Using Ensemble Method to Forecast Relative and Absolute Humidity

Digvijay Patil
Computer Engineering and Technology,
Dr. Vishwanath Karad MIT World Peace University,
Pune, India

Atharv Ganla
Computer Engineering and Technology,
Dr. Vishwanath Karad MIT World Peace University,
Pune, India,

Swarad Gat
Computer Engineering and Technology,
Dr. Vishwanath Karad MIT World Peace University,
Pune, India,

***Abstract*—When it comes to the adverse influences of pollution, hurricanes, global warming, evolving weather patterns, and changing climatic conditions, humidity is an important feature of air quality. The main factors that come into play when deciding the humidity of a particular area are relative humidity (RH), which is the amount of water vapor present in air, and absolute humidity (AH), which is the amount of water found in a parcel of air. Since precise humidity prediction is important, the aim of this study is to use Machine Learning and Deep Learning techniques to predict the values of Relative and Absolute Humidity. The RH and AH values are predicted using two simple algorithms. The values are also combined, and the final figures are forecasted and analyzed. In the problem of forecasting relative and absolute humidity, the findings demonstrate the utility of the ensemble model, which outperforms the use of single model prediction.**

Keywords*—Deep Learning, Regression, Ensemble, Air Quality, Artificial Intelligence

I. INTRODUCTION

Humidity refers to the amount of moisture in the air. It is therefore an important aspect of the hydrological cycle which affects both weather and climate. The Clausius–Clapeyron equation shows that saturation vapor pressure (thus, absolute humidity) increases exponentially with increase in atmospheric temperature. Climate scientists relate the Clausius–Clapeyron equation with anthropogenic global warming and indicate that the water holding capacity of the atmosphere increases by approximately 7% *per* 1 °C warming causing extreme rainfall and severe floods. Various fields, such as hydrology, ecology, agriculture and medicine, often use humidity but in the form of relative humidity. The relative humidity, commonly expressed as a percentage, is the ratio of the actual water vapor content of air to the water vapor content of saturated air at the same temperature. [1]

With the aid of Artificial Intelligence and Machine Learning, we hope to predict the values of Relative Humidity and Average Humidity in this study. Extreme Gradient Boosting and Fully Connected Neural Network are two of the most well-known algorithms used to make the prediction. A neural network is a system of algorithms that attempts to identify underlying associations in a set of data using a method that mimics how the human brain works. XGBoost is a decision-tree-built ensemble Machine Learning algorithm based on gradient boosting. When compared to individual algorithm forecasts, ensemble design aids in obtaining more accurate results.

II. BACKGROUND

Polluted air is one of the most serious issues of our day. Not only does it impact climate change but it also affects the health of an individual. Air pollution causes pollutants like Particle Matter (PM), particles of variable but very small diameter, reach the internal organs of respiration of an individual via inhalation, resulting in respiratory and cardiovascular diseases, reproductive and central nervous system dysfunctions, and also cancer.

Air pollution consists of a mix of varied substances in varied amounts, in different physical and chemical states. Some of these substances like nitrogen oxide, Sulphur dioxide, Volatile Organic Compounds (VOCs), dioxin and polycyclic hydrocarbons are gravely harmful to human beings. Even Ozone, which when present in the Stratosphere plays an important role in blocking the harmful UV rays of the Sun, when found in high concentrations in regions closer to the ground is harmful for the health of living organisms. In this work, we are predicting the relative and the absolute humidity. The amount of water vapor in the air is referred to as humidity. It is the most variable characteristic of the atmosphere and plays a big role in influencing the weather of a place. High humidity increases the rate of harmful or toxic chemicals in the air. Dust mites are also attracted to our homes, lowering air quality. On the other hand, bacterial and viral organisms thrive in low and high humidity's. The conclusion is that humidity has a huge say in the amount of pollution and the bacterial concentration in an area. Thus, predicting the humidity will help us keep track of the air quality.

Thus, predicting the humidity in air can help us massively on many levels, as it affects the health of human beings and also has an impact on nature in general.

III. DATA PRE-PROCESSING & ANALYSIS

The analysis was done on a dataset obtained from the UCI Machine Learning Repository. [2] The dataset included several attributes defining the type of particles, elements and compounds present in the air. The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensory Device. Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Methane Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer.[2]

The entire dataset contains 15 features out of which 2 of the features were the prediction features. The input parameters consist of certain Hydrocarbons and Nitrogen Oxides and their derivatives. These values were averaged over the values recorded for an hour. The concentrations were recorded in parts per billion (ppb) or microg/m3. Temperature was another crucial factor that was recorded in degree Celsius.

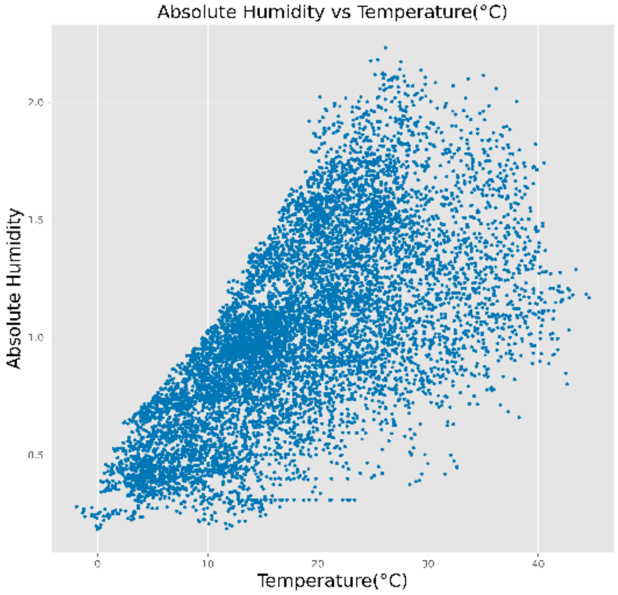

Fig. 1, Graph of Absolute Humidity (AH) vs Temperature I

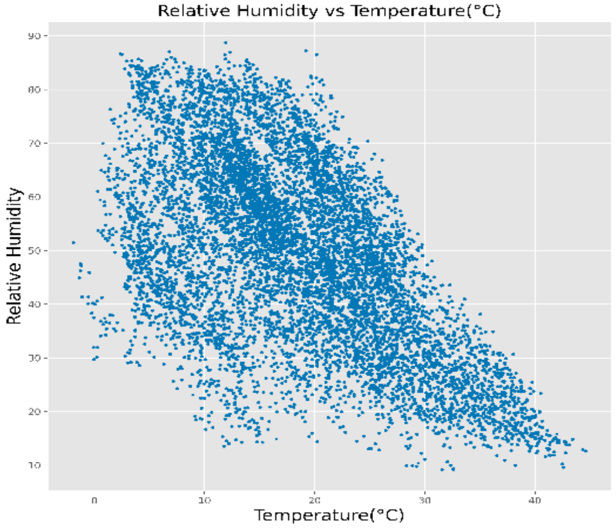

Fig. 2, Graph of Relative Humidity (RH) vs Temperature I

The output features or the features that are being predicted by the model are Relative Humidity (RH) and Absolute Humidity (AH). The RH value is measured in percentage (%) and the AH value is measured in Kg/Kg.

Analysis of the data showed that many values were having value as -200 indicating that data is not present for those columns. Columns having more than 60% of the values as -200 were removed. Another major factor that was dropped from the data frame was time at which the data was recorded. A graph of Absolute Humidity vs Time of record was plotted. An observation was made that in the 24-hour interval no significant difference in the values were noted at every hour of observation. A similar observation was made for Relative Humidity vs Time. Thus, the Time column was dropped.

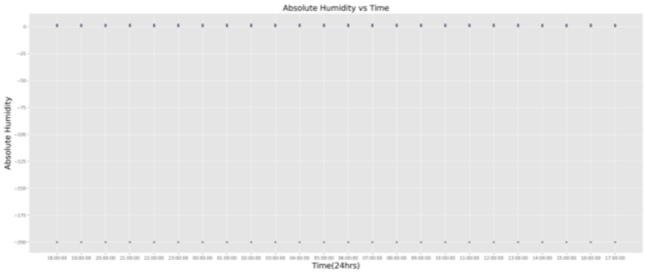

Fig. 3, Graph of Absolute Humidity (AH) vs Time (Hrs)

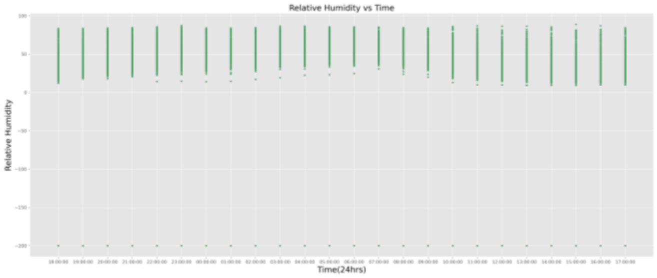

Fig. 4, Graph of Relative Humidity (RH) vs Time (Hrs)

The data frame is normalized to the range of 0 to 1 of all the columns of the dataset by the MinMaxScaler provided by the SciKitLearn framework. This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.[3]

IV. METHODOLOGY

A technique identified as Ensemble is the approach that is used for predicting the two values of humidity. Ensemble approaches are learning algorithms that use a weighted vote of their predictions to classify new data points after creating a sequence of classifiers. Ensemble methods are models that are created in multiples and then combined to produce better results. In most cases, ensemble models yield more reliable results than a single model.

The following are the two models used in the architecture:

- XGBoost: Extreme Gradient Boosting
- FCNN: Fully Connected Neural Network

A. *XGBoost*

The eXtreme Gradient Boosting algorithm is one of the two models that is used for the prediction of the output factors. The algorithm's implementation was designed to maximize compute time and memory resources. To train the model, one of the design goals was to make the most of available resources. The most important factor in XGBoost's success is its scalability across all scenarios. The system runs more than ten times faster than existing popular solutions on a single machine and scales to billions of examples in distributed or memory-limited settings.[4]

The two important reasons why gradient boosting is used here is:

1) *Execution Speed*: To make the necessary calculations fast, XGBoost uses CPU cache to store calculated *gradients* and *Hessians* (cover) to calculate the gain in each split
2) *Model Performance*: The model tries to make many regression trees in such a way that the first tree predicts the output values, the second tree tries to find the optimized gradients for the previous tree and so on. Thus, each tree contributes in the prediction of the output values.

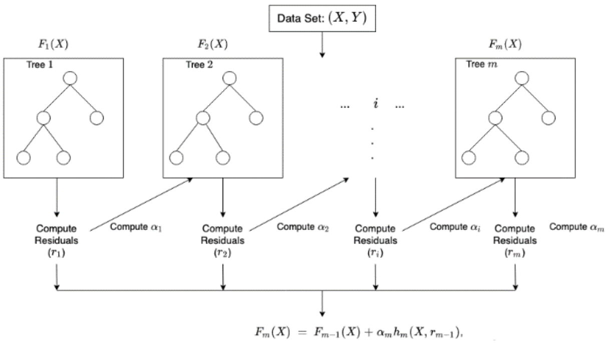
Data Set: (X, Y)
$F_1(X)$
$F_2(X)$
$F_m(X)$
Tree 1
Tree 2
Tree m
... i ...
Compute Residuals (r_1)
Compute α_1
Compute Residuals (r_2)
Compute α_2
Compute Residuals (r_i)
Compute α_i
Compute Residuals (r_m)
Compute α_m
$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1}),$

Fig. 5, XGBoost Algorithm Architecture

In our model the number of regression trees that were derived by the algorithm were 1200 for Relative Humidity value prediction and 1000 for Absolute Humidity Value Prediction. The maximum depth of each regression tree was 5 for both the predictions.

B. FCNN

A fully connected neural network (FCNN) is used as the second model for predicting the values of the Relative and Absolute Humidity.

FCNNs are typically depicted as networks of interconnected "neurons" that can compute values from inputs by feeding data into the network, which is made up of multiple neurons that each make a prediction. Each neuron has an activation function, which gives an output value in a certain range usually being either (-1, 1) or (0, 1). Neural networks offer a number of advantages, including requiring less formal statistical training, ability to implicitly detect complex nonlinear relationships between dependent and independent variables, ability to detect all possible interactions between predictor variables, and the availability of multiple training algorithms. [5]

The architecture that is used contains 1 input layer, 3 hidden layers and one output neuron. All layers are dense layers, as it is a fully connected model. Input layer contains 32 neurons, with activation function as ReLU. Second and third dense layers, which are part of the hidden layer, contain 64 neurons with the same activation function. Fourth layer, which is the last layer of the hidden layers, has 32 neurons. Finally, we have the output layer which contains one neuron. This neuron is responsible for estimating the relative and absolute humidity values. It gives a continuous set of values. Keras was used for building the neural network. The ‘Adam’ optimizer was used and the loss was calculated using the ‘Mean Squared Error’ method, since the target variable is a continuous set of values.

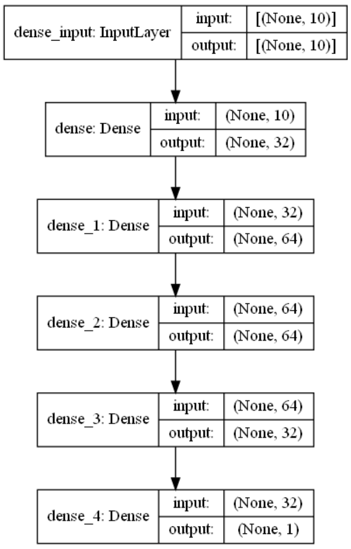

Fig. 6, Fully Connected Neural Network Architecture that is used for the prediction of AH and RH

C. *Ensemble*:

The predicted results from the XGBoost and FCNN are stored in a data frame where a comparison of predictions of both models and the expected values is made. The analysis shows that the data points that were predicted out of the expected range by XGBoost were predicted in range by FCNN and vice versa. Thus, an approach of average ensemble is used where the output values from both the models are averaged using a simple average formula for Relative Humidity (1) & Absolute Humidity (2) and the output of the averaging is considered as the final predicted values of Relative Humidity and Average Humidity.

$$RH\_Avg = (RH\_XGB + RH\_FCNN) / 2 \qquad (1)$$

$$AH\_Avg = (AH\_XGB + AH\_FCNN) / 2 \qquad (2)$$

The prediction accuracy of the models is calculated by determining relational coefficient also known as R2 Scores. R2(R Squared) Scores were chosen over other accuracy measurements since R-squared values vary from 0 to 1 and are usually expressed as percentages from 0% to 100%. The movements of a dependent variable are fully explained by movements of the independent variable when the R-squared is 100 percent (s). What constitutes a "healthy" R-Squared meaning can vary depending on the situation. Even a low R-Squared, such as 0.5, can be found high in certain areas, such as the social sciences. A decent R-Squared reading may have much higher expectations, such as 0.9 or higher.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Fig. 7, R2 Score Formula

The following is the table of comparison of the R2 scores of XGBoost, FCNN & Ensemble.

TABLE I. ACCURACY SCORES OF THE THREE MODELS IN USE

Prediction Factor	Type of Algorithm		
	XGBoost	*FCNN*	*Ensemble*
Relative Humidity (R2)	0.96597	0.91410	**0.96770**
Absolute Humidity (R2)	0.96439	0.96655	**0.97669**

[a] Prediction (R2 Scores) of the three Algorithms in use

The following figures illustrate the prediction results from the ensemble model.

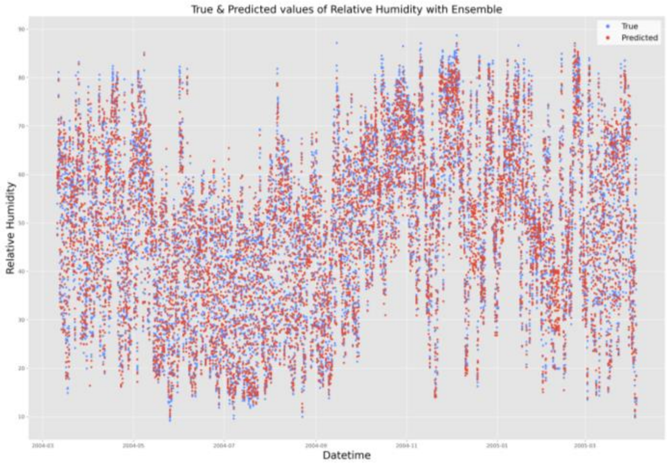

Fig. 8, Comparison of True and Predicted values of Relative Humidity at various times with Ensemble Model

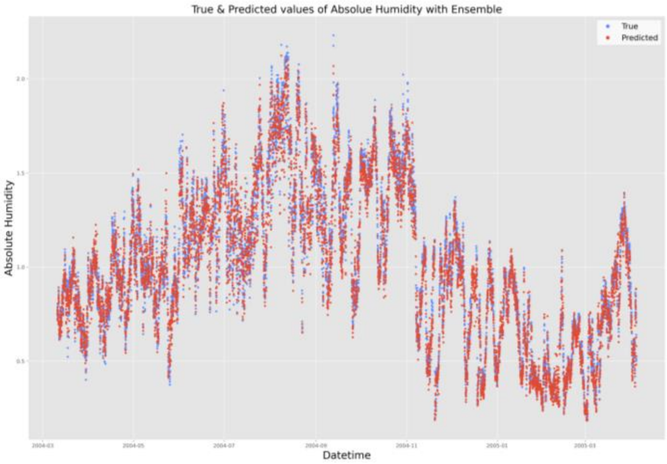

Fig. 9, Comparison of True and Predicted values of Absolute Humidity at various times with Ensemble Model

V. CONCLUSION

We can infer from the results obtained by training the models that the Ensemble model predicts Relative Humidity with a respectable accuracy of 96.77 percent and Absolute Humidity with a respectable accuracy of 97.66 percent. The prediction accuracy given by combining the two models is substantially higher than that provided by individual models.

The ability to determine humidity levels is important because it regulates air temperature by absorbing thermal radiation by both the Sun and the Earth. Furthermore, the higher the atmospheric water vapor content, the more latent energy is available for storm generation. We can calculate the values of humidity using all these models and prediction findings, and thus gain a detailed understanding of the same by further research.

REFERENCES

[1] Gunawardhana et al., 2017, L.N. Gunawardhana, G.A. Al-Rawas, S. Kazama "An alternative method for predicting relative humidity for climate change studies"Meteorol. Appl., 24 (2017), pp. 551-559, 10.1002/met.1641

[2] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sens. Actuators B Chem.*, vol. 129, no. 2, pp. 750–757, Feb. 2008, doi: 10.1016/j.snb.2007.09.060.

[3] "About us — scikit-learn 0.24.1 documentation." https://scikit-learn.org/stable/about.html#citing-scikit-learn (accessed Apr. 21, 2021).

[4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[5] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *J. Clin. Epidemiol.*, vol. 49, no. 11, pp. 1225–1231, Nov. 1996, doi: 10.1016/S0895-4356(96)00002-9.