Flood Damage Estimation using Machine Learning in GIS

Aishwarya Shaharkar[1], Yash Sonar[2], Advait Sonar[3], Dr. Chhaya Pawar[4]

[1-4]*Department of Computer Engineering, DMCE, University of Mumbai, Mumbai, India*

Abstract- Floodimpact is one of the most significant disasters in the world. More than half of the global flood damages occur in Asia. Flood disaster management in developing countries is most susceptible to and responding to current disaster situations. Generally, there are some common factors by which an area can be examined to predict floods. The area is generally examined based on certain factors. The factors have specific threshold values that are set based on the observations collected from the previous flood-affected areas. Based on these threshold values, our predictions are obtained using Machine Learning. These factors are responsible for the rise of water levels in that area. The data of a region is collected through a technique called "Satellite Remote Sensing." This technique is a powerful tool to map flooded areas. Through Satellite Remote Sensing, the probability of the occurrence flood is predicted by using the images of the area obtained. By using these images and by applying various algorithms, a conclusion is made as to whether that area is flood-prone or safe. The predictions are made based on the live status of that region.

***Keywords*: Flood, Satellite Remote Sensing, Machine Learning.**

I. INTRODUCTION

Disasters, whether man-made or natural, have become an issue of escalating concern all over the world. Over the years, numerous reports of natural disasters have been steadily on the rise and severity as well as the impact of these natural disasters to the global economy and environment has increased noticeably. The increased availability of free of cost satellite data has boosted the study of many natural and human-made processes at low cost and has boosted research in many fields [1-4]. For instance, the Sentinel satellite constellation of the Copernicus program of the European Union [5] provides synthetic aperture radar (SAR) and multispectral data with global coverage, high-frequency pass, and high spatial resolution. Other examples of free remote sensing programs are Landsat [6], and the MODIS daily satellites giving multispectral images [7]. Every year, flood events cause significant economic losses and victims. For this reason, precise flood mapping and modeling are essential for flood hazard assessment, damage estimation, and sustainable urban planning to manage flood risk properly. In such a context, satellite remote sensing is currently a low-cost tool that can be profitably exploited for flood mapping.

The frequent passes of satellites and the availability of rapid processing chains allowed the development of services providing automatic and quasi-real-time flood mapping such as, for example, the Copernicus Emergency Management Service (EMS) performed by the European Union [8], the Global Flood Detection System and the NASA Global Flood Mapping System. However, these services provide rapid mapping products that can be affected by uncertainty and are not always validated. Maps of flooded areas produced by official authorities and based on bespoke aerial photos and field surveys are highly accurate, although they are time-consuming and require higher costs to be generated while maps of flooded areas produced by remote sensing are comparatively less accurate but time-saving and cheap.

We present a multi-sensor, low-cost, and user-friendly approach for flood inundation. Specifically, we combined semi-automatic and manual approaches for flooded areas. The various machine learning algorithms used for training our model are Logistic Regression, Decision Tree, Random Forest, and XG Boost. Our research aims to define a procedure that can also be used by non-remote sensing-processing experts to map flooded areas using free satellite data. We developed and tested the presented methodology on flood events that occurred in Kerala in 2018 and Houston, Texas, in 2017 and Harvey in 2016. We used all available free satellite data services: Sentinel-1, Sentinel-2, and Landsat-8. We also focused on the factors that limit, or help, the capability of flooded area detection.

II. LITERATURE REVIEW

1) The authors have proposed a way of flood monitoring using satellite-based RGB composite imagery and index of refraction retrieval in visible and near-infrared bands. So as to map the inundation area, flooding events are investigated using unique RGB composite imagery supported by the MODIS surface reflectance (MOD09GA) data obtained from the Terra satellite, which is employed to see and analyze these events. This study proposes using an RGB combination of MODIS band 6 (1.64 μm), band 5 (1.24 μm), and band 2 (0.86 μm) data from the visible and therefore the near-infrared spectral ranges to map flood events. Consequently, this study provides a useful RGB imagery technique for mapping flood events employing a physical validation supported by the index of refraction. The technique is often applied to a spread of spectral ranges from ultraviolet to microwave wavelengths [15].

2) The authors have proposed a flood mapping strategy which consists of progressive steps. Firstly, the detection of flooded areas is performed, through semi-automatic extraction or by visual interpretation of satellite images. The flood map is later manually refined with the assistance of ancillary data like digital elevation models (DEMs), water depth (WD) models or ground photos, and web-based information. The proposed methodologies are suitable to be applied by the community involved in flooding hazard management, not necessarily experts in remote sensing processing. As case studies, they chose three flood events that recently occurred in Spain and Italy. Multispectral satellite data acquired by MODIS, Proba-V, Landsat, and Sentinel-2 and artificial aperture radar (SAR) data collected by Sentinel-1 were wont to detect flooded areas using different methodologies (e.g., Modified Normalized Difference Water

Index, SAR backscattering variation, and supervised classification) the ultimate outcome is an accurate and morphologically based flood map in vector format. within the presented case studies, they validate the methodology results with available flooded area maps provided by local authorities [16].

3) The authors have proposed Flood susceptibility assessment employing a GIS-based support vector machine model with different kernel types. In this, they used support vector machine techniques. This system in natural hazard assessment is getting extremely popular lately. It contains a training stage associated with the input and desired output values. SVM was applied using four kernel sorts of LN, PL, RBF, and SIG, and subsequently, flood probability maps were produced within the GIS environment. The flood probability index ranges from 0 to 1 as 0 represents no probability of flood occurrence and 1 shows 100% probability. so as to supply susceptibility maps and to possess a far better visual interpretation of susceptible areas, it's required to categorize the probability maps into various classes. the most important area under the success rate curve belonged to the SVM-SIG (88.89%) followed by SVM-RBF (88.02%). Furthermore, the achieved success rates for SVM-PL, SVMLN, and FR were 84.31%, 81.89%, and 72.48% respectively [17].

III. Proposed System

The Proposed system consists of Computer Vision and Machine Learning Models. Computer Vision is used to extract the pixel information from the image and machine learning is used to learn from the extracted information.

A) Computer Vision:

Computer vision is concerned with the automatic extraction, analysis, and understanding of useful information from a single image or a sequence of images. It involves the development of a theoretical and algorithmic basis to achieve automatic visual understanding. The monitoring of water levels is of extreme importance in early warning systems, and computer vision has shown to be useful [10]. Image filtration in computer vision plays a vital role in estimating water levels [11]. For example, Yu et al. [11] proposed the differencing image technique to track and detect minor changes in water level.

Thresholding:

Thresholding is the errand of each pixel in an image to either a true or false class based on the pixel's value, location, or both. The outcome of a thresholding operation is as a rule a binary image in which each pixel is assigned either a true or false value.

B) Machine Learning Models:

The required data is collected and merged to make a dataset which is split for training and testing purposes. The dataset is fed as input to the Machine Learning model and techniques for the classification and prediction of the flood also as flood-prone areas. The model is trained to pick and classify the flood-prone area on pre and through flooding areas of the flood events that occurred. The trained model is then used for the prediction of flood and classification of the flood-prone area Following models of Machine Learning are used:

1] XG Boost: XG Boost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. XG Boost provides a parallel tree boosting (also referred to as GBDT, GBM) that solves many data science problems during a fast and accurate way.

2] Random Forest Algorithm: Random forest may be a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, an excellent result most of the time. it's also one among the foremost used algorithms, due to its simplicity and variety (it is often used for both classification and regression tasks.)

3] Decision tree: Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. Decision trees learn from data to approximate a sinusoid with a group of if-then-else decision rules. The deeper the tree, the more complex the choice rules, and therefore the fitter the model. They are commonly utilized in research, specifically in decision analysis, to assist identify a technique presumably to succeed in a goal, but also are a well-liked tool in machine learning.

C) Snap Software:

We have used the SNAP architecture in our project which is right for Earth Observation processing and analysis in view of the subsequent technological innovations: Extensibility, Portability, Modular Rich Client Platform, Generic EO Data Abstraction, Tiled Memory Management, and a Graph Processing Framework [12].

D) Factors

1) NDWI Index:

Normalized Difference Water Index (NDWI), introduced in 1996, measures the moisture content in soil and plants and is determined by analogy with NDVI as:

$$NDWI = \frac{NIR - SWIR}{NIR + SWIR}$$

NIR – near-infrared range with wavelengths in the range of 0.841 – 0.876 nm. The Snap software can be used to calculate the NDWI image or use cv for the same. NDWI is used to differentiate water from dry land or instead for water mapping. Water bodies have low radiation and strong absorbability in the visible, infrared wavelengths range. NDWI uses near Infra-red and green bands of remote sensing images based on the occurrence. It can boost the water information efficiently in most of the cases [13].

2) Temperature:

An increase or decrease in temperature causes the rise or fall in rainfall, humidity, water flow, and water level. So temperature has an important impact on flood and can be one of the most important factors for the prediction of the flood. There are various ways to gather information about the temperature of an area at a particular time interval. We collected the data from an official government website.

3) Dew point:

Dew point, the temperature at which water vapor in the air naturally condenses into liquid water. If the temperature and the dew point are the same, dew forms on surfaces such as grass, trees, and cars. Dew point indicates the moisture content present in the air. If moisture is abundant in the air, the atmosphere will not have to cool off as much for that moisture to begin condensing. Dew point translates into the amount of humidity in the air, which further translates into the amount of rainfall, which affects the flood, so the dew point is also considered as an essential factor.

4) Humidity:

It is the amount of water vapor present in the air, and if the content is high, humidity increases, thereby clogging the air with water vapor, thus leaving less room for addition. High humidity is also associated with rainfall and further causes flooding. It is also blamed for all sorts of negative things, which includes rainfall, flood, hurricane. High humidity increases the water retention capacity in the soil, which adds to damages caused by the flood.

5) Wind Speed:

The wind speed, including the direction of the wind, sets the path for monsoon and rainfall. The wind speed tells how early the moist air will bring rainfall with it and also floods. It is essential to keep track of wind speed during rainfall or flood forecasting. Mostly the hurricanes and storms lead to high wind speed bringing along with-it moist air, rainfall and hence flooding occurs. For our research, we used the government official's website to collect the wind speed information before and during the flooding.

6) Pressure:

The high and low pressures determine the direction of the wind flow. A low-pressure system sucks up the moist air, which sets the direction of the wind. A large low-pressure system creates a big pocket that sucks the moisture from the high-pressure area and traps the moist air in the pocket. The trapped air has no way to escape and causes the accumulation of the moist air, which leads to the massive rainfall and hence high chance of flooding.

7) Ground Water:

Groundwater is the water retained beneath the surface of Earth in the soil pore spaces. There is a phenomenon named groundwater flooding, which occurs when the underground natural drainage system is not able to drain the rainwater, causing the water table to rise above the surface. This uprising can pose a significant threat to many rural communities in the form of flood hazards. This process is much slower than river flooding in terms of occurrence since groundwater flooding usually happens after a prolonged rainfall, which may last weeks or even months. So tracking groundwater is a crucial task since it can prove to be a silent killer.

E) Design and Methodology:

Data Source

The satellite data have been captured through Copernicus Open Access Hub. The satellite data was captured for different districts of Kerala and Texas, explicitly capturing the before and after flood events. For the said date (before flooding), the weather data was also captured through Weather Underground API. The data required for the work is collected from the USGS Earth explorer [14]. Also, the groundwater quantity for the month of the before flood event was captured through each of the district's government databases. For each of the satellite data for each of the districts, the subdivision was carried out to enhance the data availability and improve the precision.

Technology Stack

To work with the satellite form and convert it into a more consumable format, SNAP (Sentinel Application Platform) was utilized. From the same, the relevant NDWI indexed satellite image was captured. The programming language's primary choice for the research work has been Python, and Jupyter Notebook has been utilized to run all the relevant functions.

Methodology

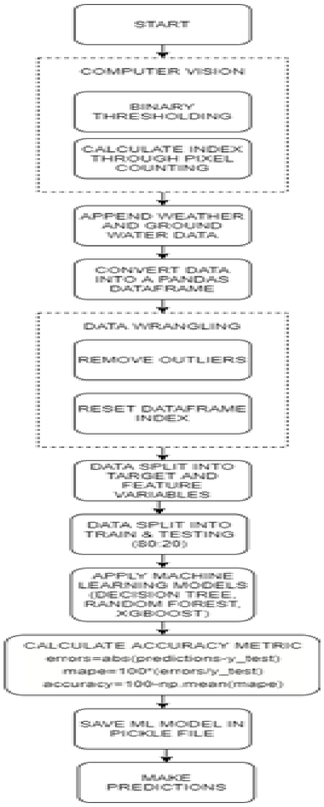

START
COMPUTER VISION
BINARY THRESHOLDING
CALCULATE INDEX THROUGH PIXEL COUNTING
APPEND WEATHER AND GROUND WATER DATA
CONVERT DATA INTO A PANDAS DATAFRAME
DATA WRANGLING
REMOVE OUTLIERS
RESET DATAFRAME INDEX
DATA SPLIT INTO TARGET AND FEATURE VARIABLES
DATA SPLIT INTO TRAIN & TESTING (80:20)
APPLY MACHINE LEARNING MODELS (DECISION TREE, RANDOM FOREST, XGBOOST)
CALCULATE ACCURACY METRIC
errors=abs(predictions-y_test)
mape=100*(errors/y_test)
accuracy=100-np.mean(mape)
SAVE ML MODEL IN PICKLE FILE
MAKE PREDICTIONS

Fig i: Flowchart

IV. Result Analysis

The research project outlines an important method that can find usefulness in both government and the private sectors. To understand it more in-depth, the complete flow and its result have been discussed. To start with, binary thresholding is done on the obtained satellite data to create a clear differentiation between the land and the water body.

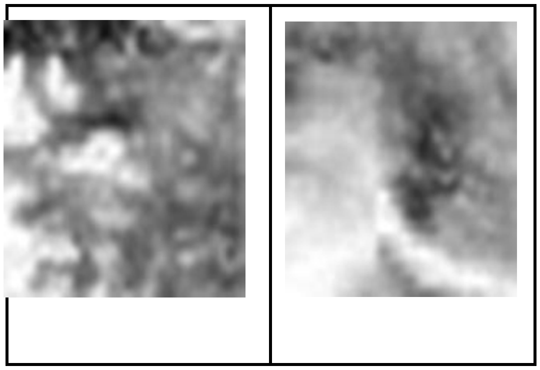


Fig ii: Before Binary Thresholding

Figure (ii) represents original satellite data before applying binary thresholding. Binary thresholding is required to extract information about land and water area by separating the land and water by different colored pixels.

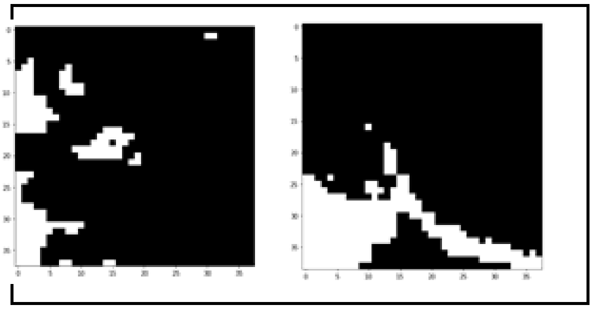
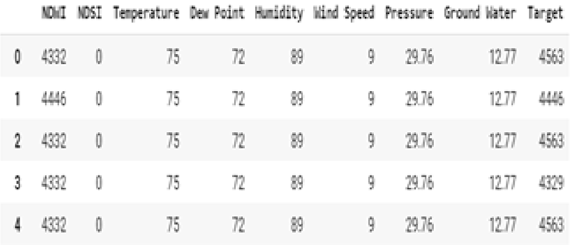


Figure iii: After Binary Thresholding

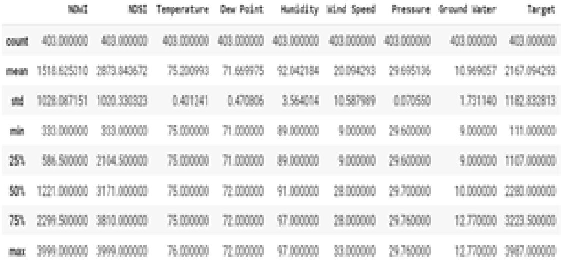
The black pixels in fig(iii) denote a soil area, whereas the white pixels denote a water body. The summation of all the pixels is calculated to obtain the respective normalized difference index. To enhance is further, a gaussian filter and other morphological filters can be applied to make it more transparent and distinctive.



	NDWI	NDSI	Temperature	Dew Point	Humidity	Wind Speed	Pressure	Ground Water	Target
0	4332	0	75	72	89	9	29.76	12.77	4563
1	4446	0	75	72	89	9	29.76	12.77	4446
2	4332	0	75	72	89	9	29.76	12.77	4563
3	4332	0	75	72	89	9	29.76	12.77	4329
4	4332	0	75	72	89	9	29.76	12.77	4563

Fig iv: Dataset

After obtaining the data from the satellite image and the district data, a data description can be created as shown in fig(iv), which will help the data wrangling task remove the outliers. It also gives a significant amount of information about the dataset being utilized and can be used to evaluate the quality of the data as a whole. It can also be used to find out if there are missing parameters.



	NDWI	NDSI	Temperature	Dew Point	Humidity	Wind Speed	Pressure	Ground Water	Target
count	403.000000	403.000000	403.000000	403.000000	403.000000	403.000000	403.000000	403.000000	403.000000
mean	1518.625310	2873.843672	75.200993	71.669975	92.042184	20.094293	29.695136	10.969057	2167.094293
std	1028.087151	1020.330323	0.401241	0.470806	3.564014	10.587989	0.070650	1.731140	1182.832813
min	333.000000	333.000000	75.000000	71.000000	89.000000	9.000000	29.600000	9.000000	111.000000
25%	586.500000	2104.500000	75.000000	71.000000	89.000000	9.000000	29.600000	9.000000	1107.000000
50%	1221.000000	3171.000000	75.000000	72.000000	91.000000	28.000000	29.700000	10.000000	2280.000000
75%	2299.500000	3810.000000	75.000000	72.000000	97.000000	28.000000	29.760000	12.770000	3223.500000
max	3999.000000	3999.000000	76.000000	72.000000	97.000000	33.000000	29.760000	12.770000	3987.000000

Fig v: Data Description

The heatmap of the entire feature column and the target column can be created to get a statistical idea of the correlation of each column with each column. It can also be used to add to the understanding of the factors that contribute the most to flooding. In fig(vi), a two-dimensional graphical representation of knowledge is presented where the individual values are contained during a matrix and are represented as colors. The seaborn python package allows the formation of annotated heatmaps, which may be altered using Matplotlib tools as per the need.

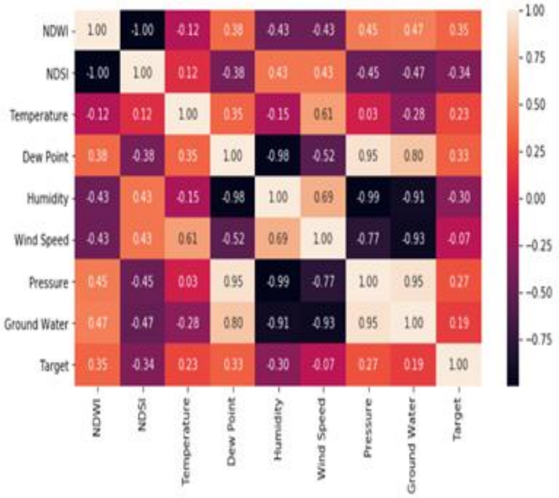


Fig vi: Correlation Heatmap Matrix

Fig (vii) represents a pair plot. It helps us to gain a more in-depth understanding of the correlation of each column with one another. It gives us an insight into the nature of the correlation, and that can be very helpful and choosing a Machine Learning model. It can also be an excellent tool to find outliers and find statistical patterns in the dataset. By default, the function will construct a grid of axes such that each numeric variable in the data will be shared in x-axis across a single column and in the y-axis across a single row. The diagonal axes will be treated differently, and the function will draw a plot to show the univariate distribution of the data for the corresponding variable in that column.

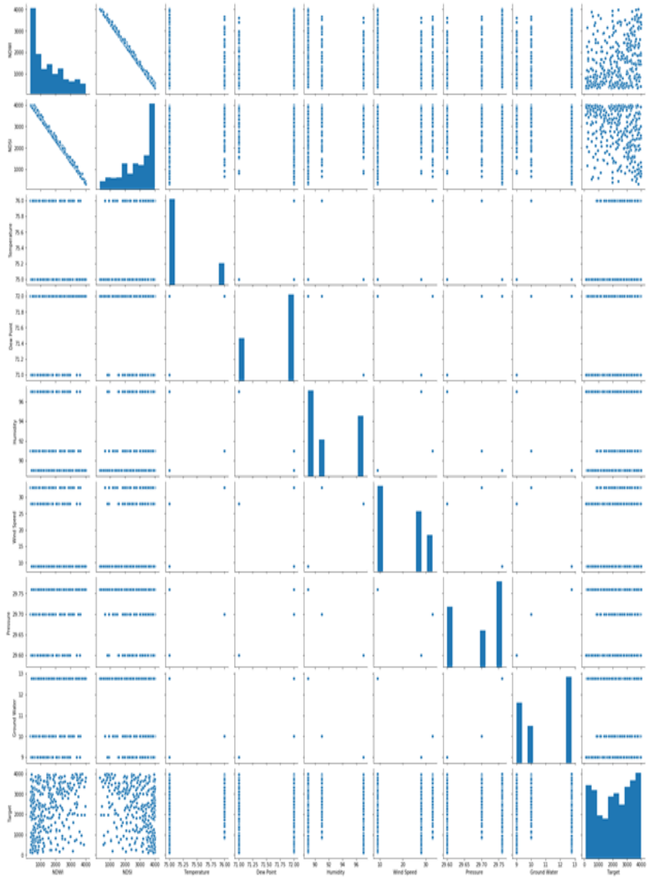


Fig vii: Pair Plot

We have chosen the following Machine Learning Model – Decision Tree, Random Forest, and XG Boost. The said machine learning models follow a very similar tree architecture design. The decision was made after observing the graphs that have been plotted previously. It was apparent that there is not a linear relationship between the feature variables and the target variables. The data's underlying structure is usually best learned through tree-based Machine Learning models and rightly so as portrayed by the accuracy metric achieved. The accuracy was calculated by using the following formula:

```
errors = abs(predictions - y_test)

mape = 100 * (errors / y_test)

accuracy = 100 - np.mean(mape)

print('Accuracy:', round(accuracy, 2), '%.')
```

Model	Decision Tree	Random Forest	XGBoost
Accuracy	54.21%	66.35%	83.52%

Table. i

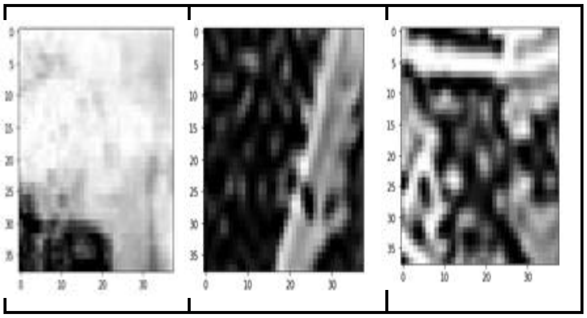


Fig viii. Satellite image

Current Water Area: 2253 Actual Future Water Area: 1203 Predicted Water Area: 1134.0	Current Water Area: 39 Actual Future Water Area: 1539 Predicted Water Area: 1846.0	Current Water Area: 540 Actual Future Water Area: 549 Predicted Water Area: 478.0

Table. ii

The user gets an option of 3 algorithms, namely XG boost, Decision Tree, and Random Forest, amongst which one of the algorithms can be chosen. Each algorithm provides variant accuracies based on the data provided. Therefore, the user can compare and choose as to which algorithm fits best for his data.

Finally, the Machine Learning model's core has been saved in a pickle file for later use, especially on the consumer-facing applications.

V. Conclusions and Future Scope

A) Summary

After considering the massive amounts of destruction caused by the effects of the flood, it becomes necessary to take immediate action in building a prediction model which will help us avoid the devastating effects and which will help the government to take necessary actions to prevent the loss of life as well as properties and money. ML is the technique used for the classification and mapping of the flood-prone area. The dataset required is collected from the satellites, which gave us the flexibility to apply different factors affecting and causing the flood. Using this model, Activities within areas identified as flood-prone can be managed to minimize flood damage to existing infrastructure. The measures are undertaken to manage activities that are grouped into two; structural and non-structural measures.

B) Conclusion

The hazardous events need to be solved by the most efficient and modern solution. Computers being more precise and accurate than humans and the Machine Learning algorithms boosting giving more accurate predictions than human intuitions, this model is undoubtedly saving much time and being more efficient than the human brain. The destruction caused by floods can be minimized and be predicted in advance by the appropriate use of these technologies. The model built is based on Machine Learning, which will help to predict flood and also show the flood-prone area with higher accuracy.

C) Future Scope

The model can further be improved by training it with more datasets that are currently unavailable to us as the data is confidential due to the privacy guidelines of the government. This model can be publicized to more users. This will smooth the process of evacuation. The government can make use of this model and save millions of dollars spent on flood relief. The population and structures within areas delineated as flood-prone are checked out during a vulnerability analysis. During the vulnerability analysis, the potential costs of flooding are evaluated as about damage to critical infrastructures like utilities, bridges, roads, buildings, and crops. Because vulnerability analysis detects the population at the absolute risk, it can also be used to determine the emergency responses that may be essential such as temporary shelters and evacuation parameters.

VI. REFERENCES

[1] Klein, T.; Nilsson, M.; Persson, A.; Håkansson, B. From Open Data to Open Analyses—New Opportunities for Environmental Applications? Environments **2017**, 4, 32.

[2] Wulder, M.A.; Masek, J.G.; Cohen, W.B.; Loveland, T.R.; Woodcock, C.E. Opening the archive: How free data has enabled the science and monitoring promise of Landsat. Remote Sens. Environ. **2012**, 122, 2–10.

[3] Turner,W.; Rondinini, C.; Pettorelli, N.; Mora, B.; Leidner, A.K.; Szantoi, Z.; Buchanan, G.; Dech, S.; Dwyer, J.; Herold, M. Free and open-access satellite data are key to biodiversity conservation. Biol. Conserv. **2015**, 182, 173–176.

[4] Li, S.; Dragicevic, S.; Castro, F.A.; Sester, M.; Winter, S.; Coltekin, A.; Pettit, C.; Jiang, B.; Haworth, J.; Stein, A. Geospatial big data handling theory and methods: A review and research challenges. ISPRS J. Photogramm. Remote Sens. **2016**, 115, 119–133.

[5] Berger, M.; Moreno, J.; Johannessen, J.A.; Levelt, P.F.; Hanssen, R.F. ESA's sentinel missions in support of Earth system science. Remote Sens. Environ. **2012**, 120, 84–90.

[6] Hansen, M.C.; Loveland, T.R. A review of large area monitoring of land cover change using Landsat data. Remote Sens. Environ. **2012**, 122, 66–74.

[7] Justice, C.O.; Vermote, E.; Townshend, J.R.; Defries, R.; Roy, D.P.; Hall, D.K.; Salomonson, V.V.; Privette, J.L.; Riggs, G.; Strahler, A. The Moderate Resolution Imaging Spectroradiometer (MODIS): Land remote sensing for global change research. IEEE Trans. Geosci. Remote Sens. **1998**, 36, 1228–1249.

[8] Copernicus Emergency Management Service. Available online: http://emergency.copernicus.eu/mapping/list-of-components/EMSR120

[9] Global Floods Detection System. Available online: http://www.gdacs.org/flooddetection/overview.aspx

[10]. Basha E.A., Ravela S., Rus D. Model-based monitoring for early warning flood detection; Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems; Raleigh, NC, USA. 5–7 November 2008. [Google Scholar]

11] Yu J., Hahn H. Remote detection and monitoring of a water level using narrow band channel. J. Inf. Sci. Eng. 2010;26:71–82. [Google Scholar]

[12] Sentinel Application Platform (SNAP). Available online: https://step.esa.int/main/toolboxes/snap/

[13] Qiao, C., Luo, J., Sheng, Y. et al. An Adaptive Water Extraction Method from Remote Sensing Image Based on NDWI. J Indian Soc Remote Sens **40,** 421–433 (2012). https://doi.org/10.1007/s12524-011-0162-7

[14] USGS Earth explorer. Available online: https://earthexplorer.usgs.gov/

[15] Ban, H.-J., Kwon, Y.-J., Shin, H., Ryu, H.-S., & Hong, S. (2017). Flood Monitoring Using Satellite-Based RGB Composite Imagery and Refractive Index Retrieval in Visible and Near-Infrared Bands. Remote Sensing, 9(4), 313. doi:10.3390/rs9040313

[16] Notti, D., Giordan, D., Caló, F., Pepe, A., Zucca, F., & Galve, J. (2018). Potential and Limitations of Open Satellite Data for Flood Mapping. Remote Sensing, 10(11), 1673. doi:10.3390/rs10111673

[17] Tehrany, M. S., Pradhan, B., Mansor, S., & Ahmad, N. (2015). Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. CATENA, 125, 91–101. doi:10.1016/j.catena.2014.10.017