TITLE PAGE:

Predicting the Accuracy Of Global Mean Temperature using Linear regression compared with K nearest neighbor Algorithm

Krishna Kanth. V1, Fahad Iqbal 2

Krishna Kanth.V1 Research Scholar,

Department of Information Technology,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India. Pincode:602105

krishnakanthv@saveetha.com

Fahad Iqbal 2

Project Guide,

Corresponding Author,

Department of Information Technology,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India. Pincode:602105

fahadiqbalkt.sse@saveetha.com

**Keywords:** Linear regression, K nearest neighbor Algorithm, OpenCV, accuracy, Anaconda

# ABSTRACT

**Aim:** The purpose of this work is to improve the prediction of the accuracy of global mean Temperature using Machine learning. **Materials and Methods:** Linear regression and K nearest neighbor Algorithm is executed with varying training and testing splits for predicting the accuracy of global mean Temperature using Machine learning. The Gpower test used is about 85% (g power setting parameters: α=0.05 and power=0.85). **Result:** Linear regression (92.3670%) has the increased accuracy over K nearest neighbor Algorithm ( 78.9700%) with a significance value of 0.001 (Two tailed, p>0.05). **Conclusion:** The accuracy of Linear regression is better when compared to accuracy of K nearest neighbor Algorithm.

Keywords: Linear regression, K nearest neighbor Algorithm, OpenCV, accuracy, Anaconda

# INTRODUCTION

We aim to use machine learning algorithms to model the mean temperature, namely Long-Short Term Memory Neural Networks (LSTM) and Random Forest Regressor (RF). For this study, we use a dataset with 30 years of radiosonde observations over the Brazilian region. In general, the results are consistent with those provided in the literature[(Brum et al. 2022)](https://paperpile.com/c/3AXnST/79pl),The proposed ensemble approach is based on three models which provide good performance in terms of model evaluation parameters like Correlation, Accuracy, R-Squared (R 2 ), Root mean square (RMSE) and Total Time to detect the predicted temperatures [(Himika et al. 2018)](https://paperpile.com/c/3AXnST/Ojlc).The purpose of this paper is to predict the most probable future global sea-level rise using advanced machine learning models. A total of 28 years' worth of sea-level rise data has been utilized for training our models using various machine learning algorithms[(Hassan et al. 2021)](https://paperpile.com/c/3AXnST/uFgL).Furthermore, the temporal and spatial variability in the modeled brightness temperatures via the SVM more closely agrees with that found in the original AMSR-E measurements. These findings suggest that the SVM is a superior alternative to the ANN for eventual use as a measurement operator within a data assimilation framework[(Forman and Reichle 2015)](https://paperpile.com/c/3AXnST/rhQq).

The research has been carried out onPredicting the accuracy of global mean Temperature using Machine learning; on an average of 44 research papers have been published in IEEE Xplore and 34 papers have been published in sciencedirect.The results were validated by in situ observations and compared with the NASA Advanced Microwave Scanning Radiometer for EOS (AMSR-E) snow water equivalent product. Satisfactory accuracy was achieved for different ecoregions with regard to daily, monthly, the Pearson correlation coefficient R ranged from 0.75 to 0.85)[(Xu et al. 2022)](https://paperpile.com/c/3AXnST/oZSx).

In this algorithm, the global precipitation measurement (GPM) product has been employed to train QPE prediction model. The real-time multiband infrared brightness temperature from Himawari-8, combined with the spatiotemporally matched numerical weather prediction (NWP) data from the global forecast system, have been used as predictor variables for QPE[(Min et al. 2019)](https://paperpile.com/c/3AXnST/e7jF). Considering that the sounding below the precipitation level becomes unreliable, the precipitation-affected observations were removed from the training dataset by means of a pre-screening test based on BT. The results show an overall ability of the algorithm to retrieve T and WV vertical profiles in line with expectations.[(Di Paola et al. 2018)](https://paperpile.com/c/3AXnST/SoBQ)

The research gap identified from the existing system is poor accuracy. This study is to improve the accuracy of classification by incorporating Linear regression and comparing its performance with K nearest neighbor Algorithm. The proposed model improves prediction of the accuracy of global mean Temperature using Machine learning.

# MATERIALS AND METHODS

The research work was done in the Soft Computing Lab, Department, college name. Sample size has been calculated using Gpower software by comparing both the controllers. Two groups are selected for comparing the process and their result is derived. In each group, 10 sets of samples and 10 samples in total are selected for this work. Two algorithms Linear regression and K nearest neighbor Algorithm are implemented using technical Analysis software. Sample size is determined as 10 for each group using GPower 3.1 software (gpower setting parameters: α=0.05 and power=0.85).

The proposed work is designed and implemented with the help of Python OpenCV software. The platform to assess deep learning was Windows 10 OS. Hardware configuration was an Intel core i7 processor with a RAM size of 4GB. System sort used was 64-bit. For implementation of code, java programming language was used. As for code execution, the dataset is worked behind to perform an output process for accuracy.

**Linear regression**

**Description**

The machine learning algorithm Linear Regression is based on supervised learning. It does a task called regression. In regression, the independent variables are used to model a target prediction value. It is mostly used to figure out how different things relate to each other and make predictions.

**Pseudo code:**

Require: Training data D, number of epochs e, learning rate ŋ, standard deviation σ

Ensure: Weights ω0 , ω1,.....ωk

Initialise weights ω0 , ω1,.....ωk from standard normal distribution with

zero mean and standard deviation σ

for epoch in 1...e do

for each (x, y) **∈** D in random order do

**y ‘**←ω0 + Σki=1ωixi

if (**y ‘**>1 and y = 1) or (**y ‘**<-1 and y=-1) then continue

ω0← ω0 - η 2(**y ‘ -y)**

**For i in 1…..k do**

ωi← ωi - η 2(**y ‘ -y)**xi

end for

end for

return ω0 , ω1,.....ωk

**K nearest neighbor Algorithm**

**Description:**

K Nearest Neighbor uses a similarity measure to classify fresh data or instances. It classifies a data point based on its neighbors' classifications. KNN works by determining the distances between a query and all the examples in the data, selecting the specified number (K) closest to the query, then voting for the most frequent label (in classification) or averaging the labels (in the case of regression).

**Pseudo code:**

fori <1 to n do Visited [i] <false

Initialize the list Path with s

Visited [s]<← true

Current s

for i2to ndo

Find the lowest element in row current and unmarked column j containing the

element.

Current ←j

Visited [j]←true

Add j to the end of list Path

Add s to the end of list Path

return Path

# Statistical Analysis

SPSS software is used for statistical analysis of Linear regression and K nearest neighbor Algorithm. Independent variables are image, objects, distance, frequency, modulation, amplitude, volume, decibels. Dependent variables are images and objects. Independent T test analysis is carried out to calculate accuracy for both methods.

# RESULTS

The proposed Linear regression and K nearest neighbor Algorithm were run at different times in Anaconda Navigator with a sample size of 10. Table 1 represents the predicted accuracy and loss of Linear regression.Table 2 represents the predicted accuracy and loss of K nearest neighbor Algorithm. These 10 data samples are used for each algorithm along with their loss values to calculate statistical values that can be used for comparison. From the results, it is observed that the mean accuracy of Linear regression was 92.3670% and K nearest neighbor Algorithm was 78.9700%. Table 3 represents mean accuracy values for Linear regression and K nearest neighbor Algorithm. Mean value of Linear regression is better when compared with the K nearest neighbor Algorithm with a standard deviation of 2.53458 and 3.04122 respectively. Table 4 shows the Independent sample T test data of Linear regression and K nearest neighbor Algorithm with the significance value obtained is 0.393 (Two tailed, p<0.05). Figure 1 denotes the comparison of Linear regression and K nearest neighbor Algorithm in terms of mean accuracy and loss.

Mean, standard deviation and standard error mean for Linear regression are 92.3670, 2.53458 and .80150 respectively. Similarly for K nearest neighbor Algorithm, the mean, standard deviation and standard error mean are 78.9700, 3.04122 and 0.96172 respectively. On the other hand, the loss values of Linear regression for mean, standard deviation and standard error mean are 7.6330, 2.53458 and 0.80150 respectively. For K nearest neighbor Algorithm, the loss values of K nearest neighbor Algorithm for mean, standard deviation and standard error mean are 21.0300, 3.04122 and 0.96172 respectively.

The group statistics value along with mean, standard deviation and standard error mean for the two algorithms are also specified. The graphical representation of comparative analysis, means of loss between two algorithms of Linear regression and K nearest neighbor Algorithm are classified. This indicates that Linear regression is significantly better with 92.3670% accuracy when compared with K nearest neighbor Algorithm classified accuracy of 78.9700%.

# DISCUSSION

In the given study, the significance value obtained is 0.001 (Two tailed, p>0.05) which implies that Linear regression appears to be better than K nearest neighbor Algorithm. Accuracy analysis of the Linear regression is analyzed as 92.3670% whereas the accuracy of Linear regression is 78.9700%.

This study models the weighted mean temperature (T m) over the West Pacific to accurately estimate precipitable water vapour (PWV) from ground-based GPS receivers. PWV accuracy depends on T m. 2011 Radiosonde (RS) data from 15 sites from 20°N to 20°S latitude and 95°E to 156°E longitude is used to develop a T m model. Then, T m and T s were analysed linearly. Validation will be done with the global T m model. This model outperforms global T m by 4.5%. Results showed that a 1 Kelvin weighted mean temperature increase increases PWV by 0.13 mm.[(Sheaffer )](https://paperpile.com/c/XWKhR8/elwp)

These large changes will have a greater impact on regional climatic extremes and ecosystem changes than mean global warming rates. Coral reef ecosystem collapse is likely in the fastest-warming regions. Due to algae and filter feeder competition, corals may survive in areas warming slowly due to increased cold water upwelling, but only as marginal coral communities. Pelagic fisheries would move from productive coastal upwelling zones to remote offshore ones. Changes in global warming, ocean-atmosphere CO/sub 2/ fluxes, marine biodiversity, and primary productivity are implied.[(Root 2015)](https://paperpile.com/c/XWKhR8/LJvk)

A summer maize production high-temperature heat damage assessment model was created using remote sensing and ground temperature data. The model's accuracy coefficients R 2 are above 0.8, and the root mean square error (RMSE) fluctuates within 2 ° C. Summer maize production in North China suffered high-temperature heat damage in 2017 and 2018. These damaged areas are mostly in southeast Hebei, most of Henan, and western Shandong. Results match statistics.[(Brown and Caldeira )](https://paperpile.com/c/XWKhR8/q7uS)An earlier study suggested estimating latent heat of evapotranspiration (ET). However, the influence of soil moisture (SM) on ET was not well considered, so this paper incorporates the Diurnal land surface temperature (T s ) Range (DTsR). ET from 2001–2006 at twelve U.S. sites validates the improved method. The site has grassland, native prairie, cropland, deciduous forest, and evergreen forest. All sites have a correlation coefficient of 0.92 between measured and predicted 16-day daytime-average ET, a bias of -1.9 W m -2, and an RMSE of 28.6 W m -2. We calculated global monthly ET from 1986 to 1995 at a spatial resolution of 1degtimes1deg from the International Satellite Land Surface Climatology Project (ISLSCP) Initiative II global interdisciplinary monthly dataset and compared it to the fifteen land surface model simulations of the Global Soil Wetness Project-2. The bias, RMSE, and correlation coefficient of the 118-month global daily ET comparison are 4.5 W m -2, 19.8 W m-2, and 0.82, respectively.[(Clarke 2017)](https://paperpile.com/c/XWKhR8/fhGS)

The limitations of this study is that it takes a very long time to train Linear regression, especially with large datasets. The future scope of this study is that the system should be expanded to include a larger number of objects with lesser time consumption in training the data set.

# CONCLUSION

The accuracy value of the Linear regression is 92.3670% whereas the accuracy value of K nearest neighbor Algorithm is 78.9700%. Based on the analysis, Linear regression (92.3670%) performs better than K nearest neighbor Algorithm ( 78.9700%)

# DECLARATIONS

Conflicts of Interests

No conflict of interest in this manuscript.

# Authors Contribution

Author SA was involved in data collection, data analysis and manuscript writing. Author CPL was involved in conceptualization, data validation and critical reviews of manuscripts.

# Acknowledgement

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences (formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. Infysec Solution, Chennai
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

# REFERENCES

W. Suparta and A. Iskandar, "Modeling of weighted mean temperature over the Western Pacific region to estimate GPS PWV," 2013 IEEE International Conference on Space Science and Communication (IconSpace), 2013, pp. 190-193, doi: 10.1109/IconSpace.2013.6599462.

T. J. Goreau and R. L. Hayes, "Global change in ocean circulation from satellite sea surface temperature records: implications for the future of coral-reefs, fisheries, and climate change," Oceans 2003. Celebrating the Past ... Teaming Toward the Future (IEEE Cat. No.03CH37492), 2003, pp. 754 Vol.2-, doi: 10.1109/OCEANS.2003.178406.

L. Yang, J. Song, L. Han, X. Zhu and X. Wang, "Monitoring and Risk Assessment of High-Temperature Heat Damage for Summer Maize Based on Remote Sensing Data," IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, 2020, pp. 2909-2912, doi: 10.1109/IGARSS39084.2020.9323336.

K. Wang and S. Liang, "An improved method for estimating global evapotranspiration based on satellite determination of surface net radiation, vegetation index, temperature, and soil moisture," IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium, 2008, pp. III - 875-III - 878, doi: 10.1109/IGARSS.2008.4779489.

[D. Brum, V. F. Rofatto, L. Gonzaga, R. De Oliveira Pena, L. F. Sapucci and M. R. Veronez, "Mean Tropospheric Temperature Estimation Using Deep Learning and Ensemble Methods," IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, 2022, pp. 6658-6661, doi: 10.1109/IGARSS46834.2022.9883387.](https://paperpile.com/c/3AXnST/79pl)

[Himika, S. Kaur and S. Randhawa, "Global Land Temperature Prediction by Machine Learning Combo Approach," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1-8, doi: 10.1109/ICCCNT.2018.8494173.](https://paperpile.com/c/3AXnST/Ojlc)

[K. M. A. Hassan, M. A. Haque and S. Ahmed, "Comparative Study of Forecasting Global Mean Sea Level Rising using Machine Learning," 2021 International Conference on Electronics, Communications and Information Technology (ICECIT), 2021, pp. 1-4, doi: 10.1109/ICECIT54077.2021.9641339.](https://paperpile.com/c/3AXnST/uFgL)

[B. A. Forman and R. H. Reichle, "Using a Support Vector Machine and a Land Surface Model to Estimate Large-Scale Passive Microwave Brightness Temperatures Over Snow-Covered Land in North America," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 9, pp. 4431-4441, Sept. 2015, doi: 10.1109/JSTARS.2014.2325780.](https://paperpile.com/c/3AXnST/rhQq)

[X. Xu, X. Liu, X. Li, Q. Shi, Y. Chen and B. Ai, "Global Snow Depth Retrieval From Passive Microwave Brightness Temperature With Machine Learning Approach," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-17, 2022, Art no. 4302917, doi: 10.1109/TGRS.2021.3127202.](https://paperpile.com/c/3AXnST/oZSx)

[M. Min et al., "Estimating Summertime Precipitation from Himawari-8 and Global Forecast System Based on Machine Learning," in IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 5, pp. 2557-2570, May 2019, doi: 10.1109/TGRS.2018.2874950.](https://paperpile.com/c/3AXnST/e7jF)

[F. Di Paola et al., "Retrieval of Temperature and Water Vapor Vertical Profile from ATMS Measurements with Random Forests Technique," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 2018, pp. 6014-6017, doi: 10.1109/IGARSS.2018.8518198.](https://paperpile.com/c/3AXnST/SoBQ)

**TABLES AND FIGURES**

**Table1.** Accuracy and Loss Analysis of Linear regression

| **Iterations** | **Accuracy(%)** | **Loss(%)** |
| --- | --- | --- |
| 1 | 88.4 | 11.6 |
| 2 | 89.7 | 10.3 |
| 3 | 90.8 | 9.2 |
| 4 | 91.01 | 8.99 |
| 5 | 92.1 | 7.9 |
| 6 | 91.8 | 8.2 |
| 7 | 93.6 | 6.4 |
| 8 | 94.8 | 5.2 |
| 9 | 95.4 | 4.6 |
| 10 | 96.06 | 3.94 |

**Table2.** Accuracy and Loss Analysis of K nearest neighbor Algorithm

| **Iterations** | **Accuracy(%)** | **Loss(%)** |
| --- | --- | --- |
| 1 | 74.8 | 25.2 |
| 2 | 75.1 | 24.9 |
| 3 | 76.4 | 23.6 |
| 4 | 77.4 | 22.6 |
| 5 | 78.2 | 21.8 |
| 6 | 79.9 | 20.1 |
| 7 | 80.5 | 19.5 |
| 8 | 81.9 | 18.1 |
| 9 | 82.1 | 17.9 |
| 10 | 83.4 | 16.6 |

**Table 3.** Group Statistical Analysis of Linear regression and K nearest neighbor Algorithm. Mean, Standard Deviation and Standard Error Mean are obtained for 10 samples. Linear regression has higher mean accuracy and lower mean loss when compared to K nearest neighbor Algorithm.

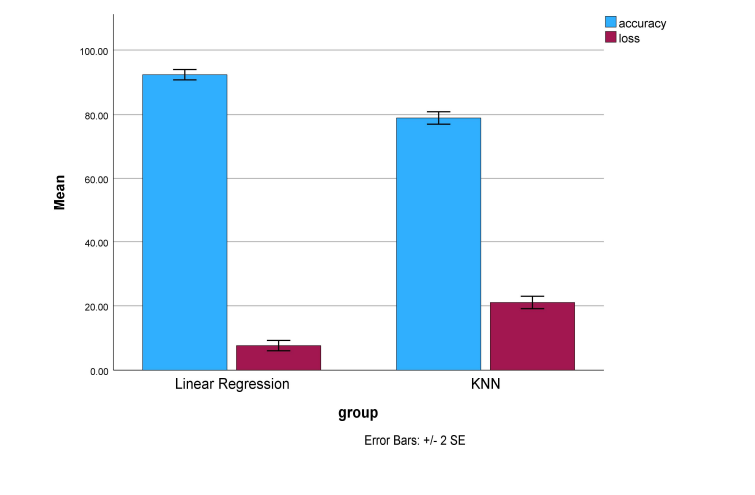
|  | **Group** | **N** | **Mean** | **Std. Deviation** | **Std. Error Mean** |
| --- | --- | --- | --- | --- | --- |
| **Accuracy** | Linear regression | 10 | 92.3670 | 2.53458 | .80150 |
| K nearest neighbor Algorithm | 10 | 78.9700 | 3.04122 | 0.96172 |
| **Loss** | Linear regression | 10 | 7.6330 | 2.53458 | 0.80150 |
| K nearest neighbor Algorithm | 10 | 21.0300 | 3.04122 | 0.96172 |

**Table 4.** Independent Sample T-test: Linear regression is insignificantly better than K nearest neighbor Algorithm with p value 0.001 (Two tailed, p<0.05)

|  | | **Levene’s test for equality of variances** | | **T-test for equality means with 95% confidence interval** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **f** | **Sig.** | **t** | **df** | **Sig. (2-tailed)** | **Mean difference** | **Std.Error difference** | **Lower** | **Upper** |
| **Accuracy** | **Equal variances assumed** | .765 | 0.393 | 10.701 | 18 | 0.001 | 13.39700 | 1.25192 | 10.76681 | 16.02719 |
| **Equal Variances not assumed** | 10.701 | 17.434 | 0.001 | 13.39700 | 1.25192 | 10.76067 | 16.03333 |
| **Loss** | **Equal variances assumed** | .765 | 0.393 | - 10.701 | 18 | 0.001 | - 13.39700 | 1.25192 | -16.02719 | -10.76681 |
| **Equal Variances not assumed** | - 10.701 | 17.434 | 0.001 | - 13.39700 | 1.25192 | - 16.03333 | -10.76067 |

**Table 5.** Comparison of the Linear regression and K nearest neighbor Algorithm with their accuracy

| **CLASSIFIER** | **ACCURACY(%)** |
| --- | --- |
| **Linear regression** | 92.3670 |
| **K nearest neighbor Algorithm** | 78.9700 |



**Fig 1.** Comparison of Linear regression and K nearest neighbor Algorithm. Classifier in terms of mean accuracy and loss. The mean accuracy of Linear regression is better than K nearest neighbor Algorithm. Classifier; Standard deviation of Linear regression is slightly better than K nearest neighbor Algorithm. X Axis: Linear regression Vs K nearest neighbor Algorithm Classifier and Y Axis: Mean accuracy of detection with +/-2SE.