IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⵣ Impact Factor 7.39 ⵣ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11798

FLOOD PREDICTION USING DIFFERENTIAL ML METHODS

K Manohar Prakul[1], Bhargav DR[2], Merin Meleet[3]

Student, Information Science and Engineering, RV College of Engineering, Bangalore, India[1]

Student, Information Science and Engineering, RV College of Engineering, Bangalore, India[2]

Assistant Professor, Information Science and Engineering, RV College of Engineering, Bangalore, India[3]

Abstract: One of the most catastrophic natural disasters that are extremely difficult to model are floods. Research on improving flood prediction models has helped to lower risks, recommend policy changes, reduce the number of fatalities, and lessen property damage from floods. Over the past two decades, research has shown that machine learning (ML) methods have greatly advanced prediction systems, offering better performance and cost-effective solutions. These methods include logistic reasoning, decision trees, support vector classification, KNN classifiers, and random forest classifiers. The process of modelling the likelihood of a discrete result given an input variable is known as logistic regression. Each internal node of a decision tree, which resembles a flowchart, represents a "test" on an
Flood alerts, flood reduction, and flood avoidance are all possible benefits of using machine learning (ML) models for flood prediction. Due to their cheap computational requirements and predominance of observational data, machine-learning (ML) approaches have grown in popularity as a result. The goal of this study was to develop a machine learning model that can forecast floods in the Nigerian state of Kebbi using historical rainfall data from the previous 33 years (33). This model may then be applied to other high-risk flood-prone states in Nigeria. This paper assessed and compared the accuracy, recall, and receiver operating characteristics (ROC) scores of three machine learning algorithms: decision trees, logistic regression, and support vector classifiers (SVR). Compared to the other two techniques, logistic regression produces higher accuracy outcomes.

I. INTRODUCTION

project is used to forecast when floods would occur. To evaluate the precision of two distinct machine learning models, the project was completed. Reasoning belongs to the There are 5 forms of lesioning: Decision trees, support vector classification, KNN classifiers, and Random Forest classifiers are all examples of logical reasoning. The process of modelling the likelihood of a discrete result given an input variable is known as logistic regression. Each internal node of a decision tree, which resembles a flowchart, represents a "test" on an attribute. Finding a hyper plane in an N-dimensional space (N is the number of features) that clearly classifies the data points is the goal of the support vector machine algorithm. The broad category of ensemble-based learning techniques includes random forest classifiers. They are easy to put into practise. It is clear that machine learning is now having a significant impact across many industries. This involves study into different events based on information from previous events. It is the capacity to gain knowledge by experience, which is only feasible when we receive some fresh, accurate, and comprehensive information. Therefore, in order to use ML, we must gather the necessary data. The data must then be pre-processed before being used for subsequent actions or functionality. To use the data as effectively as possible, it is essential to keep informed about what is currently available or existing on the system. To acquire the accuracy depending on the available data, we can use or experiment with various algorithms. Flood prediction is a crucial factor because

II. LITERATURE SURVEY

A literature survey is a review of previously published materials on the report's subject. This contains (in this order): The literature survey should be set up so that the progression of ideas in that topic is rationally (and historically) represented. Peer-reviewed literature in the most prestigious topic areas are examined in this assessment to determine the state of the art of machine learning (ML) algorithms for flood prediction. The articles that included performance evaluation and comparison of ML methods were given priority to be included in the review in order to determine which ML approaches perform better in specific applications among those that were found through search queries utilising the search strategy. In addition, four other quality measures for each article were taken into account when choosing one, including source normalised is frequently used to predict floods, particularly to anticipate flash floods or short-term floods [61]. However, it has been demonstrated that rainfall forecasts are insufficient for precise flood prediction. For instance, estimates of the soil moisture in a catchment, in addition to rainfall, are used to predict streamflow in a long-term flood prediction scenario [62]. Although accurate precipitation forecasting is crucial, other factors related to flood resource variables were taken into account in the [63]. In order to include the most efficient flood resource factors in the search queries, this literature review's technique does as follows. A mixed Deep Learning (DL) and Fuzzy Logic (FL) based algorithm is used in an urban flood estimating and checking platform developed by Karyotis et al. [8] as a part of a UK Newton Fund project in Malaysia. This model collects real-time data using inexpensive sensors. Artificial Neural Organizations (ANN) are a reliable solution for time forecasting problems, and DL [9] uses them for both supervised and unsupervised training. FL, which is based on the concept of fuzzy sets, is capable of handling "fractional truth." For the Deep Learning Model, a data window of 200 data points is the most suitable size. Additionally, it was observed that the likelihood of a flood increases when the intensity of the rain is strong, the storm span is long, and the soil absorption is extremely low. A technique to forecast floods has been created by Dola et al. [10] utilising machine learning models. Predictions of rainfall for the following month are made using rainfall statistics from previously available data. Rainfall forecasting is possible for both short- and long-term events. The Indian Meteorological Department provided the data. There are two independent datasets that include average rainfall information for each month and district from 1951 to 2000; the information after that is from 1901 to 2015 and includes average rainfall information for each state. IoT is now used by this Low Cost IoT based Flood Monitoring System to calculate how long it would take for the flood to reach land. ML algorithms are used to estimate the rainfall's severity.Linear regression, support vector machines, and artificial neural networks are the algorithms utilized for the same. The several IoT devices employed include IoT Gecko, water-float sensors, and rain-drop sensors.A buzzer beeps and sends out a warning of an impending flood as the water level increases. The dataset from the previous three months is used in the linear regression model to forecast the amount of precipitation for the following month.

III DATA PREPARTION

IoT is used by this Low Cost IoT based Flood Monitoring System to calculate how long it would take for the flood to reach land. ML algorithms are used to estimate the rainfall's severity.and is also for the regression and the the the
Linear regression, support vector machines, and artificial neural networks are the algorithms used for the same. The several IoT devices employed include IoT Gecko, water-float sensors, and rain-drop sensors.A buzzer beeps and sends out a warning of an impending flood as the water level increases. The dataset from the previous three months is used in the linear regression model to forecast the amount of precipitation for the following month. The ML lifecycle's last stage is data preparation. To get insights or generate predictions, the data is first gathered from numerous sources, cleaned of any junk, and then translated into real-time machine learning projects. In order to generate the data sets, accurately process the data, and make accurate predictions, machine learning also aids in the discovery of patterns in data. Data pre-processing, data splitting, and other processes for data preparation in machine learning will be covered in this topic, "Data Preparation in Machine Learning." So let's begin with a brief overview of machine learning's data preparation. Data preparation is the act of cleaning and converting raw data so that machine learning algorithms can generate correct predictions. Even though ML's most challenging stage is data preparation, real-time projects benefit from its reduced process complexity. The following problems have been identified during the machine learning process of data preparation:Data missing: Incomplete or missing data is a common problem in most datasets. Records occasionally have blank cells, values (such as NULL or N/A), or a particular character, such as a question mark, etc., in place of the proper data.Anomalies or outliers: When data comes from unidentified sources, ML algorithms are sensitive to the range and distribution of values. These values can negatively impact the performance of the entire machine learning training system When data comes from unidentified sources, ML algorithms are sensitive to the range and distribution of values. These numbers have the potential to ruin both the model's performance and the overall machine learning training system. Therefore, it is crucial to find these outliers or anomalies using techniques like visualisation.

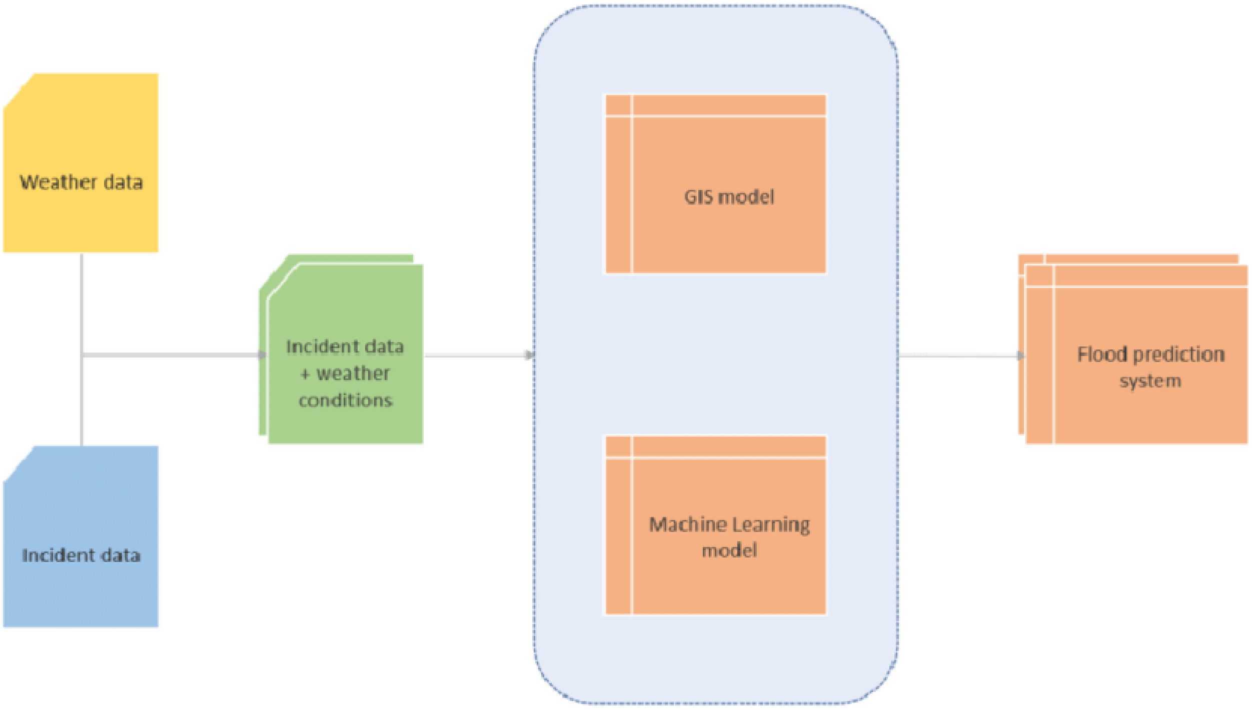
Unstructured data format: Data must be extracted and converted to a different format from a variety of sources. Therefore, before to launching an ML project, always speak with subject-matter experts or import data from established sources. Limited Features: Data from a single source always has a limited number of features; therefore, it is necessary to import data from additional sources to enhance the features or to create multiple features into datasets.feature engineering knowledge: By adding new material to ML models, features engineering improves model performance and prediction accuracy. The initial data set had nine input variables after combining the weather and flood reporting data.

DOI: 10.17148/IJARCCE.2022.11798

Variable	Type	Description
temp	Continuous	Temperature in degrees Celsius.
hum	Continuous	Relative humidity ratio in %.
precip	Continuous	Precipitation in mm.
sun	Continuous	Sun exposure in W/m².
wind_speed	Continuous	Wind speed in m/s.
wind_dir	Continuous	Wind angular direction.
station	Categorical (nominal)	Name of the weather station.
datetime	Categorical (interval)	Timestamp.
target	Categorical (binary)	Flood occurrence flag.

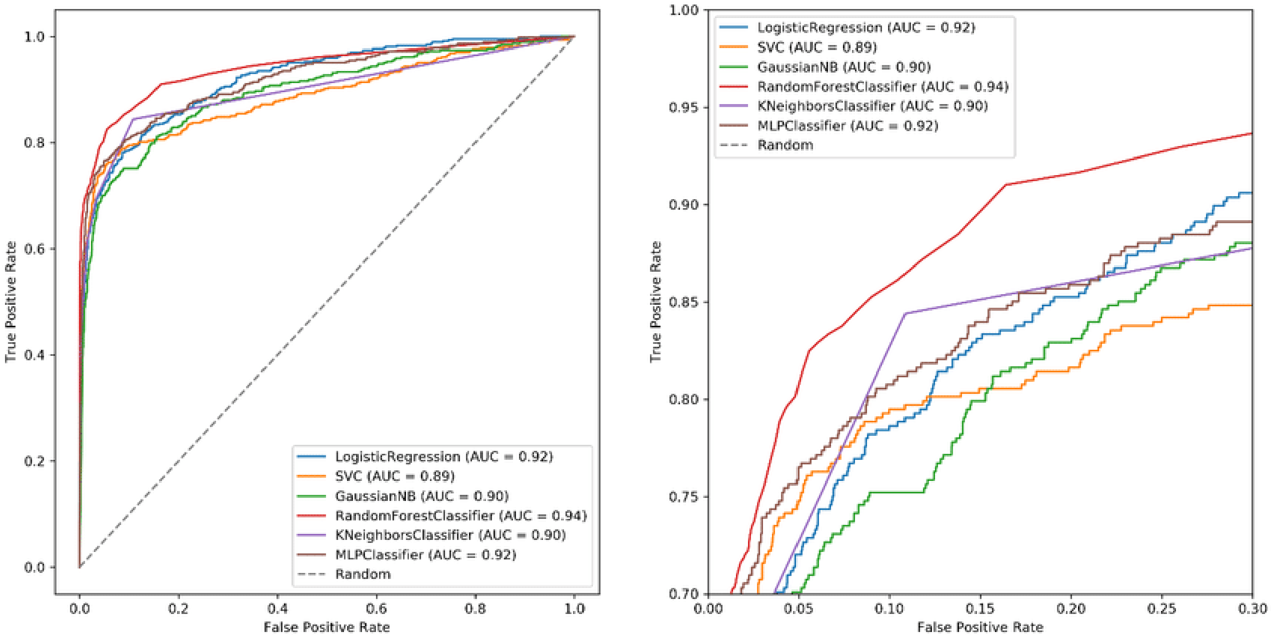
Data quality issues, such as missing values, abnormalities, and outliers were found and treated in two processes in order to clean and prepare the data for modelling. The data set's extreme values were first eliminated and replaced with null values. Second, taking into account the geospatial and temporal aspects, the following method was employed to address missing values and minimise noise caused by imputation:based on the L2 distance (in a two-dimensional Euclidian space), use observations from the adjacent weather station;. Utilize measurements derived from current observations by performing a forward and backward temporal search; For the two closest neighbours (k=2), use k-nearest neighbours (KNN); One more step was taken to prevent throwing away valuable information because the variable for wind direction still had over 241 observations with missing data. This involved utilising a random forest to impute the missing data points. According to Doreswamy et al. [27], utilising a Random Forest for imputation produced comparatively low error when compared to other common ML techniques.Following the data cleaning procedures, all variables were filtered in accordance with the flood occurrence flag (variable "goal"), enabling comparison of behaviours and results. This allowed us to perform a bivariate analysis to find patterns between each weather variable.

IV PIPELINE

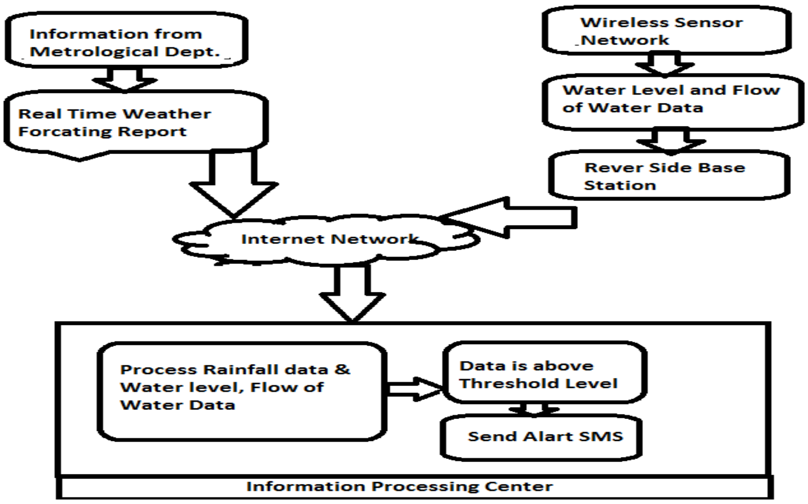


The data flow for the suggested flood prediction system is shown. The Lisbon City Council made the information utilised in this project, which included records of emergency calls to the fire department and local meteorological measures, public. The Incremental Spatial Autocorrelation approach, which can choose the best distance based on its capacity to maximise the z-score for the Gi* statistic, can be used to optimise a parameter. Using this method, the ideal separation was discovered to be 324 metres, producing the map that is displayed. The spatial prevalence of floods could be detected and quantified by the Hot Spot analysis, and when combined with the flood prediction scores given by the ML classifier, it could be utilised to factor in the danger of flood for crucial areas.

DOI: 10.17148/IJARCCE.2022.11798



A. Model architecture


Information from Metrological Dept.
Real Time Weather Forcating Report
Wireless Sensor Network
Water Level and Flow of Water Data
Rever Side Base Station
Internet Network
Process Rainfall data & Water level, Flow of Water Data
Data is above Threshold Level
Send Alart SMS
Information Processing Center

The historical records of flood incidents, in addition to the real-time cumulative data from a number of rain gauges or other sensing devices over varied return periods, are frequently used to create the ML prediction model. The dataset's conventional sources include rainfall and water level measurements made using ground rain gauges or more recently with satellites, multisensor systems, and/or radars [62]. Even Nevertheless, remote sensing is a desirable technique for instantaneously gathering higher-resolution data. Additionally, compared to rain gauges, weather radar readings frequently offer a more trustworthy dataset due to their high resolution [63]. As a result, creating a prediction model based on a radar rainfall dataset was said to generally offer superior accuracy [64]. Whether utilising a dataset based on radar To build and assess the learning models, the historical dataset of hourly, daily, and/or monthly values is partitioned into separate sets. The separate sets of data are trained, validated, verified, and tested in order to do this. In-depth descriptions of the methodology for flood modelling and the ML modelling workflow can be found in the literature [48][65]. The fundamental process for creating an ML model is depicted in Figure 2. ANNs [66], neuro-fuzzy [67], adaptive neuro-fuzzy inference systems (ANFIS) [68], support vector machines (SVM) [69], wavelet neural networks (WNN) [70], and multilayer perceptron (MLP) [71] are some of the main ML algorithms used to predict floods. The background and a brief overview of these core ML algorithms are provided in the subsections that follow.

DOI: 10.17148/IJARCCE.2022.11798

B. METHADOLOGY

Rational regression

The training of logistic regression is very effective and easier to implement and analyse.It doesn't make any assumptions about how classes are distributed in feature space.Logistic regression should not be employed if there are less data than features because this could result in overfitting.It creates linear borders.

Decision tree

Decision trees take less work to prepare the data during pre-processing than other methods do.Data normalisation is not necessary for a decision tree.A slight change in the data can result in a big change in the decision tree's structure, which can lead to instability.When compared to other algorithms, a decision tree's calculations may become far more complicated.

Support vector classification

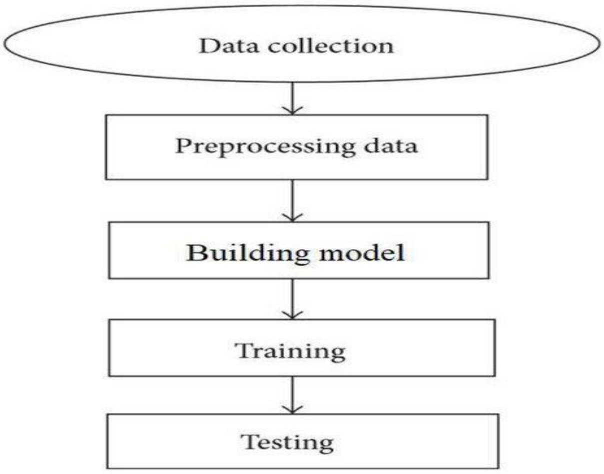
One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it is largely employed in Machine Learning Classification issues. The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary. SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors, which are used to represent these extreme instances, form the basis for the SVM method. Take a look at the diagram below, where two distinct categories are identified using

Random forest classification

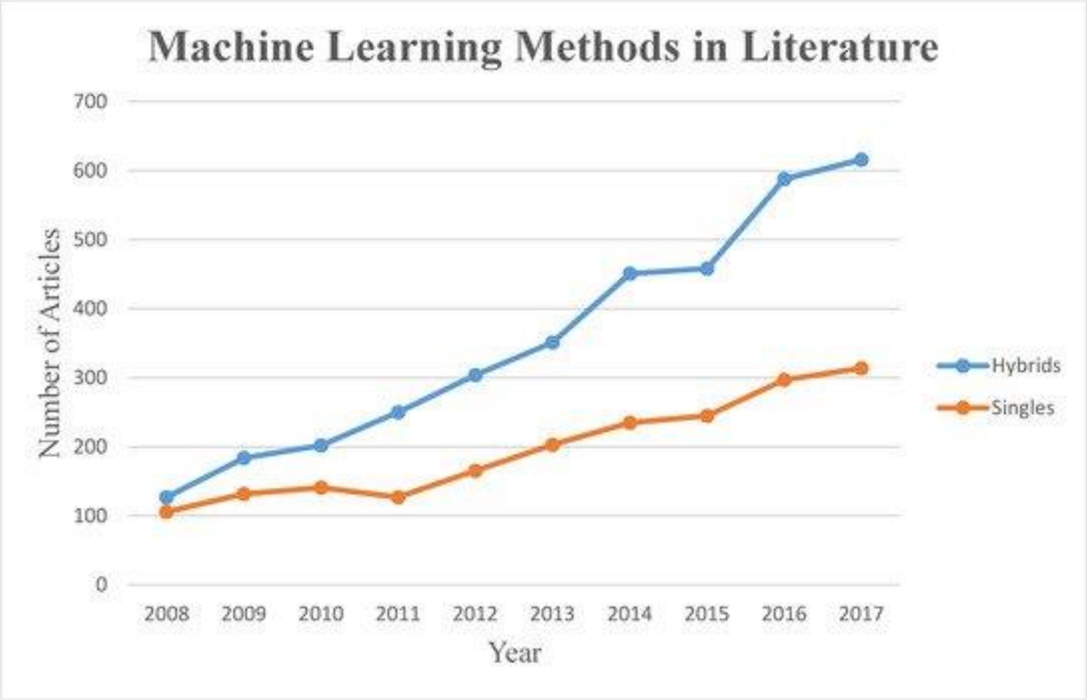
A random forest is a meta estimator that employs averaging to increase predicted accuracy and reduce overfitting after fitting numerous decision tree classifiers to distinct dataset subsamples. If bootstrap=True (the default), the size of the sub-sample is determined by the max samples argument; otherwise, each tree is constructed using the entire dataset.

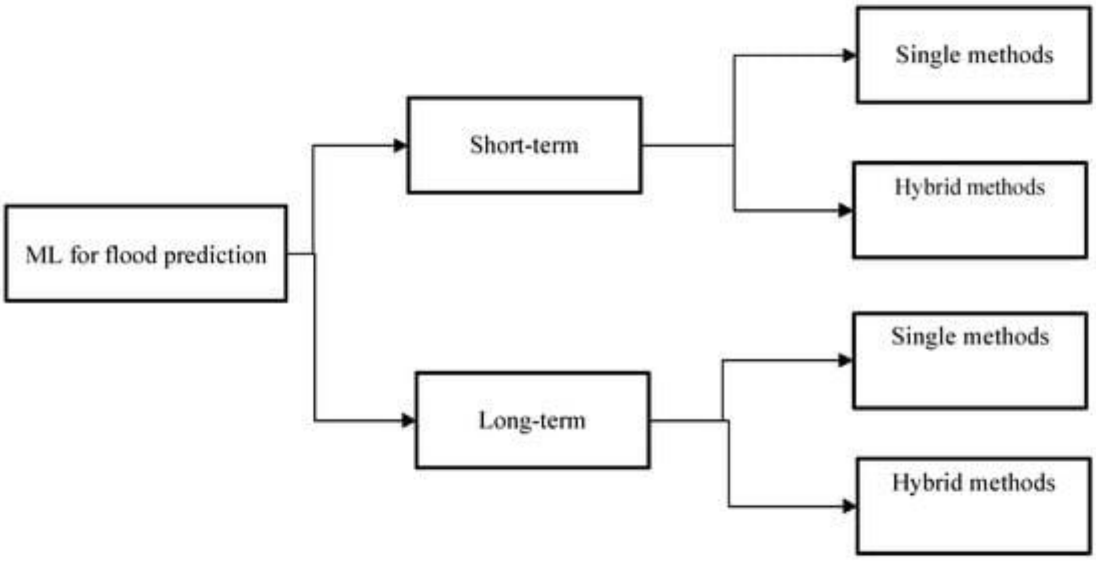
Knn classification

One of the simplest machine learning algorithms, based on the supervised learning method, is K-Nearest Neighbor. The K-NN algorithm places the new case in the category that is most comparable to the extant categories by assuming similarity between the new case/data and existing cases. A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. This means that utilising the K-NN method, fresh data can be quickly and accurately sorted into a suitable category. Although the K-NN approach is most frequently employed for classification problems, it can also be utilised for regression. Since K-NN is a non-parametric technique, it makes no assumptions about the underlying data.



DOI: 10.17148/IJARCCE.2022.11798



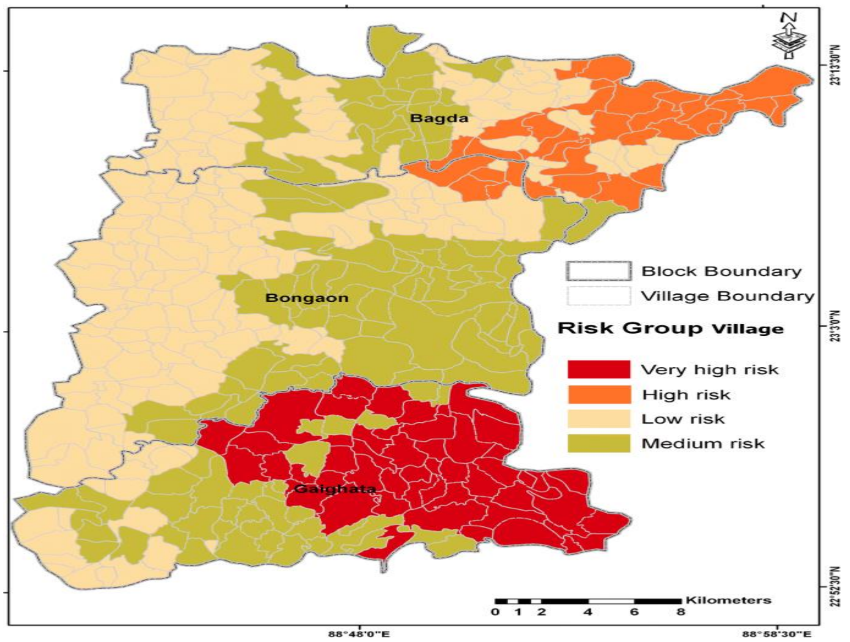


C TESTING AND SIMULATION

Additional procedures were needed once the best classifier was identified in order to reliably determine where a flood will occur if the necessary conditions are satisfied. The classifier has a limited representation of the spatial dimension because it was trained primarily on meteorological data from three stations, and because predictions are made at the station level. Additionally, it was noted that some places are more prone to flooding than others, which may indicate that the ML model may have overlooked other underlying causes and effects. According to recommendations made in the literature [13] [26], the application of GIS is suggested for this purpose in order to identify geographical heterogeneity and create a spatial vulnerability indicator. Finally, by identifying the areas most vulnerable to flooding, this indication might be utilised to refine the forecasts obtained in the previous stage.

The Gi* statistic is used in the Hot Spot analysis to locate adjacent locations that have a higher occurrence of floods based on their spatial link. This study employs historical georeferenced data. The 100x100m grid utilised for this analysis's spatial unit has a total of 8,986 cells and was defined at the city level. According to recommendations made in the literature [13] [26], the application of GIS is suggested for this purpose in order to identify geographical

heterogeneity and create a spatial vulnerability indicator. Finally, by identifying the areas most vulnerable to flooding, this indication might be utilised to refine the forecasts obtained in the previous stage.
The Gi* statistic is used in the Hot Spot analysis to locate adjacent locations that have a higher occurrence of floods based on their spatial link. This study employs historical georeferenced data. The 100x100m grid utilised for this analysis's spatial unit has a total of 8,986 cells and was defined at the city level. The clustering of geographical data points is determined by a distance parameter that is conditional on the measured spatial relationship. The Incremental Spatial Autocorrelation approach, which can choose the best distance based on its capacity to maximise the z-score for the Gi* statistic, can be used to improve this value. Using this approach, the ideal separation was discovered to be 324 metres, producing the map that is displayed.



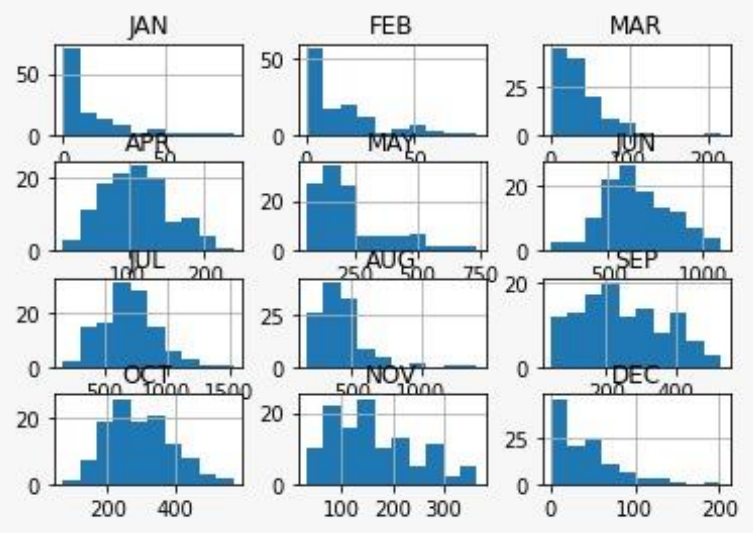
The spatial prevalence of floods could be located and measured using the Hot Spot analysis, and coupled with the flood prediction scores provided by the ML classifier, it could be utilised to factor in the danger of flooding for important places. In order to give each grid cell a single metric to represent a risk index, the predicted scores from the classification model were combined with the p-value to reflect the statistical significance for each grid cell to be identified as a hot spot. The weighted average of the two scores is calculated by this function (ranging between 0 and 1), This, as expressed below, can be changed to favour either the ML model (RF) or the GIS model (HS):
$Flood\ Risk\ Index = weightRF * scoreRF + weightHS * (1 - pvalueHS)$
The weights were left unadjusted (i.e. set to 0.5) for the trials conducted for this research, yielding a simple average of these scores.

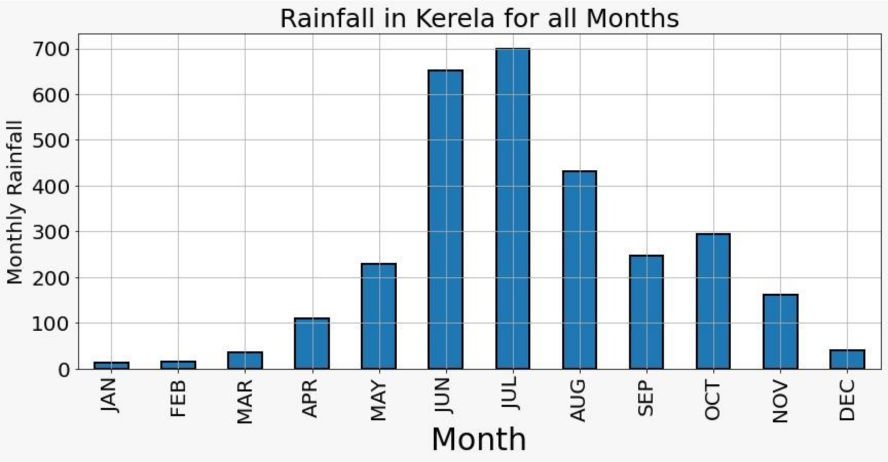
V. CONCLUSION

With minimal sensing data, the method developed during this project has produced encouraging results for flood prediction. The purpose of this work is to show that by integrating ML , it is possible to identify the important predictors and circumstances for a flooding scenario by using a small amount of data.From a machine learning (ML) perspective, non-linear models were better able to detect floods, and among those, a Random Forest model performed the best, with a Matthews Correlation Coefficient of 0.77. The 2-hour window moving average for rainfall was discovered to be the most significant flood predictor, as was determined throughout the modelling phase. Based on the hot spots noted in the historical data, the GIS model was able to modify the sensitivity of the predictions made by the ML model. Hot spots were identified as areas with a higher Flood Risk Index because they were thought to be more likely to flood.The model sensitivity was 0.84 when the weights for the two models were equal. However,A larger false positive rate resulted from the enhanced sensitivity, demonstrating the need for additional research.alterations to the score weights and/or model threshold.The geographical representation was also enhanced by the GIS model, as the ML model wasThe integration of both models generate forecasts down to the weather station level.able to calculate a risk index for each cell in the 100 m2. Regarding the data set's limitations, although depending solely on meteorological conditions and lacking geographical diversity for these conditions (because this information was only gathered from threestations), this method was consistently capable of accurately forecasting floods.confidence. Additionally, its robustness suggests that it might be repeated in a variety of settings andwhen backed by a high-resolution geographical dataset and increased performance additional weather stations.

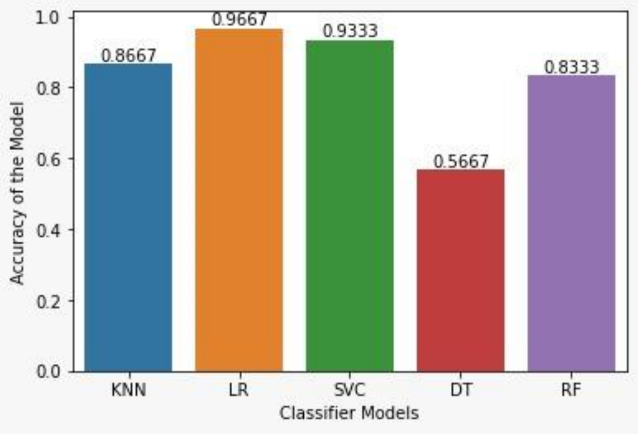
VI RESULTS

We now recognise that decision trees, out of the five models (KNN, SVC, DT, RFC, and LR), are the least accurate one because they are unstable, which means that even a slight change in the data might result in a significant change in the structure of the ideal decision tree. Since 75% of the floods that were predicted actually happened, this model proves to be very helpful for further research and the implementation of flood avoidance or early warning systems. Floods are the most dangerous of all natural disasters, causing enormous damage to human life, the environment, agriculture, and the economy. Governments are therefore under pressure to produce accurate and reliable maps of flood hazard areas and to make additional arrangements for workable flood management. flood forecasting software is suspicious expectation and additional misbehaviour inquiry, approach recommendations, and further departure demonstrating protest enrollment. The current most destructive disaster is flooding. It can destroy a whole network in the event of a massive flood. Building a flood expectation framework is essential as a tool to foresee and reduce the flood risk. It serves as an example of how important it is to inform residents to take immediate action, such evacuating rapidly to a higher, safer location. For logistic regression to provide sufficient numbers in the two categories of the reaction variable, large sample sizes are required. The size of the example needed increases with the number of illustrative components. Regression into logic is a useful tool for showing dependence. on at least one informative component, where the last one can be either clear-cut or endless, of a parallel reaction variable. The attack of the following model can be assessed using a variety of approaches.





DOI: 10.17148/IJARCCE.2022.11798



VII. ACKNOWLEDGEMENT

We thank **Prof. Merin Meleet**, Assistant Professor, Department of Information Science and Engineering, a debt of gratitude for her unwavering encouragement, insightful recommendations, and essential counsel during the course of our project work and for her assistance in the creation of this thesis. We sincerely appreciate **Dr. Sagar B.** M.'s assistance and encouragement as professor and department head of information science and engineering at RVCE.We really appreciate **Dr. K. N. Subramanya**, our beloved principal, for his support of our project endeavour.We appreciate the assistance of the whole teaching team and technical personnel of the RVCE department of information science and engineering. Finally, we would like to take this occasion to thank our family and friends for providing continuous backup assistance during the project.

REFERENCE

[1] Habitat III, United Nations (2017). 21st Issue Paper: Smart Cities. 142-149 in Housing and Sustainable Urban Development Conference of the United Nations.
https://habitat3.org/wpcontent/uploads/Habitat-III-Issue-Papers-report.pdf.

[2] C. Heinzlef and D. Serre (2018). evaluating and mapping the urban flood resilience in relation to cascading impactsthrough important infrastructure networks 30, 235–243, International Journal of Disaster Risk Reduction. https://doi.org/10.1016/j.ijdrr.2018.02.018.

[3] L. Feyen and R. Dankers (2008). Assessment of the influence of climate change on flood risk in Europe using higHresolution climate simulations. Atmospheres, 113, Journal of Geophysical Research (D19). https://doi.org/10.1029/2007JD009719.

[4] D. Serre & C. Heinzlef (2020). Urban resilience: From a constrained vision of urban engineering to a more extensive, long-term implementation. 11, 100075, Water Security. https://doi.org/10.1016/j.wasec.2020.100075.

[5] (2018). Assessing Critical Infrastructure Network Flood Resilience at the Neighborhood Scale Using DS3 Model Testing. (pg. 207–220) in Urban Disaster Resilience and Security. Cham Springer. https://doi.org/10.1007/978-3-319-68606-6 13.

[6] B. Robert, Y. Hémond, C. Heinzlef, and D. Serre (2020). Applying urban resilience measures to mitigate climate change and its risks: some developments in Canada and France from theory to practise. Cities,

[7] V. Becue, D. Serre, and C. Heinzlef (2019). Application in Avignon, in the region of Provence Alpes Côte d'Azur, of operationalizing urban resilience to floods in embanked lands. 181–193 Safety Science.

[8] The representation and forecasting of the Indian Ocean dipole in the POAMA seasonal forecast model by Mei Zhao and Harry H. Hendon. 2009, 135(4), 337–352, 10.1002/qj.370, Quarterly Journal of the Royal Meteorological Society.

[9] Deva K. Borah; Comprehensive evaluation and comparison of the hydraulic techniques of storm event watershed models. 25, 3472–3489, 10.1002/hyp.8075, Hydrological Processes (2011).

[10]Carmelina Costanzo, Francesco Macchione, and Pierfranco Costabile a storm event watershed model based on 2D fully dynamic wave equations for surface runoff. 27, 554-569, 10.1002/hyp.9237, Hydrological Processes (2013).

[11]Experimental validation of two-dimensional depth-averaged models for predicting rainfall-runoff from precipitation data in metropolitan settings by Luis Cea, M. Garrido, and Jerónimo Puertas. 2010, 88–102, 10.1016/j.jhydrol.2009.12.020; Journal of Hydrology;

[12]Pilar Garca-Navarro, Javier Fernández Pato, Daniel Caviedes-Voullième, and infiltration parameter analysis and calibration. 1016/j.jhydrol.2016.03.02, Journal of 2016 536, 496–513.