# Global Land Temperature Prediction by Machine Learning Combo Approach

Himika
*Computer Science and Engineering Department*
*Thapar Institute of Engineering and Technology*
Patiala, India
himikaa19@gmail.com

Shubhdeep Kaur
*Computer Science and Engineering Department*
*Thapar Institute of Engineering and Technology*
Patiala, India
shubhdeepk1211@gmail.com

Sukhchandan Randhawa
*Computer Science and Engineering Department*
*Thapar Institute of Engineering and Technology*
Patiala,India
sukhchandan.95@gmail.com

*Abstract*—The Global Land Surface Temperature is the radiative skin temperature of ground, depending on factors, which includes the albedo, the vegetation covers and the soil moisture. To predict the changes in temperature in a particular region is becoming increasingly important to capture the future trends in that region. Machine Learning is a specialized branch of Artificial Intelligence (AI), which gives computers the power to learn and make predictions from the data, without being explicitly programmed. In this work, Ensemble Approach for Global Land Temperatures (EAGLT) is proposed. This approach will help to predict the temperature, which is of great requirement as the problem of global warming is increasing day by day. Temperatures are collected from different cities and prediction is done using this approach. 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. Cross Validation is performed on the best performing models to check the robustness of these selected models.

***Keywords—Classification, Ensemble Approach, Global Land Temperatures, Machine Learning, Regression.***

## I. INTRODUCTION

Global Historical Climate Network relies on global land temperatures which are divided on various sections such as temperatures by major cities, states and countries. It is very essential for the climate department to determine the changing trend in temperature. The temperature of various places has either drastically gone down or up. All these various changes in these temperature ranges are of a great concern and need to be recognized by the Global Historical Climate Network department. To estimate about the global average temperatures and to compile them into a single dataset is a tedious task. The strategies which are applied to record the temperature vary according to the locations. Regions like deserts have few temperature measurement stations, so large areas have to be taken into account whereas in mountainous regions, the observations which have to be gathered is different. The methodology behind this research work is to design an ensemble approach, whi makes use of various existing regression and dual use machine learning models. The ensemble approach employs a supervised learning algorithm that enables to train the model and make predictions. So, the proposed model will enable to find a drastic change in the temperature of the major cities. The changing trend of the temperature needs to be studied using machine learning as it is an issue of major concern.

In this research work, a novel Ensemble Approach for Global Land Temperatures (EAGLT) is proposed. Three best machine learning models are chosen from a set of 15 models which can be used to predict the change in temperatures of the major cities. The models have been evaluated on the basis on Correlation, $R^2$, RMSE, Accuracy and Total Time for a given dataset. After analyzing the performance of machine learning models, top three models having the highest accuracy are chosen. The chosen models are Decision Tree, Variable Ridge Regression and Conditional Inference Tree. An ensemble model of these three models is built that provides the overall accuracy of 83.07%. It can be inferred through experimental results and graphical analysis that various temperature monitoring organizations can rely on the proposed approach for making accurate predictions.

## II. RELATED WORK

The number of researchers have studied and presented solutions for the prediction of weather conditions in terms of temperature. Coumou and Robinson [1] have discussed how there is an incredible rise in the temperature due to the amount of heat varying on monthly and seasonally basis. Coupled Model Intercomparison Project (CMIP5) climate model has been discussed which is efficient enough to produce and evaluate spatial patterns which are observed due to heat extremes. The results show that there is an increase in temperatures during the years, considerably in tropical regions. The model CMIP5 has been successful in accurately capturing the observed rise in those land regions where there is an extreme of heat.

R. Rohde *et al*. [2] gives the estimate about the earth surface land temperatures on average basis ranging from the year span of 1753 to 2011.Climate change is a key factor that is used to analyze the change in earth temperatures. To analyze the temperature rise, a large sample is collected and

analyzed over large number of cities. The temperature change has been recorded for about 14.4 million mean monthly temperatures and over 44,455 different locations. It is seen that there is an incredible increase in both the maximum and minimum temperatures.

K. K. Goldewijk *et al*. [3**]** accords a tool which estimates about the global change studies on long term basis. This tool is an extension to History Database of the Global Environment (HYDE) keeping in mind about various agricultural and demographic factors. Information is gathered through satellite about various factors such as historical population, cropland and pasture statistics. Multiple allocation algorithms have been applied to generate spatial maps. It has been concluded that the humans have shifted into those areas which are efficient enough to have an effect over the Earth's temperature. It is also concluded that excessive usage of land has also an adverse effect over the biogeochemical cycles globally which has ultimately changed the climate leading to rise in temperatures.

P. Havlik *et al*. [4] have targeted the increase of areas which are dedicated to bio fuel production. There has been an increase in content of the greenhouse gases such as Carbon Monoxide, Oxides of Sulphur such as Sulphur Dioxide and Sulphur Trioxide, Oxides of Nitrogen due to the *Indirect Land Use Change* (iLUC). Due to the increase in biofuel production, the cost of basic irrigation commodities has also risen and lead to problems like deforestation and inflation of the crop price etc. GLOBIOM is a model which depicted and focused on global forest, agriculture and biomass sectors. The results depicted that the production of biofuels from the existing forest resources have created a negative iLUC factor. On the contrary, the results depicted by the first generation biofuels have resulted into a positive iLUC factor.

F. Ji *et al*. [5] have discussed about the changing trends in land surface air temperature over multiple years. The problem of global warming has increased at a very high rate over the past 100 years and need to be checked. The proposed methodologies are Multidimensional Ensemble Empirical Model Decomposition1 (MEEMD) and Ensemble Empirical Mode Decomposition (EEMD). While dealing with multidimensional spatial temporal data, various components having the similar timescale components are connected to form a coherent structure of the timescale.

M.G.Donat *et al*.[6] described how World Meteorological Organization (WMO) have been aiming to fill the data gaps using an approach that provides a clear picture on the global level regarding the extremities of temperature. Expert Team on Climate Change Detection (ETCCDI) conducted a number of workshops in multiple regions over the span of years which helped to build a database named as HadEX, which provides complete illustration regarding precipitation extremes over the later years of twentieth century. It further continues and discusses about the dataset named as GHCNDEX created by National Climatic Data Center (NCDC)'s and Global Historical Climatology Network (GHCN) which contains all the updated global dataset of the climate extremes. It also showcased the application of the dataset, which helps in figuring out the problems related to uncertainty by comparing the available existing datasets. It is also useful in monitoring various climatic changes.

The structure of the paper is as follows: Section 2.1 explicates the proposed approach i.e. Ensemble Approach for Global Land Temperatures (EAGLT) along with the dataset used, feature selection and the approach followed to make accurate predictions in the proposed work. Section 3 presents simulation parameters and analysis of existing machine learning models. Section 4 presents the experimental results of the models and compares their performance graphically on the basis of evaluation parameters. This section also explicates the ensemble approach applied and the cross validation performed. Finally, Section 5 concludes the proposed work and describes its future scope.

## III. Proposed Approach: Ensemble Approach for Global Land Temperatures (EAGLT)

In this section, the proposed methodology along with data set used is presented.

.

### *A. Data set and its features*

In this dataset, the repackaged data from a newer compilation has been put together by the Berkeley Earth, which is affiliated with Lawrence Berkeley National Laboratory. The Berkeley Earth Surface Temperature Study combines 1.6 billion temperature reports from 16 pre-existing archives. It is nicely packaged and allows slicing of data into interesting subsets. The raw data comes from the Berkeley Earth data page. The various files have been included in this dataset are:

- Global Average Land Temperatures by Country.
- Global Average Land Temperatures by State.
- Global Average Land Temperatures by Major City.
- Global Average Land Temperatures by City.

This research work only considers Global Land Temperatures by Major City dataset. Dataset includes different features as explained in Table I.

TABLE I. FEATURES INCLUDED IN THE DATASET.

| ***Feature*** | ***Description*** |
|---|---|
| Average Temperature | Temperature of various locations |
| Latitude | Latitude of the corresponding location |
| Longitude | Longitude of the corresponding location |
| Average Temperature Uncertainty | Target predicted variable chosen and it tells about the uncertainty in the temperatures |

### *B. Proposed Methodology*

The Global land temperatures dataset selected is divided into two datasets namely *Training* and *Testing dataset*. In this work, 50% of the data is used for the purpose of training the models and the remaining 50% is used for testing the models to increase the accuracy. The selected target variable is a regression value representing the uncertainty in temperature. The step wise methodology followed in Ensemble Approach for Global Land Temperatures

(EAGLT) is depicted in Fig.1. Data Scrubbing is done before the data is actually given to the models for training. It involves the removal of incomplete or redundant data that might degrade the performance of the models. Feature Selection is another technique, followed to extract relevant features from the dataset. The chosen dataset has limited features, all of them are kept for required prediction, so feature selection is not required in the proposed work. Ensemble approach has been applied to obtain better prediction performance than any other model. The proposed ensemble approach is used to merge the best three models to attain good prediction accuracy.

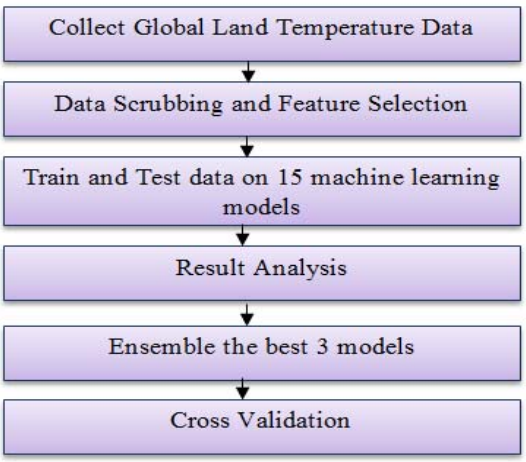



Fig. 1. Methodology Used

This dataset is run on 15 different regression models to determine the accuracy of prediction made. In this approach, top three machine learning models have been chosen with the highest accuracy among the 15 existing models. Cross validation of the ensemble model is then performed by running the top three models multiple times, so as to evaluate the estimator's performance.

## IV. Simulation Setup

Firstly, the percentage of data that is to be chosen for training and testing is to be selected. In this work, training percentage is set to be 50%. Therefore, testing is done on the remaining 50% of the data. The data is then being fed to the model. The target variable for which the prediction has to be made is set as the last variable i.e. the 4$^{th}$ variable. The variable has regression values depicting the uncertainty in the temperatures. Secondly, associated libraries required for a particular model in given in table II are installed in *R*, to build the model. Classification and Regression Training (Caret) package has been used in the models that simulates the process of model building. Thirdly, model is trained by setting the method argument value associated with a particular model. Then model prediction is performed. Finally, the ensemble model is evaluated on the basis of parameters as mentioned in section 5.

### A. Analysis of existing Machine Learning Models

Machine Learning models are used to conduct predictive analysis on the real world data. These models are trained by providing the training data. The learning algorithm maps the input data to the target variable. The output is a machine learning model that captures the patterns in the training data. The model thus obtained can be used to make predictions on the data for which the target is not known. There are different machine learning models available to deal with different set of problems. The *Classification* models, deal with the data in which the target variable takes class labels. The *Regression* models, deal with the data in which target variable takes continuous values. The dual use models, deal with both set of data. The various machine-learning models, which have been used to analyze the global land temperatures dataset, are described in Table II.

TABLE II. Various Machine Learning Models.

| ***Model Name*** | ***Method Argument Value*** | ***Type*** | ***Packages*** | ***Tuning Parameters*** |
|---|---|---|---|---|
| Bagged MARS | bagEarth | Dual Use | Earth | nprune, degree |
| Conditional Inference Tree | Ctree | Dual Use | Party | Mincriterion |
| Decision Tree | Class | Dual Use | Rpart | None |
| EARTH | bagEarth | Regression | Earth | Nprune |
| Independent Component Regression | Icr | Regression | fastICA | n.comp |
| k Nearest Neighbour | Knn | Dual Use | - | K |
| Least Angle Regression | Lars | Regression | Lars | Fraction |
| MARS | bagEarth | Dual Use | Earth | nprune, degree |
| Neural Network | Nnet | Dual Use | Nnet | size, decay |
| Non Negative Least Square | Nnls | Regression | Nnls | None |
| Non Convex Penalized Quantile Regression | Rqnc | Regression | rqPen | lambda, penalty |
| Principal Component Analysis | Pcr | Regression | Pls | Ncomp |
| Projection Pursuit Regression | Ppr | Regression | None | Nterms |
| Relaxed Lasso | Relaxo | Regression | Relaxo,plyr | Lambda, phi |
| Variable Ridge Regression | Foba | Regression | Foba | k, lambda |

### B. Proposed Work Taxonomy

In this proposed work, the used different model evaluation parameters are used for analyzing regression data used in this work:

- **Root Mean Square Error (RMSE)**

Root mean square error gives the standard deviation of the model prediction error. It is the difference between the values predicted by a model and the values actually observed. A smaller value of RMSE indicates better model performance. RMSE is given by (1):

$$RMSE = \sqrt{1/n \sum_{i=1}^{n} (y_i - y_j)^2} \quad (1)$$

where $n$ denotes the total number of observations, $y_i$ denotes the actual values and $y_j$ denotes the predicted values.

- **Correlation(r)**

The correlation coefficient is a normalized measurement of how the two variables are linearly related. Correlation between actual and predicted values is taken out in model prediction. So given pairs of values for variables X and Y, designated (x,y), correlation(r) is given by (2):

$$corr(X,Y) = \frac{E[(X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y} \quad (2)$$

If the correlation coefficient is close to 1, it indicates that the variables are positively linearly related and the scatter plot falls almost along a straight line with positive slope. For -1, it indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope. And for zero, it indicates a weak linear relationship between the variables.

- **R- Squared ($R^2$)**

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the *Coefficient of Determination*, or the *Coefficient of Multiple Determination* for multiple regression. It is square of coefficient of correlation. Here, $y_i$ denotes the observed values of the dependent variable, $\overline{y}$ as its mean, and $y_i$ as the fitted value. So, the coefficient of determination is (3):

$$R^2 = \frac{\sum(\hat{y}_i - \overline{Y})^2}{\sum(y_i - \overline{Y})^2} \quad (3)$$

- **Accuracy**

A measure of how efficiently the model is able to predict the values of a target class. It is calculated by comparing the predicted values with the actual ones. It can be calculated as (4):

$$\text{Accuracy} = (\text{round}(\text{mean}(\text{Actual-Predicted}) \leq 1)),4)*10 \quad (4)$$

This indicates that an absolute value is calculated depicting the difference between the actual and predicted value. Then the average of all the values is calculated and multiplied by a factor of 100 to calculate the accuracy in percentage.

- **Total Time**

Total Time means time taken by the currently running R processes in terms of CPU time (in seconds) to build, train and test the model on the chosen dataset. It is given by (5):

$$Total_{time} = procTime(\ ) - startTime() \quad (5)$$

## V. RESULTS

The results are obtained after building, training and testing various machine learning models to assess the performance of the models, various evaluation parameters are used namely Correlation, RMSE, R-Squared, Accuracy and Total Time are used. The detailed statistical and graphical analysis of various machine learning models helps to determine the models with the best prediction accuracy:

| Sr. No | Model Name | R | R | RMSE | Accuracy | Total Time |
|---|---|---|---|---|---|---|
| 1 | Bagged Mars | 0.0978 | 0.00956 | 0.83 | 71.89 | 295.19 |
| 2 | Conditional Inference Tree | -0.1071 | 0.01147 | 0.66 | 81.14 | 7.41 |
| 3 | Decision Tree | -0.19 | 0.04 | 0.67 | 83.28 | 0.64 |
| 4 | Earth | 0.02 | 0 | 0.67 | 81.83 | 1.07 |
| 5 | Independent Component Regression | 0.1139 | 0.01297 | 0.7 | 82.61 | 9.94 |
| 6 | K Nearest Neighbour | 0.1001 | 0.01002 | 0.74 | 75.33 | 17.74 |
| 7 | Least Angle Regression | 0.0499 | 0.00249 | 0.68 | 82.82 | 2.09 |
| 8 | MARS | 0.0229 | 0.00052 | 0.67 | 81.83 | 24.13 |
| 9 | Neural Network | -0.23 | 0.05 | 2.41 | 11.92 | 0.97 |
| 10 | Non Negative Least Square | -0.25 | 0.06 | 1.02 | 64.55 | 1.45 |
| 11 | Non Convex Penalized Quantile Regression | 0.1275 | 0.01626 | 0.65 | 79.71 | 126.52 |
| 12 | Principal Component Analysis | 0.1131 | 0.01279 | 0.7 | 82.63 | 2.56 |
| 13 | Projection Pursuit Regression | 0.1588 | 0.02522 | 0.65 | 75.46 | 40.31 |
| 14 | Relaxed Lasso | -0.2734 | 0.07475 | 0.73 | 71.96 | 154.72 |
| 15 | Variable Ridge Regression | 0.0499 | 0.00249 | 0.68 | 82.82 | 2.71 |

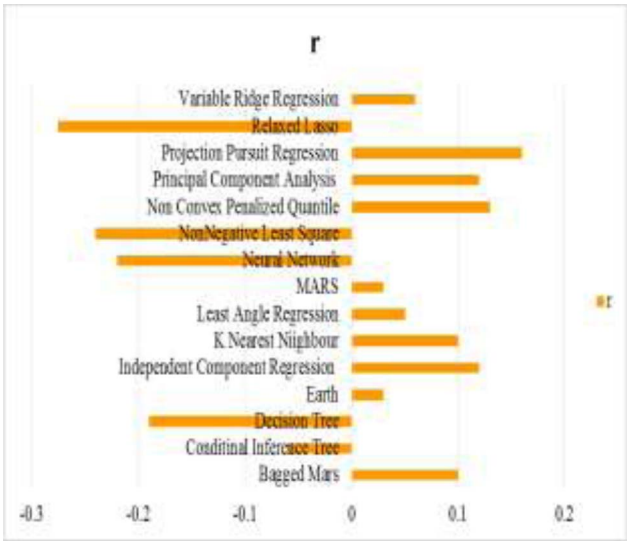
### A. *Performance Comparison*

The results of 15 models which are built, trained and tested on the chosen global land temperature dataset are captured and are depicted in Table III. It shows the comparative performance of all the models in the prediction of temperatures on RMSE, Correlation, $R^2$ and Accuracy. The performance results show that the Projection Pursuit Regression has the highest value of *correlation (r),* that is, 0.1588 which implies that the actual and predicted values in this model are most positively linearly related among all the models. On the other hand, Relaxed Lasso showing the highest value of negative correlation (r), that is, -0.2734 shows the highest value of r-squared. Decision Tree model also taken the minimum time to train and test the dataset out of all the models.

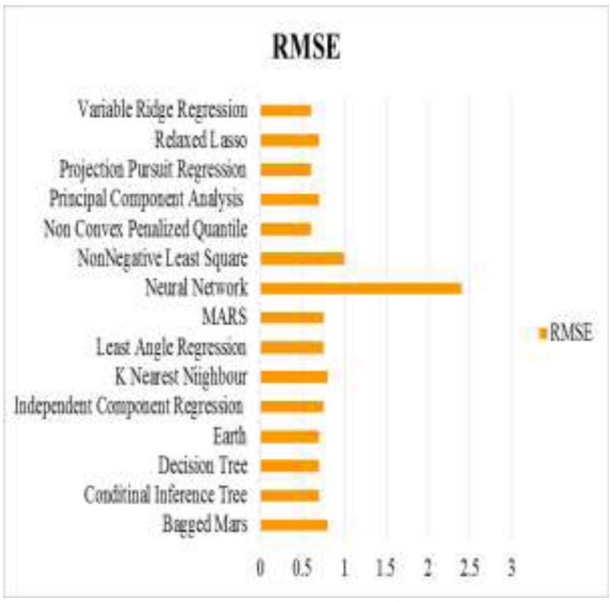
Non Convex Penalized Quantile Regression and Projection Pursuit Regression have the lowest RMSE value, which is 0.65. This indicates better model performance as the root mean square error between actual and predicted values is lowest. Decision Tree model has the highest accuracy

followed by Variable Ridge Regression and Conditional Inference Tree owing to their low RMSE value and good correlation between the actual and predicted values. Bagged MARS model takes the maximum time to build, train and test the data but the value of RMSE is low thereby giving average accuracy results. On the contrary, Neural Network model takes very less time to build, train and test but gives a high RMSE value. Hence, the prediction accuracy of this model is lowest. The model turns out to be the poorest performance of all the models.

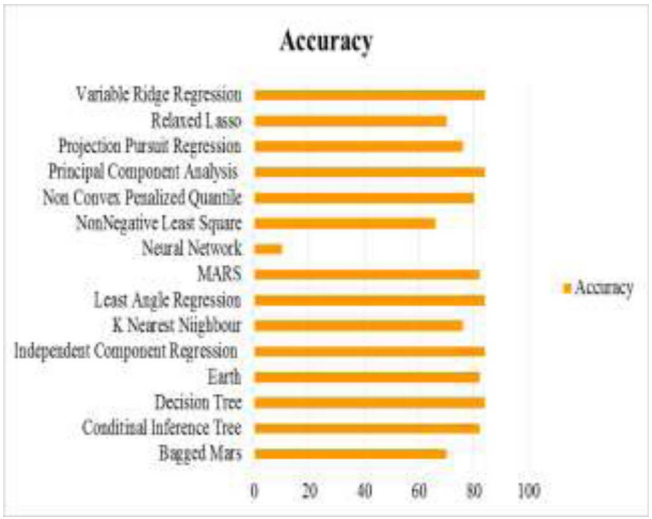
The performance comparison of the machine learning models, using the evaluation parameters Correlation (r), RMSE, Accuracy and Total Time have been shown through bar graphs as depicted in Fig. 2(a-d). It is observed that the value of correlation is maximum for Projection Pursuit Regression and the lowest is for Relaxed Lasso model. The second highest value is that of Non Convex Penalized Quantile Regression. While the value of correlation for all the other models lies approximately in the range of -0.3 to 0.2. In the case of RMSE, Neural Network has the highest value and the second highest value is that of the Non Negative Least Square model. Bagged MARS model has the third highest value. Therefore, the accuracy of these models is lowest. The RMSE value for other models falls in the range 0.65 to 0.74. It can be seen that the Total Time taken by the models varies in a vast range of values. Total Time taken by the Bagged MARS is abnormally the highest among the existing 15 machine models. The lowest time taken to build and train the model and test the dataset is by Decision Tree model.



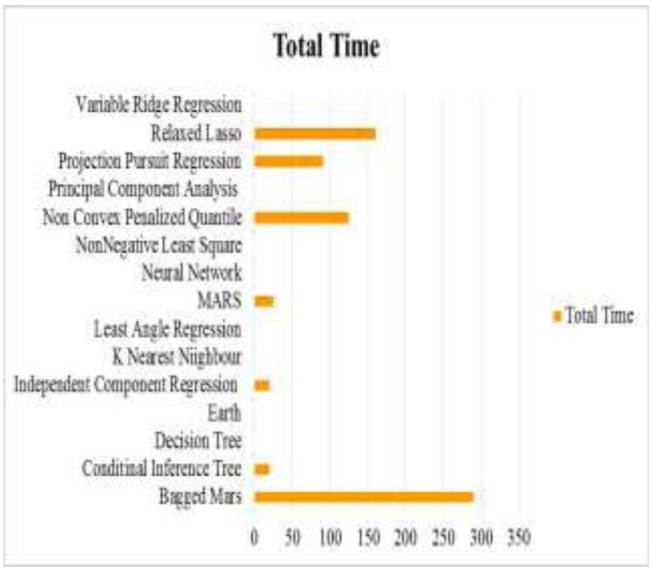

(a) Correlation(r)




(b) RMSE




(c) Accuracy




(d) Total Time

Fig. 2(a-d). Performance Comparison of models using different evaluation parameters.

### *B. Ensemble Approach*

The Ensemble approach considers top three models with the highest accuracy among all the models used to analyze the data set. The selected list of top three models gives the overall accuracy for the given data set. The best three models chosen from the above given models on the basis of their predictive accuracy are mentioned in Table IV.

TABLE IV: TOP THREE MODELS ON THE BASIS OF THEIR PREDICTIVE ACCURACY.

| ***Model Name*** | ***Accuracy*** |
|---|---|
| Decision Tree | 83.28 |
| Variable Ridge Regression | 82.82 |
| Conditional Inference Tree | 81.14 |

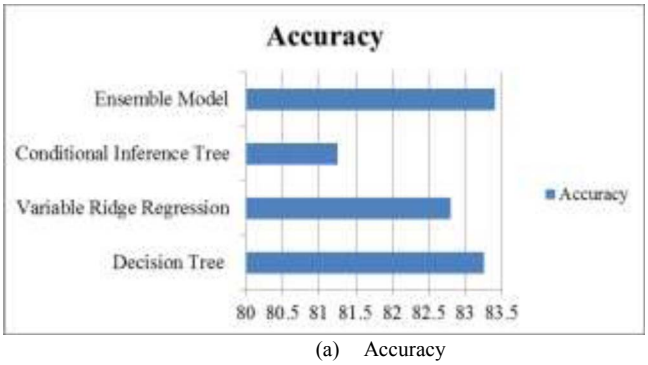
The evaluation parameters calculated using ensemble approach are given below:

**Accuracy:** The actual and predicted value of all the three models are taken into account. The average of the predicted values of these models is calculated row wise, which is considered as an ensemble value. Actual and ensemble values are compared row-wise. If the difference between the predicted values is less than 1, then it is assigned a binary value of 1 else 0. The average of these binary values is calculated which provides us the overall accuracy of the ensemble model. Accuracy calculated using this ensemble approach comes out to be 83.07%.

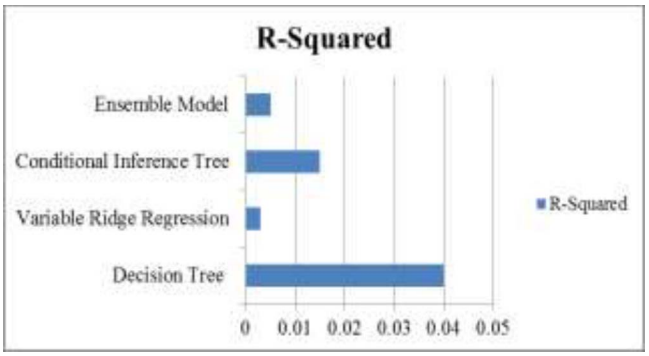
**Correlation:** Correlation is calculated using CORREL function in excel spreadsheet in which the actual and ensemble values are taken into account. The value of correlation is -0.0543, which implies that the actual and ensemble values are linearly - negative related

**R-Square:** R-Square is the square of correlation coefficient, which is equal to 0.0029.

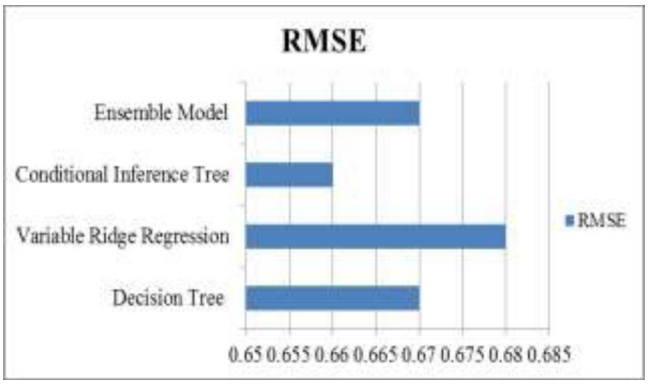
**RMSE:** Difference between the actual values and ensemble values are taken into account and RMSE is calculated. The value is equal to 0.67. The value of RMSE is low, hence the performance of the model is good. The comparison of top three models with ensemble model using evaluation parameters is depicted in Fig. 3.



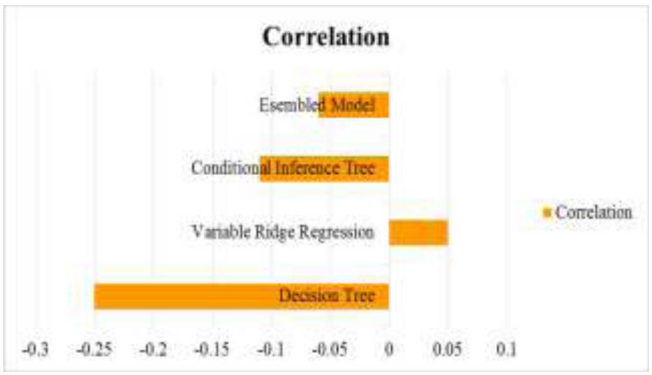

(a) Accuracy




(b) R-Squared




(c) RMSE




(d) Correlation

Fig.3: Comparison of top three models with ensemble model using evaluation parameters.

### *C. Cross Validation*

Cross Validation is a model evaluation method that gives an indication of how well the system makes new predictions for data that it has not already seen. Random permutations of ensemble model, also known as Shuffle and Split method, involves running the top three models multiple times, by shuffling the data every time keeping the training and testing ratio same in every iteration, that is 50% each.

On the basis of the given accuracy statistics as depicted in Table V, it is noticed that the accuracy of all the three models in all the iterations varies in the range of ± 5. Thus, the models chosen for predicting temperatures demonstrate a good accuracy range and can be used to further make accurate predictions when fed with unseen data.

TABLE V CROSS VALIDATION OF ENSEMBLE MODELS INCLUDED IN THE DATASET.

| *Runs* | *Decision Tree* | *Variable Ridge Regression* | *Conditional Inference Tree* |
|---|---|---|---|
| 1 | 89.14 | 90.19 | 88.80 |
| 2 | 89.52 | 90.08 | 89.5 |
| 3 | 90.05 | 90.06 | 90.95 |
| 4 | 90.29 | 89.79 | 89.41 |
| 5 | 90.02 | 89.70 | 89.18 |
| 6 | 90.17 | 90.38 | 89.41 |
| 7 | 90.65 | 89.49 | 89.85 |
| 8 | 89.79 | 90.55 | 89.14 |
| 9 | 89.77 | 90.53 | 89.49 |
| 10 | 90.57 | 89.85 | 89.20 |

### *D. Scatter Plot*

The scatter plot depicted in Fig.4 elucidates that over the number of iterations the Decision Tree Model has the highest predictive accuracy among the three models. The model Variable Ridge Regression shows an average mark of accuracy, whereas model Conditional Inference Tree reach the lowest mark of accuracy when cross-validated.

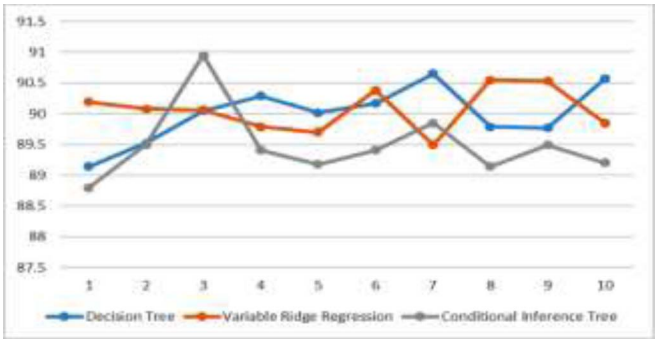



Fig. 4 Scatter Plot for ensemble model

## VI. CONCLUSION AND FUTURE SCOPE

With the increase in temperature, the problem of global warming is increasing day by day. There is an urgent and constant need to predict the uncertainty in temperature. In this paper, an Ensemble Approach for Global Land Temperatures (EAGLT) approach is proposed to predict the uncertainty in temperatures which depict the rise or fall in temperature at particular regions. It will allow various weather forecast agencies to track the change in temperatures which will be very beneficial for them. This study utilized real world global land temperature by major city dataset. The purpose of this work is to propose uncertainty in predicted temperatures on the basis of various features viz average temperature, latitude and longitude of a major city. So, 15 existing machine learning models are built, trained and tested to conduct prediction analysis for temperatures accurately. The models are evaluated on the basis of Correlation, R-Squared, RMSE, Total Time and Accuracy. Ensemble approach has been applied over the top three models among the 15 models to determine the overall Accuracy, RMSE, R-Squared and Correlation. The top three models selected with highest accuracy are Decision Tree, Variable Ridge Regression and Conditional Inference Tree. The cross validation results of the top three models elucidated that Decision Tree model has the highest accuracy rate among the three models when run multiple times. The combined value of accuracy considering the top three models using the ensemble approach comes out to be 83.07%. Thus an ensemble model is proposed with the capacity to predict the uncertainty in temperatures accurately.

The current work takes into consideration only a section of dataset. In future work, various other datasets, included in this major dataset can be considered. This approach can be applied to datasets covering various other countries, states and cities. In addition to this, other machine learning approaches can be applied in analyzing different models and comparing the results.

## REFERENCES

[1] D. Coumou and A. Robinson, “Historic and future increase in the global land area affected by monthly heat extremes,” *Environ. Res. Lett.*, vol. 8, no. 3, p. 34018, 2013.

[2] R. Rohde *et al.*, “A new estimate of the average Earth surface land temperature spanning 1753 to 2011,” *Geoinformatics Geostatistics An Overv.*, vol. 1, no. 1, p. 2, 2013.

[3] K. Klein Goldewijk, A. Beusen, G. Van Drecht, and M. De Vos, “The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years,” *Glob. Ecol. Biogeogr.*, vol. 20, no. 1, pp. 73–86, 2011.

[4] P. Havlík *et al.*, “Global land-use implications of first and second generation biofuel targets,” *Energy Policy*, vol. 39, no. 10, pp. 5690–5702, 2011.

[5] F. Ji, Z. Wu, J. Huang, and E. P. Chassignet, “Evolution of land surface air temperature trend,” *Nat. Clim. Chang.*, vol. 4, no. 6, pp. 462–466, 2014.

[6] M. G. Donat, L. V. Alexander, H. Yang, I. Durre, R. Vose, and J. Caesar, “Global land-based datasets for monitoring climatic extremes,” *Bull. Am. Meteorol. Soc.*, vol. 94, no. 7, pp. 997–1006, 2013.

[7] Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67.

[8] Milborrow, M. S. (2016). Package earth.

[9] Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, **15**(3), 651–674.

[10] Hothorn, T. and Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in r. *Journal of Machine Learning Research*, **16**, 3905–3909.

[11] D. M. Magerman, “Statistical Decision-Tree Models for Parsing ∗.”

[12] Dziewonski, Adam M., and Don L. Anderson. "Preliminary reference Earth model." *Physics of the earth and planetary interiors* 25.4 (1981): 297-356.

[13] C.-J. Lu, T.-S. Lee, and C.-C. Chiu, “Financial time series forecasting using independent component analysis and support vector regression,” *Decis. Support Syst.*, vol. 47, no. 2, pp. 115–125, 2009.

[14] Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, **4**(2), 1883.

[15] Efron, Bradley, et al. "Least angle regression." *The Annals of statistics* 32.2 (2004): 407-499.

[16] Haykin, Simon, and Neural Network."A comprehensive foundation." *Neural Networks* 2.2004 (2004).

[17] Slawski, M., Hein, M., *et al.* (2013). Nonnegative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics*, **7**, 3004–3056.

[18] Y. Wu and Y. Liu, “Variable selection in quantile regression,” *Stat. Sin.*, vol. 19, pp. 801–817, 2009.

[19] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, 1987.

[20] Friedman, Jerome H., and Werner Stuetzle. "Projection pursuit regression." *Journal of the American Statistical Association* 76.376 (1981): 817-823.

[21] Meinshausen, Nicolai. "Relaxed lasso." *Computational Statistics & Data Analysis* 52.1 (2007): 374-393.
[22] R. R. Hocking, “A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression,” *Biometrics*, vol. 32, no. 1, pp. 1–49, 1976.