World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(Research Article)

Classification of cyclones using machine learning techniques

Sasmita Kumari Nayak *

Associate Professor, Computer Science and Engineering, Centurion University of Technology and Management, Odisha, India.

World Journal of Advanced Research and Reviews, 2023, 20(02), 433–440

Publication history: Received on 18 September 2023; revised on 25 October 2023; accepted on 27 October 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.20.2.2156

Abstract

In this article, we provide a method for identifying and categorizing cyclones, both tropical and extratropical. The method is designed with the goal of producing a global labeled dataset for cyclones, and it is based on a set of rigorous criteria. The heuristics are defined from date, time, pressure, wind speed, wind directions, latitude and longitudes. Numerous researchers have confirmed that machine learning, a kind of artificial intelligence, can offer a fresh approach to overcoming the limitations of cyclone classification, whether employing a pure data-driven model or enhancing numerical models with machine learning. This article introduces progress based on machine learning in genesis classification, track records, intensities, and extreme weather forecasts associated with tropical as well as extratropical cyclones (such as strong winds and rainstorms and their disastrous impacts). The challenges of cyclones in recent years and successful cases of machine learning methods in these aspects are summarized and analyzed.

Keywords: Classification; Cyclone; Machine Learning; Dataset; Features; Random Forest

1. Introduction

When a cyclone hits land, it may cause severe economic damage and casualties. It is vitally desirable to have a precise understanding of the cyclone development in order to provide the public with a timely warning since any tropical disturbance of adequate scale has the potential to grow abruptly into a cyclone over the warm water [1], [9]. In order to enhance the initial selection of satellite data to be utilized in a range of categorization and forecasting applications, the High Performance Computing Group of NOAA Earth System Research Laboratory (ESRL) is studying uses of machine learning [2].

Python is an approachable and well-known programming language in the field of Machine Learning. Here, we may examine the differences between supervised and unsupervised learning, as well as the connections between statistical modeling and machine learning, and compare them. We can investigate a number of well-liked techniques, such as Classification, Regression, Clustering, and Dimensional Reduction, as well as well-liked models, such Train/Test Split, Root Mean Squared Error (RMSE), and Random Forests.

The following section will describe the motivation for the work followed by data, methods for deriving classifications, and constructing the models in Section 3. Section 4 compares the performance of four ML models used for accuracy detection models (i.e., Decision Tree, Random Forest, Naïve Bayes and SVM). A summary of the study is given in the Section 5.

* Corresponding author: Sasmita Kumari Nayak; Email:nayaksasmita484@gmail.com

2. Motivation

Extreme weather conditions have a major effect on people's everyday lives and economies. Particularly expensive cyclones, hurricanes mostly harm coastal communities [2]. Therefore, a robust, accurate, and time-efficient deep learning algorithm may be trained using a well-tailored labeled dataset of cyclones and extra-tropical cyclones, which can subsequently be used by forecasters for both classification and prediction purposes.

Specialists claim Python is the most popular language for AI and ML among all. As AI and ML are implemented across many channels and sectors, large organizations invest in these domains, and the need for specialists in ML and AI develops proportionately. One of the key causes Python is the most often used programming language for AI is the fantastic selection of libraries. Python libraries offer fundamental components so that developers don't always have to write them from scratch. Python's versatility enables programmers to select the programming paradigms with which they are most comfortable or even mix these paradigms to effectively address a variety of challenges.

To analyze and study the types of cyclones from the collected data and classify them into various segments according to their features using python programming for better understanding.

3. Methodology

We used 66 years (1949-2015) of track data from Kaggle website to classify and differentiate types of cyclones according to their different features using machine learning python programming.

3.1. Machine learning

The main objective of machine learning, which consists of a number of computer programs, is to create mathematical models using statistics in order to draw conclusions from samples. Learning is the execution of a computer program that uses training data or experiences to maximize the model's parameters given a model that defines specific parameters. The model may describe data-derived knowledge; forecast the situation of the future, or do both.

A model is developed via ML algorithm training on a training data set. The ML algorithm produces a prediction based on the model when fresh input data is introduced. The accuracy of the forecast is assessed, and if it is deemed acceptable, the ML algorithm is used. If the accuracy is not satisfactory, the ML algorithm is again trained using an expanded training set.

3.2. Data pre-processing

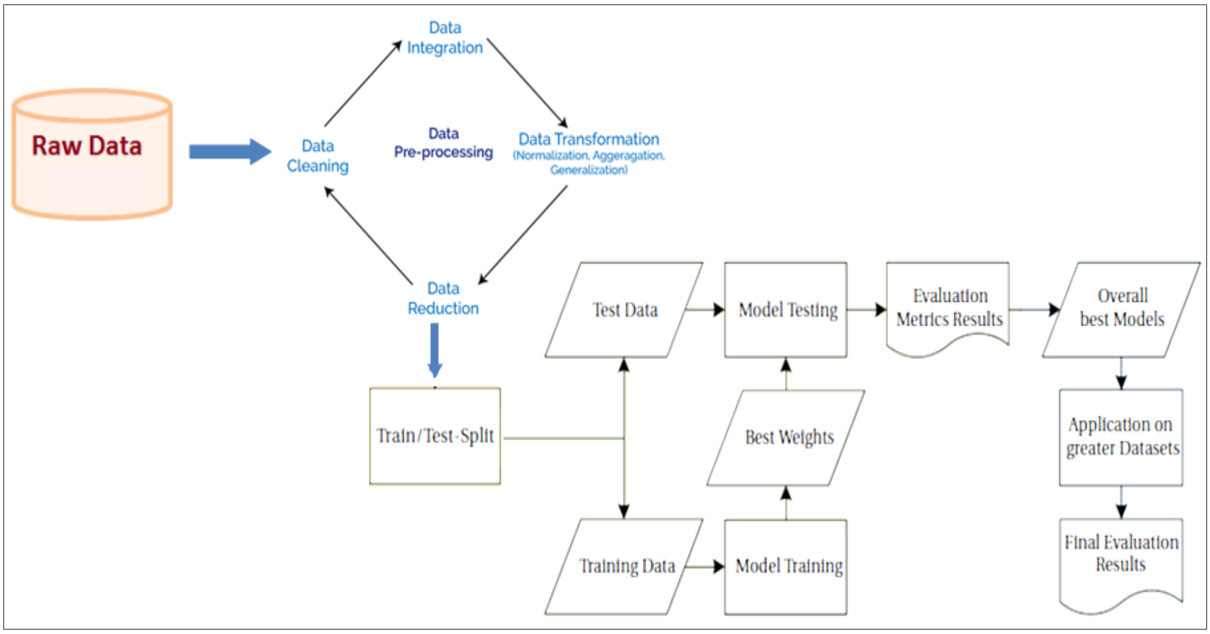

Figure 1 Practical Process of Machine Learning [3]

Data preprocessing is the process by which the data is altered or encoded to put it in a state where the computer can parse it with ease. In other words, the algorithm can now quickly comprehend the data's characteristics. It is a process used to turn a raw data set into a clean data set. Every time data is obtained from various sources, it is done so in a raw manner that makes analysis impossible. Hence, it needs to be cleaned first.

When working on a machine learning topic, dealing with data quality concerns takes up more than half of our time since data is frequently gathered from several sources that are typically not very reliable and that too in various forms. It is just unreasonable to anticipate flawless data. Human mistake, measurement instrument limitations, or defects in the data collecting process might all cause issues. There are a few ways to handle them.

3.2.1. Missing values

Missing values are quite common in datasets. Regardless of how it happened, missing values must be taken into account. It may have happened during data collection or it might have been because of a data validation rule. Some, techniques to avoid the missing values like eliminate rows with missing data, manual filling of data, fill the mean or median or mode of each column data etc.

3.2.2. Inconsistent values

We are aware that data may include numbers with discrepancies. We have most likely already encountered this problem. In the 'Address' box, for instance, the 'Phone number' is present. The information may have been misunderstood when being scanned from a handwritten form or it could have been a case of human error. As a result, it is always advisable to do data evaluation, such as determining what the data type of the characteristics should be and if it is the same for all of the data objects.

3.2.3. Duplicate values

Data items that are identical to one another may be present in a dataset. It may occur, for example, if a single user submits a form many times. The method of removing duplicates is frequently referred to as deduplication. The majority of the time, duplicates are eliminated in order to avoid biasing machine learning algorithms in favor of a certain data object.

3.2.4. Feature sampling

A highly popular technique for choosing a portion of the dataset we are investigating is sampling. When memory and time are limited, dealing with the entire dataset might often end up being too costly. We can use a sampling technique to assist us shrink the amount of the dataset so that a more effective but more expensive machine learning approach may be used.

The important thing to remember is that sampling should be done in a way that ensures the sample produced has roughly the same characteristics as the original dataset, indicating that the sample is representative. For this, the appropriate sample size and sampling approach must be chosen. It has two main variations as well :

Sampling without Replacement: As each item is selected, it is removed from the set of all the objects that form the total dataset

Sampling with Replacement: Items are not removed from the total dataset after getting selected. This means they can get selected more than once.

3.2.5. Dimensionality reduction

Dimension is conceptually the number of geometric planes that the dataset occupies, which may be so many that a pencil and paper cannot be used to display it. The complexity of the dataset increases as the number of these planes increases.

3.3. Data analysis

Investigation of the content and composition of data, will help to inform how the project should proceed and whether the data can answer the questions we are aiming to address or not. Data analysts are in charge of deciphering data, applying statistical methods to the outcomes, and producing periodic reports. They create and execute statistically efficient and high-quality data analytics, data gathering methods, and other techniques. Additionally, they are in charge of managing databases and collecting data from primary or secondary sources. In addition, they locate, examine, and evaluate patterns or trends in large, complicated data sets. To identify and fix code errors, data analysts examine reports, printouts, and performance indicators. They can clean and filter data by doing this.

In this paper, first we fed the dataset into the program and then searched for the missing values. Then we changed the date and time format from object to datetime format for easier interpretation. Next we replaced the missing values to fill them appropriately. At last, we created two new columns named Latitude Hemisphere and Longitude Hemisphere to change the directions of wind from N, S, E, W to 1, 2, 3, 4 respectively to read the samples as integers, not as objects.

3.4. Model construction

3.4.1. Decision tree

The decision tree method, which continually divides the data sample into subdivisions based on decision rules and resembles tree branches, has been used in many recent remote sensing investigations for both classification and regression. The benefit of the decision tree is that it gives visible if-then rules with the relative relevance of predictors, facilitating simple interpretation and physical insights to the categorization rules. Through training, it chooses an empirically best set of predictions on its own. Overall accuracy, which measures the percentage of samples that are properly categorized [4], [7], [12] is calculated to assess the degree of fitting for 1000 separate training datasets.

The above formula is used for attribute selection measures (ASM) for information gain using Entropy from the dataset. Information gain is the difference between the original information requirement and the new requirement.

3.4.2. Random forests

It is an ensemble method based on classification and regression trees (CART), which was initially created to address the well-known issue of DT that is dependent on the configuration of training data. It has a long history of use in remote sensing applications for both regression and classification work. In order to create several independent decision trees, RF uses two randomization methods [11], [12]. The ultimate choice is then decided using a majority voting (or weighted voting) technique. The mean drop in accuracy when the variable is permuted to random values, as quantified by RF, indicates the relative relevance of a given variable. The crucial predictor is more important than other variables. To fit the data into the RF method, this study employed the R programming language. Since the RF model includes.

3.4.3. Naive-bayes algorithm

It is a classification method built on the Bayes Theorem and predicated on the idea of predictor independence. An assumption made by a Naive Bayes classifier is that the existence of one feature in a class has no bearing on the presence of any other features.

Bayes' Theorem is given by:

$$\mathbf{P}\ (\mathbf{A}|\mathbf{B}) = \frac{\mathbf{P}(\mathbf{B}|\mathbf{A})\ .\ \mathbf{P}(\mathbf{A})}{\mathbf{P}(\mathbf{B})}$$

Here, P (A): Probability of hypothesis being true, known as prior probability.

P (B): Probability of the evidence.

P (A|B): Probability of evidence given that hypothesis is true.

P (B|A): Probability of the hypothesis given that the evidence is true.

This principle only applies to conditional probability. The likelihood that something will happen provided that something else has already happened is known as the conditional probability. We can determine the likelihood of an occurrence using the conditional probability and our prior information.

3.4.4. Support vector machines (svm)

In recent years, SVM has become one of the most popular machine learning models for remote sensing applications because it determines the best hyperplane for data classification. SVM makes good use of a kernel function to change the data dimension into a higher one in order to find an ideal hyperplane. The radial basis function proved to be the best kernel function in our test among the linear, polynomial, and radial basis functions. To get the optimum performance in identifying the development of tropical cyclones, the kernel and penalty parameters employed in the SVM model were automatically changed during the data training process [10]. Each predictor variable was linearly scaled to the range from 0 to 1 before to data training to account for the magnitude difference.

SVM does not instantly reveal the information on the relative value of predictors, in contrast to other machine learning techniques. As an alternative, the F-score test is used to pinpoint the key SVM-based traits that distinguish the genesis of tropical cyclones. A test's accuracy is measured by the F score, also known as the F1 score or the F measure, which is the weighted harmonic mean of the test's recall and precision [5-8], [11], [12].

4. Results and Discussion

The dataset contains events occurred in between 1949 to 2015 from the smallest to the biggest cyclones in the Pacific Oceans having 26137 samples and 22 attributes such consisting of Name, Timestamp, Status, Latitude and Longitude of the location, Maximum and minimum Pressure, Low Wind Directions, Moderate Wind Directions, High Wind Directions etc. for better accuracy of the data. We have collected the data from Kaggle website.

For our simulation, we chose to take help of Google Colab for python programming. Google Colab is a free python programming platform, can be accessed via our browser in the desktop. We used ML programing codes for visualization as well as ML models for accuracy predictions.

Using python programming in our dataset, we found out of 26137 samples, there are more than 10000 events turned out to be Tropical storms, 6733 Hurricanes and so on. We also visualized the cyclones according to their yearly frequencies, monthly frequencies, category wise frequencies etc.

4.1. Cyclone classification visualizations

Using ML coding, we classified the data into different categories of cyclones according to their features and frequencies. Figure 3 differentiates the cyclones according to their status and classifies them with their number of counts and finds which type of cyclone has occurred the most. From this figure we got that Tropical Storms (TS) type of cyclones occurs most of the times, followed by Tropical Depression (TD) and Hurricanes (HU).The occurrence of Subtropical Depression (SD) and Subtropical Storm (SS) are lowest.

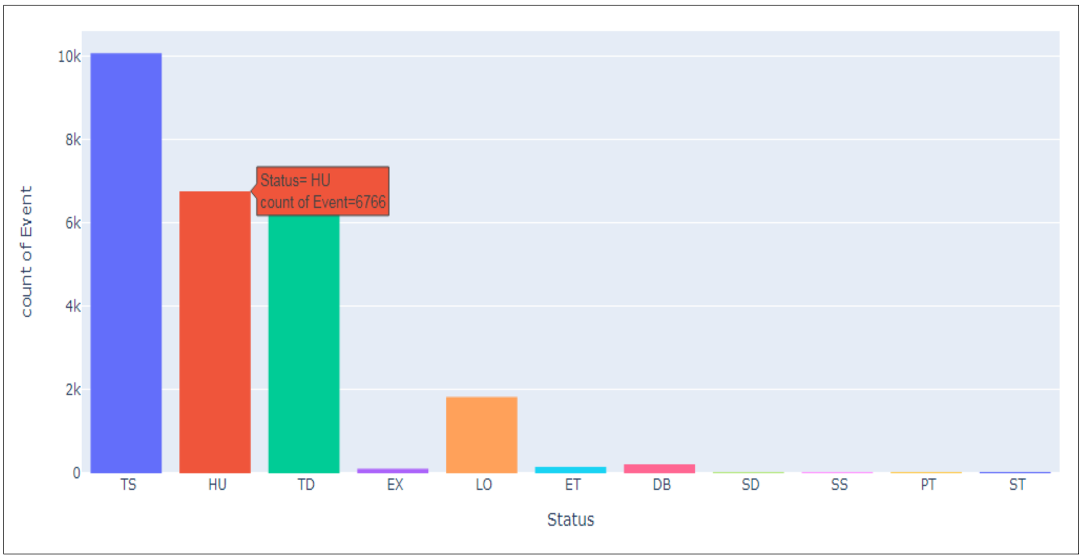

Figure 2 Classification of Cyclones by their Counts

Figure 3, visualizes the top ten cyclones according to their frequencies From the diagram, we observed that the cyclones that have taken place the most number of times at 46.1% are remained Unnamed whereas Fernanda at 6.7% is the second most followed by John at 6.2%.

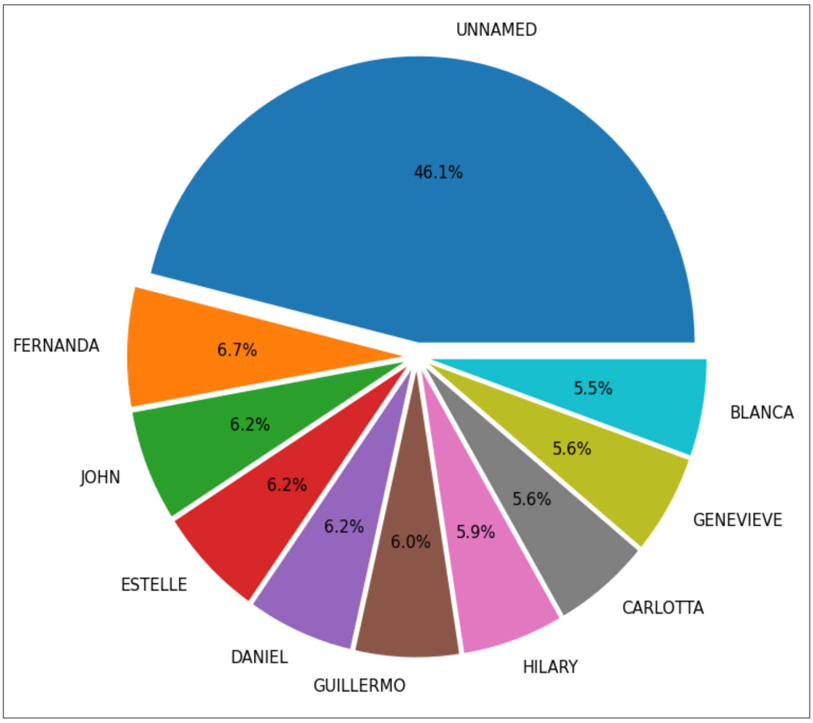

Figure 3 Top Ten Cyclones by their Frequencies

In Figure 4, we have classified the cyclones according to their frequencies month wise where we found that the frequencies of cyclones occurred in the month of August are maximum, followed by September and July

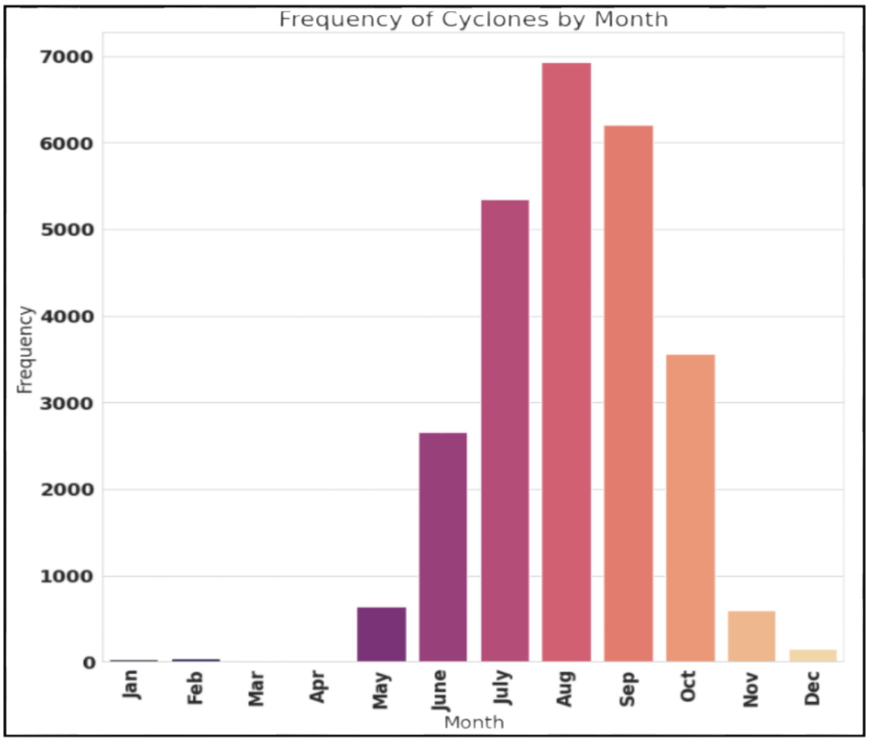

Figure 4 Frequency Of Cyclones By Month

In Figure 5, we observed the yearly increase and decrease of Cyclones according to their frequencies. We found that the average frequencies of the cyclones were maximum in the 1990's, after that they decreased a little bit but again moving towards peak in the 2010's.

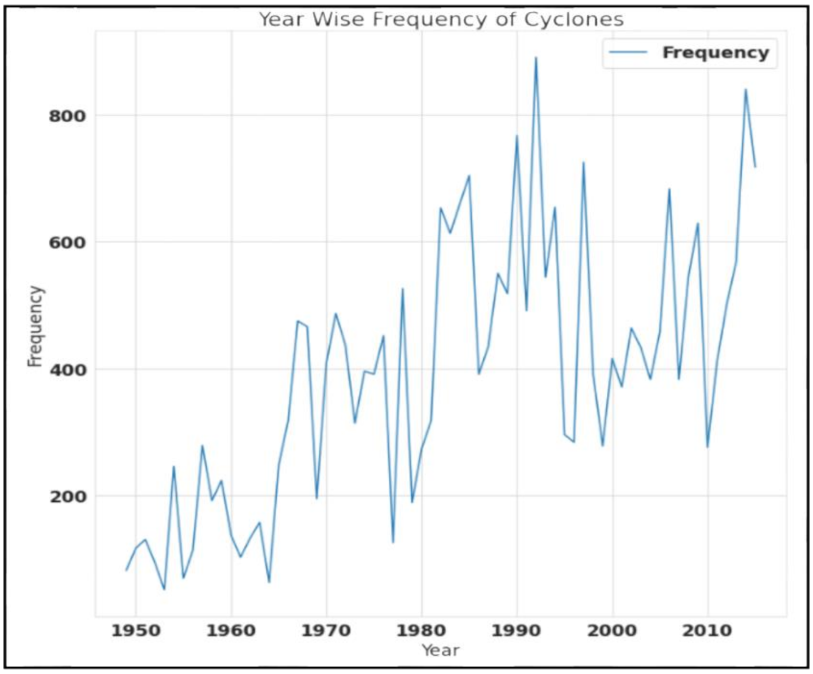

Figure 5 Year Wise Frequency Of Cyclones

4.2. Finding the most accurate ml model

We have performed Decision Tree, Random Forest, Naive – Bayes and SVM model for our collected data to found out which model has better accuracy and the results we found were given in the below table .

Table 1 Results of different ML Models

Result	Decision Tree	Random Forest	Naive-Bayes	SVM
Accuracy (%)	96.5106827	98.0871782	92.6769747	92.8232593
Recall (%)	96.5106827	98.0871782	92.6769747	92.8232593
Precision (%)	96.3996987	98.0397278	94.4453425	94.6456327

From the above table, we can easily conclude that Random forest ML model has the best accuracy in comparison with other models. The accuracy score of Random forest as well as Recall score and Precision score also found out to be maximum.

The Random forest model can be used for both classification and regression and can also handle missing values. It has a main advantage of identifying the most important features from the training dataset.

5. Conclusion

For more than a century, meteorologists have been concerned about cyclones. Numerous academics have carried out in-depth research on important topics, including the structure, dynamics, and forecasting methods. Machine learning is developed from statistical techniques that may automatically identify pertinent rules for detection, analysis, prediction, etc. from enormous amounts of data. The use of machine learning to solve the fundamental issues with cyclones offers a fresh perspective on how to solve several obstacles in this area.

It can be concluded that, Random forest model has best accuracy of 98.087% in comparison with other ML models for our data. The Random forest model can be used for both classification and regression and can also handle missing values. It has a main advantage of identifying the most important features from the training dataset.

References

[1] Kim, M., Park, M. S., Im, J., Park, S., & Lee, M. I. (2019). Machine learning approaches for detecting tropical cyclone formation using satellite data. *Remote Sensing*, *11*(10), 1195.

[2] Bonfanti, C., Trailovic, L., Stewart, J., & Govett, M. (2018, July). Machine Learning: Defining Worldwide Cyclone Labels for Training. In *2018 21st International Conference on Information Fusion (FUSION)* (pp. 753-760). IEEE.

[3] Sasmita Kumari Nayak, Swati Sucharita Barik, Mamata Beura," Weather Forecasts Based on Rainfall Prediction Using Machine Learning Methodologies," Adalya Journal 9 (6), Page No: 72 – 80, ISSN NO: 1301-2746.

[4] Chen, R., Zhang, W., & Wang, X. (2020). Machine learning in tropical cyclone forecast modeling: A review. *Atmosphere*, *11*(7), 676.

[5] Jena, T. R., Barik, S. S., & Nayak, S. K. (2020). Electricity consumption & prediction using machine learning models. *Acta Tech. Corviniensis-Bull. Eng*, *9*, 2804-2818.

[6] Park, M. S., Kim, M., Lee, M. I., Im, J., & Park, S. (2016). Detection of tropical cyclone genesis via quantitative satellite ocean surface wind pattern and intensity analyses using decision trees. *Remote sensing of environment*, *183*, 205-214.

[7] Nayak, S. K. (2020). Analysis and high accuracy prediction of coconut crop yield production based on principle component analysis with machine learning models. *International Journal of Modern Agriculture*, *9*(4), 359-369.

[8] Nayak, S. K., Barik, S. S., & Beura, M. (2020). Analysis of Infectious Hepatitis Disease with High Accuracy Using Machine Learning Techniques. *TEST Engineering & Management*, *83*, 83.

[9] Chen, Z., Yu, X., Chen, G., and Zhou, J. (2018). "Cyclone Intensity Estimation Using Multispectral Imagery from the FY-4 Satellite," in 2018 International Conference on Audio, Language and Image Processing (ICALIP) (IEEE), 46–51. doi:10.1109/icalip.2018.8455603

[10] Huang, C.-C., Fang, H.-T., Ho, H.-C., and Jhong, B.-C. (2019). Interdisciplinary Application of Numerical and Machine-Learning-Based Models to Predict Half-Hourly Suspended Sediment Concentrations during Typhoons. *J. Hydrology* 573, 661–675. doi:10.1016/j.jhydrol.2019.04.001

[11] Jena, T. R., Barik, S. S., & Nayak, S. K. (2020). Electricity consumption & prediction using machine learning models. Acta Tech. Corviniensis-Bull. Eng, 9, 2804-2818.

[12] Nayak, S. K., Beura, M., Siddique, M., & Mishra, S. P. Analysis of Indian Food Based on Machine learning Classification Models.