Weather Forecasting using an Ensemble Data Mining Model

Jose Jimenez Espada

April 20th, 2025

1 Introduction

Weather forecasting has an impact on public safety, infrastructure, and financial support. Weather forecast models are preemptive in attempting to protect the public as effectively as possible, as well as preparing for better infrastructure and impact. Natural disasters from weather cost the world billions, if not trillions, of dollars in damages. Preventing this is a near impossible task; however, predicting it would be useful. It is also important to have relevant and accurate models, since they also can impact how towns and cities prepare. False positives are costly failures for large cities due to congestion, surge pricing, and over-preparation. False negatives could underestimate what these cities and towns are up against, thus causing disaster. The goal of this paper is not only to experiment and discover, but also to solve and enhance a problem scientist and technologies still get wrong. The vision of this paper is to enhance current technologies and see what was not seen before. Although current technologies are already predicting weather, this experiment will hopefully point out what causes the emergence of natural disasters. This will also be useful for predicting normal weather patterns.

2 Related Works

The references the paper will use already have attempted weather forecasting using their methods. This paper will also take inspiration from other papers that do not attempt to solve weather patterns, and this paper will learn the Machine Learning (ML) methods from the other papers. For example, image contrast enhancement; although this has nothing to do with weather forecasting, the methods in this paper will help with the results of this paper by taking their models as an approach to this experiment's model [4]. The simplest models attempted by other papers have been successful, such as the linear regression model [1]. Although it is a simple model, it was effective for the reference. Most of the references refer to ML methods, and this paper will take a combined approach from these references, in hopes of getting a much better result.

2.1 Fujiwhara Effect

The Fujiwhara effect explains a phenomena, where storms instead of harming each other or unifying into a megastorm(s), it is when two or more storms feed off each other [10]. This phenomena explains why storms exists and coexists. If this were not correct, storms would just merge into a megastorm, but the meterology of the earth works in a seperate clusters, and every storm has its own system. The idea behind this concept is that one storm may cause another, making weather even more predicatable, once this is fully understood.

3 Method

This paper will involve an ensemble method, which will consist of a Deep Learning Neural Network, Linear Regression model, a multiple dimension heuristic histogram model, and K-Nearest Neighbor, will be factored into an ensemble model [5, 8]. The ensemble model will be a Perceptron Neural Network that will take into account one processing element allowing to achieve a combined result. In other words, this ML approach has two layers, the inner layer involves the four ML methods, and the outer layer involves the combination of the inner layer (weighted output) giving the best approximation. The inner layer will give its prediction, some methods results may conflict others, and that is why the ensemble approach will be used. The result of this algorithm will be a large classifier, it will explain what the weather will be wether if it is raining, sunny, hurricane, etc. This will be numerically demonstrated where 0 is the low-end value, and 9 on the high end. These values are translated to classify what type of weather is occurring.

Numerical Value	Weather Classifier
0	Sunny
1	Partly Cloudy
2	Cloudy
3	Light Rain
4	Light Snow
5	Heavy Rain
6	Heavy Snow
7	Hail
8	Tornado
9	Hurricane

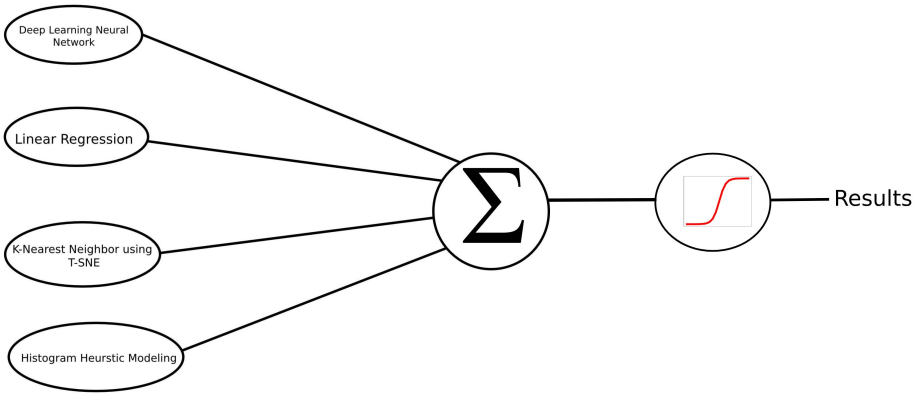

Figure 1: The "Ensemble" Neural Network (the first layer) Diagram

4 Experiment

4.1 Dataset

The data entry set will originate from valuable sources such as the National Oceanic and Atmospheric Administration (NOAA) [9] and kaggle datasets. The experiment will run on my device(s) which is limited in processing power and optimization. Because of this, the data will be based on only a specific area. The data will be primarily based in the United States as a whole. For training reasons, we will also use Australian weather datasets to adjusts our models.

The experiment used five datasets. One dataset contains Atlantic hurricane and tropical storm statistics from the year 2000 to 2023, containing 126 data entry sets [12]. The experiment also used a dataset containing every single recorded tornado that occurred in the United States since 1950 up to 2022, and contains about 622,000 recorded events and their stats [13]. The last two datasets are on classical weather classifying outputs containing over 1,000,000 recorded events [14,15]. The experiment will need more information on the location of the events recorded, so, we also used an api from openweathermap.org [11], we extracted average temperature readings, humidity, and altitude.

For testing and training purposes our dataset will be split using a 75/25 ratio, where 75 percent of the dataset, randomized is training, and 25 percent of it is used for testing.

4.2 Model Structure

The outer layer will be a weighted approach of all the methods combined. This can be well managed and demonstrated by a single layer Perceptron Neural Network using a single processing element. The initial values of weights will be

set to 1, and so will the bias. The activation function of the outer layer will be a relu activation function.

$$P_E = \Sigma wp + b^T \tag{1}$$

Equation for the "Ensemble" processing element and the Deep Neural Network processing element(s)

The parameters of the ensemble network are the following. Average temperature, Average humidity, Altitude, Yesterday's weather, the day before yesterday, the weather two days ago, the weather three days ago, the weather four days ago, the weather five days ago, and the weather six days ago. Notice how one part of the parameters consists of setting (location and time), and the other part consists of the recent past weather conditions in that setting. With the three parameters of the setting, Average temperature, Average humidity, and Altitude, these three variables when used simultaneously perfectly describe a unique climate.

The first part of the inner layer is the Deep Learning Neural Network [3]. This is different from the outer layer of the experiment, because this inner layer has multiple layers of calculations [2]. As mentioned in the parameter subsection, the neural network will take the parameters given into the system, and use it as part of the calculation, and it will result in a classifier. Equation (1) is still relevant for this part, except that there are multiple processing elements. The number of layers in the experiment will be based on a validation set. The network will be two layers deep, which contains one hidden layer, and one output layer.

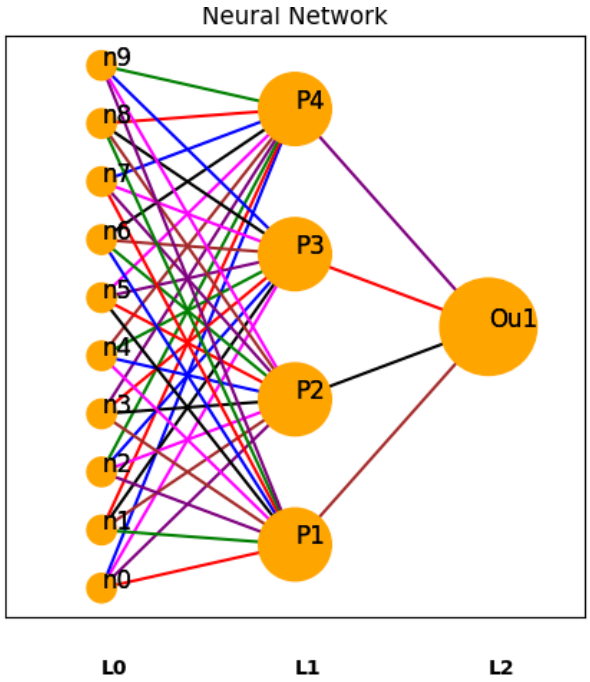

Figure 2: The Inner Deep Neural Network Diagram

The second part of the inner layer is the histogram approach [4]. This part of the inner layer, uses the heuristics of a histogram. The training will be based on the complexity of the bins. A single bin can represent single or multiple dimensions at the same time. The amount of bins will be found using the validation set. The histogram will not classify, but approximate the result. At the same time, another factor to this that will aid our analysis, is **streak analysis**. This method helps the system understand both maximum and minimum values, and hypothetically, these values are predictable. The streak analysis tracks and models world or local records broken throughout the years. A world record is rarely broken, thus we may track how frequently the record is broken. This concept enhances the heuristic by creating predictable bounds to the model.

The third part of the inner layer is the Linear Regression [1]. This may not always work due to lack of correlation [6]. Because of lack of correlation, this part of the layer will be effective only when the correlation percentage threshold is met. The "correlation percentage threshold" is defined by the equation below(4).

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} \tag{2}$$

Equation for the correlation coefficient

$$-1 \leq \rho_{x,y} \leq +1 \tag{3}$$

The correlation coefficient can only be within these range of values

$$cp = |\rho_{x,y}| \tag{4}$$

Equation for the "correlation percentage"

There are other terminology for correlation percentage, but for this paper, we will use this expression.

The fourth part of the inner layer is based on K-Nearest Neighbor [7]. This alone is not enough since it involves multiple dimensions, and thus needs a dimensionality reduction method. In this experiment, we will use "feature crossing," our parameters are large enough that when combined, the parameter will stand out more compared to other data points. This will help classify the category of weather that will occur, while taking account of all the leading variables.

As a result, the "Ensemble" method will unify all results in a weighted manner, giving a conclusive result on the weather. The experiment would have to iterate per every segment of the day it is interested in. For example, if we care about weather occurring at 3 pm, 4 pm, and 5 pm, the algorithm would have to calculate per every hour of each segment.

5 Results

5.1 Deep Learning Neural Network

The deep learning neural network went throughout various data entry points and adjustments. Due to the complexity of the project, and demand the calculation, the network went through out 200 epochs. It is found that five hidden processing elements in the neural network is correct balanced approach to this classification problem.

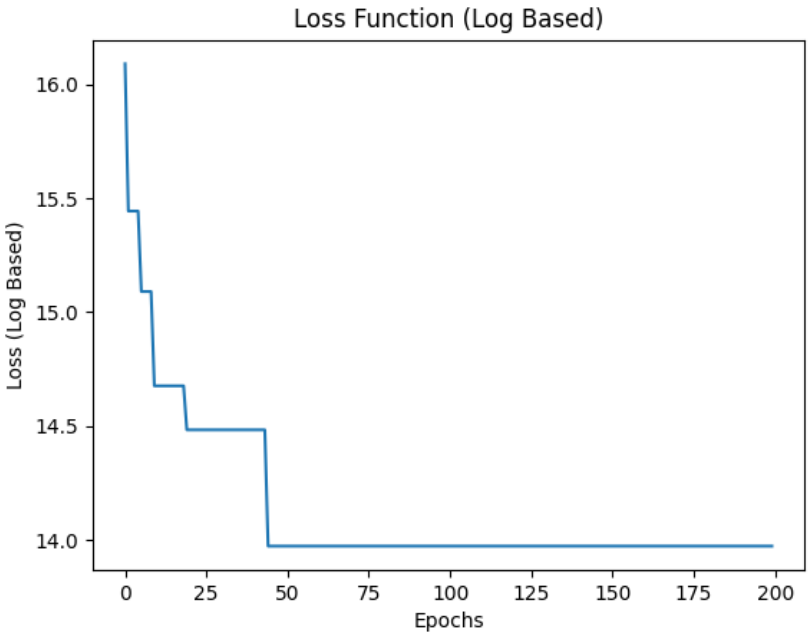

Figure 3: Streak Analysis Plot

The loss of network function is large due to the amount of epoch training given in the dataset. The accuracy of this layer of the ensemble resulted in 45.45%.

5.2 Histogram Approach

We have plotted the datasets into a histogram to find out the correct bin size. We find that having half of selected bins gave us useful information when it comes to the heuristics. Combining windspeed, precipitation, altitude, humidity, temperature gave us a widespread of information.

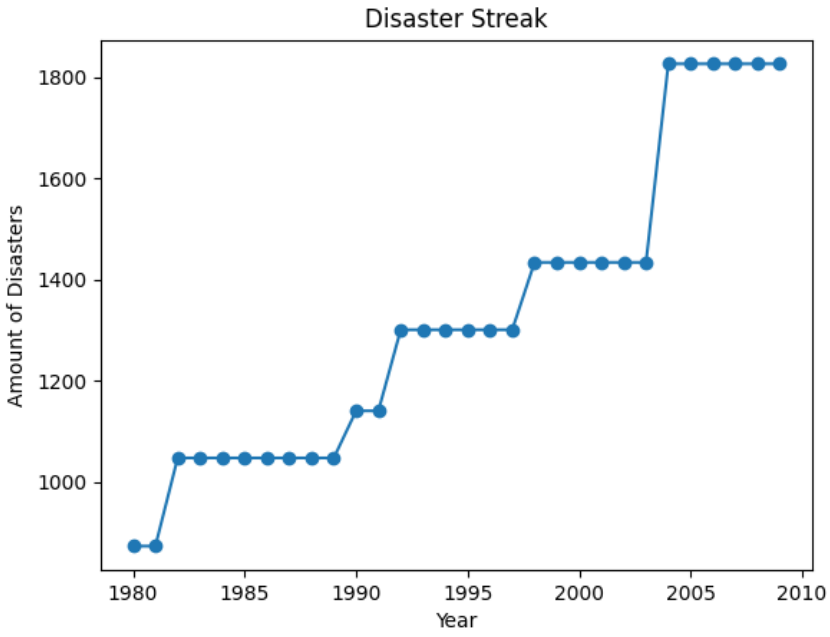

Figure 4: Streak Analysis Plot

The streaks analysis aided the heurstic by indicating if it overestimated or underestimated. Luckily in the experiment, this occurrence is rare. This heuristic method also assisted in showing where the mean of future data patterns will likely end up next.

5.3 Linear Regression

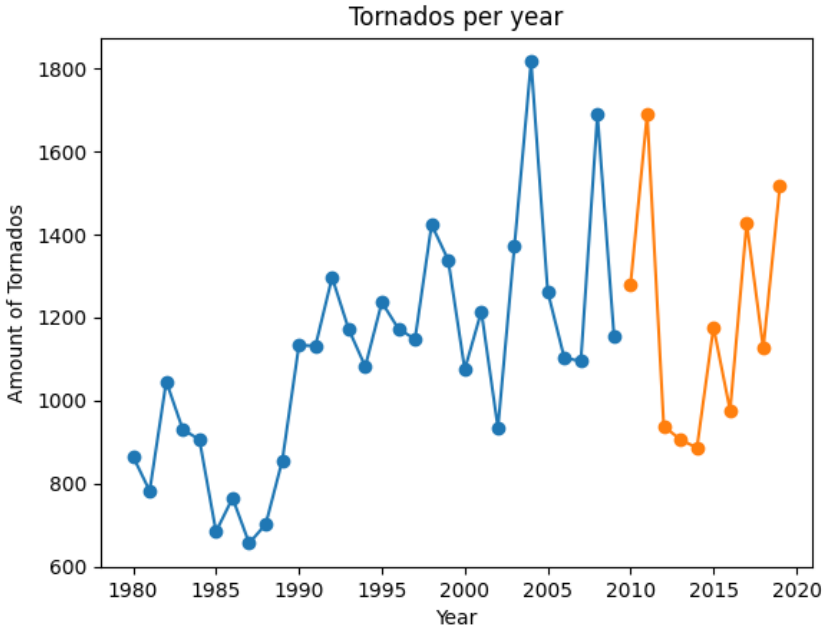

Figure 5: Tornados per year

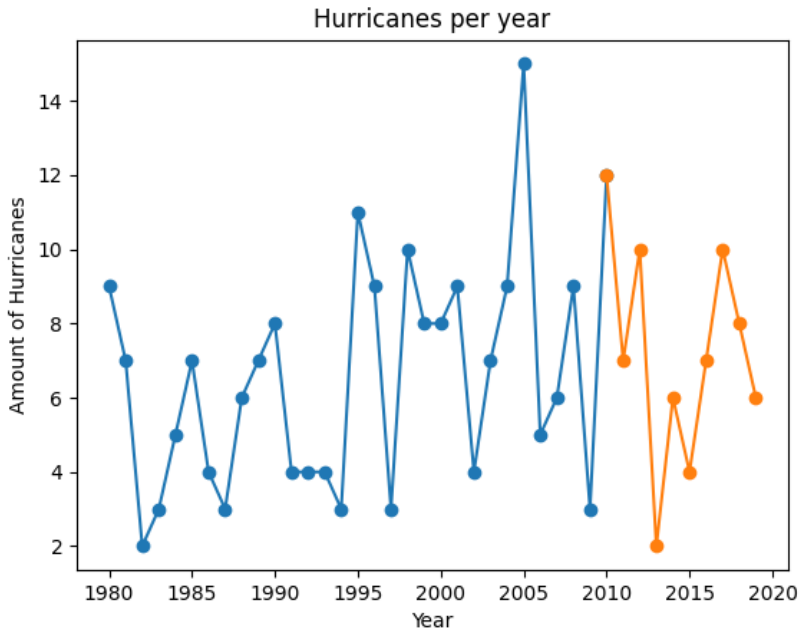

Figure 6: Hurricanes per year

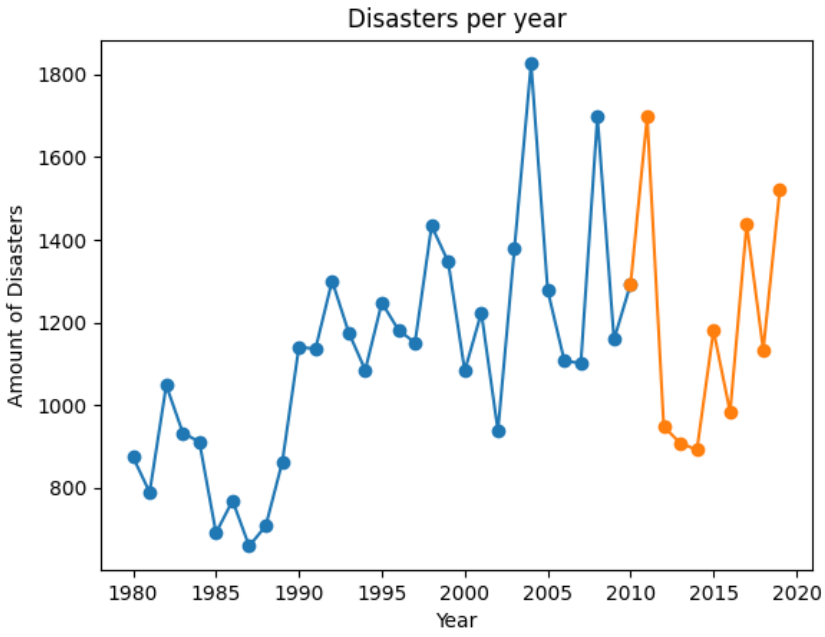

Figure 7: Disasters per year

In the experiment, we assumed there is a linear trend in weather patterns. Other studies have shown that this is the case[1]. In our analysis, we have seen that disaster occurence do increase on average on a yearly basis by about 6.5 storms. That means for every 15 years, 100 more storms occur on average on a yearly basis.is different

5.4 K-Nearest Neighbor

We also conducted two studies of K-Nearest Neighbor uniquely tied to feature crossing, and multiple axis for visualization.

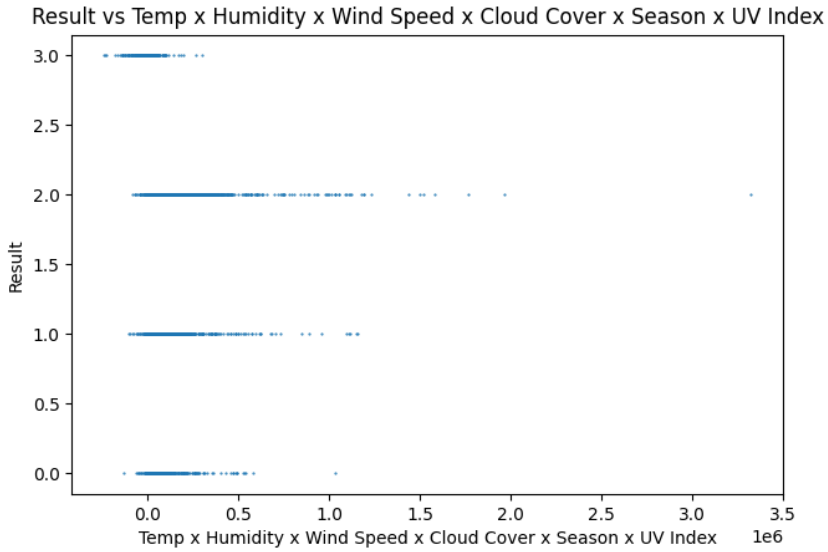

Figure 8: Result vs Temp x Humidity x Wind Speed x Cloud Cover x Season x UV Index

In figure 8, we can see that there are three different types of results. Based on the table shown earlier (Numerical value and Weather Classifier), we can see that this chart is generally predicting either sunny days, rainy, snowy, or disastrous day (hurricane or tornado). Even though datapoints on this image scale look parallel to other datapoints, they are not. In other words, every datapoint has a uniquely calculated and assigned value that no other datapoint has otherwise.

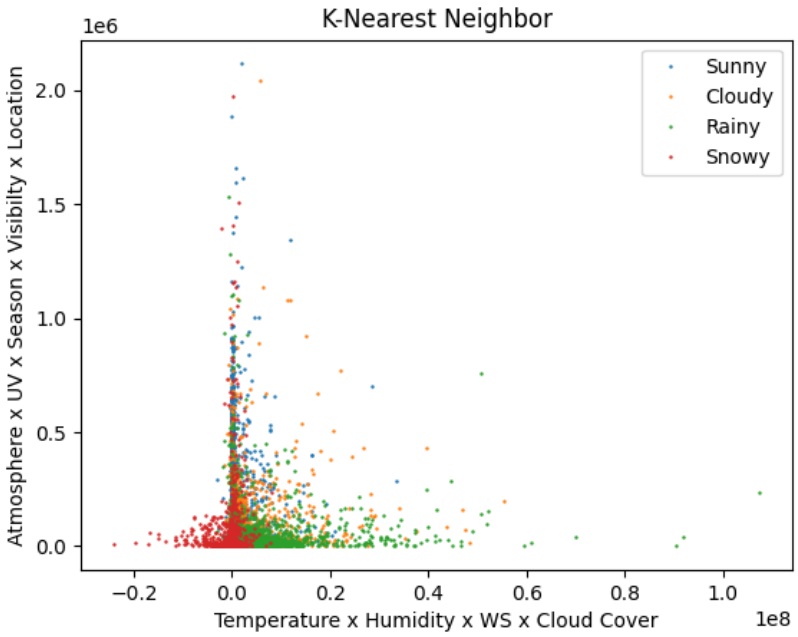

Figure 9: K-Nearest Neighbor Multiple Feature Crossing Axises

The same would go for our other study here with multiple axises. We cannot visually identify unique and separable clusters, since most data points are group closely to each other. However, we do see that every data point in the second analysis does cluster with one another, and no datapoint is ever the same.

5.5 Outer Layer (Ensemble)

When combined, our results improved slightly and uniquely. For the most part the network consisted more of multiple heuristics rather than more discrete values. The classifications did improve to 1800 / 3300 testing points or 54.54%.

6 Future Proposal

This experiment has shown us how complex the weather may be. This experiment needed more training and testing sets to further enhance the model. This method is designed to handle massive amounts of complexity, yet due to limitations, it is not delivered. The experiment should be able to go through more trainings to further diminish the loss function and improve accuracy.

Although the model is very sophisticated and designed for complexity, it is hungry for more data. The parameter selection should be expanded to handle

and process more historical datasets at a time, instead of the six recent classifications of the past. Improving the feature selection to further identify the climate of the setting is also suggested.

7 Conclusion

In this experiment, we tackled a major issue occuring in the real world. This experiment attempted to create a forecasting tool to try to predict disastrous weather from occurring before it made landfall. The experiment used what is called an ensemble model to try to tackle the complexity of the weather, by classification. The ensemble method contained other models that were part of the system. Deep learning neural networks, histogram heuristics, linear regression, and k-nearest neighbor.

References

[1] Mark Holmstrom, Dylan Liu, Christopher Vo. Machine Learning Applied to Weather Forecasting. *Stanford CS229 Machine Learning Course Project Report*. Stanford University, December 15, 2016.

[2] Thomas G. Dietterich. Ensemble Learning. *The Handbook of Brain Theory and Neural Networks, Second edition*. Oregon State University, 2002.

[3] Yann LeCun, Yoshua Bengio, Geoffrey Hilton. Deep Learning. *Nature*. 2015.

[4] Bin Xiao, Yunqiu Xu, Han Tang, Xiuli Bi, Weisheng Li. Histogram Learning in Image Contrast Enhancement. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Long Beach, CA, USA, 2019, pp. 1880-1889, doi: 10.1109/CVPRW.2019.00239.

[5] Tilmann Gneiting and Adrian E. Raftery. Weather Forecasting with Ensemble Methods. *Science*. 248-249(2005).DOI:10.1126/science.1115255

[6] S. Kothapalli and S. G. Totad. A real-time weather forecasting and analysis. *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. Chennai, India, 2017, pp. 1567-1570, doi: 10.1109/ICPCSI.2017.8391974.

[7] Kramer, O. (2013). Dimensionality Reduction with Unsupervised Nearest Neighbors. *Dimensionality Reduction with Unsupervised Nearest Neighbors*. Intelligent Systems Reference Library, vol 51. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-38652-7_2

[8] M. Wattenberg, F. Viégas, and I. Johnson. How to Use t-SNE Effectively. *Distill*. vol. 1, no. 10, Oct. 2016, doi: https://doi.org/10.23915/distill.00002.

[9] National Oceanic and Atmospheric Administration. *Noaa.gov*. 2019. www.noaa.gov.

[10] K. Yogi, "Feature cross a deep dive with practical examples," Medium, *medium.com* (accessed Apr. 23, 2025).

[11] "Fujiwhara effect," Wikipedia, *wikipedia.org* (accessed Apr. 23, 2025).

[12] "Weather Historical Stats API." openweathermap.org

[13] Middlehigh, "North American Hurricanes from 2000." kaggle.com

[14] NOAA, "Tornados [1950 - 2022]." kaggle.com

[15] Ruthvik Srinivas Deekshitulu, "Weather_Data." kaggle.com

[16] Ananth R, "WEATHER PREDICTION." kaggle.com