataset. The value k should be chosen carefully as poor choice of k may result of high variance or high bias. In this experiment our $k = 5$ or 10 since it has been experimentally proven that these numbers do not cause very high variance or bias [9].

3.2 Experimental results

Monthly temperature change was obtained from Kaggle, which is an online platform that enables different data scientists to collaborate on different data science projects and challenges. The dataset covers the years 1961 to 2020. The data set which contained temperature change from all countries was appropriate for this study because it contained data of interest to our research which is Zambia. The dataset has two main features relevant to the study and one target namely 'Area', 'Months', 'Year Code' and 'Value' respectively. All data with the country code 251 representing Zambia was used to select the three relevant features specific to Zambia. The snippet below shoes the first five rows in table one.

Table 1: Dataset Features and targets

	Area	Months	Year Code	Value
227885	Zambia	January	1961	0.166
227886	Zambia	January	1962	-0.281
227887	Zambia	January	1963	-0.297
227888	Zambia	January	1964	0.294
227889	Zambia	January	1965	-0.591

The dataset was later normalized and preprocessed to remove all null values by dropping the rows that contained null values. The months column was encoded using the Label Encoder from Sklearn library were changed to appropriate number format as shown below.

Table 2:Showing the month column label encoded

	month	year	temp_change
227885	5	1961	0.166
227886	5	1962	-0.281
227887	5	1963	-0.297
227888	5	1964	0.294
227889	5	1965	-0.591
...	...	...	...
228900	12	2016	1.300
228901	12	2017	0.629
228902	12	2018	0.888
228903	12	2019	1.310
228904	12	2020	1.304

The scatter plot was produced to show the relationship between the progression of years from 1961 to 2020 and the temperature change during those year. Figure 1 and figure 2 show the scatter plots of temperatures against months and years respectively.

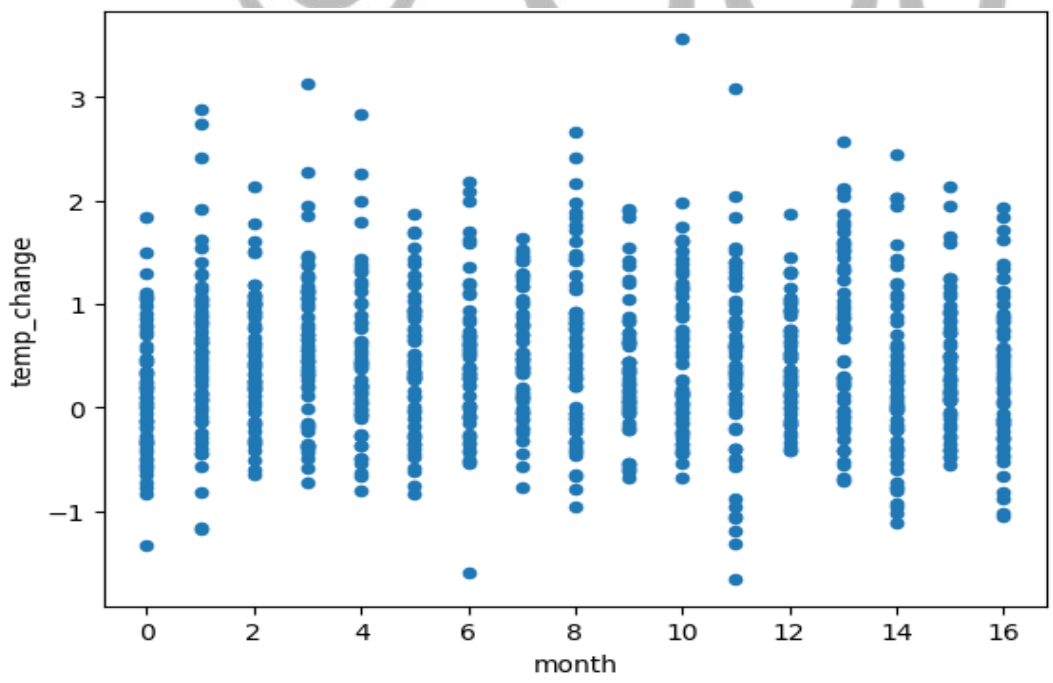

Figure 2: Temperature change plot against month

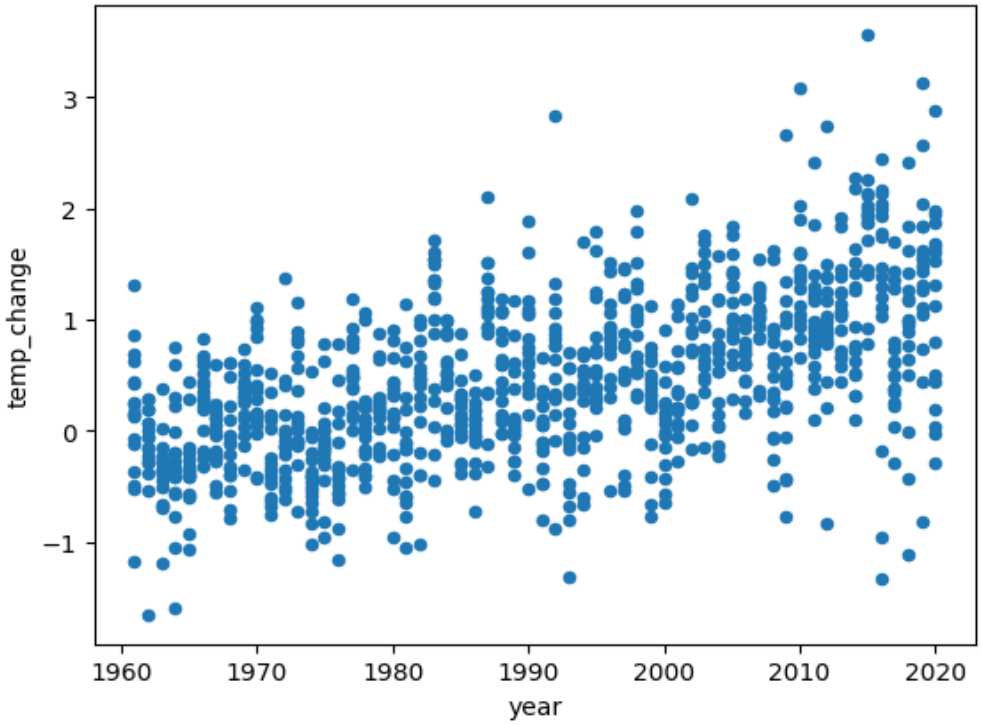

Figure 3: Temperature change against year

3.3 Evaluation Metrics and results

Before being used on the baseline models, the dataset was divided into 80% training set and 20% test set. The data dataset was later randomly again divided into 60% training set, 20% test set and 20% validation and used on the Deep Neural Network. The Metrics on the machine learning and Deep learning models are summarized in the table below.

Table 3: Model Metrics and Scores

	Metric	**%Score**
Linear Regression	score	27
Ridge Regression	score	27
KNN	score	19
Random Forest	score	35
Deep Neural Network	loss	47

3.4 Discussion and Results

After data preprocessing was finished, the first baseline model to be used on the data was Linear Regression and it gave a prediction of score 33% on the training data and 27% on the test data. This was expected as a linear model does not perform very well on non-linear data. The second model was a Ridge Regression performed exactly as a Linear Regression model as can be seen in table 1. KNN model showed signs of overfitting as there was a big difference between the prediction score on the training set data of 65% and that of test set data of 19%. This showed that the model was not able to generalize well from the training set data to test set data. The final baseline model was Random Forest which also performed very well on training dataset with a score of 90% and a score of 36% on test set. This also showed that the model was overfitting and will most likely give incorrect results. The deep neural network DNN was constructed with two input neurons and a dense of 13 layers of neurons. The prediction performance was 52%.

4.0 Conclusion

The study was to show the performance difference between the four baseline machine learning models i.e. Linear Regression, Ridge Regression, KNN, Rand forest and the Deep Neural Network. We are to considered prediction performance on the unseen data and found that the DNN model performed better than all the other baseline models. However, the results were not the best because ideally a DNN model does not do well on non-linear data. Other techniques can be used to improve these results like transfer learning as proposed by many other scholars.

References

[1] N. Kimura, K. Ishida and D. Baba, "Surface Water Temperature Predictions at a Mid-Latitude Using DNN and Transfer learning," *MDPI,* pp. 1-14, 2021 April 2021.

[2] G. Bing, M. Langguth, Y. Ji and A. Mozaffari, "Temperature forecasting by deep learning methods," Geoscienfic Model Development, Julich, 2022.

[3] L. Wang, B. Xu, C. Zhang, G. Fu, X. Chen and Y. Zheng, "Surface water temperature predictions in large-deep reservoirs using a long short-term memory model.," *Ecological Indicators,* pp. 1-12, 14 December 2022.

[4] A. Muller and S. Guido, Introduction to Machine Learning with python: A guide for data scintists, Califonia: O'reilly Media, 2017.

[5] s. Hanoon, N. A. Ahmed, N. Zaini and R. Arif, "Developing Machine Learning Algorithms for meteorological temeperature and humidity forcasting at at terengganu sate," Scientific Reports, Terengganu, 2021.

[6] S. Roy, "Forecasting The Air Temperature at a Weather Station Using Deep Neural Networks," in *9th International Young Scientist Conference on Computational Science (YSC 2020)*, New York, 2020.

[7] S. Lee, L. Yung-seop and L. Youngdoo, "Forecasting Daily Temperatures with Different Time Interval Data Using Deep Neural Networks," *Applied Sciences,* vol. 10, no. 1609, pp. 1-24, 2020.

[8] D. Fister, J. Perez-Aracil, C. Pelaez-Rodriguez, D. Ser and S. Salcedo-Sanz, "Accurate long-term air temperature prediction with Machine Learning models and data reduction techniques," *Aplied Soft Computing,* pp. 1-20, 13 February 2023.

[9] M. Apaydin, S. Yumus, A. Dirgemenci and K. Omer, "Evaluation of air temperature with machine learning regression methods using Seoul City meteorological data," *Pamukkale University Journal of Engineering Sciences,* vol. 28, no. 5, pp. 737-747, 2022.

[10] W. Jaharabi, A. M. Hassain and R. Tahmis, "Predicting Temperature of Major Cities Using Machine Learning and Deep Learning," in *n/a*, Bangladesh, 2023.

[11] S. O. Ajani, J. Member, A. Esther, D. U. Daniel and H. Yushin, "Greenhouse Micro-Climate Prediction Based on Fixed Sensor Placement: A machine learning Aproach.," *MDPI,* pp. 1-14, 10 July 2023.

[12] H. Xiao, L. Yongbin, Z. Xiang, W. Jun and Z. Qian, "Prediction of Apple Slices Drying Kinetic during Infrared-Assisted-Hot Air Drying by Deep Neural Networks," *MDPI,* pp. 1-16, 02 November 2022.