IEC
Institute of Electronics and Computer

Journal of Artificial Intelligence and Systems, 2022, 4, 22-32
https://iecscience.org/journals/AIS
ISSN Online: 2642-2859

Performance Evaluation of Machine Learning Models for Weather Forecasting

Iliyas Ibrahim Iliyas[1, *], Andra Umoru[1], Adati Elkanah Chahari[2], Mustapha Mallam Garba[1]

[1] Department of Mathematic Science, University of Maiduguri, Borno State, Nigeria

[2] Department of General Studies Education, Federal College of Education, Yola, Adamawa State, Nigeria

[*] Corresponding author's E-mail: iliyasibrahimiliyas@unimaid.edu.ng

How to cite this paper: Iliyas Ibrahim Iliyas, Andra Umoru, Adati Elkanah Chahari, Mustapha Mallam Garba (2022). Performance Evaluation of Machine Learning Models for Weather Forecasting. Journal of Artificial Intelligence and Systems, 4, 22–32. https://doi.org/10.33969/AIS.2022040102.

Received: August 24, 2022
Accepted: October 4, 2022
Published: October 12, 2022

Abstract

Temperature is used to indicate variability and climate changes that indicate the process which is been carried out within the ecosystem and its services. The lack of knowledge about temperature affects human lives in terms of agriculture, transportation, mining, etc. temperature forecasting is used to predict atmospheric conditions based on parameters that caused the temperature to change. This study aims to explore the use of machine learning models for the prediction of temperature, evaluate the performance of these models, and use the model to predict temperature. In this study we explore the use of four different machine learning algorithms for forecasting weather temperature, the algorithms are: Ridge, Random Forest, Linear Regression, and Decision tree. We divided the dataset into training and testing sets, The models were tested on 1000 testing sets based on RMSE score with Decision Tree having the best score of 0.036, Random Forest: 0.208 while Logistic Regression and Ridge had the lowest score of 0.759 respectively.

Keywords

Regression, Atmosphere, Models, Forecasting, Temperature

1. Introduction

Temperature is used to indicate variability and climate changes that indicate the process which is been carried out within the ecosystem and its services. Which has a vital role in terms of shaping situations for the being and growth of organisms around us and the ecological determination of temperature and watercourse alteration within the environment, and nature, which occurs daily, seasonal, annually, or even for decades. The reason why temperature changes occur is when there are changes from the surface heat which exchanges within the atmosphere and a turbulent mix of various temperatures (Graf, Zhu, & Sivakumar, 2019). Necessary knowledge about the temperature is crucial in each passion of humans like in agriculture, tourism, power generation, airport system, and the mining industry (Jaseena & Kovoor, 2020). Issues tend to occur when we don't know the temperature, like in agricultural sectors; farmers might make bad decisions on their farms when they can't predict the temperature for the upcoming days, secondly, airplanes cannot fly in the air when the climate is bad, which requires knowledge of the temperature to avoid bad timing, in mining sectors, when miners do not know the temperature which can be used to oversee the earth's surface, this can lead to unexpected floods or drought. Also, in engineering problems like ocean engineering (Raza & Jothiprakash, 2014).). A temperature forecast is used for the prediction of the atmosphere condition of a determined location based on some weather parameters/features; these parameters are based on information gathered in the atmosphere of a given location (Jaseena & Kovoor, 2020). Forecasting is important because the temperature conditions variate rapidly and continuously, which when predicted accurately, can help our daily life's in areas like agriculture for determining crop planting or crop yield, for travel to determine the conducive climate for planes to travel in the air. The world depends on prediction since our world suffers since weather is continuous and has side effects on the environment. That is the reason why accurate prediction with zero error is necessary for our day-to-day activities (Jakaria et al, 2020).

Temperature forecasting systems are mostly categorised into three depending on the type of model or the methodology applied for forecasting, we have statistical models, artificial intelligence (AI), and hybrid models. A statistical model is used on a linear dataset, examples of statistical models are ARIMA, ARMA, and its variants. AI model is the one that uses deep learning or machine learning algorithms for forecasting, which are used on a nonlinear dataset, examples of this model are: Artificial Neural Network, Convolutional Neural Network, Support Vector Machine, Recurrent Neural Network, Ridge, Random Forest, Long Short-Term Memory Network, Logistic Regression, and Decision Tree are few of the popular artificial intelligence models used in forecasting temperature. A hybrid model is a model which comprises the combination of more than one model that is used for the enhancement of the performance of models for temperature forecasting (Jaseena, et al, 2020). In this study, we compare the use of four different machine learning models in forecasting temperature from a dataset that was obtained from Kaggle called: Weather History, the models used in this study are: Random Forest, Linear Regression, Decision Tree, and Ridge.

2. Machine Learning Models

2.1. Linear regression

Linear regression is a type of supervised learning method which is used for modelling continuous variables and making a prediction. It creates a straight hyperplane in between data points to determine classes of data points (Ray, 2019). It is a statistical model that observes the linear relationship between either two or more variables (independent) and a dependent variable, it means that when the independent variable(s) increase or decrease, the outcome (dependent) shall also increase or decrease (Ray, 2019).

2.2. Support Vector Machine

Support vector machine can be used in both regression and classification problems. In the support vector machine, a decision boundary has been drawn as a hyperplane method where a set of data points belonging to a certain group are separated by the hyperplane. These data points can be linear separable or not but there exists a mathematical function named kernel which is used to separate them into member of a class, SVM is aimed at classifying data points based on examples in the training dataset (Ray, 2019).

2.3. Ridge

Ridge is a regression algorithm for regularization, feature selection, and classification, it is used to determine the feature subset because it is important for classification problems even if the variables are highly correlated. It can be used for handling missing values on attributes/features, ridge regression uses a special type of estimator for shrinkage of coefficients named: ridge estimator. Ridge regression is an L2 type of regularization that adds an L2 penalty. L2 penalty is calculated as the square of the magnitude of the coefficients. Ridge regression has a cost function that is updated by summing the penalty values (Deepa, 2021).

2.4. Decision Tree

A decision tree is a type of supervised learning which is used for regression and classification problems where the splitting of data is continuously done on a certain parameter. Each decision is represented as leaves, classification problems are categorised in a form of Yes/No while for regression, decisions are continuous (Ray, 2019). A decision tree is used to classify a target input through learning simple chained decision rules based on previous input features, it resembles an upside-down tree, including the first decision rule on top while the subsequent decision rules are distributed under the first decision like branches (Nishat, 2021).

2.5. Random Forest

Random forest is a type of learning algorithm that works by more than one decision tree during training time and provides an output class. It can be used for classification and regression too, it used the utilization of de-correlated trees by building a multitude of decision trees on a bootstrapped sample of training data, the process is known as bagging. Bootstrapping is used for filtering a few numbers of attributes out of all attribute columns, it increases bias and decreases the variance (Ray, 2019).

There are various models used for temperature forecasting namely: SVM, RF, DT, LSTM, and LR are some of them. hybrid and ANN models proposed by Graf et al (2019) for forecasting water temperature, they applied these models on four mother wavelets which include Discrete Meyer, Daubechies, Symlet, and Haar, the dataset was a study from the daily water temperature of Warta River in Poland, from eight different river gauges and daily air temperature from seven metrological stations. Their proposed hybrid model performs better than the ANN model and concluded that from the used four mother wavelets, discrete Mayer at level 10 performed better than Daubechies and Haar while the lowest among them is Haar in terms of accuracy.

Sharma et al (2021) explored prediction models for solar power generation based on national weather service forecasting with machine learning algorithms. Using SVM with multiple kernel functions and linear least squares. They evaluated the model accuracy using solar internet reading and history data from NWS for 11 months, their study shows that SVM and linear least-squares performed better than the simple model based on sky condition forecast and pas predict future model.

Holmstrom et al (2016) forecast maximum and minimum temperature for seven days, from weather data for the previous two days, the model sued by these authors are linear regression model and a variation used on a functional regression model, the linear regression performs better than the functional regression and professional model based on root mean square score from their results and concluded that linear regression model has a low bias, high variance and the reason for the low bias is because of the choice of data used that is for the period of only past two days which might be too short.

Hewage et al (2020) forecast weather system from local weather stations with two states of art techniques namely: Long short-term memory (LSTM) and Temporal Convolutional Network and they showed results that TCN performs better than other classical machines learning techniques and LSTM.

Cramer et al (2017) applied and compared seven machine learning techniques, Markov chain extended, Genetic programming, Radial Basis Neural Networks, Support Vector Regression, M5 Model trees, K-Nearest and M5 Rules in forecasting rainfall from 43 cities with different climate features. They provided a summary of the correlation between the diverse cities and predicted accuracy, the machine learning techniques were evaluated by determining the prediction error and concluded that Radial Basis Function performs better than the rest of the remaining techniques.

Athiyarath et al (2020) provided a comparative analysis of different algorithms for forecasting weather and provided the usefulness and limitations of various time-series data from different domains namely: appliances energy prediction dataset online retail dataset and solar flare dataset. The authors used regression, Longest Shortest-Term Memory LSTM, Convolutional Neural Network CNN, weighted M VFTS, ARIMA, and Convolutional Bidirectional Long Short Term Memory Networks CBLSTM algorithms. They concluded that the performance of models can be determined based on short, mid, or long-time space, For short-range CNN, and weighted MVFTS, CBLSTM performs quite better, while for the midterm and long-term, MVFTS and ARIMA, CBLSTM provided significant outcome. They outlined that linear did not perform quite well in all the areas that were examined in the time spaces whereas CBLSTM and MVFTS performed tremendously.

Jakaria et al (2020) presented a weather forecasting technique that used historical data gotten from multiple weather stations with SVM, Extra-tree Regression (ETR), Multi-Layer Perception (MLPR), and Ridge Regression on a dataset from around Nashville: Chattanooga, Knoxville, Jackson, Paducah, Bowling Green, Atlanta, Tupelo, Florence, and Birmingham, they evaluated the performance of the model based on single and ten city situation and find out that with one city situation, the models have shown high RMSE which gives rise to the need of considering neighbouring cities for achieving less RMSE. There is a need for the exploration of machine learning without excluding traditional machine learning models which is the reason for this study and to highlight the performance of these models with a large amount of data. This study explores the use of four machine learning models for the prediction of temperature and those models were evaluated based on the RMSE, MSE, and MAE scores and plotted a graph of actual and predicted temperature for the 1000 data points.

3. Materials and Methods

3.1. Dataset

The weather dataset was taken from Kaggle website namely: weather history, using the link: https: www.kaggle.com/dataset/muthuj7/weather-dataset consisting of 11 parameters namely: Summary, Precip Type, Temperature, Apparent Temperature, Humidity, Wind Speed (km/h), Wind Bearing (degress), Visibility, Loud Cover, Pressure (millibars), Daily Summary, the data has 96453 data points, out of the total data points, 95453 was used for training and 1000 was used for testing the models.

3.2. Data Preprocessing

The dataset consists of categorical values in two columns namely: summary and precipe type which was converted into numbers using the dummies function in python for the algorithm to execute, there were no missing values in the dataset.

3.3. Performance Metrics

To evaluate our models on the dataset, we used MSE to show the average squared difference between the model's predicted value and the actual value from the dataset.

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{r}_t - \overline{\boldsymbol{r}}_t\right)^2}$$

T = represents no/length of the training sets;
$\underline{\boldsymbol{r}_t}$ =is the predicted temperature;
$\boldsymbol{r}_t$ =is the actual temperature;
While the RMSE will show the squared root of the average squared difference between the model's predicted value and the actual value from the dataset.

$$\sum_{t=1}^{T}\left(\boldsymbol{r}_t - \overline{\boldsymbol{r}}_t\right)^2$$

T = represents no/length of the training sets;
$\underline{\boldsymbol{r}_t}$ =is the predicted temperature;
$\boldsymbol{r}_t$ =is the actual temperature;
And MAE shows the distance of the actual values from the dataset to the model's prediction values, to account for the negative errors gotten.

$$\frac{1}{\boldsymbol{n}}\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}}\left|\boldsymbol{y}^{real} - \boldsymbol{y}^{pred}\right|$$

4. Results and Discussion

This section provided the results and discussion of the five models applied in the prediction of weather temperature. The graph shows actual and model predictions of 1000 sets.

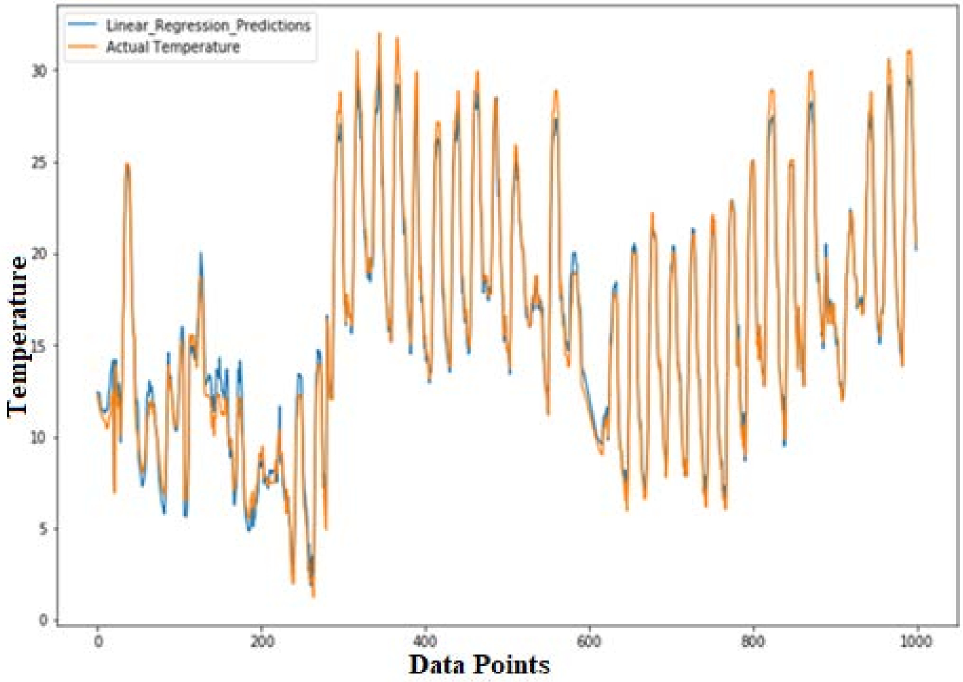

Figure 1 Linear Regression Prediction

Figure 1 showed the plotted graph of linear regression prediction of testing data points and the actual dataset, the testing set includes 1000 data points that were split from the total number of records from the dataset.

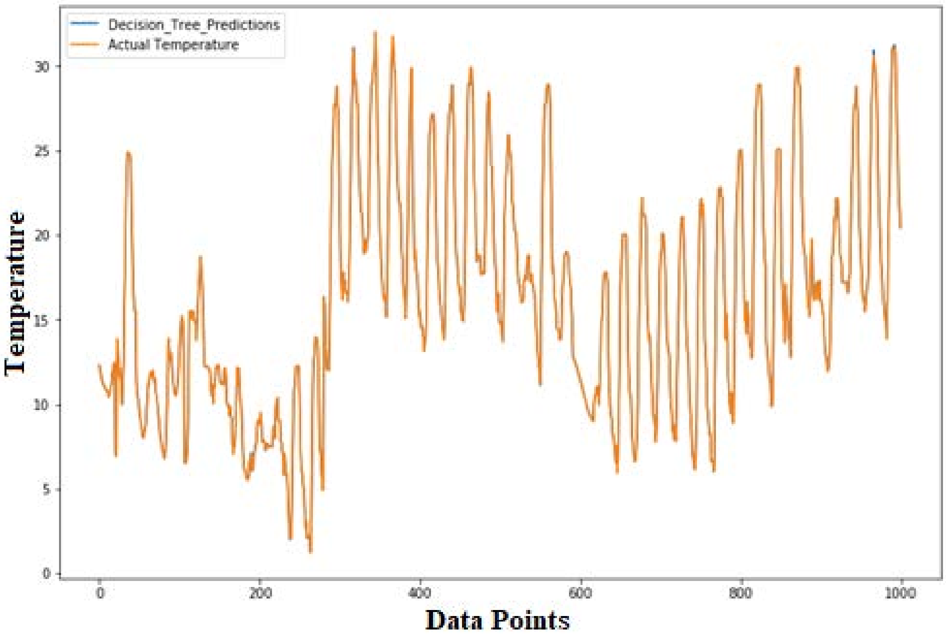

Figure 2 Decision Tree Result

Figure 2 highlighted the plotted graph of Decision tree prediction and the actual dataset selected as the testing sets.

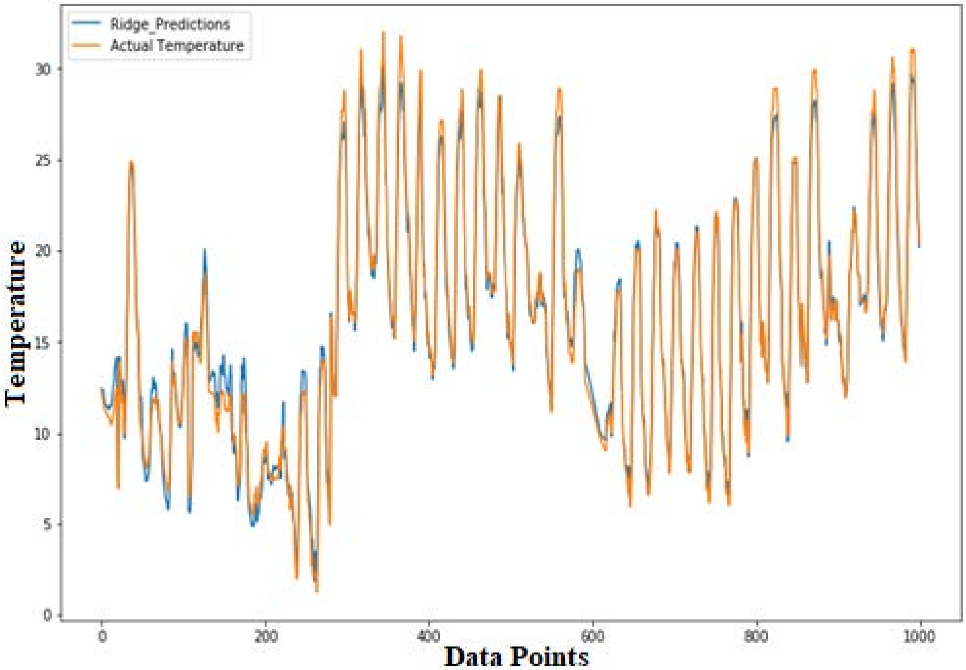

Figure 3 Ridge Result

Figure 3 highlighted the plotted graph of Ridge prediction and the actual dataset selected as the testing sets.

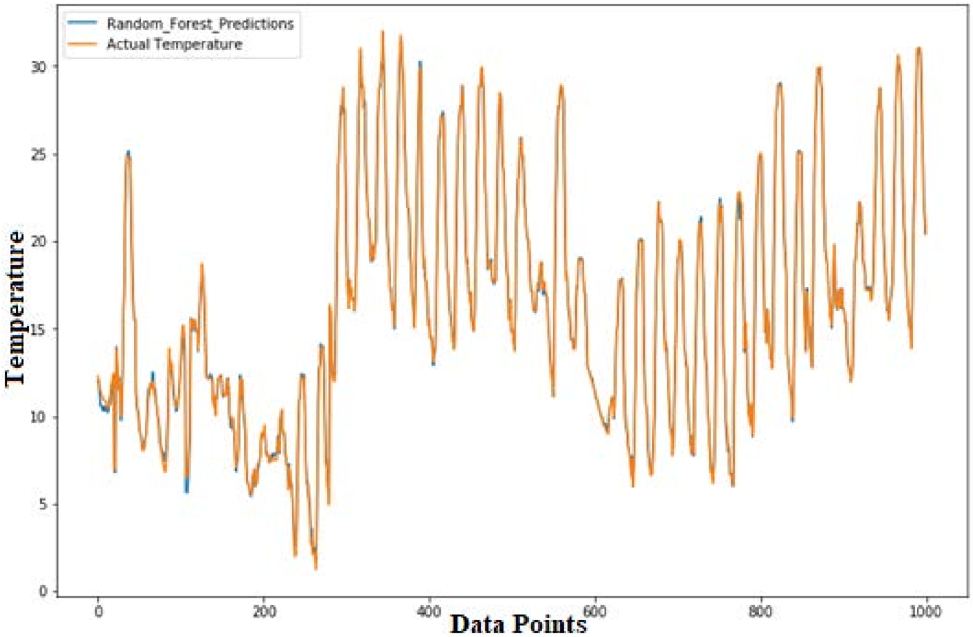

Figure 4 Random Forest Result

Figure 4 highlighted the plotted graph of Random Forest prediction and the actual dataset selected as the testing sets.

After training those models on the training set dataset, we evaluated the models in predicting 1000 data points and we plotted a graph where red lines represent the actual dataset and the blue lines represent the model's prediction. The plotted graph in Figure 1 represented the actual and predicted temperature by the Linear Regression model, Figure 2 showed the plotted graph for the actual and predicted temperature of the Decision Tree, and Figure 3 showed the plotted graph of both actual and predicted temperature using Ridge model while Figure 4 showed the plotted graph for both actual and predicted temperature by Random Forest. Based on the result highlighted in the figures, the Decision tree was able to predict almost the same temperature as the actual temperature in the dataset with other models showing a reasonable accuracy with a few deviations.

Table 1 RMSE score of the four models

Performance Metrics	DT	LR	RF	Ridge
RMSE	0.038	0.759	0.226	0.575
MSE	0.001	0.576	0.051	0.576

MAE	0.010	0.604	0.051	0.604

After splitting the data, Table 1 highlighted the results of the model's performance based on RMSE, MSE, and MAE metrics, DT shows a better RMSE score of 0.038 followed by RF:0.226, Ridge:0.575, LR:0.759. While in terms of MSE, DT: 0.001, LR: 0.576, RF: 0.051, Ridge: 0.576 and in terms of MAE, DT: 0.010, LR: 0.604, RF: 0.051 and 0.604. The table indicated that DT outperforms other models that were used.

5. Conclusion

This study explored the use of four machine learning models for the prediction of temperature, performance evaluation of these models, and temperature prediction made by these models. The four machine learning models are: Ridge, Decision tree, Logistic Regression, and Random Forest, the dataset was split into training and testing sets where we trained the model on the training sets and evaluated the performance of the models on the testing set, using RMSE, MSE and MAE metrics and the models were tested to predict the temperature of testing sets based on plotted graphs which compared the actual and predicted temperature. Based on the graphs plotted, both models showed similar trends, but DT shows better trend similarity, and based on the performance metrics, DT outperforms all the models in terms of both RMSE, MSE, and MAE. This model can be used for real-life applications to improve our daily life activities and, help researchers in choosing machine learning algorithms for predicting temperature. The study only limits the use of machine learning models but can be tested with deep learning models and there is a need for the use of more than one dataset from different locations.

References

[1] Athiyarath, S., Paul, M., & Krishnaswamy, S. (2020). A comparative study and analysis of time series forecasting techniques. SN Computer Science, 1(3), 1-7. https://doi.org/10.1007/s42979-020-00180-5

[2] Cramer, S., Kampouridis, M., Freitas, A. A., & Alexandridis, A. K. (2017). An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. Expert Systems with Applications, 85, 169-181. https://doi.org/10.1016/j.eswa.2017.05.029

[3] Deepa, N., Prabadevi, B., Maddikunta, P. K., Gadekallu, T. R., Baker, T., Khan, M. A., & Tariq, U. (2021). An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier. The Journal of Supercomputing, 77(2), 1998-2017. https://doi.org/10.1007/s11227-020-03347-2

[4] Graf, R., Zhu, S., & Sivakumar, B. (2019). Forecasting river water temperature time series using a wavelet–neural network hybrid modelling

approach. Journal of Hydrology, 578, 124115. https://doi.org/10.1016/j.jhydrol.2019.124115

[5] Hewage, P., Behera, A., Trovati, M., Pereira, E., Ghahremani, M., Palmieri, F., & Liu, Y. (2020). Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. Soft Computing, 24(21), 16453-16482. https://doi.org/10.1007/s00500-020-04954-0

[6] Holmstrom, M., Liu, D., & Vo, C. (2016). Machine learning applied to weather forecasting. Meteorol. Appl, 1-5.

[7] Jakaria, A. H. M., Hossain, M. M., & Rahman, M. A. (2020). Smart weather forecasting using machine learning: a case study in Tennessee. arXiv preprint arXiv:2008.10789. https://doi.org/10.48550/arXiv.2008.10789

[8] Jaseena, K. U., & Kovoor, B. C. (2020). Deterministic weather forecasting models based on intelligent predictors: A survey. Journal of King Saud University-Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2020.09.009

[9] Nishat, M. M., Dip, R. R., Faisal, F., Nasrullah, S. M., Ahsan, R., Shikder, M. F., & Hoque, M. A. (2021). A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms. EAI Endorsed Transactions on Pervasive Health and Technology, 18(e6). https://doi.org/10.4108/eai.13-8-2021.170671

[10] Ray, S. (2019). A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (35-39). IEEE. doi: 10.1109/COMITCon.2019.8862451

[11] Raza, K. & Jothiprakash, V. (2014). Multi-Output ANN Model for Prediction of Seven Meteorological Parameters in a Weather Station. Journal of The Institution of Engineers (India): Series A, Springer, 95(4): 221-229. http://doi.org/10.1007/s40030-014-0092-9

[12] Sharma, N., Sharma, P., Irwin, D., & Shenoy, P. (2011). Predicting solar generation from weather forecasts using machine learning. In 2011 IEEE international conference on smart grid communications (SmartGridComm) (528-533). https://doi: 10.1109/SmartGridComm.2011.6102379.