International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 10 | Oct 2022 www.irjet.net p-ISSN: 2395-0072

URBAN FLOOD SUSCEPTIBILITY MAP OF CHENNAI - GIS AND RANDOM FOREST METHOD

Deepak K R[1] Sabu P[2] Muhammed Yousuf S[1] Cyril George Thottumkal[3]

[1] *Students (M Tech), Department of Civil engineering, College of Engineering Trivandrum, Kerala, India.*
[2] *Assistant Professor, Department of Civil Engineering, College of Engineering, Trivandrum, Kerala, India.*
[3] *Research Associate, International Center for Technological Innovations, Alleppey, Kerala, India.*

Abstract: Natural disasters like floods are causing massive damage to natural and human resources, especially in coastal areas. Hence, identifying the susceptible areas to flood is an important task for every country to prevent such dangerous consequences. GIS and remote sensing techniques have provided many ways to control and manage floods like flood forecasting, providing spatial information etc. Concerning social, economic, and environmental perspectives, flood is one of the most devastating disasters in Chennai corporation for the past many years. Urban Flood susceptibility mapping is done using Random Forest (RF) model. Urban Flood Susceptibility map was generated using this model by considering eleven different independent variables (land use/land cover, elevation, slope, aspect, NDVI, NDBI, Topographic Wetness Index, Stream Power Index, rainfall, drainage density, distance from river) and past flood and non-flood locations. In total, 300 historic flood locations and 300 non-flood locations were collected from past flood episodes data. The independent variables and historic flood and non-flood locations were combined to generate the database for flood susceptibility mapping. From this, 85% of the data is taken for training the model and 15% of the data set is used for testing and validation. The Random Forest model showed a prediction accuracy of 95.5%. The output obtained from the Random Forest model was used to map the urban flood susceptibility for Chennai Corporation, producing the best outcomes and has classified the regions into low and high susceptible zones. Hence, the RF model is well suitable for predicting the urban flood susceptibility zone. The findings of this analysis will assist decision-makers in carrying out effective flood management in the future.

1. INTRODUCTION

Most individuals worldwide are exposed to several types of natural disasters. One of the most important natural disasters on the globe among them is flooding. Waterlogging and floods are caused by sudden, extensive, and persistent rainfall. It is known that 13% of Asia's population lives in low elevation coastal areas that are severely exposed to weather events like floods and 140 million people are affected by a flood event. This is especially true in the region of South India due to the depression over the southwest Bay of Bengal and owing to a strong El Nino. South India's Chennai is a significant coastal low-lying metropolis. The previous floods in Chennai in 1976, 1985, 1996, 1998, 2005, 2008, 2010, 2015, and 2021 resulted in massive damages to property, infrastructures, and human health. Major rivers and drainage management systems failing are the causes of these flood occurrences. Control and preventive measures should be taken for managing the damages caused by floods to agriculture, infrastructure and other natural resources (Hamid et al., 2007). Therefore, the flood susceptibility study is significant for early warning systems and mitigation of upcoming flood scenarios. The application of machine learning (ML) techniques to determine a natural hazard's vulnerability has advanced quickly in recent years. The following three sections can be used to describe the general research goals of this type of study. They are the building of a spatial database, a framework for assessing hazards' susceptibility based on ML and variables connected to hazards, and the acquisition of a hazard susceptibility map. The employment of benchmark ML techniques with great computing efficiency and very basic structures, such as Decision Tree, Support Vector Machine, etc. These studies now choose to assess the susceptibility to natural hazards utilising state-of-the-art ML techniques with more complicated structures such as adaptive neuro-fuzzy inference systems, convolutional neural networks, and recurrent neural networks. (Zaholi et al., 2015).

Fig. 1. Consequences of flooding

Machine learning can successfully circumvent the data scarcity that many prediction models encounter, it has become widely used in the field of assessing the susceptibility to natural hazards. However, there are still existing technical issues that would render effects that are not favourable to the susceptibility assessment results. The main aim of this study is to identify and map out flood susceptible zones in the Chennai.

In this study, Chennai corporation is taken as the study area. It is one of the major metro cities in India. The instantaneous heavy, widespread and continuous rainfall leads to water logging and floods. In particular, in the region of South India due to the depression over the southwest Bay of Bengal and owing to a strong El Nino. Chennai is an important coastal low-lying city in south India. The past flood events of Chennai in 1976, 1985, 1996, 1998, 2005, 2008, 2010, 2015, and 2021 caused several damages to property, human health and infrastructure and many more. The reasons for these flood events are the failure of major rivers and the drainage management system. To controlling of floods and suggest preventive measures are necessary to reduce the probable damages to agriculture, infrastructure, and other natural resources. Therefore, the flood susceptibility study is significant for the early warning system and mitigation of upcoming flood episodes in Chennai corporation.

The objective of this study is to generate an urban flood susceptibility map of Chennai corporation using the Random Forest method. Floods are the most frequent naturally occurring disasters and can cause great damage to human life and property. The factors affecting floods include factors like rainfall and cyclone frequency in that area and hazard inducing environmental factors such as elevation, land use factors etc. The disaster inducing factors are less predictable but geographic analysis can be used to predict the risk distribution of hazard inducing environmental factors. Thus, the susceptibility to a flood event and the impact of the same are different for different regions based on these parameters. By developing a systematic procedure for flood susceptibility assessment and creating the model for the same, the flood risk of any geographical area can be computed by following this methodology. The input parameters and environmental constraints in which the model is created must be maintained. And by using this model, the study is limited to detecting the urban flood susceptibility zones using Random Forest method.

2. MATERIALS AND METHODS

The satellite data, rainfall data, Digital Elevation Model, Flood inventory map, and Study area map collected for this study are discussed below in detail.

Satellite Data:

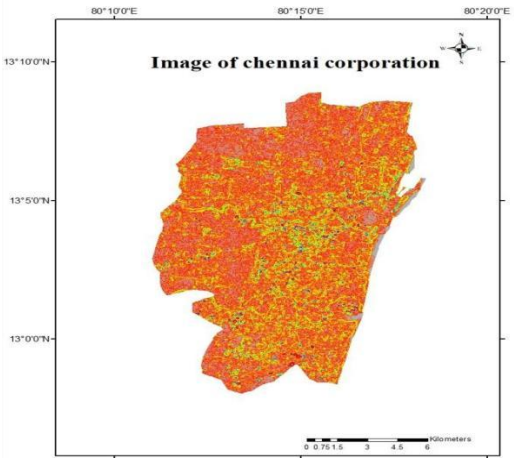

Fig. 2. Satellite image of Chennai corporation

The satellite data of the study area dated 12/01/2022 was downloaded from USGS earth explorer. Landsat OLI/TIRS image was downloaded. It consists of 11 bands and the spatial resolution for the bands is 30m for multispectral, 15m for panchromatic and 100m for Thermal Infra Red bands and radiometric resolution is 16 bits.

Study Area Map: Chennai Corporation is located in the north-eastern part of Tamil Nadu on the south-eastern coast of India and is confined with a latitude of $12^0$50' to $13^0$15'N and longitude of 80°50' to 80°20' E. It encloses an area of 176 sq. km and accounts for a high population density of 17,000 per square kilometer. The region was occupied by recently constructed urban infrastructure to the tune of 60%. Lithologically, the inland region is made up of river sediments and charnockite, whereas the coastline section is made up of marine sediments. The research area gets cyclonic and depression-related severe rainfall that regularly causes floods. The map of the study area is digitalised from the online map in Arc-GIS and this shape file is then used to clip the satellite image.

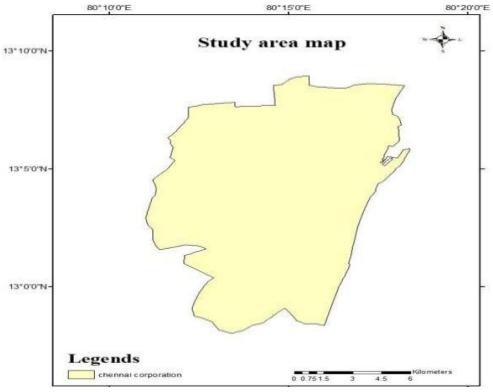

Fig. 3. Boundary map of the study area

Flood Inventory Map: The flood inventory map is the historical database of flood-prone regions, compiled from previous documents. This sort of map is very helpful in forecasting possible floods. Since its precision depends considerably on the temporal and spatial scale of flood reporting. For this analysis, the flood inventory map was constructed using 300 flood points and 300 non-flood points in the region, derived from flood statistics data of 2015 & 2021 Which is obtained from Bhuvan. By assigning a value of 0 to the non-flood sites and a value of 1 to flood sites, it can generate binary training & testing datasets. Randomly 85% of the data collection was chosen to operate the model and the remaining 15% of the data was used for model validation.

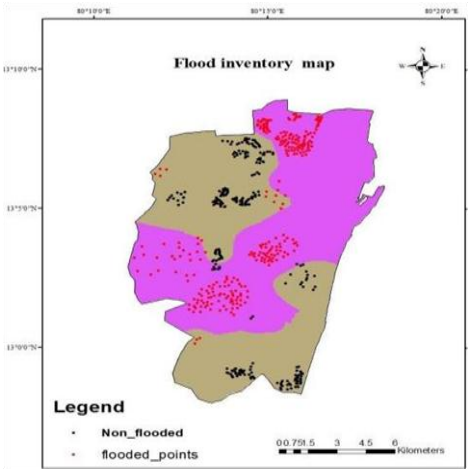

Fig. 4. Flood inventory map

Digital Elevation Model: Topographic factors are important for modelling flood studies, which will directly and indirectly influence the hydrological characteristics of the study area. At first, a Digital Elevation Model (DEM) was prepared from the JAXA DEM with 30m spatial resolution for the study area in the ArcGIS environment. The topographic factors are derived from DEM, such as slope, elevation, aspect, topographic wetness index (TWI), and stream power index (SPI) in ArcGIS.

Rainfall data:

Table 1. Details of data collected

Sl no	Data	Details	Agency
1	Landsat8 (OLI/TIRS)	*Spatial resolution* Band (1to7) = 30m Band8 = 15m Band9 = 30m Band10,11 = 100m	NASA (USGS)
		Spectral resolution Band 1– (0.43 –0.45) Band2– (0.45–0.51) Band3– (0.53–0.59) Band 4– (0.64–0.67) Band 5– (0.85–0.88) Band6– (1.571.65) Band7– (2.11–2.29) Band8– (0.5– 0.6) Band9– (1.36–1.38) Band10 – (10.6–11.9) Band11– (11.5-12.51) *Radiometric resolution* 16bits	
2	Study area map	Chennai corporation	ArcGIS online map
3	Digital Elevation Model	Spatial resolution=30m	JAXA
4	Flood inventory map	Past flood data	Bhuvan
5	Rainfall data	Average annual rainfall	IMD

Rainfall is considered to be one of the most influencing factors. Heavy rainfall is responsible for increasing the underground hydrostatic level and water pressure. In addition, heavy rainfall from the upstream point in a shorter period generally has a high potential for maximum flood events. Areas having high average annual rainfall are assumed to have a high risk of flooding. For this study, Rainfall data for a period of 30 years (1992 – 2021) of the study area from five rain gauge stations have been collected from Indian Meteorological Department (IMD). A rain gauge (also known as udometer, pluviometer, barometer, and hyetometer) is an instrument used by meteorologists and hydrologists to gather and measure the amount of liquid precipitation over an area in a predefined area, over a period of time. It is used for determining the depth of precipitation (usually in mm) that occurs over a unit area and thus measuring rainfall amount.

Software: Microsoft Excel, Jupyter Notebook, and ArcGIS were the applications used in this investigation. A geographic information system (GIS) for using maps and geographic data is called ArcGIS. It is utilised for a variety of tasks, including the creation and usage of maps, the gathering and analysis of geographic data, the sharing and discovery of geographic information, the use of maps and geographic data in a variety of applications, and the management of geographic data in databases. RF machine learning model is applied in Jupyter notebooks, and the output from this model is utilised to create maps showing how susceptible cities are to flooding.

3. METHODOLOGY

The study area experiences heavy rainfall associated with depressions and cyclones which lead to frequently occurring floods. In the present study, flood susceptibility analysis was performed using GIS along with machine learning using Random Forest models. Flood is a dynamic natural hazard that is caused by many natural and anthropogenic impacts. Thus, it is very important to consider the application of machine learning techniques to provide a complete overview of the flood susceptibility of a study area. The description of the theoretical background of each model and also the manner in which this model was applied to evaluate flood susceptibility is presented below. The different data required for this study have been collected and several flood conditioning factors are also considered for the flood susceptibility mapping of the study area. The detailed methodology of the study is shown below.

3.1 Flood susceptibility factors:

The selection of the urban flood conditioning factors varies from one study area to another. Eleven flood susceptibility factors for the flood susceptibility modelling in the area were chosen. These are elevation, slope, rainfall, aspect, NDVI, NDBI, LULC, drainage density, river proximity, Topographic Wetness Index, and Stream Power Index. These variables are essential in determining and delineating flood-prone regions. These factors were then transformed into the raster format.

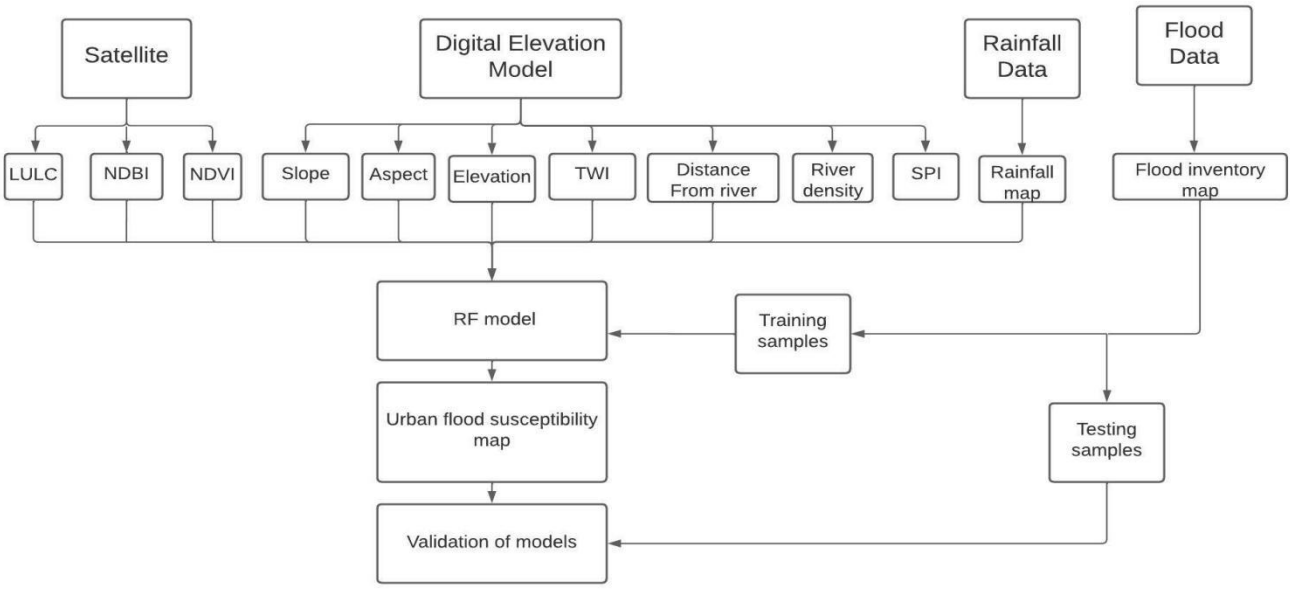

Fig. 5. Flowchart of methodology

Elevation: For determining flood susceptibility of a region, elevation was identified as one of the significant factors. In comparison, lower altitudes are more commonly linked with areas' susceptibility to floods. The susceptibility and severity of floods are inversely linked to the height. It is therefore suggested that the areas that are strongly linked with very low elevation can undergo more extreme flooding than areas at a higher elevation. For this analysis, an elevation map with a spatial resolution of 30 m was prepared from JAXA DEM. The elevation values of the Chennai corporation range from -40m to 64m.

Aspect: Aspect is significant in predicting flood susceptibility of the region. This is the highest direction down the slope. For this analysis, the aspect map was built using Arc GIS tools from the Digital Elevation Model.

Aspect (in degrees)	Label
-1-0	Flat
0–22.5	North
22.5–67.5	North East
67.5- 112	East
112–157.5	South East
157.5–202.5	South
202.5–247.5	South West
247.5–292.5	West
292.5–337.5	North West
337.5–359.9	North

Table 2. Aspect classes

River proximity: Distance from River is a significant parameter for assessing flood susceptibility. Stream flow increases because of heavy and stormy runoff in a drainage system and as it reaches stream capacity limits, it will transform into a flood.

Therefore, closer to the river, the susceptibility to flooding becomes more and more distant from the river, and the susceptibility to flooding becomes less. Distance from river or river proximity map was prepared using the Euclidean distance tool in Arc GIS.

Drainage density: The drainage density is determined by the drainage length per unit area. Concentrated flow occurs during the runoff in a drainage system, and then Stream Power Index (SPI): SPI is defined as the erosive potential and surface runoff rate. The higher the SPI displays the higher surface runoff capacity and lower the SPI describes the lower surface runoff rate. Throughout this way, heavy rainfall in the lower SPI region will contribute to the flooding. Throughout this analysis, the following equation (1) was calculated using DEM in the Arc GIS.

$$SPI = As * \tan \beta \qquad (1)$$

Where As represents the specific catchment area in square meters and β represents the slope in degrees.

Topographic Wetness Index: Topographic Wetness Index (TWI) is a significant parameter used to estimate a region's susceptibility to flooding. TWI closely regulates spatial spreading and depletion of surface runoff. The Flow Direction Tool from Hydrology section of Spatial Analyst toolbox is used to create a flow direction. This assigns a value for each pixel based on the direction in which water flows along the slope in that pixel considered. This ultimately determines the destination of water flowing across the land. DEM is given as an input to the tool and a flow direction map is created. Similarly, in the Hydrology section, the Flow accumulation tool is used to create a flow accumulation map. The Flow Accumulation tool calculates the flow into each cell by identifying the upstream cells that flow into each downslope cell. In other words, each cell's flow accumulation value is determined by the number of upstream cells flowing into it based on landscape topography. The flow accumulation map is then reclassified. This flow accumulation layer is considered as the upstream contributing area. Now TWI is calculated using the following computational formula in the Raster calculator tool in the map algebra toolbox. The TWI was determined from the DEM in the GIS environment in this research work and is represented in the following equation (2).

where the channelized flow exceeds the channel potential and the surplus water discharges into the surrounding region which causes the region to flood. A drainage density map is prepared by using line density tool in ArcG

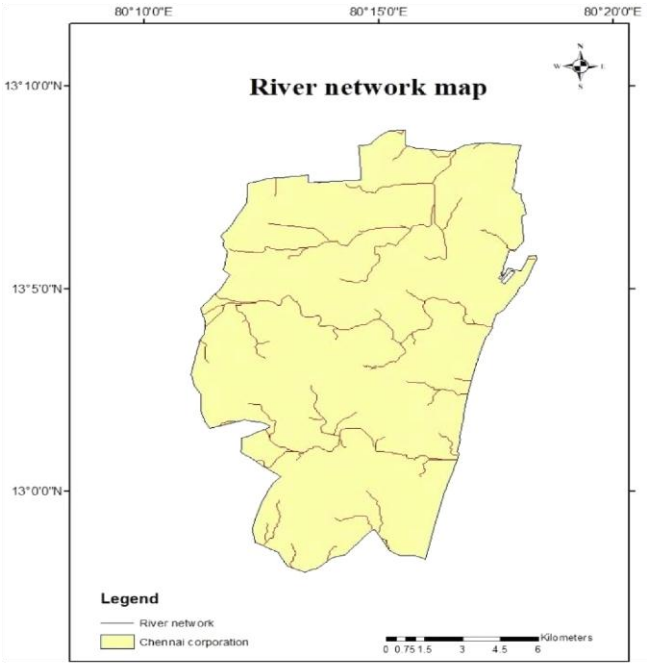

Fig. 6. River network map

$$TWI = \ln (As / \tan (\beta)) \qquad (2)$$

Where, As represents the specific catchment area, the slope is in degree. Many studies show that TWI and flood susceptibility are positively correlated to each other.

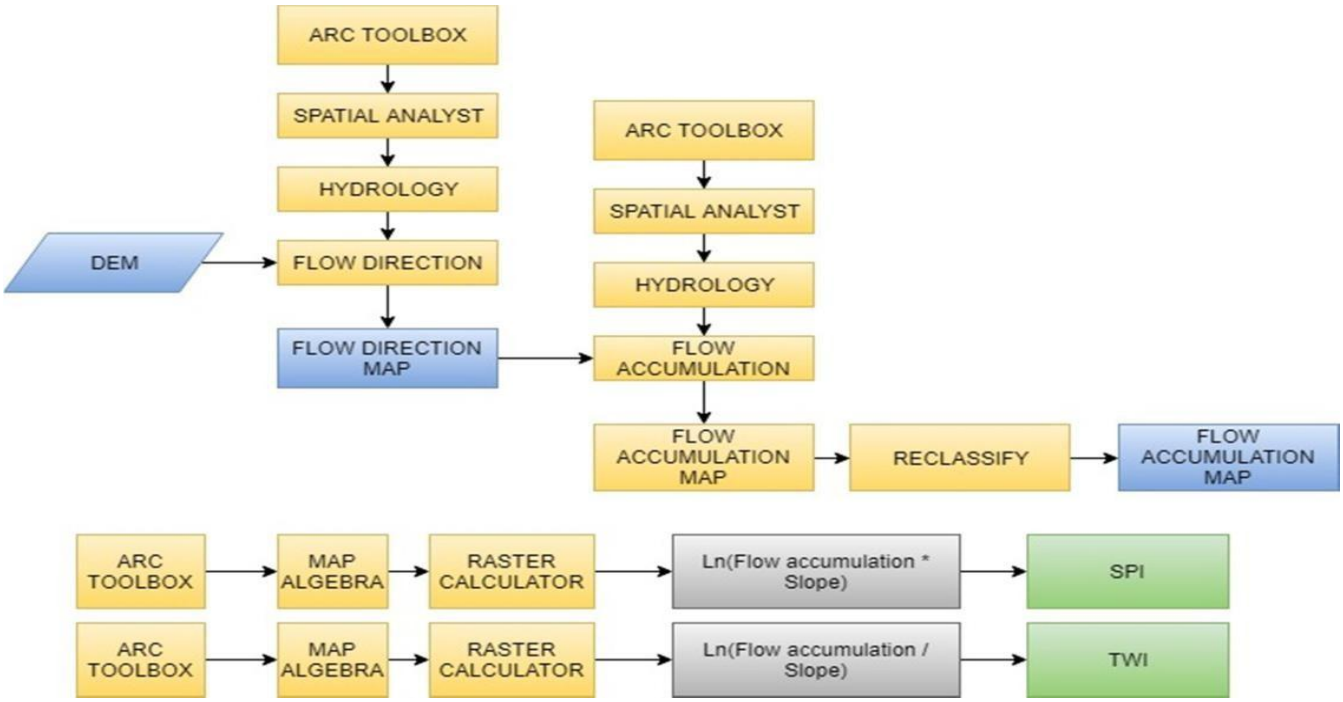

Fig. 7. Flowchart for creation of SPI& TWI

Land Use and Land Cover (LULC): In this study, LULC classification is done by the maximum likelihood supervised classification method in Arc GIS. In supervised classification, pixel classification is done and training samples corresponding to each class are taken by comparing with the google map image. Consideration was given to four types of land use and land cover in which 495 training pixel samples were taken for a waterbody, 672 for vegetation, 586 for the built-up area and 715 for barren land, and are reclassified and labelled as given below.

Class	Label	Pixel count
Waterbody	1	6785
Vegetation	2	73782
Built up areas	3	99546
Barren land	4	13477

Table 3. LULC Classification

Slope: The magnitude and intensity of water accumulation and water percolation are defined by the slope of the region. It indicates steepness at each cell of a raster surface. This adversely affects the storm too. The lower slope indicates Flat terrain and a higher slope indicates steep terrain. The inclination of the slope is calculated in degrees. In this analysis, the slope map in the Arc GIS framework was prepared from the 30 m JAXA DEM and using the slope tool in the Arc GIS toolbox.

Rainfall: One of the main determining elements of the flood is rainfall. The hydrostatic level and water pressure underground rise as a result of heavy rainfall. Areas with high yearly rainfall averages are thought to have a significant risk of flooding. The average annual rainfall of 5 rain gauge stations is determined to prepare the rainfall distribution map for this study. The data is then imported into ArcGIS as a point feature, and the spatial interpolation is carried out using an inverse distance weighted tool.

Normalized Difference Vegetation Index (NDVI): The NDVI is one of the extensively used factors for determining flood susceptibility. It indicates the vegetative cover of the region. Areas having lower or decreasing NDVI values (0.1 or below) denote non-vegetated features such as barren land, snow cover, sand, etc. Moderate values represent shrub and grassland (0.2 to 0.3) while high values indicate forests (0.6 to 0.8). NDVI was calculated by raster calculator in ArcGIS. The NDVI index for the study area is calculated using equation (3) using bands 4 and 5. The NDVI value normally ranges from -1 to +1.

$$NDVI = (NIR - RED) / (NIR + RED) \quad (3)$$

Normalized Difference Built-up Index (NDBI): The NDBI is used to analyse the built-up areas. It uses the SWIR and NIR bands to analyze the built up areas.

NDBI = (SWIR - NIR) / (SWIR + NIR) (4)
Its value generally ranges from -1 to +1.

3.2 Preparation of thematic maps:
The thematic maps of the eleven flood susceptibility factors such as slope, elevation, aspect, rainfall, Stream Power Index, Topographic Wetness Index, Land use Land cover, Distance to the river, Drainage density, NDBI, NDVI was prepared using the tools in Arc GIS. The values of the points of each factor corresponding to points in the flood inventory map were taken using the extract values to points tool in Arc GIS. These values were collected in excel format and have been given input to Random Forest Model.

4. RANDOM FOREST MODEL

RF, a highly effective combination of tree predictors, was systematically proposed by Breiman in 2001. Random Forest is a supervised learning algorithm. Random Forests are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. Random Forests grows many classification trees. Each tree is grown as follows:

1. Randomly select "k" features (parameters; here hazard inducing indices) from total "m" features. Where k < m

2. Among the "k" features, calculate the node "d" using the best split point.

3. Split the node into daughter nodes using the best split.

4. Repeat 1 to 3 steps until "l" number of nodes has been reached.

5. Build a forest by repeating steps 1 to 4 for "n" number times to create "n" number of trees.

Once the trees are created the model can be used in prediction of results. To perform prediction using the trained random forest, the algorithm uses the below pseudo code.

1. Takes the test features and uses the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)

2. Calculate the votes for each predicted target.

3. Consider the highly voted predicted target as the final prediction from the random forest algorithm.
This concept of voting is known as majority voting (Abdu et al., 2021). Thus, here if a pixel is analysed, the class (flooded and non flooded) which acquires the greater number of votes is assigned to that pixel.

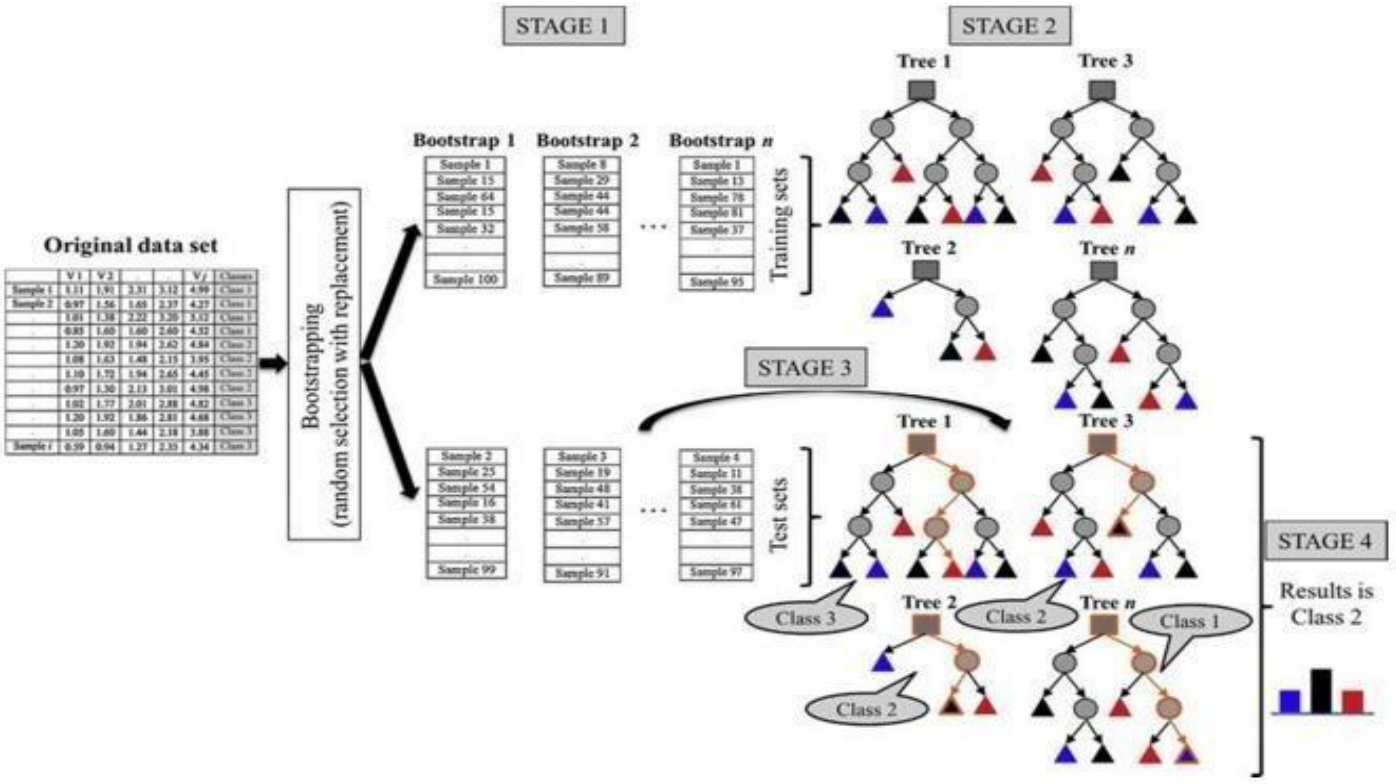

Fig. 8. Classification in Random Forest model

The algorithm is written in python language. The data from seven raster layers and the training data are provided to the algorithm. Additionally, the number of trees to be created and the depth up to which the pruning is to be done is also given to the programme. The depth of trees was kept constant and the number of trees was varied and accuracy was checked for each case to get an optimum number of trees for the model. Similarly, the optimum number of trees was taken constant and the depth was varied and accuracy was checked to get an optimum depth level (Gang et al., 2021). Once the training and testing samples are provided the accuracy between trained and tested values is checked. If the accuracy is satisfied the entire data set of whole study area is given as input. The model classifies each pixel to any of the three classes and the output is stored in .csv format. This output .csv file was added to ArcGIS software and was plotted as a raster layer. This gives the urban flood susceptibility map of the study area. To get a smooth and continuous raster, these points are interpolated using IDW interpolation tool in Spatial Analyst toolbox. Validation of the model is done by using confusion matrix.

5. RESULTS AND DISCUSSION

5.1 THEMATIC MAP: Thematic Maps were prepared using ArcGIS. Thematic Maps of NDVI, NDBI, Rainfall, Slope, Elevation, Aspect, Topographic Wetness Index, Stream Power Index, Land Use Land Cover, Drainage Density and River Proximity were shown below.

NDVI map: The NDVI map was prepared using the equation (3) and the map corresponding to the flood inventory map has also prepared.

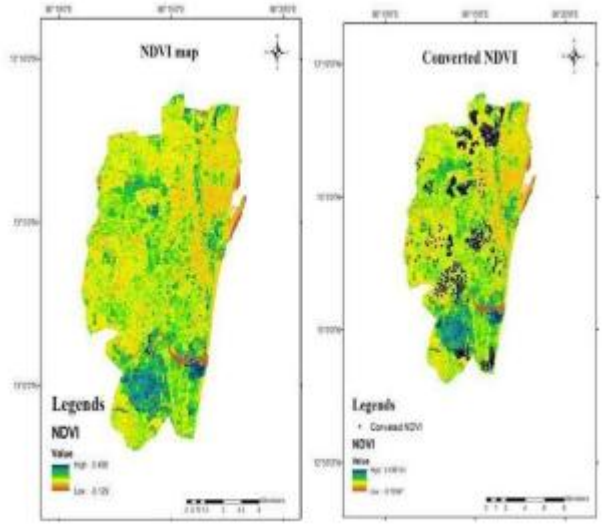

Fig. 9. NDVI map & Converted NDVI map

NDBI map: The Normalised Differential Built-up Index map was prepared using the equation (4).

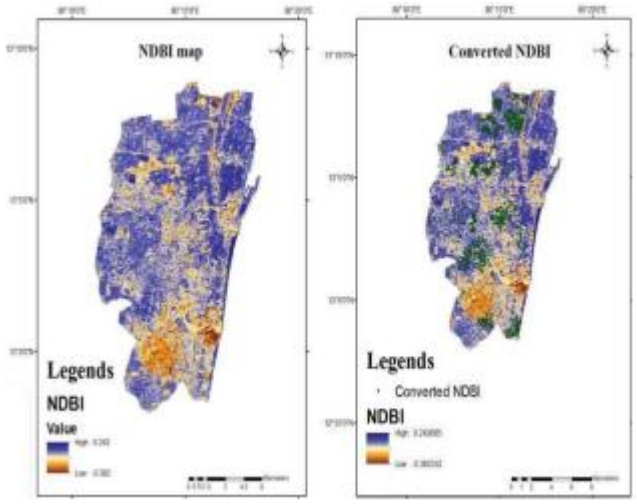

Fig. 10. (a) NDBI map (b) Converted NDBI map

Rainfall map: The Rainfall map was prepared using an inverse distance weighted tool in Arc GIS.

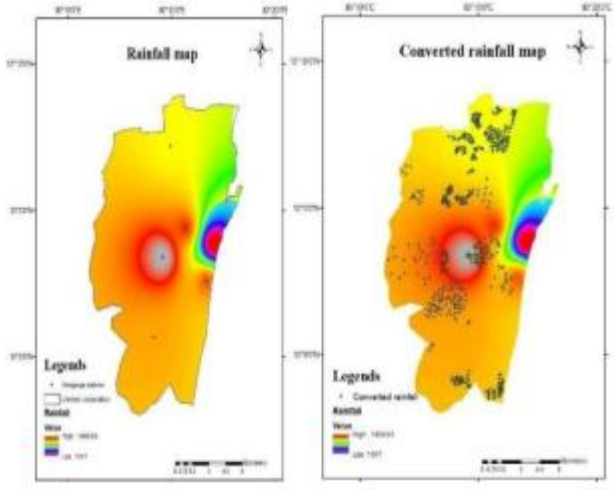

Fig. 11. (a) Rainfall map (b) Converted Rainfall map

Slope map: The Slope map was prepared using the inverse distance weighted tool in Arc GIS.

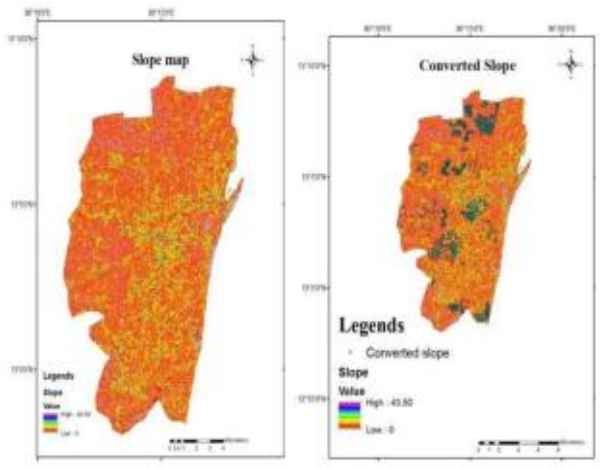

Fig. 12. (a)Slope map (b)Converted slope map

Elevation map: The elevation map is prepared from the Digital Elevation Model.

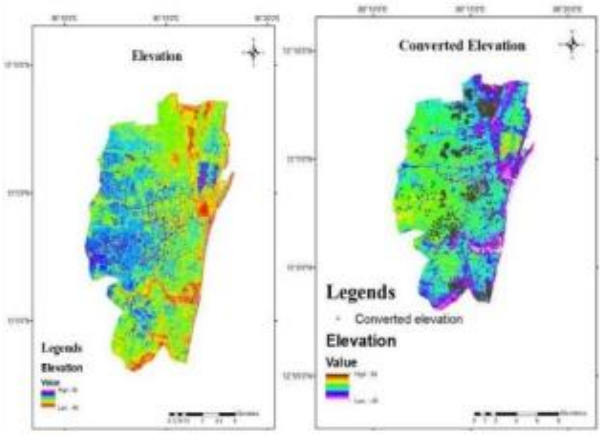

Fig. 13. (a) Elevation map (b) Converted elevation

Aspect map: The Aspect map was prepared in Arc GIS.

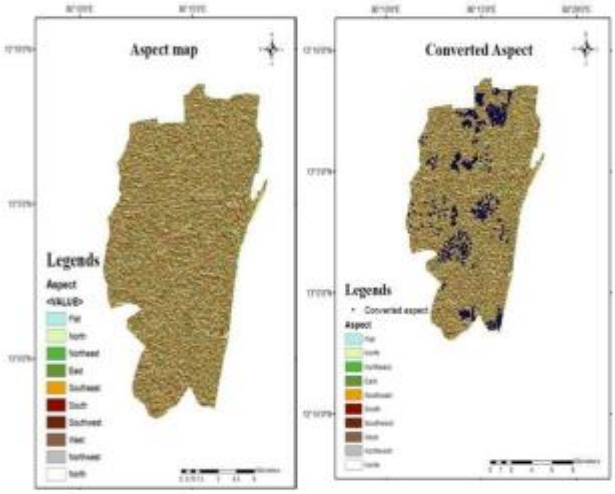

Fig.14. (a) Aspect map (b) Converted Aspect map

Topographic Wetness Index map: Topographic Wetness Index map was prepared using equation (2).

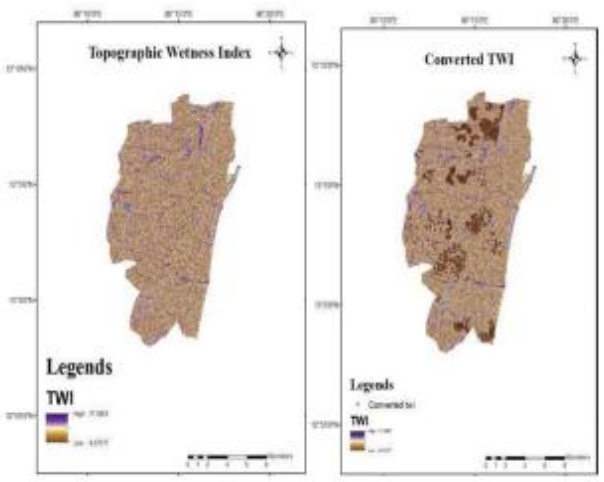

Fig. 15. (a) TWI map (b) Converted TWI map

Stream Power Index map: Stream Power Index can be calculated using the equation (1).

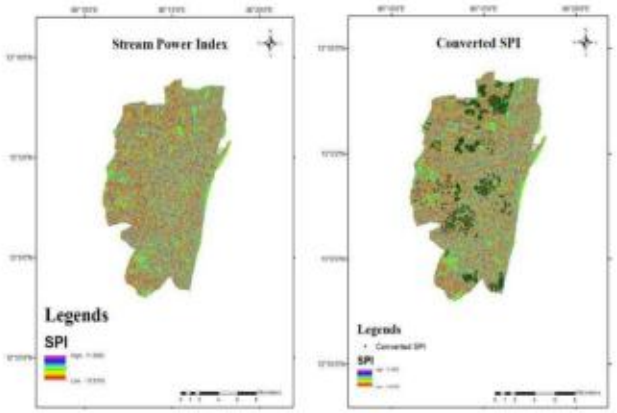

Fig. 16. (a) Stream Power Index map (b) Converted SPI map

LULC:

Land cover map: The LULC map was prepared using supervised classification in ArcGIS.

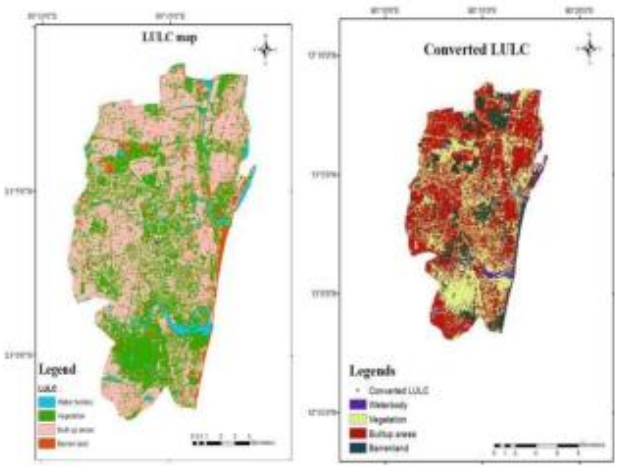

Fig.4.9 (a) LULC map (b) Converted LULC map

Drainage density map:

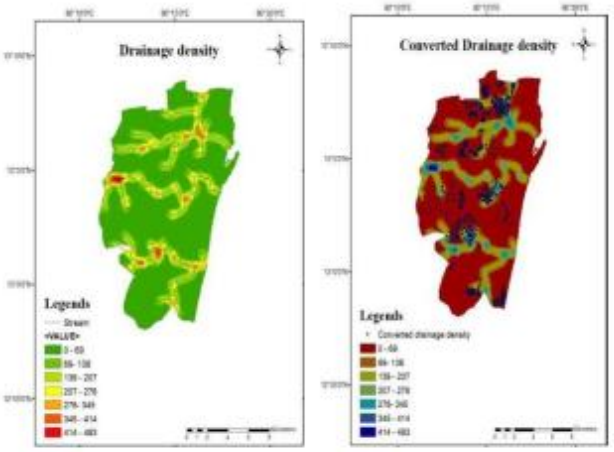

Fig. 18. (a) Drainage density map (b) Converted Drainage density map.

River proximity map:

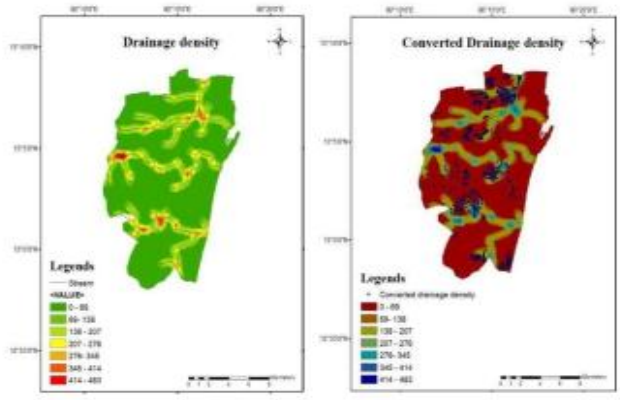

Fig. 19. (a) River proximity map (b) Converted river proximity map.

5.2 Analysis of RF Model:

From this analysis done on the study area, many results were obtained and are discussed here. The maps generated in Arc GIS were used to study the geological and hydrological factors and their spatial distribution over the Chennai corporation has been discussed in detail. Also, the prediction made by the Random Forest model and the final susceptibility map obtained is discussed.

Preparation and analysis of data sets: Different raster maps were created for all indices which are considered to be flood inducing and were analysed to obtain the following output map with spatial distribution and intensity variation across the Chennai corporation.

```
import numpy as np
import pandas as pd
import  matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics
import plotly.express as px
```

Fig. 20. Important libraries of RF model

The algorithm is written in python language. The above mentioned are the necessary libraries which are imported for the data analysis. Pandas and numpy are necessary for reading the dataset and matplotlib and seaborn are used for data visualisation and plotting and others are necessary machine learning libraries. The entire data set has been converted into excel format to give input to the Random Forest model. These data comprise of 600 datasets of flood conditioning factors and the Flooded (1) and non-flooded (0) points.

```
df = pd.read_excel('Final Dataset.xlsx')
```

```
df.head(10)
```

	OBJECTID	Latitude	longitude	Aspect	Elevation	LULC	NDBI	NDVI	Rainfall	Slope	SPI	River Proximity	Drainage Density	TWI	Indicator
0	1	13.116170	80.233959	147.4970	15	3	-0.008767	0.132600	1031.39	0.656959	-11.292500	0.010158	0.0	-3.72037	0
1	2	13.116712	80.235429	292.0550	14	3	-0.021392	0.145293	1028.23	0.000000	-11.908400	0.009070	0.0	-3.02722	0
2	3	13.115468	80.235646	240.9780	14	3	0.014757	0.093144	1029.69	0.734494	-11.189400	0.010454	0.0	-3.83194	0
3	4	13.114539	80.234459	180.7020	15	3	0.053367	0.105875	1032.41	1.313740	-3.038840	0.009425	0.0	4.46846	0
4	5	13.117398	80.234520	253.2760	12	3	-0.027218	0.169776	1029.23	0.734494	-2.056200	0.009056	0.0	8.10592	0
5	6	13.118121	80.236480	68.4652	13	3	0.000940	0.134071	1026.13	0.929039	-11.606300	0.007334	0.0	-3.37380	0
6	7	13.116342	80.237221	239.9680	13	3	0.027568	0.122004	1026.60	1.354160	-1.248370	0.009097	0.0	6.71824	0
7	8	13.117749	80.238321	331.4270	11	3	0.032463	0.093843	1024.26	3.346180	0.487033	0.007416	0.0	7.99468	0
8	9	13.116437	80.239557	120.7140	12	3	0.034302	0.067233	1023.51	1.393410	0.543532	0.008488	0.0	7.93579	0
9	10	13.115050	80.239771	219.4350	13	3	0.055260	0.092280	1024.82	1.354160	-2.604480	0.008430	0.0	4.84356	0

Fig. 21. Dataset for RF model

5.3 Training and Testing of RF model:

The entire data set has been labelled into X and Y in which X indicates the values of flood causing factors and Y indicates the values of flood and non flood points. And 85% of the entire data set is classified into training and 15% into the testing phase. The data set has been fed into the model for training and testing in which x train and x test indicate the flood causative factors and y train and y test indicate the indicator values.

```
X = df.drop(columns = ['Indicator', 'OBJECTID'])
y = df['Indicator']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y , test_size=0.15)
```

Fig. 22. Splitting of the dataset in the RF model

```
%time model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

Fig. 23. Training and Testing in RF model

After the data set has been splitted for training and testing. The model has been trained by giving x train as input and y train as output, that is the flood causative factors and the indicator values and during the testing phase the prediction is done by giving the input as x test variables only and indicator values are obtained as output by the mode pred. The obtained y_pred variables have been compared with the y_test values to check the accuracy of prediction.

```
test_set = X_test.reset_index(drop = True)[['Latitude','longitude']]
test_set['Predicted Value'] = y_pred
```

```
test_set.head()
```

	Latitude	longitude	Predicted Value
0	13.125320	80.264638	1
1	13.120527	80.264755	1
2	13.062452	80.256213	1
3	13.123852	80.231252	0
4	13.117159	80.248570	0

5.4 Preparation of Urban Flood Susceptibility map using RF model:

The final urban flood susceptibility map of Chennai corporation has been prepared using Random Forest machine learning model. The predicted output obtained from the Random Forest model is fed into the Arc GIS and has been interpolated using the Inverse Distance Weighted tool and is classified into two zones low (46.04%) and high susceptibility zones (53.96%).

Class	RF (sq. km)	RF (%)
Low	80.10	46.04
High	93.89	53.96

Table 4. Classification of the map from RF model

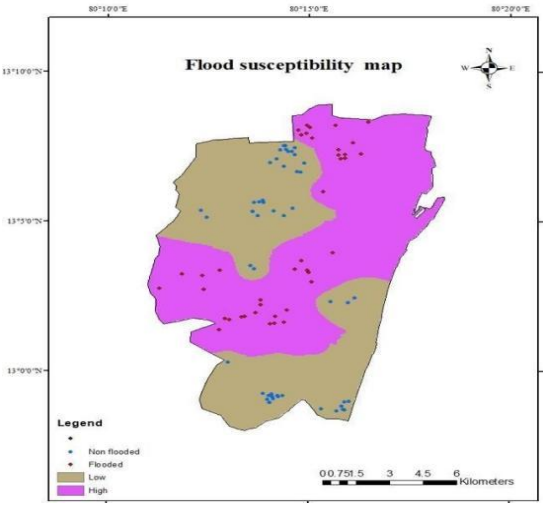

Fig. 25. Urban flood susceptibility map

5.5 Validation of RF model:

Model validation is a significant part of evaluating the model or its application accuracy. For this, confusion matrix is taken as the best method for validation or determining accuracy of the model. A confusion matrix is a table which is used to define the performance of the classification algorithm. The confusion matrix is defined by its four basic characteristics. That is True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP defines the number of pixels or points which have been correctly classified as non flood points. TN defines the number of pixels or points which have been correctly classified as flooded points. FP defines the number of misclassified pixels or points with flood but are actually non flooded. FN represents the number of misclassified pixels or points with non flooded but is actually flooded. The confusion matrix with actual values on x axis and predicted values on y axis are presented in a tabular format below.

```
sns.heatmap(metrics.confusion_matrix(y_test, y_pred), annot = True,
<AxesSubplot:>
```

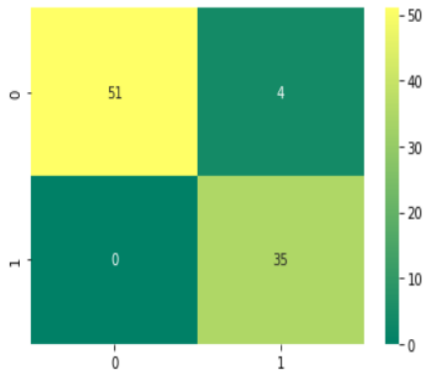

Fig. 26. Confusion matrix from RF model

Here 51 times model predicted 0 (non flooded) when actual values are 0. Then 4 times model predicted 0 when the actual value is 1(flooded). 0 times model predicted 1 when actual values are 0. 35 times the model predicted 1 when actual values are 1. Diagonal elements show the correctly classified points or pixels whereas non-diagonal elements show incorrectly classified pixels or points. So, the accuracy of the model can be calculated by the sum of diagonal elements to the sum of diagonal and non diagonal elements.

Accuracy = (TP + TN / TP + TN +FP + FN)
= (51+ 35 / 51 + 4 + 0 + 35)
= 0.955

The accuracy obtained as a result of classification by the novel Random Forest (RF) machine learning algorithm is 0.955 (95%).

6. CONCLUSION

The urban flood susceptibility map has been classified into low and high susceptible regions. The method using the Random Forest model produced the best outcomes. The prediction accuracy is 0.955 (95.5%) for the Random Forest model. The findings of this analysis will assist decision- makers in carrying out effective flood management in the future. There are also many other important factors that regulate flood susceptibilities such as the influence of dams and other hydraulic structures which regulates the flood and were not considered in this study. Also, parameters such as spatial variation of typhoon frequency, soil texture, etc. can be incorporated in further studies.

7. REFERENCES

1. Abdu R M, Swapan T, and Sushanta M, 2021, 'Flood susceptibility modelling using advanced ensemble machine learning models, International Journal of Geoscience frontiers, Elsevier, 12, pp.10105.

2. Chukwuma E, and Okonkwo C, 2020, 'A GIS based flood vulnerability modelling of Anambra State using an integrated IVFRN-DEMATEL-ANP model', Journal of Hydraulics and Water Resources Engineering, Heliyon, 6, pp.9.

3. Gang, Z, Bo P and Zongxue X, 2021, 'Assessment of urban flood susceptibility using semi-supervised machine learning model', Journal of Science of environment, Elsevier, 659, pp.940 – 949.

4. Hamid D, Ali T and Omid R, 2021, 'A hybridized model based on neural network and swarm intelligence-grey wolf algorithm for spatial prediction of urban flood inundation', Journal of Hydrology, Elsevier, 603, pp.126854.

5. Sadhan M., Subodh C P, and Indrajith J, 2020, 'Prediction of highly flood prone areas by GIS based heuristic and statistical model in a monsoon dominated region of Bengal Basin', Society and Environment, Elsevier, 19, pp.100343.

6. Yichen Z Y, Hui L, Ruishan C and Zhenhuan L, 2020, 'A GIS-Based Approach for Flood Risk Zoning by Combining Social Vulnerability and Flood Susceptibility: A Case Study of Nanjing, China', Environmental Research and Public Health, MDPI, 8, pp.134.

7. Zening W, Yanxia S, Huiliang W, and Meimei W, 2019, 'Assessing urban flood disaster risk using Bayesian network model and GIS applications', Journal of Geomatics, Natural hazards and risk, Taylor and Francis, 216, pp.2163-2184.

8. Zhaoli W, Chengguang L, and Xiaohong C, 2015, 'Flood hazard risk assessment model based on random forest', Journal of Hydrology, Elsevier, 527, pp.1130– 1.